Linear Models in Statistics

Lecture notes

by

Andrea Kraus, Ph.D.

These lecture notes are intended for students of the 3rd year bachelor course *M5120 Linear Models in Statistics I*. This is a one-semester course offering an overview of linear statistical models as the fundamental tool of statistical analysis. The students encounter theory, software implementation, applications and interpretation. After the course the students are expected to recognize the situations that can be addressed by linear models, formulate and implement the model, and interpret the results. At the same time, the students are made aware of the limitations of the model and should be able to recognize and possibly avoid problems in a given situation.

The lecture notes are based primarily on the following texts:

Julian J. Faraway (2014). *Linear Models with R.* Second edition.

Chapman& Hall/CRC.

Simon N. Wood (2006). *Generalized Additive Models; An introduction with R.*

Chapman& Hall/CRC.

Jiří Anděl (2005). *Základy matematické statistiky.*

Matfyzpress.

I would like to thank Professor Victor Panaretos for kindly sharing his teaching experience and course materials for his course on Regression at the Swiss Federal Institute of Technology in Lausanne. I would also like to thank my students for their feedback on various bits and pieces of these notes.

Contents

Chapter 1

Introduction

1.1 Motivation

Is eating chocolate good for our health?

Effects of chocolate

- it has been suggested that chocolate consumption
  - is beneficial to cardiovascular health (effects on “bad” cholesterol, blood pressure, stroke, . . .)
  - lowers the risk of diabetes
  - improves cognitive function & reduces memory decline
  - . . .
- but it has also been suggested that chocolate consumption
  - leads to obesity (risk for cardiovascular problems, diabetes)
  - leads to dental problems
  - decreases bone density
  - . . .
- should be eaten in moderation . . .

It’s an uncertain world . . .

- How much of
  - chocolate and other goodies is good for our health?
  - levels of bacteria, fertilizers, chemicals, . . . is safe?

- What is the right size for
    - the height of a dam?
    - insurance premium?
    - mortgage interest?
- What is
    - the average salary?
    - public opinion on ...?
    - results in upcoming elections?

Sources of uncertainty

- we do not fully understand the phenomenon
    - human body
    - nature
- we do not know the future
    - occurrence and size of a flood
    - occurrence and size of insurance claims
    - level of inflation
- we do not collect complete data
    - average salary
    - public opinion
- measurement error, human factor, ...

Statistics is all around us

- statistics is used to quantify the uncertainty
- Strategy
    1. build a mathematical model, i.e. define
        - what is known
        - what is uncertain
    2. build a probabilistic model for what is uncertain
    3. use probability calculus to draw conclusions

4. "translate" back to the original problem (interpret the results)

- uncertainty at the beginning ⇢ imperfect answers at the end
- statistics is used for quantifying uncertainty,

  not for getting rid of it

Notation

- random variable X, Y
  - (náhodná veličina)
- random vector/matrix $\mathbf{X}$, $\mathbf{Y}$
  - (náhodný vektor/matice)
- density/probability mass function f
  - (hustota/pravděpodobnostní funkce)
- parameters θ, β, normal distribution $\mathrm{N}(\mu, \sigma^2)$
  - (parametry, normální rozdělení)
- expectation $\mathsf{E}\, X$, $\mathsf{E}\, \mathbf{X}$
  - (střední hodnota)
- variance/covariance/variance-covariance matrix

  $\mathsf{Var}\, X$, $\mathsf{Cov}(X, Y)$, $\mathsf{Var}\, \mathbf{X}$
  - (rozptyl/kovariance/kovarianční matice)

$$\mathsf{Var}\, \mathbf{X} = \begin{pmatrix} \mathsf{Var}\, X_1 & \mathsf{Cov}(X_1, X_2) & \dots & \mathsf{Cov}(X_1, X_n) \\ \mathsf{Cov}(X_1, X_2) & \mathsf{Var}\, X_2 & \dots & \mathsf{Cov}(X_2, X_n) \\ \dots & \dots & \dots & \dots \\ \mathsf{Cov}(X_1, X_n) & \mathsf{Cov}(X_2, X_n) & \dots & \mathsf{Var}\, X_n \end{pmatrix}$$

Statistician's TODO list

1. identify right questions
2. collect relevant data $x_1, \dots, x_n$
3. think of them as realisations of random variables $X_1, \dots, X_n$

   with distributions (densities/frequency functions) $f_1, \dots, f_n$

where f_i is in fact $f_i(x, \theta)$

4. estimate θ/make inference about θ
5. use the results to answer the questions

Example

1. Does consuming [amount of] chocolate decrease blood pressure [type, measurement]?
   - ~~Is chocolate good for our health?~~
2. design a trial, collect participants' blood pressures $x_1, \dots, x_n$
3. suppose e.g. that $X_i \sim \mathrm{N}(\mu_i, \sigma^2)$
   - μ_i: function of eating [amount of] chocolate, age, gender, ...
   - e.g. $\mu_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k}$
   - $x_{i,1} = \begin{cases} 1 & \text{if the person eats [amount of] chocolate} \\ 0 & \text{otherwise} \end{cases}$
4. test $H_0 : \beta_1 \geq 0$ versus $H_1 : \beta_1 < 0$
5. if we reject H_0 in favour of H_1 at $\alpha\%$ level, we have shown that at $\alpha\%$ level consuming [amount of] chocolate is associated with a lower blood pressure [type, measurement]
   - if we do not reject H_0 in favour of H_1 at $\alpha\%$ level, we have not shown that at $\alpha\%$ level consuming [amount of] chocolate is associated with a lower blood pressure [type, measurement]

Linear model

- model: $Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k} + \varepsilon_i$
  - ▷ matrix notation: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
  - ▷ assumptions: $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$, $\mathsf{Var}\,\boldsymbol{\varepsilon} = \sigma^2\mathbf{I}$
    - ∗ then $\mathsf{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$, $\mathsf{Var}\,\mathbf{Y} = \sigma^2\mathbf{I}$
  - ▷ we often assume that $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$
    - ∗ then $\mathbf{Y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$
- parameter: $\boldsymbol{\theta} = (\beta_0, \ldots, \beta_k, \sigma^2)^\top = (\boldsymbol{\beta}^\top, \sigma^2)^\top$
  - ▷ estimation (point, interval)
  - ▷ testing
  - ▷ interpretation

1.2 Statistics

Example

1. Does consuming [amount of] chocolate decrease blood pressure [type, measurement]?
   - ~~Is chocolate good for our health?~~
2. design a trial, collect participants' blood pressures $x_1, \ldots, x_n$
3. suppose e.g. that $X_i \sim \mathrm{N}(\mu_i, \sigma^2)$
   - μ_i: function of eating [amount of] chocolate, age, gender, ...
   - $\mu_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k}$
   - $x_{i,1} = \begin{cases} 1 & \text{if the person eats [amount of] chocolate} \\ 0 & \text{otherwise} \end{cases}$
4. test $H_0 : \beta_1 \geq 0$ versus $H_1 : \beta_1 < 0$
5. if we reject H_0 in favour of H_1 at $\alpha\%$ level, we have shown that at $\alpha\%$ level consuming [amount of] chocolate is associated with a lower blood pressure [type, measurement]
   - if we do not reject H_0 in favour of H_1 at $\alpha\%$ level, we have not shown that at $\alpha\%$ level consuming [amount of] chocolate is associated with a lower blood pressure [type, measurement]

Linear model

- model: $Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k} + \varepsilon_i$
  - ▷ matrix notation: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
  - ▷ assumptions: $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$, $\mathsf{Var}\,\boldsymbol{\varepsilon} = \sigma^2\mathbf{I}$
    - ∗ then $\mathsf{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$, $\mathsf{Var}\,\mathbf{Y} = \sigma^2\mathbf{I}$
  - ▷ we often assume that $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$
    - ∗ then $\mathbf{Y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$
- parameter: $\boldsymbol{\theta} = (\beta_0, \ldots, \beta_k, \sigma^2)^\top = (\boldsymbol{\beta}^\top, \sigma^2)^\top$
  - ▷ estimation (point, interval)
  - ▷ testing
  - ▷ interpretation

Parameter estimation

- we observe data with a distribution depending on a parameter
- we would like to use the data to estimate the value of the parameter
- estimator is a function of data (only!)
- for a one-dimensional parameter θ
  - point estimator $\hat{\theta}$
  - confidence interval $(\hat{\theta}_L, \hat{\theta}_U)$
- for a vector parameter $\boldsymbol{\theta}$
  - point estimator $\hat{\boldsymbol{\theta}}$
  - confidence region

Methods of point estimation

1. method of moments
   - "equate" theoretical and empirical moments
     - $\widehat{\mathsf{E}\,Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$
     - $\widehat{\mathsf{E}\,Y^2} = \frac{1}{n}\sum_{i=1}^{n} Y_i^2$
     - ...
2. maximum likelihood estimation
   - maximize the likelihood with respect to θ
   - likelihood
     - probability of observing the data at hand under a given model
   - very popular thanks to certain asymptotic optimality properties
3. other methods exist and we will see some

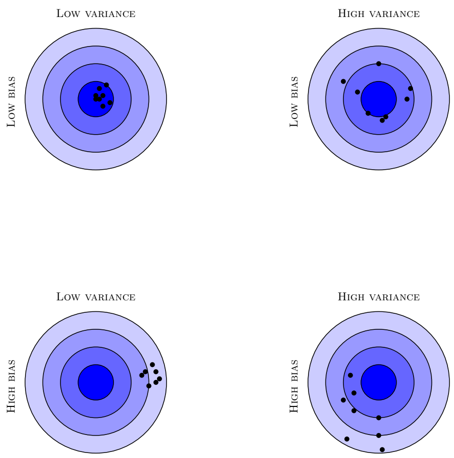
Maximum likelihood estimation

- $Y_1, \ldots, Y_n \overset{\text{ind.}}{\sim} f_i(y, \boldsymbol{\theta})$
- likelihood $L(y_1, \ldots, y_n; \boldsymbol{\theta}) = \prod_{i=1}^{n} f_i(y_i; \boldsymbol{\theta})$
- log-likelihood $\ell(y_1, \ldots, y_n; \boldsymbol{\theta}) = \sum_{i=1}^{n} \log\{f_i(y_i; \boldsymbol{\theta})\}$
- MLE $\hat{\boldsymbol{\theta}}_{\text{MLE}} = \operatorname{argmax}_{\boldsymbol{\theta}} \ell(y_1, \ldots, y_n; \boldsymbol{\theta})$

- usual computation
  - ▷ score function $\mathbf{U}(y_1, \ldots, y_n; \boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log\{f_i(y_i; \boldsymbol{\theta})\}$
  - ▷ score equation $\mathbf{U}(y_1, \ldots, y_n; \boldsymbol{\theta}) = \mathbf{0}$
  - ▷ find the solution $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ of the score equation
  - ▷ observed Fisher information matrix $\mathbf{J}(y_1, \ldots, y_n; \boldsymbol{\theta}) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}} \log\{f_i(y_i; \boldsymbol{\theta})\}$
  - ▷ show that $\mathbf{J}(y_1, \ldots, y_n; \hat{\boldsymbol{\theta}}_{\text{MLE}})$ is positive definite
  - ▷ Fisher information matrix (under regularity conditions) $\mathbf{I}(y_1, \ldots, y_n; \boldsymbol{\theta}) = \mathsf{E}_\theta\, \mathbf{J}(y_1, \ldots, y_n; \boldsymbol{\theta})$

Properties of estimators

- parameter θ is a number but estimator $\hat{\theta}$ is a random variable
  - ▷ $\hat{\theta}$ has a distribution
  - ▷ important distribution summaries: $\mathsf{E}\,\hat{\theta}, \mathsf{Var}\,\hat{\theta}$



- ideally, estimation improves with sample size
- let $\hat{\theta}_n$ be an estimator of θ based on n data points
- we define desirable properties for the sequence $\{\hat{\theta}_n\}_{n \in \mathbb{N}}$

- for a sequence of estimators $\hat{\theta}_n$ of a parameter θ

  1. unbiasedness
     - ▷ $\mathsf{E}_\theta\, \hat{\theta}_n = \theta \;\forall \theta$
  2. consistency
     - ▷ $\hat{\theta}_n \to \theta$ as $n \to \infty$ in $\mathsf{P}_\theta\; \forall \theta$ or a.s.

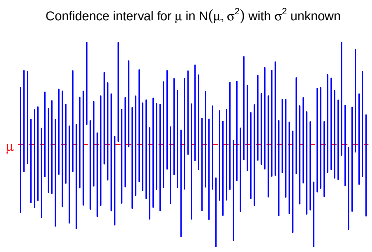
3. usual asymptotic normality
   - $\sqrt{n}(\hat\theta_n - \theta) \to \mathrm{N}(0, V(\theta))$ as $n \to \infty$ in distribution
4. efficiency
   - "small" $\mathsf{Var}\,\hat\theta$

Properties of MLE

- under regularity conditions
  - consistency
    - $\hat\theta_{\mathrm{MLE},n} \to \theta$ a.s. as $n \to \infty$
  - asymptotic normality
    - $\sqrt{n}(\hat\theta_{\mathrm{MLE},n} - \theta) \to \mathrm{N}(0, V(\theta))$ as $n \to \infty$ in distribution
  - asymptotic efficiency
    - $V(\theta)$ is the smallest possible
  - bias
    - $\hat\theta_{MLE}$ is often biased, with bias decreasing with n

Interval estimation

- parameter θ is a number but estimator $\hat\theta$ is a random variable
- confidence interval $(\hat\theta_{\mathrm{L}}, \hat\theta_{\mathrm{U}})$ is a pair of random variables
- $(1-\alpha)\,\%$ confidence interval satisfies that
  - $\mathsf{P}_\theta\{\theta \in (\hat\theta_{\mathrm{L}}, \hat\theta_{\mathrm{U}})\} = 1 - \alpha\ \forall\theta$
- note that randomness is in the borders, not in θ



- properties
  - coverage $1 - \alpha$
  - length $\hat\theta_{\mathrm{U}} - \hat\theta_{\mathrm{L}}$

- ▷ ideally: a short interval with high coverage

Testing hypotheses

- ∘ we observe data with a distribution depending on a parameter
- ∘ we would like to use the data to answer questions about the parameter
  - ▷ is $\theta > 0$, $\theta < 0$, $\theta = 1$, ...?
- ∘ to do so, we can test hypotheses about the parameter
  - ▷ $H_0 : \theta \geq 0$ vs. $H_1 : \theta < 0$
  - ▷ $H_0 : \theta = 1$ vs. $H_1 : \theta \neq 1$
  - ▷ ...
- ∘ testing has two possible results
  1. we reject H_0 in favour of H_1
     - ▷ we can say we have shown H_1 (at the level α)
  2. we do not reject H_0 in favour of H_1
     - ▷ we can say we have not shown H_1 (at the level α)
     - ▷ !!!we cannot say we have shown H_0!!!
- ∘ the roles of H_0 and H_1 are not symmetric

- ∘ testing has two possible results
  1. we reject H_0 in favour of H_1
  2. we do not reject H_0 in favour of H_1
- ∘ we can reach a wrong conclusion in two ways
  1. when H_0 is true and we reject H_0 in favour of H_1
     - ▷ "$\mathsf{P}_{H_0}(\text{reject } H_0) = \alpha$"
     - ▷ α: type I. error, level of the test
  2. when H_1 is true and we do not reject H_0 in favour of H_1
     - ▷ "$1 - \mathsf{P}_{H_1}(\text{reject } H_0) = \beta$"
     - ▷ β: type II. error
     - ▷ $1 - \beta$: power of the test
- ∘ often impossible to keep both errors low at the same time

- when choosing a test, we keep the level α fixed and try to maximize the power β
- the roles of H_0 and H_1 are not symmetric

- the roles of H_0 and H_1 are not symmetric
- it is important to choose a good H_0, H_1 pair
- testing
  - $H_0 : \theta \geq 0$ vs. $H_1 : \theta < 0$
  - $H_0 : \theta \leq 0$ vs. $H_1 : \theta > 0$
  - $H_0 : \theta = 0$ vs. $H_1 : \theta \neq 0$

  answer different questions

1.3 Data analysis in practice

Example: `fev` data

- from: `http://www.statsci.org/data/general/fev.html`
- question: association between the FEV[l] and Smoking,

  corrected for Age[years], Height[cm] and Gender
- data:

FEV	Age	Height	Gender	Smoking
1.708	9	144.8	Female	Non
1.724	8	171.5	Female	Non
1.720	7	138.4	Female	Non
1.558	9	134.6	Male	Non
...	...	...	...	...
3.727	15	172.7	Male	Current
2.853	18	152.4	Female	Non
2.795	16	160.0	Female	Current
3.211	15	168.9	Female	Non

Getting the data to `R`

- data `fev.txt`
- for `*.txt` files:
  - `read.table(...)`
- for `*.csv` files (from Excel)

  - ▷ `read.csv(...)`

- ∘ read the data and look at them

```
> fev <- read.table("fev.txt", header=TRUE)
>
> class(fev)
[1] "data.frame"
> dim(fev)
[1] 654   6
> names(fev)
[1] "ID"     "Age"    "FEV"    "Height" "Sex"    "Smoker"
>
> fev[1:3, ]
   ID Age   FEV Height    Sex Smoker
1 301   9 1.708   57.0 Female    Non
2 451   8 1.724   67.5 Female    Non
3 501   7 1.720   54.5 Female    Non
>
> fev <- fev[, -1]
```

Before we fit a model to data

- ∘ before we do the analysis, we need to
  - ▷ get to know the variables
  - ▷ get to understand the relationships among the variables
  - ▷ identify possible problems for the analysis
  - ▷ possibly spot obvious mistakes in data

⇒ first step in applied data analysis: descriptive statistics

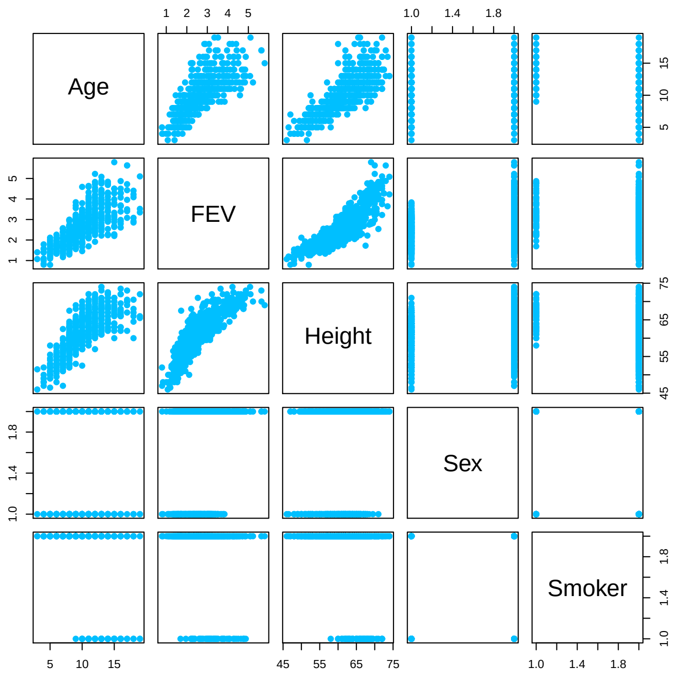
  - ▷ informal data descriptions (no model, no inference)
    - ∗ numerical and graphical
    - ∗ their choice depends on the type of variable(s) of interest

First look at the variables

```
> summary(fev)
      Age              FEV            Height           Sex
 Min.   : 3.000   Min.   :0.791   Min.   :46.00   Female:318
 1st Qu.: 8.000   1st Qu.:1.981   1st Qu.:57.00   Male  :336
 Median :10.000   Median :2.547   Median :61.50
 Mean   : 9.931   Mean   :2.637   Mean   :61.14
 3rd Qu.:12.000   3rd Qu.:3.119   3rd Qu.:65.50
 Max.   :19.000   Max.   :5.793   Max.   :74.00
     Smoker
 Current: 65
 Non    :589
```

First look at the relationships between the variables

```
> pairs(fev, col="deepskyblue", pch=19)
```



1.4 Descriptive statistics

1.4.1 Types of variables

Types of variables

1. in mathematical statistics:
   - continuous (uncountably many possible values)
   - discrete (at most countably many possible values)
2. in applied statistics:
   - quantitative
   - categorical
     - ▷ nominal
     - ▷ ordinal
3. in R:

- numeric
- - ▷ factor
  - ▷ ordered factor

Quantitative variable

- distribution
- characteristics of location
  - ▷ mean
  - ▷ maximum, minimum
  - ▷ quantiles, in particular quartiles and median

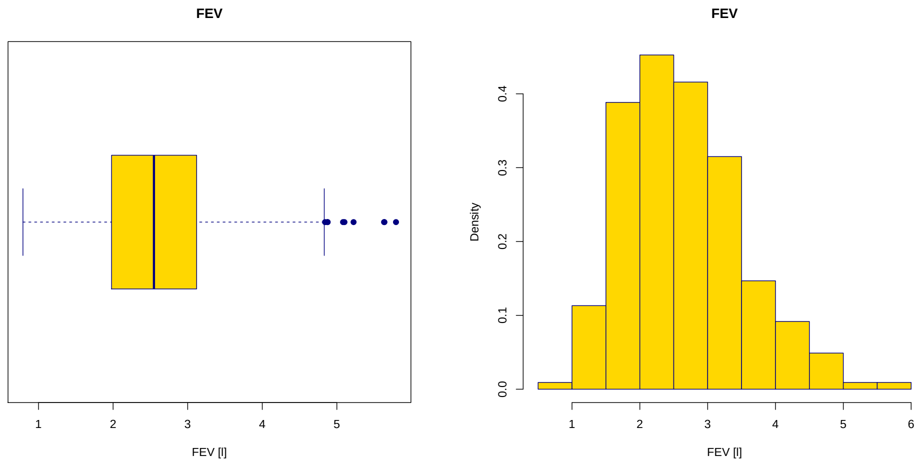
```
> summary(fev$FEV)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.791   1.981   2.548   2.637   3.118   5.793
```

- characteristics of dispersion
  - ▷ standard deviation
  - ▷ interquartile range

```
> sd(fev$FEV)
[1] 0.8670591
> IQR(fev$FEV)
[1] 1.1375
```

Graphics for quantitative variable

```
> hist(fev$FEV, , freq=FALSE,
+ main="FEV", xlab="FEV [l]",
+ col="gold", border="navyblue")
>
> boxplot(fev$FEV, horizontal=TRUE,
+ main="FEV",  xlab="FEV [l]",
+ col="gold", border="navyblue", pch=19)
```



Categorical variable

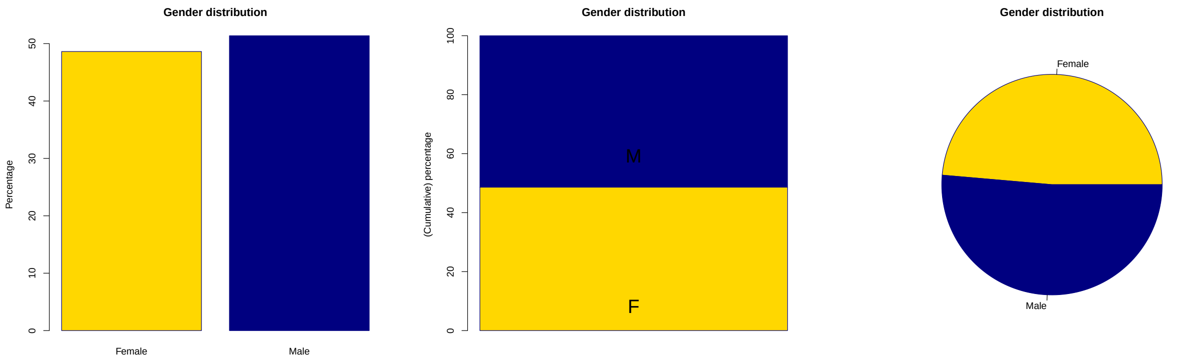
- distribution
  - counts of observations per category
  - percentage of observations per category
  - cumulative percentage of observations per category (for ordinal variables)
- characteristics
  - modus

```
> summary(fev$Sex)
Female   Male
   318    336
> prop.table(table(fev$Sex))
   Female      Male
0.4862385 0.5137615
> cumsum(prop.table(table(fev$Sex)))
   Female      Male
0.4862385 1.0000000
> # not so interesting for a nominal variable
```

Graphics for categorical variable

```
> barplot(100*prop.table(table(fev$Sex)),
+         main="Gender distribution", ylab="Percentage",
+         col=c("gold", "navyblue"), border="navyblue")
>
> barplot(100*matrix(prop.table(table(fev$Sex)), nrow=2, ncol=1),
+         main="Gender distribution", ylab="(Cumulative) percentage",
+         col=c("gold", "navyblue"), border="navyblue")
```

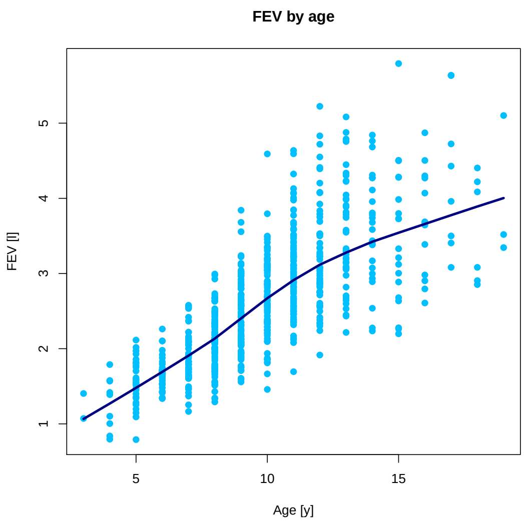
```
> text(x=0.7, y=4, labels="F",
+      cex=2, pos=3)
> text(x=0.7, y=55, labels="M",
+      cex=2, pos=3)
>
> pie(summary(fev$Sex),
+     main="Gender distribution",
+     col=c("gold", "navyblue"), border="navyblue")
```



1.4.2 Relationships between variables

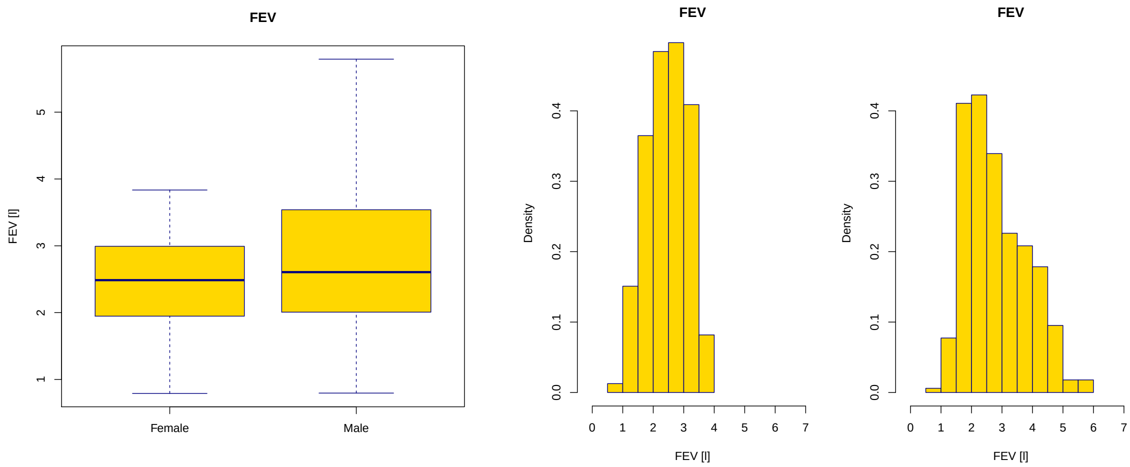
Quantitative vs quantitative

```
> plot(fev$FEV~fev$Age,
+      main="FEV by age",
+      ylab="FEV [l]", xlab="Age [y]",
+      pch=19, col="deepskyblue")
> lines(lowess(fev$FEV~fev$Age),
+       lwd=3, col="navyblue")
```



Quantitative vs categorical

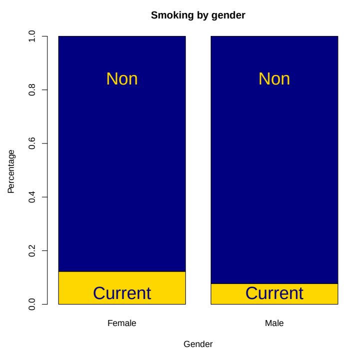
```
> boxplot(fev$FEV~fev$Sex,
+         main="FEV",  ylab="FEV [l]",
+         col="gold", border="navyblue", pch=19)
>
> par(mfrow=c(1, 2))
> hist(fev$FEV[fev$Sex=="Female"], freq=FALSE,
+      xlim=c(min(fev$FEV)-1, max(fev$FEV)+1), ylim=c(0, 0.48),
+      main="FEV", xlab="FEV [l]",
+      col="gold", border="navyblue")
> hist(fev$FEV[fev$Sex=="Male"], freq=FALSE,
+      xlim=c(min(fev$FEV)-1, max(fev$FEV)+1), ylim=c(0, 0.48),
+      main="FEV", xlab="FEV [l]",
+      col="gold", border="navyblue")
```



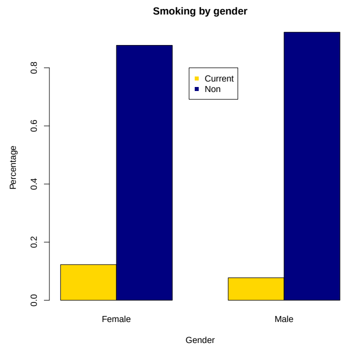
Categorical vs categorical

```
> table(fev$Smoker, fev$Sex)
         Female Male
  Current     39   26
  Non        279  310
>
> prop.table(table(fev$Smoker, fev$Sex),
+            margin=1
+            )
            Female      Male
  Current 0.6000000 0.4000000
  Non     0.4736842 0.5263158
>
> prop.table(table(fev$Smoker, fev$Sex),
+            margin=2
+           )
             Female       Male
  Current 0.12264151 0.07738095
  Non     0.87735849 0.92261905
```

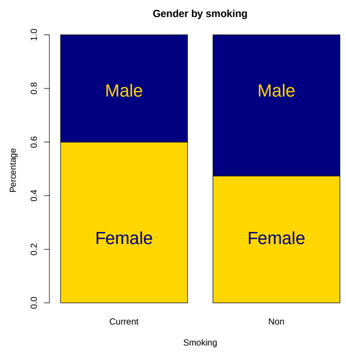
```
> prop.table(table(fev$Smoker, fev$Sex),
+            margin=2
+           )
              Female       Male
  Current 0.12264151 0.07738095
  Non     0.87735849 0.92261905
>
> barplot(height=prop.table(table(fev$Smoker, fev$Sex),
+                    margin=2), beside=F,
+         ylab="Percentage", xlab="Gender", main="Smoking by gender",
+         col=c("gold", "navyblue")
+         )
> text(x=1.2*c(0:2)+0.7, y=0, labels="Current",
+      col="navyblue", cex=2, pos=3)
> text(x=1.2*c(0:2)+0.7, y=0.8, labels="Non",
+      col="gold", cex=2, pos=3)
```




```
> prop.table(table(fev$Smoker, fev$Sex),
+            margin=2
+           )
              Female       Male
  Current 0.12264151 0.07738095
  Non     0.87735849 0.92261905
>
> barplot(height=prop.table(table(fev$Smoker, fev$Sex),
+                           margin=2), beside=T,
+         ylab="Percentage", xlab="Gender", main="Smoking by gender",
+         col=c("gold", "navyblue")
+ )
>
> legend(x=3.3, y=0.8, legend=c("Current", "Non"),
+        col=c("gold", "navyblue"), pch=15)
```

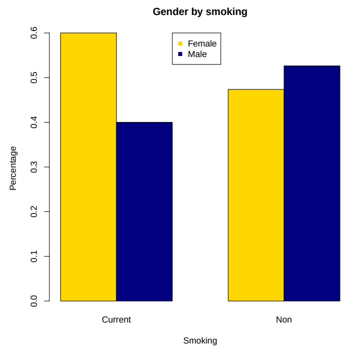



```
> prop.table(table(fev$Sex, fev$Smoker),
+           margin=2
+ )
          Current       Non
  Female 0.6000000 0.4736842
  Male   0.4000000 0.5263158
>
> barplot(height=prop.table(table(fev$Sex, fev$Smoker),
+                     margin=2), beside=F,
+         ylab="Percentage", xlab="Smoking", main="Gender by smoking",
+         col=c("gold", "navyblue")
+ )
> text(x=1.2*c(0:2)+0.7, y=0.2, labels="Female", col="navyblue",
+      cex=2, pos=3)
> text(x=1.2*c(0:2)+0.7, y=0.75, labels="Male", col="gold",
+      cex=2, pos=3)
```




```
> prop.table(table(fev$Sex, fev$Smoker),
+           margin=2
+ )
          Current       Non
  Female 0.6000000 0.4736842
  Male   0.4000000 0.5263158
>
> barplot(height=prop.table(table(fev$Sex, fev$Smoker),
```

```
+                      margin=2), beside=T,
+         ylab="Percentage", xlab="Smoking", main="Gender by smoking",
+         col=c("gold", "navyblue")
+         )
>
> legend(x=3, y=0.6, legend=c("Female", "Male"),
+        col=c("gold", "navyblue"), pch=15)
```



Chapter 2

Linear algebra essentials

2.1 The problem

2.1.1 Linear model

Chocolatey example

- Does consuming [amount of] chocolate decrease blood pressure [type, measurement]?
- collect blood pressures $x_1, \dots, x_n$
- suppose that $X_i \sim \mathrm{N}(\mu_i, \sigma^2)$
  - ▷ $\mu_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_k x_{i,k}$
  - ▷ $x_{i,1} = \begin{cases} 1 & \text{if the person eats [amount of] chocolate} \\ 0 & \text{otherwise} \end{cases}$
  - ▷ $x_{i,j}$, $j \in \{2, \dots, k\}$: age, gender, BMI, ...
- test $H_0 : \beta_1 \geq 0$ versus $H_1 : \beta_1 < 0$ to answer the question

Linear model

- $Y_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_k x_{i,k} + \varepsilon_i$, $i \in \{1, \dots, n\}$
  - ▷ Y_i: outcome, response, output, dependent variable
    - ∗ random variable, we observe a realization y_i
    - ∗ (odezva, závisle proměnná, regresand)
  - ▷ $x_{i,1}, \dots, x_{i,k}$: covariates, predictors, explanatory variables,
    input, independent variables
    - ∗ given, known

        * (nezávisle proměnné, regresory)
    - $\beta_0, \dots, \beta_k$: coefficients
        * unknown
        * (regresní koeficienty)
    - ε_i: random error
        * random variable, unobserved

- $\varepsilon_i \overset{\text{iid}}{\sim} (0, \sigma^2)$, $i \in \{1, \dots, n\}$
    - $\mathsf{E}\, \varepsilon_i = 0$: no systematic errors
    - $\mathsf{Var}\, \varepsilon_i = \sigma^2$: same precision
- we often assume that $\varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$, $i \in \{1, \dots, n\}$

Linear model in the matrix form

- $Y_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_k x_{i,k} + \varepsilon_i$, $i \in \{1, \dots, n\}$
- let

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,k} \\ 1 & x_{2,1} & \dots & x_{2,k} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n,1} & \dots & x_{n,k} \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_k \end{pmatrix}, \ \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}$$

- then $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 \mathbf{I})$ and often $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I})$
    - $\mathbf{X}$: design matrix
        * (regresní matice, matice plánu)
- let $p = k + 1$
- then $\underbrace{\mathbf{Y}}_{n \times 1} = \underbrace{\mathbf{X}}_{n \times p} \underbrace{\boldsymbol{\beta}}_{p \times 1} + \underbrace{\boldsymbol{\varepsilon}}_{n \times 1}$
- we assume that $n > p$ (and often think about $n \to \infty$, p fixed)

Example: `bloodpress` data

- from `sites.stat.psu.edu/~lsimon/stat501wc/sp05/data/`
- association between the mean arterial blood pressure[mmHg] and age[years], weight[kg], body surface area[m^2], duration of hypertension[years], basal pulse[beats/min], stress

- data:

BP	Age	Weight	BSA	DoH	Pulse	Stress
105	47	85.4	1.75	5.1	63	33
115	49	94.2	2.10	3.8	70	14
...	...	...	...	...	...	...
110	48	90.5	1.88	9.0	71	99
122	56	95.7	2.09	7.0	75	99

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

$$\begin{pmatrix} 105 \\ 115 \\ \dots \\ 110 \\ 122 \end{pmatrix} = \begin{pmatrix} 1 & 47 & 85.4 & 1.75 & 5.1 & 63 & 33 \\ 1 & 49 & 94.2 & 2.10 & 3.8 & 70 & 14 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 48 & 90.5 & 1.88 & 9.0 & 71 & 99 \\ 1 & 56 & 95.7 & 2.09 & 7.0 & 75 & 99 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \dots \\ \beta_6 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_{19} \\ \varepsilon_{20} \end{pmatrix}$$

2.1.2 Task for this chapter

Design matrix

- model:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,k} \\ 1 & x_{2,1} & \dots & x_{2,k} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n,1} & \dots & x_{n,k} \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \dots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}$$

- design matrix:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,k} \\ 1 & x_{2,1} & \dots & x_{2,k} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n,1} & \dots & x_{n,k} \end{pmatrix} = (\mathbf{1} \mid \mathbf{x}_{,1} \mid \mathbf{x}_{,2} \mid \dots \mid \mathbf{x}_{,k}) = \begin{pmatrix} \mathbf{x}_{1,} \\ \mathbf{x}_{2,} \\ \dots \\ \mathbf{x}_{n,} \end{pmatrix}$$

  - ▷ k covariates and $\mathbf{1}$ are the p columns of $\mathbf{X}$
  - ▷ n observations are the n rows of $\mathbf{X}$

Matrix algebra in a linear model

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
- coefficient vector $\boldsymbol{\beta}$
  - ▷ fixed but unknown
  - ▷ $p \times 1$ matrix
  - ▷ $\boldsymbol{\beta}^\top$ defines a mapping $\boldsymbol{\beta}^\top : \mathbb{R}^p \mapsto \mathbb{R}$

$$\mathbf{x}_{i,} \in \mathbb{R}^p \rightsquigarrow \mathsf{E}\,\mathbf{Y}_i \in \mathbb{R}$$

- design matrix $\mathbf{X}$
  - ▷ fixed and known
  - ▷ $n \times p$ matrix
  - ▷ defines a mapping $\mathbf{X} : \mathbb{R}^p \mapsto \mathbb{R}^n$

$$\boldsymbol{\beta} \in \mathbb{R}^p \rightsquigarrow \mathsf{E}\,\mathbf{Y} \in \mathbb{R}^n$$

  - ▷ idea: when estimating $\boldsymbol{\beta}$, how about choosing $\widehat{\boldsymbol{\beta}}$ so that $\mathbf{X}$ maps $\hat{\boldsymbol{\beta}}$ as close to $\mathbf{Y}$ as possible?

2.2 Linear mapping

Linear mapping from $\mathbb{R}^p$ to $\mathbb{R}^n$

- function $f : \mathbb{R}^p \mapsto \mathbb{R}^n$ such that
  - ▷ $f(\mathbf{x}+\mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y})$ ... additivity
  - ▷ $f(\alpha\mathbf{x}) = \alpha f(\mathbf{x})$ ... homogeneity
- described by an $n \times p$ matrix $\mathbf{A}$: $f(\mathbf{x}) = \mathbf{A}\mathbf{x}$

↪ idea:

  - ▷ $\forall\ \mathbf{x} \in \mathbb{R}^p$ can be written as $\mathbf{x} = \sum_{i=1}^p c_i \mathbf{v}_i$,

    where $\mathcal{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_p\}$ is a basis of $\mathbb{R}^p$

  ↪ $f(\mathbf{x})$ is determined by $\{f(\mathbf{v}_1), \dots, f(\mathbf{v}_p)\}$

    because $f(\mathbf{x}) = f(\sum_{i=1}^p c_i \mathbf{v}_i) = \sum_{i=1}^p c_i f(\mathbf{v}_i)$

  - ▷ $\forall\ \mathbf{y} \in \mathbb{R}^n$ can be written as $\mathbf{y} = \sum_{i=1}^n c_i \mathbf{w}_i$,

    where $\mathcal{W} = \{\mathbf{w}_1, \dots, \mathbf{w}_n\}$ is a basis of $\mathbb{R}^n$

  ↪ just need to write each $f(\mathbf{v}_i)$ in terms of $\mathcal{W}$

  - ▷ free choice of $(\mathcal{W}, \mathcal{V}) \to$ various $\mathbf{A}$'s representing the same f

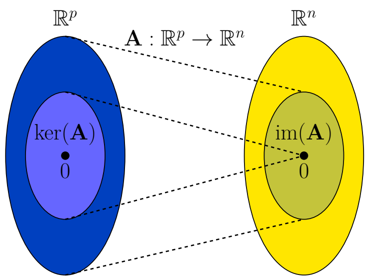
- operations $f_1 \circ f_2$, $f_1 + f_2$, αf
  - ▷ result into linear mappings

    represented by $\mathbf{A}_1\mathbf{A}_2$, $\mathbf{A}_1 + \mathbf{A}_2$, $\alpha\mathbf{A}$

2.2.1 Associated subspaces

Kernel and image

- kernel (nullspace)
  - ▷ $\ker(\mathbf{A}) = \ker(f) = \{\mathbf{x} \in \mathbb{R}^p;\ \mathbf{A}\mathbf{x} = \mathbf{0}\}$
    $= \{\mathbf{x} \in \mathbb{R}^p;\ f(\mathbf{x}) = \mathbf{0}\}$
  - ▷ subspace of $\mathbb{R}^p$, $\dim(\ker(\mathbf{A}))$: nullity of $\mathbf{A}$
- image (range, column space)
  - ▷ $\mathsf{im}(\mathbf{A}) = \mathsf{im}(f) = \{\mathbf{y} \in \mathbb{R}^n;\ \exists \mathbf{x} \in \mathbb{R}^p : \mathbf{A}\mathbf{x} = \mathbf{y}\}$
    $= \{\mathbf{y} \in \mathbb{R}^n;\ \exists \mathbf{x} \in \mathbb{R}^p : f(\mathbf{x}) = \mathbf{y}\}$
  - ▷ subspace of $\mathbb{R}^n$, $\dim(\mathsf{im}(\mathbf{A}))$: rank of $\mathbf{A}$
- schematically



- rank nullity theorem: $\dim(\ker(\mathbf{A})) + \dim(\mathsf{im}(\mathbf{A})) = p$

Four fundamental subspaces associated to A

- kernel and image of $\mathbf{A}$
  - ▷ column space of $\mathbf{A}$: $\mathsf{im}(\mathbf{A}) = \{\mathbf{y} \in \mathbb{R}^n;\ \exists \mathbf{x} \in \mathbb{R}^p : \mathbf{A}\mathbf{x} = \mathbf{y}\}$
    - ∗ $\dim(\mathsf{im}(\mathbf{A})) = \mathsf{rank}(\mathbf{A})$
  - ▷ kernel of $\mathbf{A}$: $\ker(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^p;\ \mathbf{A}\mathbf{x} = \mathbf{0}\}$
    - ∗ $\dim(\ker(\mathbf{A})) = p - \mathsf{rank}(\mathbf{A})$
- kernel and image of $\mathbf{A}^\top$
  - ▷ column space of $\mathbf{A}^\top$: $\mathsf{im}(\mathbf{A}^\top)$: coimage of $\mathbf{A}$
    - ∗ $\dim(\mathsf{im}(\mathbf{A}^\top)) = \mathsf{rank}(\mathbf{A}^\top) = \mathsf{rank}(\mathbf{A})$
    - ∗ row space of $\mathbf{A}$
  - ▷ kernel of $\mathbf{A}^\top$: $\ker(\mathbf{A}^\top)$: cokernel, left nullspace of $\mathbf{A}$
    - ∗ $\dim(\ker(\mathbf{A}^\top)) = n - \mathsf{rank}(\mathbf{A})$: corank of $\mathbf{A}$

2.2.2 Orthogonality

Inner product on $\mathbb{R}^n$

- dot product:
  - ▷ $< \mathbf{x}, \mathbf{y} >= \mathbf{y}^\top \mathbf{x} = \sum_{i=1}^n x_i y_i$
- associated norm (length of $\mathbf{x}$):
  - ▷ $||\mathbf{x}|| = \sqrt{< \mathbf{x}, \mathbf{y} >} = \sqrt{\sum_{i=1}^n x_i^2}$
- angle θ between $\mathbf{x}$ and $\mathbf{y}$:
  - ▷ $\cos\theta = \frac{<\mathbf{x},\mathbf{y}>}{||\mathbf{x}||\,||\mathbf{y}||}$
- orthogonality for $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{y} \neq \mathbf{0}$:
  - ▷ $< \mathbf{x}, \mathbf{y} >= 0$
- orthogonal complement $W^\perp$ of a subspace W of $\mathbb{R}^n$:
  - ▷ $W^\perp = \{\mathbf{y} \in \mathbb{R}^n ; < \mathbf{x}, \mathbf{y} >= 0 \text{ for every } \mathbf{x} \in W\}$
    - ∗ $W^\perp$ is a subspace of $\mathbb{R}^n$
    - ∗ $W^\perp \cap W = \{\mathbf{0}\}$
    - ∗ $\dim(W) + \dim(W^\perp) = n$
- orthogonality between fundamental subspaces associated to $\mathbf{A}$
  - ▷ $\ker(\mathbf{A}) = (\mathsf{im}(\mathbf{A}^\top))^\perp$ (in $\mathbb{R}^p$)
  - ▷ $\ker(\mathbf{A}^\top) = (\mathsf{im}(\mathbf{A}))^\perp$ (in $\mathbb{R}^n$)

Orthogonal columns

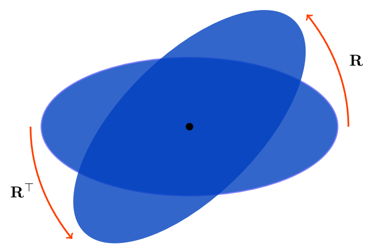
- matrix with orthogonal columns:
  - ▷ $\mathbf{U} = (\mathbf{u}_{,1} \,|\, \mathbf{u}_{,2} \,|\, \ldots \,|\, \mathbf{u}_{,p})$
    $< \mathbf{u}_{,i}, \mathbf{u}_{,j} >= 0$ for $i \neq j$
- matrix with orthonormal columns:
  - ▷ $\mathbf{U} = (\mathbf{u}_{,1} \,|\, \mathbf{u}_{,2} \,|\, \ldots \,|\, \mathbf{u}_{,p})$
    $< \mathbf{u}_{,i}, \mathbf{u}_{,j} >= 0$ for $i \neq j$
    $||\mathbf{u}_{,i}|| = 1$ for $i \in \{1, \ldots, p\}$
  - ▷ $\mathbf{U}^\top \mathbf{U} = \mathbf{I} \Rightarrow$ mapping $\mathbf{U} : \mathbf{x} \mapsto \mathbf{U}\mathbf{x}$ preserves

- ∗ inner product
- ∗ norm
- ∗ angles
- ∗ distances

Orthogonal matrix

- ∘ square matrix $\mathbf{R}$ with orthonormal columns (and rows)
  - ▷ $\mathbf{R}^\top \mathbf{R} = \mathbf{R}\mathbf{R}^\top = \mathbf{I}$
    i.e. $\mathbf{R}^\top = \mathbf{R}^{-1}$
- ∘ $\mathbf{R}^{-1} = \mathbf{R}^\top$ is also an orthogonal matrix
- ∘ product of orthogonal matrices is also an orthogonal matrix
- ∘ geometrically
  - ▷ change of orthonormal basis (coordinate transformation)
  - ▷ mapping $\mathbf{R}$: $\mathbf{x} \mapsto \mathbf{R}\mathbf{x}$ … rotation
    - ∗ preserves the origin
    - ∗ preserves angles
    - ∗ preserves distances
    - ∗ proper rotation

      if $\det \mathbf{R} = 1$

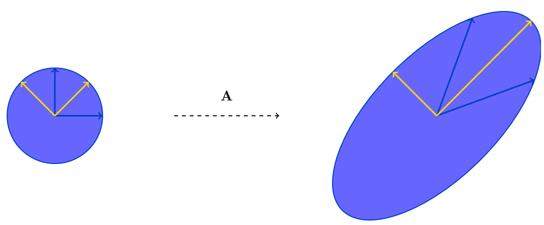


2.3 Matrix decompositions

2.3.1 Eigen-decomposition

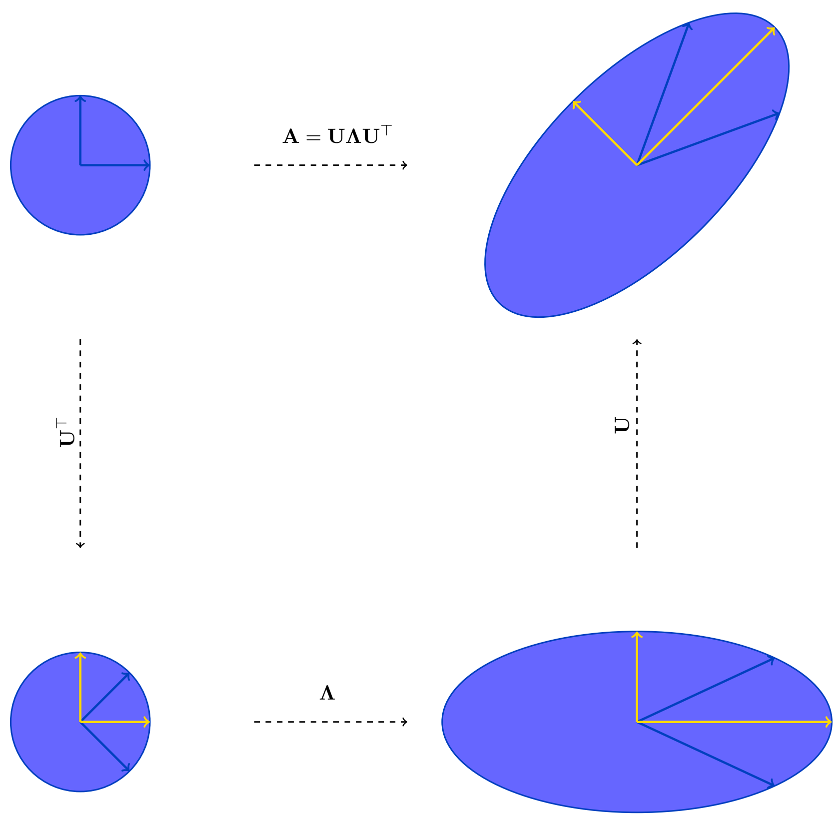
Spectral decomposition (eigen-decomposition)

- ∘ let $\mathbf{A}$ be a symmetric $p \times p$ matrix
  - ▷ eigenvalues $\lambda_1, \ldots, \lambda_p$
  - ▷ eigenvectors $\mathbf{u}_{\cdot,1}, \ldots, \mathbf{u}_{\cdot,p}$
  - ▷ everything real
  - ▷ $\mathbf{A}\mathbf{u}_{\cdot,i} = \lambda_i \mathbf{u}_{\cdot,i}$
  - ▷ f elongates/shrinks $\mathbf{u}_{\cdot,i}$ by λ_i



- eigen-decomposition: $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, where
  - ▷ $\mathbf{U} = (\mathbf{u}_{\cdot,1} \,|\, \mathbf{u}_{\cdot,2} \,|\, \ldots \,|\, \mathbf{u}_{\cdot,p})$
    - ∗ $\mathbf{U}$ is $p \times p$ orthogonal matrix
  - ▷ $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_p\}$
    - ∗ $\mathbf{\Lambda}$ is $p \times p$ diagonal matrix
  - ▷ convention
    - ∗ λ_i is the i^{th} largest eigenvalue of $\mathbf{A}$
    - ∗ $\mathbf{u}_{\cdot,i}$ is the eigenvector corresponding to λ_i

Geometry for $\mathbf{A} \succ 0$

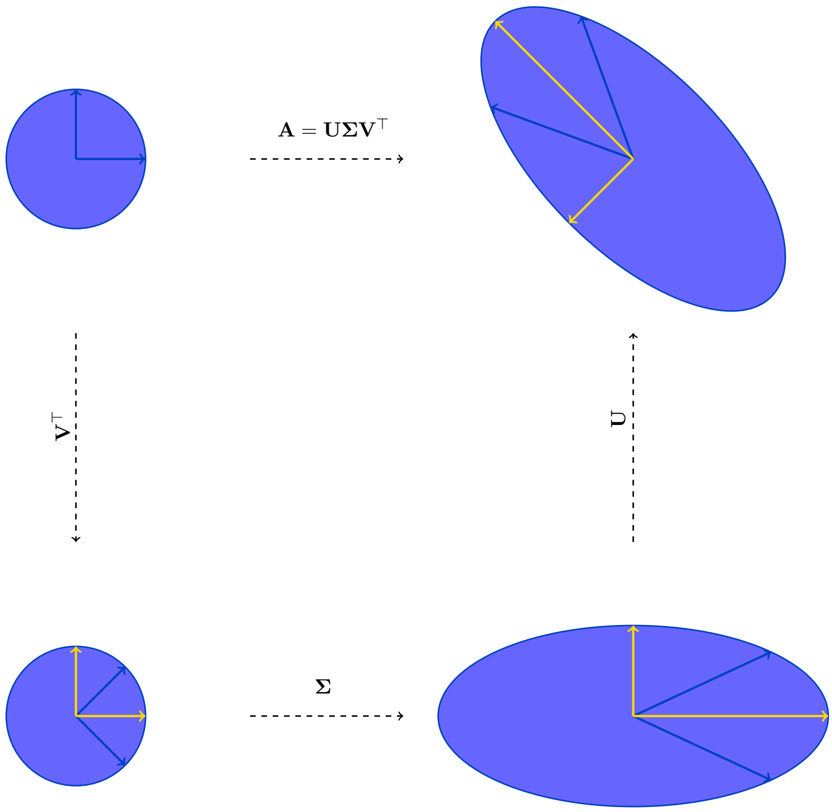


2.3.2 Singular value decomposition

Singular value decomposition (SVD)

- let $\mathbf{A}$ be an $n \times p$ $(n \geq p)$ rectangular matrix
  - ▷ singular values $\sigma_1, \dots, \sigma_p$
  - ▷ left singular vectors $\mathbf{u}_{\cdot,1}, \dots, \mathbf{u}_{\cdot,p}$
  - ▷ right singular vectors $\mathbf{v}_{\cdot,1}, \dots, \mathbf{v}_{\cdot,p}$
  - ▷ $\mathbf{A}\mathbf{v}_{\cdot,i} = \sigma_i \mathbf{u}_{\cdot,i}$ & $\mathbf{A}^\top \mathbf{u}_{\cdot,i} = \sigma_i \mathbf{v}_{\cdot,i}$
  - ▷ everything real, singular values non-negative
- SVD: $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where
  - ▷ $\mathbf{U}$ is $(n \times n)$ orthogonal
    - ∗ first p columns of $\mathbf{U}$: $\mathbf{u}_{\cdot,1}, \dots, \mathbf{u}_{\cdot,p}$
  - ▷ $\mathbf{\Sigma}$ $(n \times p)$ diagonal
    - ∗ $\sigma_1, \dots, \sigma_p$ on the diagonal of $\mathbf{\Sigma}$
  - ▷ $\mathbf{V}$ is $(p \times p)$ orthogonal
    - ∗ columns of $\mathbf{V}$: $\mathbf{v}_{\cdot,1}, \dots, \mathbf{v}_{\cdot,p}$
  - ▷ convention : singular values in the descending order

Geometry for a square A



SVD and spectral decomposition

- SVD: rectangular matrix $\mathbf{A}$ ($n \times p$, $n \geq p$)
  - singular values and vectors $(\sigma_1, \mathbf{u}_{.,1}, \mathbf{v}_{.,1}), \ldots, (\sigma_p, \mathbf{u}_{.,p}, \mathbf{v}_{.,p})$
  - $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$, where
    - $\mathbf{U}$ is $(n \times n)$ and $\mathbf{V}$ is $(p \times p)$, both orthogonal
    - $\boldsymbol{\Sigma}$ $(n \times p)$ diagonal with non-negative diagonal
- Spec. dec.: square symmetric matrix $\mathbf{A}$ $(p \times p)$
  - eigenvalues and eigenvectors $(\lambda_1, \mathbf{u}_{.,1}), \ldots, (\lambda_p, \mathbf{u}_{.,p})$

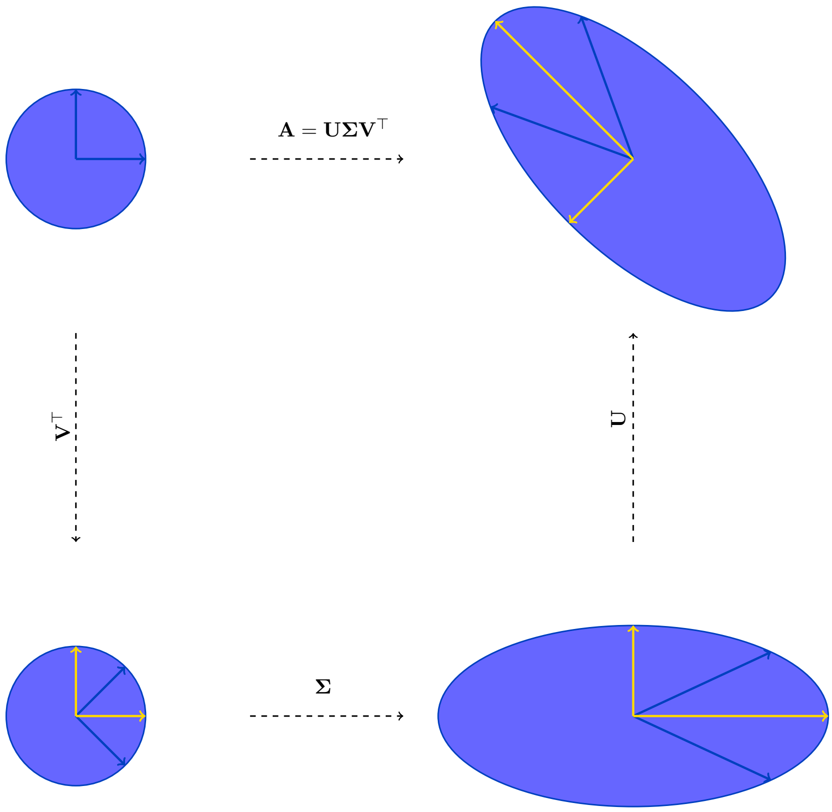
- ▷ $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, where
  - ∗ $\mathbf{U}$ is $(p \times p)$ orthogonal
  - ∗ $\mathbf{\Lambda}$ $(p \times p)$ diagonal
- ∘ for a square symmetric $\mathbf{A}$, $\mathbf{A} \succeq 0$: $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$
- ∘ for a rectangular matrix $\mathbf{A}$ $(n \times p,\ n \geq p)$, $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$
  - ▷ $\mathbf{A}^\top\mathbf{A} = \mathbf{V}\mathbf{\Sigma}^\top\mathbf{\Sigma}\mathbf{V}^\top$ $(p \times p) \Rightarrow \mathbf{v}_{.,i}$'s are eigenvectors of $\mathbf{A}^\top\mathbf{A}$
  - ▷ $\mathbf{A}\mathbf{A}^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^\top\mathbf{U}^\top$ $(n \times n) \Rightarrow \mathbf{u}_{.,i}$'s are eigenvectors of $\mathbf{A}\mathbf{A}^\top$
  - ▷ σ_i's are square roots of non-zero λ_i's of $\mathbf{A}^\top\mathbf{A}$ and $\mathbf{A}\mathbf{A}^\top$

Reduced SVD's

- ∘ SVD: rectangular matrix $\mathbf{A}$ $(n \times p,\ n \geq p)$
  - ▷ singular values and vectors $(\sigma_1, \mathbf{u}_{.,1}, \mathbf{v}_{.,1}), \ldots, (\sigma_p, \mathbf{u}_{.,p}, \mathbf{v}_{.,p})$
  - ▷ $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where
    - ∗ $\mathbf{U}$ is $(n \times n)$ and $\mathbf{V}$ is $(p \times p)$, both orthogonal
    - ∗ $\mathbf{\Sigma}$ $(n \times p)$ diagonal with non-negative diagonal
- ∘ if $n > p$
  - ▷ $\mathbf{A} = \left(\mathbf{U}_1 \,|\, \mathbf{U}_2\right) \begin{pmatrix} \mathbf{\Sigma}_1 \\ \hline \mathbf{0} \end{pmatrix} \mathbf{V}^\top$
  - ↪ $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}^\top$, where
    - ∗ $\mathbf{U}_1$ $(n \times p)$ with orthogonal columns
    - ∗ $\mathbf{\Sigma}_1$ is $(p \times p)$
    - ∗ thin SVD: $\mathbf{A} = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}^\top$
  - ▷ if $r = \mathsf{rank}(\mathbf{A}) < p$, $\mathbf{\Sigma}_1$ is $(r \times r)$ ... compact SVD
- ∘ SVD writes $\mathbf{A}$ as a sum of multiples of rank-one matrices:

  $\mathbf{A} = \sum_{i=1}^{p} \sigma_i \mathbf{u}_{.,i}\mathbf{v}_{.,i}^\top = \sum_{i=1}^{r} \sigma_i \mathbf{u}_{.,i}\mathbf{v}_{.,i}^\top$

Geometry



SVD and linear mapping

- SVD: rectangular matrix $\mathbf{A}$ $(n \times p,\ n \geq p)$
  - ▷ singular values and vectors $(\sigma_1, \mathbf{u}_{\cdot,1}, \mathbf{v}_{\cdot,1}), \ldots, (\sigma_p, \mathbf{u}_{\cdot,p}, \mathbf{v}_{\cdot,p})$
  - ▷ $\mathbf{A} = \mathbf{U\Sigma V}^\top$, where
    - ∗ $\mathbf{U}$ and $\mathbf{V}$: orthonormal bases of $\mathbb{R}^n$ and $\mathbb{R}^p$ such that $\mathbf{A}$ maps the i^{th} basis vector of $\mathbb{R}^p$ to a non-negative multiple of the i^{th} basis vector of $\mathbb{R}^n$, and sends the left-over basis vectors to zero
- $\ker(\mathbf{A})$
  - ▷ spanned by the $\mathbf{v}_{\cdot,i}$ corresponding to the null σ_i

- $\mathsf{im}(\mathbf{A})$
  - ▷ spanned by the $\mathbf{u}_{\cdot,i}$ corresponding to the positive σ_i
- $\dim(\ker(\mathbf{A})) + \dim(\mathsf{im}(\mathbf{A})) = p$

2.3.3 QR decomposition

QR decomposition (factorization)

- let $\mathbf{A}$ be a $p \times p$ matrix $\rightsquigarrow$ $\mathbf{A} = \mathbf{QR}$, where
  - ▷ $\mathbf{Q}$ is $p \times p$ orthogonal
  - ▷ $\mathbf{R}$ is $p \times p$ upper triangular
- let $\mathbf{A}$ be an $n \times p$ $(n \geq p)$ matrix $\rightsquigarrow$ $\mathbf{A} = \mathbf{QR}$, where
  - ▷ $\mathbf{Q}$ is $n \times n$ orthogonal
  - ▷ $\mathbf{R}$ is $n \times p$ upper triangular
- if $n > p$
  - ▷ $\mathbf{A} = (\mathbf{Q}_1 \,|\, \mathbf{Q}_2) \begin{pmatrix} \mathbf{R}_1 \\ \hline \mathbf{0} \end{pmatrix} = \mathbf{Q}_1 \mathbf{R}_1$, where
    - ∗ $\mathbf{Q}_1$ is $n \times p$ with orthogonal columns
    - ∗ $\mathbf{R}_1$ is $p \times p$ upper triangular
  - ▷ $\mathsf{rank}(\mathbf{A}) = p \Rightarrow \mathsf{rank}(\mathbf{R}_1) = p$

2.4 Pseudoinverse

2.4.1 Moore–Penrose pseudoinverse

Moore–Penrose pseudoinverse

- let $\mathbf{A}$ be a $p \times p$ matrix, $\mathsf{rank}(\mathbf{A}) = p$
- inverse $\mathbf{A}^{-1}$ is the $p \times p$ matrix satisfying
  - ▷ $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$
  - ▷ $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$
- let $\mathbf{A}$ be an $n \times p$ $(n \geq p)$ matrix
- Moore–Penrose pseudoinverse $\mathbf{A}^{+}$ is the $p \times n$ matrix satisfying

  - ▷ $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$ (generalized inverse)
  - ▷ $\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+$ (generalized reflexive inverse)
  - ▷ $(\mathbf{A}\mathbf{A}^+)^\top = \mathbf{A}\mathbf{A}^+$
  - ▷ $(\mathbf{A}^+\mathbf{A})^\top = \mathbf{A}^+\mathbf{A}$

- ∘ $\mathbf{A}^+$ exists and is unique

Construction of $\mathbf{A}^+$

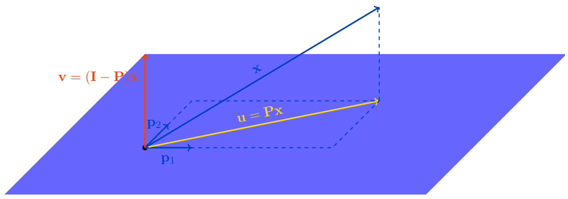
- ∘ let $\mathbf{A}$ be an $n \times p$ ($n \geq p$) matrix
  - ▷ if $\mathsf{rank}(\mathbf{A}) = p$
    - ∗ $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ (thin SVD, i.e. $\mathbf{U}$ is $n \times p$ & $\mathbf{\Sigma}$ is $p \times p$)
    - ∗ $\mathbf{A}^+ = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^\top$
  - ▷ if $\mathsf{rank}(\mathbf{A}) = r < p$
    - ∗ $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ (compact SVD, i.e. $\mathbf{U}$ is $n \times r$ & $\mathbf{\Sigma}$ is $r \times r$)
    - ∗ $\mathbf{A}^+ = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^\top$
- ∘ let $\mathbf{A}$ be a $p \times p$ symmetric matrix
  - ▷ $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ (spectral decomposition)
  - ▷ $\mathbf{A}^+ = \mathbf{U}\mathbf{\Lambda}^+\mathbf{U}^\top$, where
    - ∗ $\mathbf{\Lambda}^+$: diagonal with $1/\lambda_i$ on diagonal if $\lambda_i \neq 0$, 0 otherwise

2.5 Orthogonal projection

Orthogonal projection

Projection on $\mathbb{R}^n$

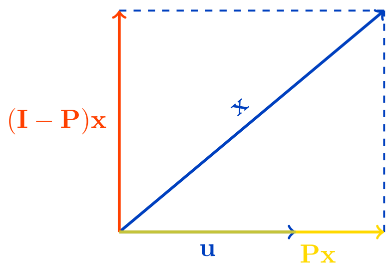
- ∘ projection: linear mapping $\mathbf{P} : \mathbb{R}^n \mapsto \mathbb{R}^n$ such that $\mathbf{P}\mathbf{P} = \mathbf{P}$
  - ▷ $\mathbf{P}$ idempotent
  - ▷ $\mathbf{P}$ is identity on $\mathsf{im}(\mathbf{P})$
- ∘ $\forall\ \mathbf{x} \in \mathbb{R}^n : \mathbf{x} = \underbrace{\mathbf{P}\mathbf{x}}_{\mathbf{u}} + \underbrace{(\mathbf{I} - \mathbf{P})\mathbf{x}}_{\mathbf{v}}$
  - ▷ $\mathbf{u} \in \mathsf{im}(\mathbf{P})$ & $\mathbf{v} \in \ker(\mathbf{P})$: unique decomposition
- ∘ $(\mathbf{I} - \mathbf{P})$ is a projection on $\ker(\mathbf{P})$ and $\ker(\mathbf{I} - \mathbf{P}) = \mathsf{im}(\mathbf{P})$



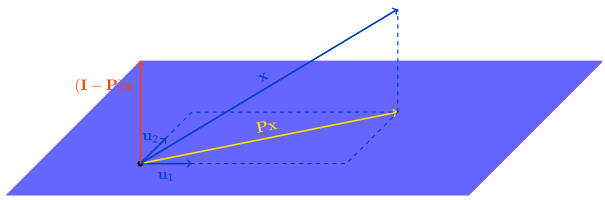
- orthogonal projection: projection with a symmetric $\mathbf{P}$
  - ▷ $\mathbf{P}$ is symmetric iff $\mathsf{im}(\mathbf{P}) = (\,\mathsf{im}(\mathbf{I}-\mathbf{P})\,)^{\perp}$
  - ▷ $\exists!\mathbf{Px} \in \mathsf{im}(\mathbf{P})$ and $||\mathbf{x}-\mathbf{Px}||^2 = \min_{\mathbf{y}\in\mathsf{im}(\mathbf{P})} ||\mathbf{x}-\mathbf{y}||^2$
  - ▷ if $\mathbf{P}$ and $\mathbf{P_1}$ are orthogonal projections and $\mathsf{im}(\mathbf{P}_1) \le \mathsf{im}(\mathbf{P})$
    - ∗ $\mathbf{PP}_1 = \mathbf{P}_1 = \mathbf{P}_1\mathbf{P}$

Construction of P

1. orth. projection on $\mathsf{im}(\mathbf{u})$:
   - $\mathbf{P} = \frac{1}{||\mathbf{u}||^2}\mathbf{uu}^\top$
   - $\mathbf{Px} = \frac{<\mathbf{u},\mathbf{x}>}{||\mathbf{u}||^2}\mathbf{u} \in \mathsf{im}(\mathbf{u})$
     - ▷ leaves $c\mathbf{u}$ unchanged
     - ▷ annihilates the complementary basis



2. $\{\mathbf{u}_1,\dots,\mathbf{u}_p\}$ orthonormal, orth. projection on $\mathsf{im}(\mathbf{U})$, $\mathbf{U} = (\mathbf{u}_1|\mathbf{u}_2|\dots|\mathbf{u}_p)$
   - $\mathbf{P} = \mathbf{UU}^\top$
   - $\mathbf{Px} = \sum_{i=1}^{p} < \mathbf{u}_i, \mathbf{x} > \mathbf{u}_i \in \mathsf{im}(\mathbf{U})$
     - ▷ leaves $\sum_{i=1}^{p} c_i\mathbf{u}_i$ unchanged
     - ▷ annihilates the complementary basis



Orthogonal projection onto a column space

- let $\mathbf{A}$ be an $n \times p$ $(n \ge p)$ matrix

1. $\mathsf{rank}(\mathbf{A}) = p$
   - columns $\mathbf{a}_{.,1},\dots,\mathbf{a}_{.,p}$ linearly independent
   - $\mathbf{P} = \mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top$

- $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ (thin SVD, i.e. $\mathbf{U}$ is $n \times p$ & $\mathbf{\Sigma}$ is $p \times p$)
  - ▷ $(\mathbf{A}^\top\mathbf{A})^{-1} = \mathbf{V}\mathbf{\Sigma}^{-2}\mathbf{V}^\top$
  - ▷ $\mathbf{P} = \mathbf{U}\mathbf{U}^\top$

2. $\mathsf{rank}(\mathbf{A}) = r < p$
   - $\mathbf{P} = \mathbf{A}(\mathbf{A}^\top\mathbf{A})^+\mathbf{A}^\top$
   - $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ (compact SVD, i.e. $\mathbf{U}$ is $n \times r$ & $\mathbf{\Sigma}$ is $r \times r$)
     - ▷ $(\mathbf{A}^\top\mathbf{A})^+ = \mathbf{V}\ \mathbf{\Sigma}^{-2}\mathbf{V}^\top$
     - ▷ $\mathbf{P} = \mathbf{U}\mathbf{U}^\top$

2.6 Application to linear regression

Application to linear regression

Estimation in linear regression (theory)

- linear regression: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$, $\mathsf{Var}\,\boldsymbol{\varepsilon} = \sigma^2\mathbf{I}$
- we want to estimate $\boldsymbol{\beta}$
- start with estimating $\boldsymbol{\mu} = \mathsf{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \in \mathsf{im}(\mathbf{X})$
  - ▷ we look for $\hat{\boldsymbol{\mu}} \in \mathsf{im}(\mathbf{X})$ closest to $\mathbf{Y}$
    - ∗ we look to minimize $||\hat{\boldsymbol{\mu}} - \mathbf{Y}||^2$

      $\Rightarrow$ $\hat{\boldsymbol{\mu}}$ is the orthogonal projection of $\mathbf{Y}$ onto $\mathsf{im}(\mathbf{X})$
  - ▷ $\hat{\boldsymbol{\mu}} = \begin{cases} \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} = \mathbf{U}\mathbf{U}^\top\mathbf{Y} & \text{if } \mathsf{rank}(\mathbf{X}) = p, \\ \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{+}\mathbf{X}^\top\mathbf{Y} = \mathbf{U}\mathbf{U}^\top\mathbf{Y} & \text{if } \mathsf{rank}(\mathbf{X}) < p, \end{cases}$
    - ∗ where $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ (thin/compact SVD)
- $\hat{\boldsymbol{\mu}} \in \mathsf{im}(\mathbf{X}) \Rightarrow \exists\ \hat{\boldsymbol{\beta}}$ such that $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$
- if $\mathsf{rank}(\mathbf{X}) = p \Rightarrow \exists!\ \hat{\boldsymbol{\beta}}$ such that $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$

Estimation in linear regression (practice in R)

- aim: minimize $||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2$ w.r.t. $\boldsymbol{\beta}$
- use that $\mathbf{X} = \mathbf{Q}\mathbf{R}$
  - ▷ $\mathbf{Q}$ is $n \times n$ orthogonal

- ▷ $\mathbf{R}$ is $n \times p$ upper triangular
- ▷ $\mathbf{Q}$ and $\mathbf{Q}^\top$ rotations

○ $||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 = ||\mathbf{Q}^\top(\mathbf{Y} - \mathbf{Q}\mathbf{R}\,\boldsymbol{\beta})\,||^2$

$$= \left|\left|\begin{pmatrix}\mathbf{Q}_1^\top \\ \hline \mathbf{Q}_2^\top\end{pmatrix}\mathbf{Y} - \begin{pmatrix}\mathbf{R}_1 \\ \hline \mathbf{0}\end{pmatrix}\boldsymbol{\beta}\right|\right|^2$$

$$= \left|\left|\mathbf{Q}_1^\top\,\mathbf{Y} - \mathbf{R}_1\,\boldsymbol{\beta}\right|\right|^2 + \left|\left|\mathbf{Q}_2^\top\,\mathbf{Y}\right|\right|^2$$

- ▷ minimize $||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 \Leftrightarrow$ minimize $||\mathbf{Q}_1^\top\mathbf{Y} - \mathbf{R}_1\boldsymbol{\beta}||^2$

○ if $\mathsf{rank}(\mathbf{X}) = p$

- ▷ $\mathbf{R}_1$ invertible
- ▷ $\hat{\boldsymbol{\beta}} = \mathbf{R}_1^{-1}\mathbf{Q}_1^\top\mathbf{Y}$

Chapter 3

Normal distribution

3.1 The problem

3.1.1 Linear model

Linear model

- $Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k} + \varepsilon_i$, $i \in \{1, \ldots, n\}$
  - Y_i: outcome, response, output, dependent variable
    - random variable, we observe a realization y_i
    - (odezva, závisle proměnná, regresand)
  - $x_{i,1}, \ldots, x_{i,k}$: covariates, predictors, explanatory variables, input, independent variables
    - given, known
    - (nezávisle proměnné, regresory)
  - $\beta_0, \ldots, \beta_k$: coefficients
    - unknown
    - (regresní koeficienty)
  - ε_i: random error
    - random variable, unobserved
- $\varepsilon_i \overset{\text{iid}}{\sim} (0, \sigma^2)$, $i \in \{1, \ldots, n\}$
  - $\mathsf{E}\, \varepsilon_i = 0$: no systematic errors
  - $\mathsf{Var}\, \varepsilon_i = \sigma^2$: same precision
- we often assume that $\varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$, $i \in \{1, \ldots, n\}$

Example: `bloodpress` data

- from `sites.stat.psu.edu/~lsimon/stat501wc/sp05/data/`
- association between the mean arterial blood pressure[mmHg] and age[years], weight[kg], body surface area[m^2], duration of hypertension[years], basal pulse[beats/min], stress
- data:

BP	Age	Weight	BSA	DoH	Pulse	Stress
105	47	85.4	1.75	5.1	63	33
115	49	94.2	2.10	3.8	70	14
...	...	...	...	...	...	...
110	48	90.5	1.88	9.0	71	99
122	56	95.7	2.09	7.0	75	99

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

$$\begin{pmatrix} 105 \\ 115 \\ \dots \\ 110 \\ 122 \end{pmatrix} = \begin{pmatrix} 1 & 47 & 85.4 & 1.75 & 5.1 & 63 & 33 \\ 1 & 49 & 94.2 & 2.10 & 3.8 & 70 & 14 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 48 & 90.5 & 1.88 & 9.0 & 71 & 99 \\ 1 & 56 & 95.7 & 2.09 & 7.0 & 75 & 99 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \dots \\ \beta_6 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_{19} \\ \varepsilon_{20} \end{pmatrix}$$

3.1.2 Task for this chapter

Normal distribution in a linear model

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
- assumptions of the normal linear model:
  - ▷ $\mathbf{X}$ fixed and known
  - ▷ $\boldsymbol{\beta}$ fixed unknown
  - ▷ $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$

    $\Rightarrow \mathbf{Y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$
- estimators of $\boldsymbol{\beta}$ and σ^2
  - ▷ functions of $\mathbf{Y}$
- test statistics concerning $\boldsymbol{\beta}$ and σ^2
  - ▷ functions of $\mathbf{Y}$

$\Rightarrow$ to make inference in normal linear model, we need to study

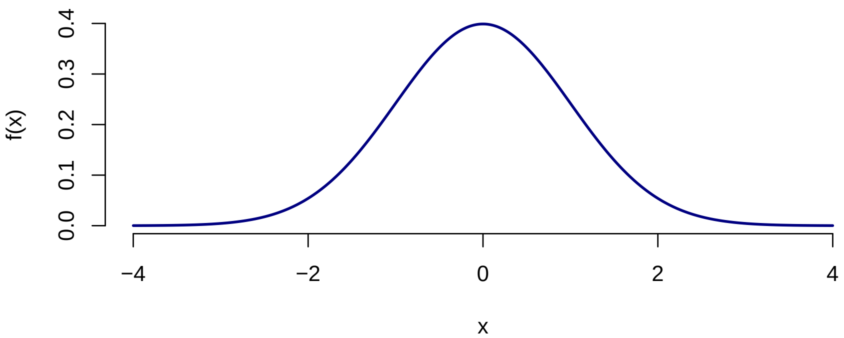
- ▷ multivariate normal distribution $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- ▷ distributions of functions of $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

3.2 Univariate normal distribution

3.2.1 Definition

Normal distribution $\mathrm{N}(\mu, \sigma^2)$

- let $\mu \in \mathbb{R}$ and $\sigma^2 > 0$
  - ▷ density $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}$
  - ▷ for the standard normal distribution ($\mu = 0$, $\sigma^2 = 1$):

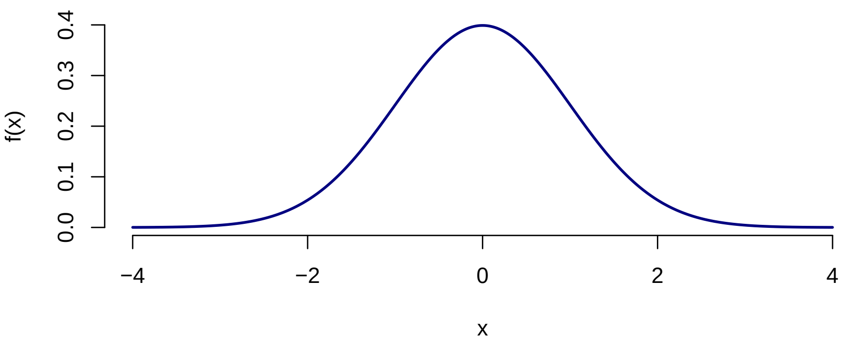


- if $\sigma^2 = 0$ then $X = \mu$ a.s.

3.2.2 Properties

Properties of $\mathrm{N}(\mu, \sigma^2)$

- $\mu \in \mathbb{R}$ and $\sigma^2 > 0$
- Let $a, b \in \mathbb{R}$, $X \sim \mathrm{N}(\mu, \sigma^2)$. Then $aX + b \sim \mathrm{N}(a\mu + b, a^2\sigma^2)$.
- Let $Z \sim \mathrm{N}(0, 1)$ and $X = \mu + \sigma Z$. Then $X \sim \mathrm{N}(\mu, \sigma^2)$.



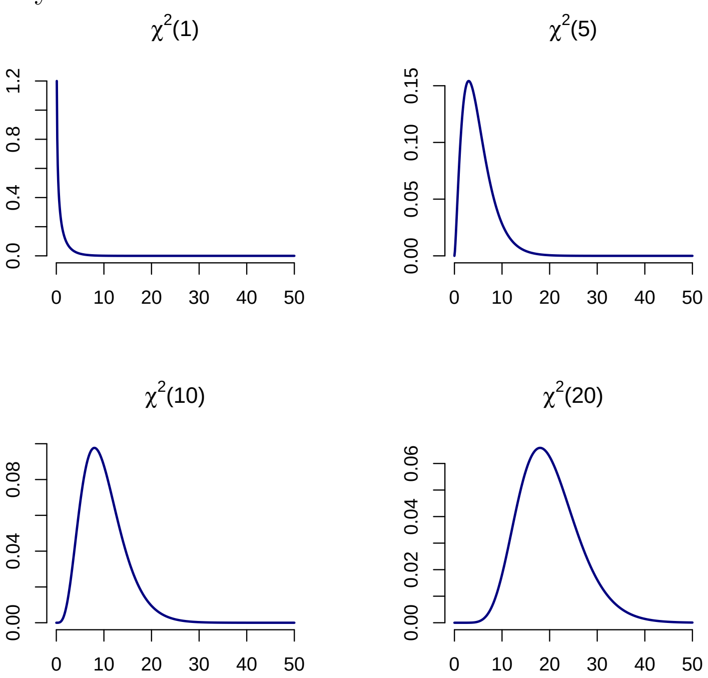
- Let $a_i, b_i \in \mathbb{R}$, $X_i \overset{\text{ind.}}{\sim} \mathrm{N}(\mu_i, \sigma_i^2)$ for $i \in \{1, \dots, n\}$.

  Then $\sum_{i=1}^n (a_i X_i + b_i) \sim \mathrm{N}(\sum_{i=1}^n (a_i \mu_i + b_i), \sum_{i=1} a_i^2 \sigma_i^2)$.

3.2.3 Related distributions

$\chi^2(n)$ distribution

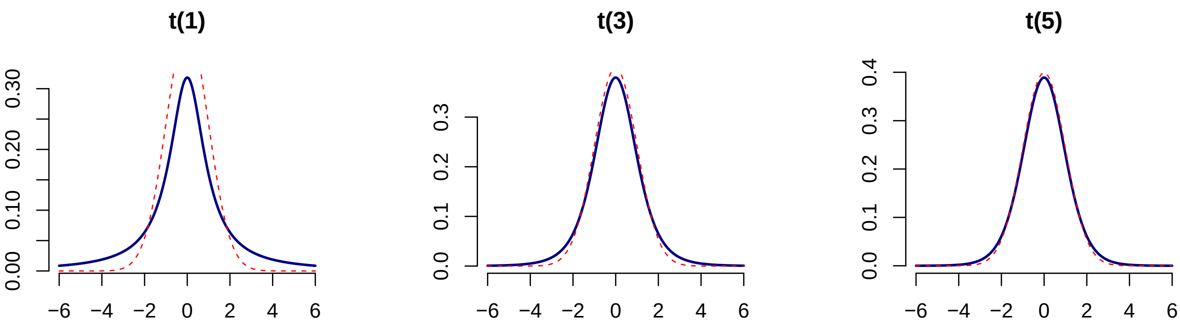
- let $Z \sim \mathrm{N}(0,1) \rightsquigarrow Z^2 \sim \chi^2(1)$
- let $Z_i \overset{\text{ind.}}{\sim} \mathrm{N}(0,1)$ for $i \in \{1, \dots, n\} \rightsquigarrow X = \sum_{i=1}^n Z_i^2 \sim \chi^2(n)$
- density



- $\mathsf{E}\,X = n, \mathsf{Var}\,X = 2n$

Student's t-distribution

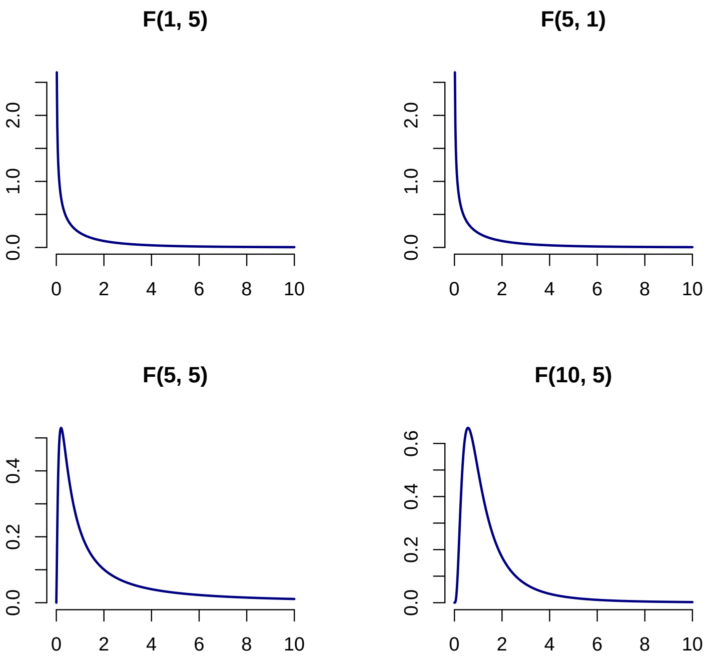
- let $Z \sim \mathrm{N}(0,1)$ and $X \sim \chi^2(n)$, $Z \perp\!\!\!\perp X$
  - ▷ $T = \frac{Z}{\sqrt{X/n}} \sim t(n)$
- density



- $\mathsf{E}\, T = 0$ for $n > 1$, $\mathsf{Var}\, T = n/(n-2)$ for $n > 2$

Fisher–Snedecor distribution

- let $X_1 \sim \chi^2(n_1)$ and $X_2 \sim \chi^2(n_2)$, $X_1 \perp\!\!\!\perp X_2$
  - $F = \frac{X_1/n_1}{\sqrt{X_2/n_2}} \sim F(n_1, n_2)$
- density



- $\mathsf{E}\, F = n_2/(n_2 - 2)$ for $n_2 > 2$

3.3 Multivariate normal distribution

3.3.1 Definition

Multivariate normal distribution $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

- $\boldsymbol{\mu} \in \mathbb{R}^n$, $\boldsymbol{\Sigma}$ is an $n \times n$ positive semidefinite matrix

Definition. *A random vector* $\mathbf{X} : (\Omega, \mathcal{A}) \mapsto (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ *has multivariate normal distribution* $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *if and only if* $\mathbf{a}^\top \mathbf{X} \sim \mathrm{N}(\mathbf{a}^\top \boldsymbol{\mu}, \mathbf{a}^\top \boldsymbol{\Sigma}\, \mathbf{a})$ *for every* $\mathbf{a} \in \mathbb{R}^n$.

- if $\mathsf{rank}(\boldsymbol{\Sigma}) = n$ then $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is non-degenerate
  - ▷ has density

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

- if $\mathsf{rank}(\boldsymbol{\Sigma}) = r < n$ then $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is degenerate
  - ▷ a.s. "lives" in a subspace of $\mathbb{R}^n$ of dimension r
  - ▷ no density w.r.t. Lebesgue measure on $\mathcal{B}(\mathbb{R}^n)$

Non-degenerate multivariate normal distribution

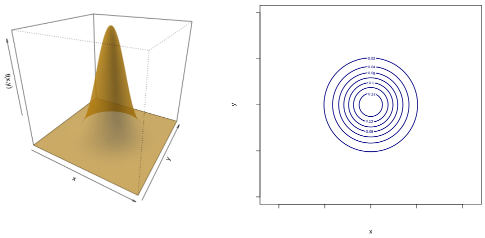
- density

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

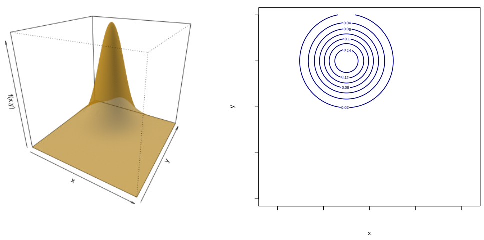
- $\boldsymbol{\Sigma}$: square symmetric positive definite matrix
  - ▷ spectral decomposition $\boldsymbol{\Sigma} = \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^\top$
  - ▷ $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0$
  - ▷ $\boldsymbol{\Sigma}^{-1} = \mathbf{U} \boldsymbol{\Lambda}^{-1} \mathbf{U}^\top$
- quadratic form $(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ can be written as
  - ▷ $(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{U} \boldsymbol{\Lambda}^{-1} \mathbf{U}^\top (\mathbf{x} - \boldsymbol{\mu}) = \{\mathbf{U}^\top (\mathbf{x} - \boldsymbol{\mu})\}^\top \boldsymbol{\Lambda}^{-1} \{\mathbf{U}^\top (\mathbf{x} - \boldsymbol{\mu})\}$
- level sets of $f(\mathbf{x})$, $I_c = \{\mathbf{x} \in \mathbb{R}^n;\ f(\mathbf{x}) = c\}$ for $c > 0$:
  - ▷ ellipsoids centred at $\boldsymbol{\mu}$
  - ▷ directions of principal axes: $\mathbf{u}_{1,}, \ldots, \mathbf{u}_{n,}$
  - ▷ lengths of principal semi-axes: $\sqrt{d\lambda_1}, \ldots, \sqrt{d\lambda_n}$

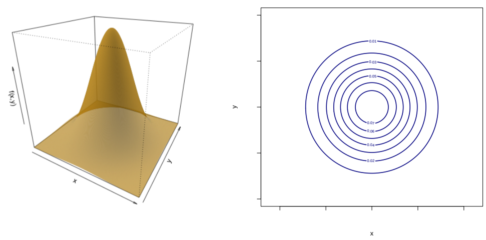
Non-degenerate bivariate normal distribution

- $$\mathrm{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$$

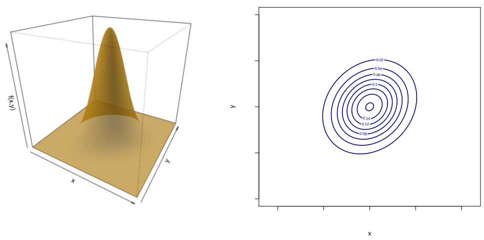


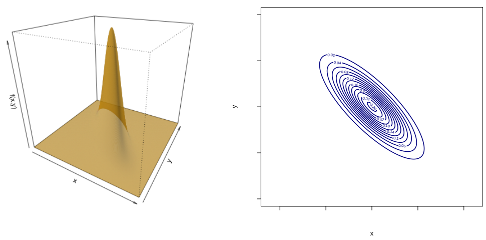
- $$\mathrm{N}\left(\begin{pmatrix}-1\\2\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right) \qquad \mathrm{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}2 & 0\\0 & 2\end{pmatrix}\right)$$





- $$\mathrm{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.2\\0.2 & 1\end{pmatrix}\right) \qquad \mathrm{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & -0.8\\-0.8 & 1\end{pmatrix}\right)$$





3.3.2 Properties

Properties of $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

- $\boldsymbol{\mu} \in \mathbb{R}^n$, $\boldsymbol{\Sigma}$ is an $n \times n$ symmetric positive semidefinite matrix

Theorem (MVN 1)**.** *Let* $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*. Then* $\mathsf{E}\,\mathbf{X} = \boldsymbol{\mu}$ *and* $\mathsf{Var}\,\mathbf{X} = \boldsymbol{\Sigma}$*.*

Theorem (MVN 2)**.** *Let* $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \mathrm{N}(0,1)$ *and* $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$*. Then* $\mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$*.*

Theorem (MVN 3). *Let* $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *and let* $\mathbf{A}$ *be an* $m \times n$ *real matrix and* $\mathbf{b} \in \mathbb{R}^m$. *Then* $\mathbf{AX} + \mathbf{b} \sim \mathrm{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$.

- proofs are given during the lectures and can also be found in Jiří Anděl: Základy matematické statistiky

$\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ seen through $\mathrm{N}(\mathbf{0}, \mathbf{I})$

- $\boldsymbol{\mu} \in \mathbb{R}^n$, $\boldsymbol{\Sigma}$ is an $n \times n$ symmetric positive semidefinite matrix

1. if $\mathsf{rank}(\boldsymbol{\Sigma}) = n$
   - spectral decomposition $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$
   - $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0$
   - $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top = \underbrace{\mathbf{U}\boldsymbol{\Lambda}^{1/2}}_{\widetilde{\boldsymbol{\Sigma}}} \boldsymbol{\Lambda}^{1/2}\mathbf{U}^\top = \widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\Sigma}}^\top$
   - let $\mathbf{Z} = \widetilde{\boldsymbol{\Sigma}}^{-1}(\mathbf{X} - \boldsymbol{\mu}) = \boldsymbol{\Lambda}^{-1/2}\,\mathbf{U}^\top(\mathbf{X} - \boldsymbol{\mu})$

     $\Rightarrow$ $\mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$ (n-dimensional), $\mathbf{X} = \boldsymbol{\mu} + \widetilde{\boldsymbol{\Sigma}}\mathbf{Z}$ and $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
2. if $\mathsf{rank}(\boldsymbol{\Sigma}) = r < n$
   - spectral decomposition $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$
   - $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$, $\lambda_{r+1} = \lambda_{r+2} = \ldots = \lambda_n = 0$
   - $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top = \underbrace{\mathbf{U}_{n\times r}}_{(\mathbf{u}_{,1}\,|\,\mathbf{u}_{,2}\,|\,\ldots\,|\,\mathbf{u}_{,r})} \underbrace{\boldsymbol{\Lambda}_{r\times r}}_{\mathrm{diag}\{\lambda_1,\lambda_2,\ldots,\lambda_r\}} \mathbf{U}_{n\times r}^\top = \underbrace{\mathbf{U}_{n\times r}\boldsymbol{\Lambda}_{r\times r}^{1/2}}_{\widetilde{\boldsymbol{\Sigma}}} \boldsymbol{\Lambda}_{r\times r}^{1/2}\mathbf{U}_{n\times r}^\top = \widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\Sigma}}^\top$
   - let $\mathbf{Z} = \widetilde{\boldsymbol{\Sigma}}^{+}(\mathbf{X} - \boldsymbol{\mu}) = \boldsymbol{\Lambda}_{r\times r}^{-1/2}\,\mathbf{U}_{n\times r}^\top(\mathbf{X} - \boldsymbol{\mu})$

     $\Rightarrow$ $\mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$ (r-dimensional), $\mathbf{X} = \boldsymbol{\mu} + \widetilde{\boldsymbol{\Sigma}}\mathbf{Z}$ and $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Density of $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

- $\boldsymbol{\mu} \in \mathbb{R}^n$, $\boldsymbol{\Sigma}$ is an $n \times n$ symmetric positive definite matrix

Theorem (MVN 4). *Let* $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *where* $\mathsf{rank}(\boldsymbol{\Sigma}) = n$. *Then* $\mathbf{X}$ *has density* $f(\mathbf{x})$ *w.r.t. Lebesgue measure on* $\mathcal{B}(\mathbb{R}^n)$ *and*

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}.$$

- a proof is given during the lectures and can also be found in Jiří Anděl: Základy matematické statistiky

Characteristic function (reminder)

Definition (Characteristic function of a random variable). *Let X be a random variable. The function $\psi_X : \mathbb{R} \mapsto \mathbb{C}$ defined by $\psi_X(t) = \mathsf{E}\exp\{\mathrm{i}\, t\, X\}$, $t \in \mathbb{R}$, is the characteristic function of X.*

Definition (Characteristic function of a random vector). *Let $\mathbf{X}$ be an n-dimensional random vector. The function $\psi_{\mathbf{X}} : \mathbb{R}^n \mapsto \mathbb{C}$ defined by $\psi_{\mathbf{X}}(\mathbf{t}) = \mathsf{E}\exp\{\mathrm{i}\, \mathbf{t}^\top \mathbf{X}\}$, $\mathbf{t} \in \mathbb{R}^n$, is the characteristic function of $\mathbf{X}$.*

Properties of characteristic function (reminder)

Theorem (ChF 1). *Let $X \sim \mathrm{N}(\mu, \sigma^2)$. Then $\psi_X(t) = \exp\left\{i\, t\, \mu - \frac{1}{2}\sigma^2 t^2\right\}$.*

Theorem (ChF 2). *Let $\mathbf{X}$ be an n-dimensional random vector and $\mathbf{X}_1$ and $\mathbf{X}_2$ its subvectors such that $\mathbf{X} = (\mathbf{X}_1^\top, \mathbf{X}_2^\top)^\top$. Then $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2$ iff $\psi_{\mathbf{X}}(\mathbf{t}) = \psi_{\mathbf{X}_1}(\mathbf{t}_1) \times \psi_{\mathbf{X}_2}(\mathbf{t}_2)$ for every $\mathbf{t} = (\mathbf{t}_1^\top, \mathbf{t}_2^\top)^\top \in \mathbb{R}^n$.*

- a proof can be found in *Petr Lachout: Teorie pravděpodobnosti (1998). Nakladatelství Univerzity Karlovy*

Characteristic function of $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

- $\boldsymbol{\mu} \in \mathbb{R}^n$, $\boldsymbol{\Sigma}$ is an $n \times n$ symmetric positive semidefinite matrix

Theorem (MVN 5). *Let $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then*

$$\psi_{\mathbf{X}}(\mathbf{t}) = \exp\left\{\mathrm{i}\, \mathbf{t}^\top \boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^\top \boldsymbol{\Sigma}\, \mathbf{t}\right\}.$$

- a proof is given during the lectures and can also be found in Jiří Anděl: Základy matematické statistiky

Subvectors of $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

- $\boldsymbol{\mu} \in \mathbb{R}^n$, $\boldsymbol{\Sigma}$ is an $n \times n$ symmetric positive semidefinite matrix

Theorem (MVN 6). *Let $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and let $k \in \{1, \ldots, n\}$. Then*

$$\begin{pmatrix} X_1 \\ X_2 \\ \cdots \\ X_k \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \\ \cdots \\ \mu_k \end{pmatrix}, \begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,k} \\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,k} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{k,1} & \sigma_{k,2} & \cdots & \sigma_{k,k} \end{pmatrix}\right).$$

- a proof is given during the lectures and can also be found in Jiří Anděl: Základy matematické statistiky

- analogous statement is true for any sub-vector of $\mathbf{X}$
- converse is not true

(In)dependence in $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

- $\boldsymbol{\mu} \in \mathbb{R}^n$, $\boldsymbol{\Sigma}$ is an $n \times n$ symmetric positive semidefinite matrix

Theorem (MVN 7)**.** *Let $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and let $k \in \{1, \ldots, n-1\}$. Denote $\mathbf{X}_1 = (X_1, \ldots, X_k)^\top$, $\mathbf{X}_2 = (X_{k+1}, \ldots, X_n)^\top$ and $\mathbf{X}_1 \sim \mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{1,1})$, $\mathbf{X}_2 \sim \mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{2,2})$. If*

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{1,1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{2,2} \end{pmatrix}$$

then $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2$.

- a proof is given during the lectures and can also be found in Jiří Anděl: Základy matematické statistiky
- $\mathbf{AX} \perp\!\!\!\perp \mathbf{BX}$ iff $\mathbf{A}\boldsymbol{\Sigma}\mathbf{B}^\top = \mathbf{0}$

3.3.3 Related distributions

Quadratic forms

- Let $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu} \in \mathbb{R}^n$, $\boldsymbol{\Sigma}$ is an $n \times n$ symmetric positive semidefinite matrix

Theorem (QF 1)**.** *Let $\mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$. Then $\mathbf{Z}^\top\mathbf{Z} \sim \chi^2(n)$.*

Theorem (QF 2)**.** *Let $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\mathsf{rank}(\boldsymbol{\Sigma}) = n$. Then $(\mathbf{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2(n)$.*

Theorem (QF 3)**.** *Let $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\mathsf{rank}(\boldsymbol{\Sigma}) = r < n$. Then $(\mathbf{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{+} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2(r)$.*

- proofs are given during the lectures and analogous statements are proved in Jiří Anděl: Základy matematické statistiky

Quadratic forms

- Let $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu} \in \mathbb{R}^n$, $\boldsymbol{\Sigma}$ is an $n \times n$ symmetric positive semidefinite matrix

Theorem (QF 4)**.** *Let $\mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$ and let $\mathbf{P}$ be an $n \times n$ projection matrix of rank r. Then $\mathbf{Z}^\top \mathbf{P}\, \mathbf{Z} \sim \chi^2(r)$.*

- a proof is given during the lectures and analogous statements are proved in Jiří Anděl: Základy matematické statistiky

Chapter 4

Linear model

4.1 The problem

4.1.1 Linear model

Linear model

- $Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k} + \varepsilon_i$, $i \in \{1, \ldots, n\}$
  - ▷ Y_i: outcome, response, output, dependent variable
    - ∗ random variable, we observe a realization y_i
    - ∗ (odezva, závisle proměnná, regresand)
  - ▷ $x_{i,1}, \ldots, x_{i,k}$: covariates, predictors, explanatory variables, input, independent variables
    - ∗ given, known
    - ∗ (nezávisle proměnné, regresory)
  - ▷ $\beta_0, \ldots, \beta_k$: coefficients
    - ∗ unknown
    - ∗ (regresní koeficienty)
  - ▷ ε_i: random error
    - ∗ random variable, unobserved
- $\varepsilon_i \overset{\text{iid}}{\sim} (0, \sigma^2)$, $i \in \{1, \ldots, n\}$
  - ▷ $\mathsf{E}\, \varepsilon_i = 0$: no systematic errors
  - ▷ $\mathsf{Var}\, \varepsilon_i = \sigma^2$: same precision
- we often assume that $\varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$, $i \in \{1, \ldots, n\}$

Example: `bloodpress` data

- from `sites.stat.psu.edu/~lsimon/stat501wc/sp05/data/`
- association between the mean arterial blood pressure[mmHg] and age[years], weight[kg], body surface area[m^2], duration of hypertension[years], basal pulse[beats/min], stress
- data:

BP	Age	Weight	BSA	DoH	Pulse	Stress
105	47	85.4	1.75	5.1	63	33
115	49	94.2	2.10	3.8	70	14
...	...	...	...	...	...	...
110	48	90.5	1.88	9.0	71	99
122	56	95.7	2.09	7.0	75	99

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

$$\begin{pmatrix} 105 \\ 115 \\ \dots \\ 110 \\ 122 \end{pmatrix} = \begin{pmatrix} 1 & 47 & 85.4 & 1.75 & 5.1 & 63 & 33 \\ 1 & 49 & 94.2 & 2.10 & 3.8 & 70 & 14 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 48 & 90.5 & 1.88 & 9.0 & 71 & 99 \\ 1 & 56 & 95.7 & 2.09 & 7.0 & 75 & 99 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \dots \\ \beta_6 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_{19} \\ \varepsilon_{20} \end{pmatrix}$$

4.1.2 Task for this chapter

Estimation in linear model

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
  - ▷ outcome $\mathbf{Y}$
    - ∗ random vector, we observe a realization $\mathbf{y}$
  - ▷ predictors $\mathbf{x}_{,1}, \dots, \mathbf{x}_{,k}$
    - ∗ vector of given (known) constants
  - ▷ coefficients $\boldsymbol{\beta}$
    - ∗ vector of unknown constants
  - ▷ error $\boldsymbol{\varepsilon}$
    - ∗ unknown random vector, we do not observe its realization
  - ▷ assumptions: $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 \mathbf{I})$
    - ∗ $\mathsf{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$: the expected value of $\mathbf{Y}$ is a <u>linear</u> function of $\boldsymbol{\beta}$
    - ∗ $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$: no systematic errors
    - ∗ $\mathsf{Var}\,\boldsymbol{\varepsilon} = \sigma^2 \mathbf{I}$: independence and same precision
- task: given the observed data $\mathbf{y}$ and known matrix $\mathbf{X}$, find estimators $\widehat{\boldsymbol{\beta}}$ (and $\widehat{\sigma^2}$) of $\boldsymbol{\beta}$ (and σ^2) with desirable properties

4.2 Estimating $\boldsymbol{\beta}$

4.2.1 Orthogonal projection

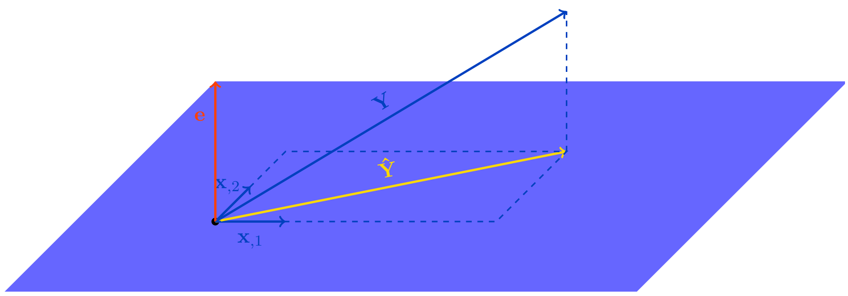
$\widehat{\boldsymbol{\beta}}$ motivated by orthogonal projection

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon}$ unknown, $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$
- idea: set $\boldsymbol{\varepsilon} \overset{!}{=} \mathbf{0}$ and solve $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$ w.r.t. $\boldsymbol{\beta}$
  - ▷ then $\underbrace{\mathbf{Y}}_{n\times 1} \overset{!}{=} \underbrace{\mathbf{X}}_{n\times p}\underbrace{\boldsymbol{\beta}}_{p\times 1}$
  - ▷ n linear equations with p unknowns and $n > p$

    $\Rightarrow$ a solution exists only if $\mathbf{Y} \in \mathsf{im}(\mathbf{X})$
- modified idea: find $\hat{\mathbf{Y}} \in \mathsf{im}(\mathbf{X})$ such that $||\mathbf{Y} - \hat{\mathbf{Y}}||^2$ is the smallest possible and solve $\hat{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta}$ w.r.t. $\boldsymbol{\beta}$
  - ▷ then $\hat{\mathbf{Y}}$ is the orthogonal projection of $\mathbf{Y}$ onto $\mathsf{im}(\mathbf{X})$
  - ▷ projection matrix onto $\mathsf{im}(\mathbf{X})$ is $\underbrace{\mathbf{H}}_{\text{hat matrix}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^+\mathbf{X}^\top$
  - ▷ solving $\hat{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta}$ is solving $\mathbf{X}\,(\mathbf{X}^\top\mathbf{X})^+\mathbf{X}^\top\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$
  - ▷ estimate $\boldsymbol{\beta}$ by $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^+\mathbf{X}^\top\mathbf{Y}$
  - ▷ but $\widehat{\boldsymbol{\beta}}$ is the unique solution of $\hat{\mathbf{Y}} = \mathbf{X}\,\boldsymbol{\beta}$ iff $\mathsf{rank}(\mathbf{X}) = p$
    - ∗ and then $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$

Geometric intuition

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon}$ unknown, $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$
- fitted model: $\underbrace{\mathbf{Y}}_{\text{observed value}} = \underbrace{\mathbf{H}\,\mathbf{Y}}_{\text{fitted value } \hat{\mathbf{Y}}} + \underbrace{(\mathbf{I} - \mathbf{H})\,\mathbf{Y}}_{\text{residual } \mathbf{e}}$



- $< \hat{\mathbf{Y}}, \mathbf{e} >= \mathbf{e}^\top\hat{\mathbf{Y}} = \mathbf{Y}^\top(\mathbf{I} - \mathbf{H})^\top\mathbf{H}\,\mathbf{Y} = 0$, i.e. $\hat{\mathbf{Y}} \perp \mathbf{e}$

4.2.2 Least squares

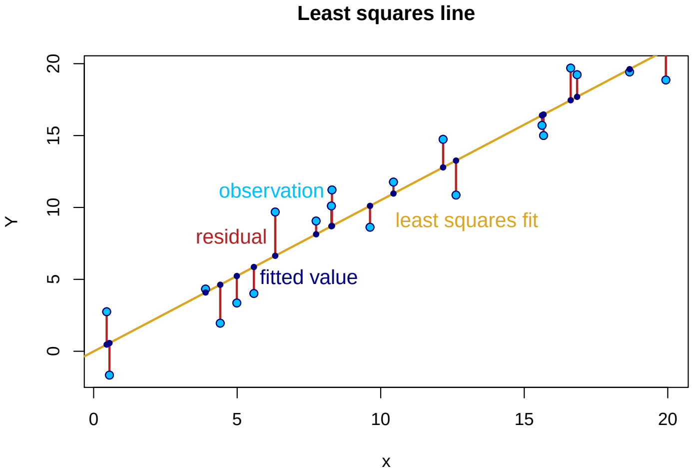
$\widehat{\boldsymbol{\beta}}$ as least squares estimator

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon}$ unknown, $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$
- idea: make the residuals as small as possible
  - ▷ minimize $||\boldsymbol{\varepsilon}||^2 = \sum_{i=1}^n \varepsilon_i^2$ w.r.t. $\boldsymbol{\beta}$

    $\rightsquigarrow$ Least Squares Estimator (LSE) $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n \varepsilon_i^2$
  - ▷ also called the OLS (Ordinary Least Squares) solution
- computation:
  - ▷ $\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$
  - ▷ $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 = \arg\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$
- look for the minimum by differentiating:
  - ▷ $\frac{\partial}{\partial \boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \overset{!}{=} 0$
  - ▷ $-2\,\mathbf{X}^\top \mathbf{Y} + 2\,\mathbf{X}^\top \mathbf{X}\,\boldsymbol{\beta} \overset{!}{=} 0$
  - ▷ $\underbrace{\mathbf{X}^\top \mathbf{X}\,\boldsymbol{\beta} \overset{!}{=} \mathbf{X}^\top \mathbf{Y}}_{\text{normal equations}}$
  - ▷ $\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \overset{?}{\succ} 0$ at $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$
  - ▷ $2\,\mathbf{X}^\top \mathbf{X} \succ 0$ for all $\boldsymbol{\beta}$
  - ▷ convex function $\Rightarrow$ minimum
- normal equations have unique solution iff $\mathsf{rank}(\mathbf{X}) = p$
  - ▷ and then $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$

Geometric intuition

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon}$ unknown, $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$
- fitted model: $\underbrace{\mathbf{Y}}_{\text{observed value}} = \underbrace{\mathbf{X}\widehat{\boldsymbol{\beta}}}_{\text{fitted value } \hat{\mathbf{Y}}} + \underbrace{(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})}_{\text{residual } \mathbf{e}}$
- least squares estimator minimizes the sum of squared vertical distances between the fitted and observed values



4.2.3 Computing $\widehat{\boldsymbol{\beta}}$

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

- we have seen two approaches give the same $\widehat{\boldsymbol{\beta}}$
- both approaches give unique $\widehat{\boldsymbol{\beta}}$ iff $\mathsf{rank}(\mathbf{X}) = p$
- both approaches would give infinitely many $\widehat{\boldsymbol{\beta}}$s if $\mathsf{rank}(\mathbf{X}) < p$
- a rank-deficient design matrix means a problem in design/model formulation
- we need to fix that problem to obtain reasonable conclusions from our model
- from now on we assume that $\mathsf{rank}(\mathbf{X}) = p$
- we will get back to (nearly) rank-deficient $\mathbf{X}$ in Chapter 9

$\widehat{\boldsymbol{\beta}}$ the way it is computed in R

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon}$ unknown, $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$
- $\widehat{\boldsymbol{\beta}}$ minimizes $||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2$ w.r.t. $\boldsymbol{\beta}$
- R uses that $\mathbf{X} = \mathbf{QR}$ (QR decomposition from Chapter 2)
  - ▷ $\mathbf{Q}$ $(n \times n)$ orthogonal
  - ▷ $\mathbf{R}$ $(n \times p)$ upper triangular
  - ▷ $\mathbf{X} = \mathbf{Q}\,\mathbf{R} = (\mathbf{Q}_1 \,|\, \mathbf{Q}_2) \begin{pmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{pmatrix} = \mathbf{Q}_1\,\mathbf{R}_1$

- ▷ R does not allow $\mathsf{rank}(\mathbf{X}) < p$

  $\Rightarrow \mathsf{rank}(\mathbf{R}_1) = p$
- ▷ $\mathbf{Q}$ and $\mathbf{Q}^\top$ are rotations

  $$* \ ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 = ||\mathbf{Q}^\top(\mathbf{Y} - \mathbf{Q}\,\mathbf{R}\,\boldsymbol{\beta})\,||^2 = \left|\left|\begin{pmatrix}\mathbf{Q}_1^\top \\ \mathbf{Q}_2^\top\end{pmatrix}\mathbf{Y} - \begin{pmatrix}\mathbf{R}_1 \\ \mathbf{0}\end{pmatrix}\boldsymbol{\beta}\right|\right|^2$$
  $$= \left|\left|\mathbf{Q}_1^\top\,\mathbf{Y} - \mathbf{R}_1\,\boldsymbol{\beta}\right|\right|^2 + \left|\left|\mathbf{Q}_2^\top\,\mathbf{Y}\right|\right|^2$$
- ▷ minimize $||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 \Leftrightarrow$ minimize $||\mathbf{Q}_1^\top\mathbf{Y} - \mathbf{R}_1\boldsymbol{\beta}||^2$
- ▷ $\hat{\boldsymbol{\beta}} = \mathbf{R}_1^{-1}\mathbf{Q}_1^\top\mathbf{Y}$ (compare with $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$)

Geometric intuition

- ∘ model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon}$ unknown, $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$
- ∘ $$\underbrace{\mathbf{Y}}_{\text{observed value}} = \underbrace{\mathbf{X}\widehat{\boldsymbol{\beta}}}_{\text{fitted value }\hat{\mathbf{Y}}} + \underbrace{(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})}_{\text{residual }\mathbf{e}}$$

  ▷ $$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{Q}\left(\mathbf{Q}^\top\mathbf{Q}\mathbf{R}\,\widehat{\boldsymbol{\beta}}\right) = \mathbf{Q}\left(\begin{pmatrix}\mathbf{R}_1 \\ \mathbf{0}\end{pmatrix}\widehat{\boldsymbol{\beta}}\right) = \mathbf{Q}\begin{pmatrix}\mathbf{R}_1\widehat{\boldsymbol{\beta}} \\ \mathbf{0}\end{pmatrix}$$

  ▷ $$\mathbf{e} = (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{Q}\left(\mathbf{Q}^\top(\mathbf{Y} - \mathbf{Q}\mathbf{R}\,\widehat{\boldsymbol{\beta}})\right) = \mathbf{Q}\left(\begin{pmatrix}\mathbf{Q}_1^\top \\ \mathbf{Q}_2^\top\end{pmatrix}\mathbf{Y} - \begin{pmatrix}\mathbf{R}_1 \\ \mathbf{0}\end{pmatrix}\widehat{\boldsymbol{\beta}}\right) = \mathbf{Q}\begin{pmatrix}\mathbf{0} \\ \mathbf{Q}_2^\top\mathbf{Y}\end{pmatrix}$$
- ∘ $\mathbf{Q}^\top$ conveniently rotates $\mathbf{Y}$ and $\mathsf{im}(\mathbf{X})$ and $\mathbf{Q}$ rotates back
- ∘ $\widehat{\mathbf{Y}} \perp \mathbf{e}$

Geometric intuition

- ∘ model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon}$ unknown, $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$
- ∘ $$\underbrace{\mathbf{Y}}_{\text{observed value}} = \underbrace{\mathbf{X}\widehat{\boldsymbol{\beta}}}_{\text{fitted value }\hat{\mathbf{Y}}} + \underbrace{(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})}_{\text{residual }\mathbf{e}} = \mathbf{Q}\begin{pmatrix}\mathbf{R}_1\widehat{\boldsymbol{\beta}} \\ \mathbf{0}\end{pmatrix} + \mathbf{Q}\begin{pmatrix}\mathbf{0} \\ \mathbf{Q}_2^\top\mathbf{Y}\end{pmatrix}$$
- ∘ see Figure 1.5 on page 20 in Simon Wood's *Generalized additive models* for a nice illustration

4.3 Quality of estimation

4.3.1 Gauss–Markov theorem

Linear transformation of a random vector

- we want to study $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$

Theorem. *Let* $\mathbf{X}$ *be an* n*-dimensional random vector with a finite variance-covariance matrix and let* $\mathbf{A}$ *be an* $m \times n$ *matrix. Then*

- $\mathsf{E}(\mathbf{A}\,\mathbf{X}) = \mathbf{A}\,\mathsf{E}\,\mathbf{X}$;
- $\mathsf{Var}(\mathbf{A}\,\mathbf{X}) = \mathbf{A}(\mathsf{Var}\,\mathbf{X})\mathbf{A}^\top$;
- $\mathsf{E}(\mathbf{X}^\top\mathbf{X}) = (\mathsf{E}\,\mathbf{X})^\top(\mathsf{E}\,\mathbf{X}) + \mathsf{tr}(\mathsf{Var}\,\mathbf{X})$.

- proof is a simple exercise

Is $\widehat{\boldsymbol{\beta}}$ a reasonable estimator?

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon}$ unknown, $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$
- $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$
- has a nice motivation but how about properties?
  - ▷ $$\mathsf{E}\,\widehat{\boldsymbol{\beta}} = \mathsf{E}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\,\mathsf{E}\,\mathbf{Y} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\,\boldsymbol{\beta} = \boldsymbol{\beta}$$

    $\Rightarrow$ unbiased
  - ▷ $$\begin{aligned}\mathsf{Var}\,\widehat{\boldsymbol{\beta}} = \mathsf{Var}\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}\right) &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\,\mathsf{Var}\,\mathbf{Y}\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right)^\top\\ &= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\,(\mathbf{X}^\top\mathbf{X})^{-1}\\ &= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\end{aligned}$$
- how good is $\mathsf{Var}\,\widehat{\boldsymbol{\beta}}$?
  - ▷ $\widehat{\boldsymbol{\beta}}$ is a linear estimator, i.e. $\widehat{\boldsymbol{\beta}} = \mathbf{AY}$ for a matrix $\mathbf{A}$
  - ▷ $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator, i.e. $\mathsf{E}_{\boldsymbol{\beta}}\,\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ for all $\boldsymbol{\beta}$
  - ▷ in fact, $\widehat{\boldsymbol{\beta}}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$, i.e. $\widehat{\boldsymbol{\beta}}$ has the smallest variance among all linear unbiased estimators of $\boldsymbol{\beta}$

Gauss–Markov theorem

Theorem (Gauss–Markov). *Let $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\mathbf{X}$ is an $n \times p$ matrix, $\mathsf{rank}(\mathbf{X}) = p$, $\boldsymbol{\beta} \in \mathbb{R}^p$, and $\boldsymbol{\varepsilon}$ is an n-dimensional random vector with $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$ and $\mathsf{Var}\,\boldsymbol{\varepsilon} = \sigma^2\mathbf{I}$. Then $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$, i.e. if $\widetilde{\boldsymbol{\beta}}$ is a linear unbiased estimator of $\boldsymbol{\beta}$ then $\mathsf{Var}\,\widetilde{\boldsymbol{\beta}} - \mathsf{Var}\,\widehat{\boldsymbol{\beta}} \succeq 0$.*

- see the blackboard for a proof
  - ▷ main steps
    - ∗ show that if $\widetilde{\boldsymbol{\beta}} = \mathbf{AY}$ then $\mathbf{AX} = \mathbf{I}$
    - ∗ show that $\mathsf{Var}\,\widetilde{\boldsymbol{\beta}} - \mathsf{Var}\,\widehat{\boldsymbol{\beta}} = \sigma^2\mathbf{A}(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})^\top\mathbf{A}^\top$

4.4 Estimating σ^2

4.4.1 Estimating σ^2

Estimating σ^2

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon}$ unknown, $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$, $\mathsf{Var}\,\boldsymbol{\varepsilon} = \sigma^2\mathbf{I}$
- fitted model: $\mathbf{Y} = \underbrace{\mathbf{H}\,\mathbf{Y}}_{\hat{\mathbf{Y}}} + \underbrace{(\mathbf{I} - \mathbf{H})\,\mathbf{Y}}_{\mathbf{e}} = \underbrace{\mathbf{X}\widehat{\boldsymbol{\beta}}}_{\hat{\mathbf{Y}}} + \underbrace{(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})}_{\mathbf{e}}$
- idea: estimate $\boldsymbol{\varepsilon}$ by $\mathbf{e}$
  - ▷ some care is needed ...
    - ∗ $\mathsf{E}\,\mathbf{e} = \mathsf{E}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathsf{E}\,\mathbf{Y} - \mathbf{X}\,\mathsf{E}\,\widehat{\boldsymbol{\beta}} = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} = \mathbf{0}$
    - ∗ $\mathsf{Var}\,\mathbf{e} = \mathsf{Var}((\mathbf{I} - \mathbf{H})\,\mathbf{Y}) = (\mathbf{I} - \mathbf{H})\,\mathsf{Var}\,\mathbf{Y}\,(\mathbf{I} - \mathbf{H})^\top = \sigma^2(\mathbf{I} - \mathbf{H})$
    - ∗ $\mathsf{rank}((\mathbf{I} - \mathbf{H})) = n - \mathsf{rank}(\mathbf{X}) = n - p < n \Rightarrow$ dependence
  - ▷ $\mathsf{E}(\mathbf{e}^\top\mathbf{e}) = (\mathsf{E}\,\mathbf{e})^\top(\mathsf{E}\,\mathbf{e}) + \mathsf{tr}(\mathsf{Var}\,\mathbf{e})$
    $$= \mathsf{tr}(\sigma^2(\mathbf{I} - \mathbf{H}))$$
    $$\overset{*}{=} \sigma^2(n - \mathsf{rank}(\mathbf{X})) = \sigma^2(n - p)$$
    - ∗ ∗: $\mathsf{tr}(\mathbf{P}) = \mathsf{rank}(\mathbf{P})$ for orthogonal projection matrices
  - ▷ $\widehat{\sigma^2} = \frac{1}{n-p}\mathbf{e}^\top\mathbf{e} = \frac{1}{n-p}\sum_{i=1}^n e_i^2 = \frac{1}{n-p}\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$
    - ∗ unbiased estimator of σ^2

4.5 Quality of model fit

4.5.1 Coefficient of determination

Sums of squares

- for $\widehat{\boldsymbol{\beta}}$ we obtain the minimal $||\mathbf{e}||^2 = ||\mathbf{Y} - \hat{\mathbf{Y}}||^2 = ||\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}||^2$
- we have seen properties of $\widehat{\boldsymbol{\beta}}$ but how about $\hat{\mathbf{Y}}$?
- a question: how close $\hat{\mathbf{Y}}$ actually is to $\mathbf{Y}$?
  - how well do the covariates in $\mathbf{X}$ explain what we see in $\mathbf{Y}$?
- an answer:
  - there is some variability in Y_is for different i
    - Total Sum of Squares TSS: $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$
    - also called SST
  - the model explains a part of the variability in Y_is
    - for different is there are different $\mathbf{x}_{i,}$s and so different $\hat{Y}_i$s
    - Explained Sum of Squares ESS: $\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$
    - also called Sum of Squares due to Regression
  - but some variability remained unexplained by the model
    - Residual Sum of Squares RSS: $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
    - also called Sum of Squared Residuals or Sum of Squared Errors

Coefficient of determination $\mathbf{R^2}$

- relationship among the sums of squares
  - TSS = RSS + ESS
    - $||\mathbf{Y} - \bar{Y}\mathbf{1}||^2 = ||\mathbf{Y} \pm \hat{\mathbf{Y}} - \bar{Y}\mathbf{1}||^2 = ||\mathbf{Y} - \hat{\mathbf{Y}}||^2 + ||\hat{\mathbf{Y}} - \bar{Y}\mathbf{1}||^2$
    - because $< \mathbf{Y} - \hat{\mathbf{Y}}, \hat{\mathbf{Y}} >= 0 =< \mathbf{Y} - \hat{\mathbf{Y}}, \mathbf{1} >$ as $\hat{\mathbf{Y}}$ is the orthogonal projection of $\mathbf{Y}$ onto $\mathsf{im}(\mathbf{X})$ and $\mathbf{1} \in \mathsf{im}(\mathbf{X})$
  - variability: total = unexplained + explained
- so how well do the covariates in $\mathbf{X}$ explain what we see in $\mathbf{Y}$?
  - coefficient of determination $R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$
    - proportion of variability explained by the model
    - $0 \leq R^2 \leq 1$ and bigger is better
  - adjusted coefficient of determination $R^2_{adj} = 1 - \frac{RSS/(n-p)}{TSS/(n-1)}$
    - an alternative that takes the number of predictors into account
    - $RSS/(n-p) = \widehat{\sigma^2}$ from the linear regression,
      $TSS/(n-1) = \widehat{\sigma^2}$ without the linear regression

Chapter 5

Normal linear model

5.1 The problem

5.1.1 Normal linear model

Normal linear model

- $Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k} + \varepsilon_i$, $i \in \{1, \ldots, n\}$
  - Y_i: outcome, response, output, dependent variable
    - random variable, we observe a realization y_i
    - (odezva, závisle proměnná, regresand)
  - $x_{i,1}, \ldots, x_{i,k}$: covariates, predictors, explanatory variables, input, independent variables
    - given, known
    - (nezávisle proměnné, regresory)
  - $\beta_0, \ldots, \beta_k$: coefficients
    - unknown
    - (regresní koeficienty)
  - ε_i: random error
    - random variable, unobserved
- $\varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$, $i \in \{1, \ldots, n\}$
  - $\mathsf{E}\, \varepsilon_i = 0$: no systematic errors
  - $\mathsf{Var}\, \varepsilon_i = \sigma^2$: same precision

Example: `bloodpress` data

- from `sites.stat.psu.edu/~lsimon/stat501wc/sp05/data/`
- association between the mean arterial blood pressure[mmHg] and age[years], weight[kg], body surface area[m^2], duration of hypertension[years], basal pulse[beats/min], stress
- data:

BP	Age	Weight	BSA	DoH	Pulse	Stress
105	47	85.4	1.75	5.1	63	33
115	49	94.2	2.10	3.8	70	14
...	...	...	...	...	...	...
110	48	90.5	1.88	9.0	71	99
122	56	95.7	2.09	7.0	75	99

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

$$\begin{pmatrix} 105 \\ 115 \\ \dots \\ 110 \\ 122 \end{pmatrix} = \begin{pmatrix} 1 & 47 & 85.4 & 1.75 & 5.1 & 63 & 33 \\ 1 & 49 & 94.2 & 2.10 & 3.8 & 70 & 14 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 48 & 90.5 & 1.88 & 9.0 & 71 & 99 \\ 1 & 56 & 95.7 & 2.09 & 7.0 & 75 & 99 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \dots \\ \beta_6 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_{19} \\ \varepsilon_{20} \end{pmatrix}$$

5.1.2 Task for this chapter

Estimation in normal linear model

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
  - ▷ outcome $\mathbf{Y}$
    - ∗ random vector, we observe a realization $\mathbf{y}$
  - ▷ predictors $\mathbf{x}_{,1}, \dots, \mathbf{x}_{,k}$
    - ∗ vector of given (known) constants
  - ▷ coefficients $\boldsymbol{\beta}$
    - ∗ vector of unknown constants
  - ▷ error $\boldsymbol{\varepsilon}$
    - ∗ unknown random vector, we do not observe its realization
  - ▷ assumptions: $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
    - ∗ $\mathsf{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$: the expected value of $\mathbf{Y}$ is a linear function of $\boldsymbol{\beta}$
    - ∗ $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$: no systematic errors
    - ∗ $\mathsf{Var}\,\boldsymbol{\varepsilon} = \sigma^2\,\mathbf{I}$: independence and same precision
- task: given the observed data $\mathbf{y}$ and known matrix $\mathbf{X}$, find estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma^2}$ of $\boldsymbol{\beta}$ and σ^2 with desirable properties

5.2 Estimating $\boldsymbol{\beta}$ and σ^2

5.2.1 Likelihood

Likelihood

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$
  - $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}) \Rightarrow \mathbf{Y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$
  - density of $\mathbf{Y}$:

$$f(\mathbf{y};\, \boldsymbol{\beta}, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$$

  - density is a function of $\mathbf{y}$ (parameters are fixed)
- likelihood:

$$L(\boldsymbol{\beta}, \sigma^2;\, \mathbf{y}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$$

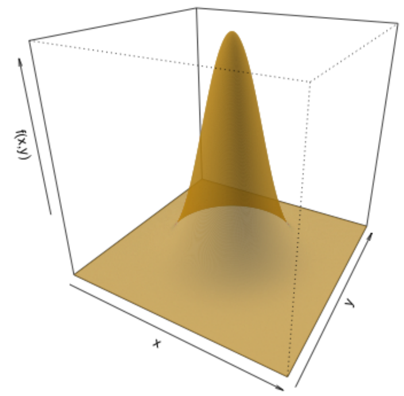
- likelihood is a function of the parameters ($\mathbf{y}$ is fixed)
- log-likelihood:

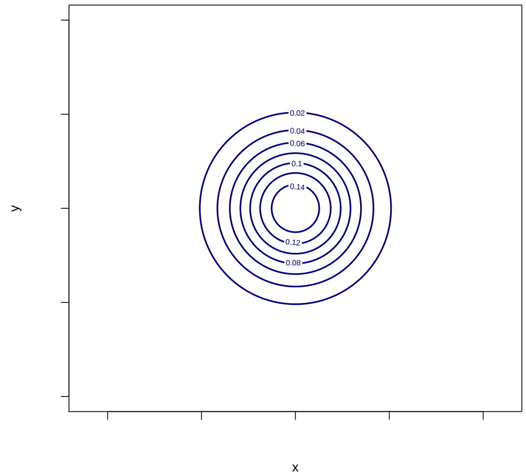
$$\ell(\boldsymbol{\beta}, \sigma^2;\, \mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \left\{\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$$

Log-likelihood

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$
- log-likelihood:

$$\ell(\boldsymbol{\beta}, \sigma^2;\, \mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \left\{\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$$





5.2.2 Matrix derivatives

Matrix derivatives: definition

- let $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$
- denominator-layout notation:

$$\frac{\partial}{\partial \mathbf{x}}\,\mathbf{y} = \begin{pmatrix} \frac{\partial}{\partial x_1} y_1 & \dots & \frac{\partial}{\partial x_1} y_m \\ \dots & \dots & \dots \\ \frac{\partial}{\partial x_n} y_1 & \dots & \frac{\partial}{\partial x_n} y_m \end{pmatrix}$$

- if $n = 1$

$$\frac{\partial}{\partial x}\,\mathbf{y} = \left(\frac{\partial}{\partial x} y_1, \dots, \frac{\partial}{\partial x} y_m\right)$$

- if $m = 1$

$$\frac{\partial}{\partial \mathbf{x}}\,y = \begin{pmatrix} \frac{\partial}{\partial x_1} y \\ \dots \\ \frac{\partial}{\partial x_n} y \end{pmatrix}$$

Matrix derivatives: useful formulae

- let $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{x} \in \mathbb{R}^n$
  - ▷ $\dfrac{\partial}{\partial \mathbf{x}}\,\mathbf{A}\,\mathbf{x} = \mathbf{A}^\top$
  - ▷ $\dfrac{\partial}{\partial \mathbf{x}}\,\mathbf{x}^\top \mathbf{A} = \mathbf{A}$

▷ $\dfrac{\partial}{\partial \mathbf{x}}\, \mathbf{x}^\top \mathbf{A}\, \mathbf{x} = (\mathbf{A} + \mathbf{A}^\top)\, \mathbf{x}$

5.2.3 Maximizing the likelihood

Score function

∘ log-likelihood

$$\ell(\boldsymbol{\beta}, \sigma^2;\, \mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \left\{\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$$

∘ score function ($\boldsymbol{\beta}$-related part):

$$\begin{aligned}
\mathbf{U}_{1:p}(\boldsymbol{\beta}, \sigma^2;\, \mathbf{y}) &= \frac{\partial}{\partial \boldsymbol{\beta}}\left(-\frac{1}{2\sigma^2}\,(\mathbf{y} - \mathbf{X}\,\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\,\boldsymbol{\beta})\right) \\
&= \frac{\partial}{\partial \boldsymbol{\beta}}\left(-\frac{1}{2\sigma^2}\left(\mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{X}\,\boldsymbol{\beta} - \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{y} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\,\boldsymbol{\beta}\right)\right) \\
&= -\frac{1}{2\sigma^2}\left(-\mathbf{X}^\top\mathbf{y} - \mathbf{X}^\top\mathbf{y} + (\mathbf{X}^\top\mathbf{X} + \mathbf{X}^\top\mathbf{X})\,\boldsymbol{\beta}\right) \\
&= \frac{1}{\sigma^2}\left(\mathbf{X}^\top\mathbf{y} - \mathbf{X}^\top\mathbf{X}\,\boldsymbol{\beta}\right)
\end{aligned}$$

Score function ctd.

∘ log-likelihood

$$\ell(\boldsymbol{\beta}, \sigma^2;\, \mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \left\{\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$$

∘ score function (σ^2-related part):

$$\begin{aligned}
U_{p+1}(\boldsymbol{\beta}, \sigma^2;\, \mathbf{y}) &= \frac{\partial}{\partial \sigma^2}\left(-\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\,(\mathbf{y} - \mathbf{X}\,\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\,\boldsymbol{\beta})\right) \\
&= -\frac{n}{2\,\sigma^2} + \frac{1}{2\sigma^4}\,(\mathbf{y} - \mathbf{X}\,\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\,\boldsymbol{\beta}) \\
&= \frac{1}{2\,\sigma^2}\left(-n + \frac{1}{\sigma^2}\,(\mathbf{y} - \mathbf{X}\,\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\,\boldsymbol{\beta})\right)
\end{aligned}$$

Score equation

∘ score equation:

$$\mathbf{U}(\boldsymbol{\beta}, \sigma^2;\, \mathbf{y}) = \begin{pmatrix} \frac{1}{\sigma^2}\left(\mathbf{X}^\top\mathbf{y} - \mathbf{X}^\top\mathbf{X}\,\boldsymbol{\beta}\right) \\ \frac{1}{2\,\sigma^2}\left(-n + \frac{1}{\sigma^2}\,(\mathbf{y} - \mathbf{X}\,\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\,\boldsymbol{\beta})\right) \end{pmatrix} = \mathbf{0}$$

- score equation for $\boldsymbol{\beta}$
  - ▷ $\frac{1}{\sigma^2}\left(\mathbf{X}^\top\mathbf{y} - \mathbf{X}^\top\mathbf{X}\,\boldsymbol{\beta}\right) \stackrel{!}{=} \mathbf{0}$
  - ▷ actually the normal equations
  - ▷ $\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$
- score equation for σ^2
  - ▷ $\frac{1}{2\,\sigma^2}\left(-n + \frac{1}{\sigma^2}\,(\mathbf{y} - \mathbf{X}\,\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\,\boldsymbol{\beta})\right) \stackrel{!}{=} 0$
  - ▷ $\widehat{\sigma^2}_{\mathrm{MLE}} = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\,\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}})^\top(\mathbf{Y} - \mathbf{X}\,\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}})$

Fisher information

- observed Fisher information matrix:

$$\mathbf{J}(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = \frac{1}{\sigma^2}\begin{pmatrix} \mathbf{X}^\top\mathbf{X} & \frac{1}{\sigma^2}\left(\mathbf{X}^\top\mathbf{y} - \mathbf{X}^\top\mathbf{X}\,\boldsymbol{\beta}\right) \\ \frac{1}{\sigma^2}\left(\mathbf{X}^\top\mathbf{y} - \mathbf{X}^\top\mathbf{X}\,\boldsymbol{\beta}\right) & -\frac{1}{\sigma^2}\left(\frac{n}{2} - \frac{1}{\sigma^2}\,(\mathbf{y} - \mathbf{X}\,\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\,\boldsymbol{\beta})\right) \end{pmatrix}$$

- $\mathbf{J}(\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}}, \widehat{\sigma^2}_{\mathrm{MLE}})$:

$$\frac{1}{\widehat{\sigma^2}_{\mathrm{MLE}}}\begin{pmatrix} \mathbf{X}^\top\mathbf{X} & \mathbf{0} \\ \mathbf{0} & \frac{n}{2\,\widehat{\sigma^2}_{\mathrm{MLE}}} \end{pmatrix} \succ 0$$

- Fisher information matrix:

$$\mathbf{I}(\boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma^2}\begin{pmatrix} \mathbf{X}^\top\mathbf{X} & \mathbf{0} \\ \mathbf{0} & \frac{n}{2\,\sigma^2} \end{pmatrix}$$

5.3 Distribution

5.3.1 Distribution of the MLE

Distribution of $\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}}$

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$
- $\mathbf{Y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\,\mathbf{I})$
- distribution of $\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$?
- MVN 3:

  Let $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and let $\mathbf{A}$ be an $m \times n$ real matrix and $\mathbf{b} \in \mathbb{R}^m$. Then $\mathbf{A}\mathbf{X} + \mathbf{b} \sim \mathrm{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$.

- $\widehat{\boldsymbol{\beta}}_{\text{MLE}} \sim \text{N}(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1})$

Distribution of $\widehat{\sigma^2}_{\text{MLE}}$

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \text{N}(\mathbf{0}, \sigma^2 \mathbf{I})$
- $\mathbf{Y} \sim \text{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$
- distribution of $\widehat{\sigma^2}_{\text{MLE}} = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\,\widehat{\boldsymbol{\beta}}_{\text{MLE}})^\top (\mathbf{Y} - \mathbf{X}\,\widehat{\boldsymbol{\beta}}_{\text{MLE}})$?
- recall that
  - ▷ $\hat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{Y} = \mathbf{X}\,(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$
  - ▷ $(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{e} = (\mathbf{I} - \mathbf{H})\,\mathbf{Y}$
    - $\mathbf{e} \sim \text{N}(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H}))$ (by MVN 3)
- QF 3:

  Let $\mathbf{X} \sim \text{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\mathsf{rank}(\boldsymbol{\Sigma}) = r < n$. Then $(\mathbf{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^+ (\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2(r)$.
- $(\mathbf{I} - \mathbf{H})^+ = (\mathbf{I} - \mathbf{H})$

  $$\Rightarrow \frac{1}{\sigma^2}\,\mathbf{e}^\top (\mathbf{I} - \mathbf{H})\,\mathbf{e} \sim \chi^2(n-p)$$
- $\mathbf{e}^\top (\mathbf{I} - \mathbf{H})\,\mathbf{e} = \mathbf{Y}^\top (\mathbf{I} - \mathbf{H})^\top (\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})\,\mathbf{Y} = \mathbf{e}^\top \mathbf{e}$
- $\frac{n}{\sigma^2}\,\widehat{\sigma^2}_{\text{MLE}} \sim \chi^2(n-p)$

Relationship between $\widehat{\boldsymbol{\beta}}_{\text{MLE}}$ and $\widehat{\sigma^2}_{\text{MLE}}$

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \text{N}(\mathbf{0}, \sigma^2 \mathbf{I})$
- $\mathbf{Y} \sim \text{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$
- $\widehat{\boldsymbol{\beta}}_{\text{MLE}} \sim \text{N}(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1})$
- $n\,\widehat{\sigma^2}_{\text{MLE}} \sim \chi^2(n-p)$
- joint distribution of $\widehat{\boldsymbol{\beta}}_{\text{MLE}}$ and $\widehat{\sigma^2}_{\text{MLE}}$?
- recall that
  - ▷ $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$
  - ▷ $(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{e} = (\mathbf{I} - \mathbf{H})\,\mathbf{Y}$

- Corollary of MVN 7:

  Let $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then $\mathbf{AX} \perp\!\!\!\perp \mathbf{BX}$ iff $\mathbf{A\Sigma B}^\top = \mathbf{0}$.

- $$(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{I}-\mathbf{H})^\top = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\,(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{0}$$

- $\widehat{\boldsymbol{\beta}} \perp\!\!\!\perp \mathbf{e}$ and $\widehat{\boldsymbol{\beta}} \perp\!\!\!\perp \widehat{\sigma^2}_{\mathrm{MLE}}$

5.4 Summary

5.4.1 Estimation in the normal linear model

Estimation in the normal linear model

Theorem. *Let* $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ *where* $\mathbf{X}$ *is an* $n \times p$ *matrix,* $\mathsf{rank}(\mathbf{X}) = p$*,* $\boldsymbol{\beta} \in \mathbb{R}^p$*, and* $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$.

Then the maximum likelihood estimators of $\boldsymbol{\beta}$ *and* σ^2 *are given by* $\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ *and* $\widehat{\sigma^2}_{\mathrm{MLE}} = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\,\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}})^\top(\mathbf{Y} - \mathbf{X}\,\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}})$.

Their distributions are $\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}} \sim \mathrm{N}(\boldsymbol{\beta}, \sigma^2\,(\mathbf{X}^\top\mathbf{X})^{-1})$ *and* $\frac{n}{\sigma^2}\,\widehat{\sigma^2}_{\mathrm{MLE}} \sim \chi^2(n-p)$*, and* $\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}}$ *and* $\widehat{\sigma^2}_{\mathrm{MLE}}$ *are independent.*

- unbiased estimator of σ^2: $\widehat{\sigma^2} = \frac{1}{n-p}(\mathbf{Y} - \mathbf{X}\,\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}})^\top(\mathbf{Y} - \mathbf{X}\,\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}})$
- its distribution: $\frac{(n-p)}{\sigma^2}\,\widehat{\sigma^2} \sim \chi^2(n-p)$ and $\widehat{\boldsymbol{\beta}} \perp\!\!\!\perp \widehat{\sigma^2}$

Chapter 6

Inference in normal linear model

6.1 The problem

6.1.1 Normal linear model

Normal linear model

- $Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k} + \varepsilon_i$, $i \in \{1, \ldots, n\}$
  - Y_i: outcome, response, output, dependent variable
    - random variable, we observe a realization y_i
    - (odezva, závisle proměnná, regresand)
  - $x_{i,1}, \ldots, x_{i,k}$: covariates, predictors, explanatory variables, input, independent variables
    - given, known
    - (nezávisle proměnné, regresory)
  - $\beta_0, \ldots, \beta_k$: coefficients
    - unknown
    - (regresní koeficienty)
  - ε_i: random error
    - random variable, unobserved
- $\varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$, $i \in \{1, \ldots, n\}$
  - $\mathsf{E}\, \varepsilon_i = 0$: no systematic errors
  - $\mathsf{Var}\, \varepsilon_i = \sigma^2$: same precision

Example: `bloodpress` data

- from `sites.stat.psu.edu/~lsimon/stat501wc/sp05/data/`
- association between the mean arterial blood pressure[mmHg] and age[years], weight[kg], body surface area[m^2], duration of hypertension[years], basal pulse[beats/min], stress
- data:

BP	Age	Weight	BSA	DoH	Pulse	Stress
105	47	85.4	1.75	5.1	63	33
115	49	94.2	2.10	3.8	70	14
...	...	...	...	...	...	...
110	48	90.5	1.88	9.0	71	99
122	56	95.7	2.09	7.0	75	99

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

$$\begin{pmatrix} 105 \\ 115 \\ \dots \\ 110 \\ 122 \end{pmatrix} = \begin{pmatrix} 1 & 47 & 85.4 & 1.75 & 5.1 & 63 & 33 \\ 1 & 49 & 94.2 & 2.10 & 3.8 & 70 & 14 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 48 & 90.5 & 1.88 & 9.0 & 71 & 99 \\ 1 & 56 & 95.7 & 2.09 & 7.0 & 75 & 99 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \dots \\ \beta_6 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_{19} \\ \varepsilon_{20} \end{pmatrix}$$

6.1.2 Task for this chapter

Inference in normal linear model

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
  - ▷ outcome $\mathbf{Y}$
    - ∗ random vector, we observe a realization $\mathbf{y}$
  - ▷ predictors $\mathbf{x}_{,1}, \dots, \mathbf{x}_{,k}$
    - ∗ vector of given (known) constants
  - ▷ coefficients $\boldsymbol{\beta}$
    - ∗ vector of unknown constants
  - ▷ error $\boldsymbol{\varepsilon}$
    - ∗ unknown random vector, we do not observe its realization
  - ▷ assumptions: $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I})$
    - ∗ $\mathsf{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$: the expected value of $\mathbf{Y}$ is a linear function of $\boldsymbol{\beta}$
    - ∗ $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$: no systematic errors
    - ∗ $\mathsf{Var}\,\boldsymbol{\varepsilon} = \sigma^2 \mathbf{I}$: independence and same precision
- task: given the observed data $\mathbf{y}$ and known matrix $\mathbf{X}$, draw conclusions about $\mathbf{Y}$ and the relationship between $\mathbf{Y}$ and $\mathbf{X}$

6.2 Estimators and distributions

6.2.1 Estimators

Point estimation in the normal linear model

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
  - ▷ $\mathbf{X}$ is an $n \times p$ matrix, $\mathsf{rank}(\mathbf{X}) = p$
  - ▷ $\boldsymbol{\beta} \in \mathbb{R}^p$
  - ▷ $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
- estimating $\boldsymbol{\beta}$
  - ▷ $\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}} = \widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \widehat{\boldsymbol{\beta}}_{\mathrm{MOM}} = \widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$
    - ∗ BLUE
    - ∗ distribution: $\widehat{\boldsymbol{\beta}} \sim \mathrm{N}(\boldsymbol{\beta}, \sigma^2\,(\mathbf{X}^\top\mathbf{X})^{-1})$
- estimating σ^2
  - ▷ $\widehat{\sigma^2}_{\mathrm{MLE}} = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^\top(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$
    - ∗ distribution: $\frac{n}{\sigma^2}\,\widehat{\sigma^2}_{\mathrm{MLE}} \sim \chi^2(n-p)$
  - ▷ $\widehat{\sigma^2} = \frac{1}{n-p}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^\top(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$
    - ∗ unbiased
    - ∗ distribution: $\frac{(n-p)}{\sigma^2}\,\widehat{\sigma^2} \sim \chi^2(n-p)$
- $\widehat{\boldsymbol{\beta}} \perp\!\!\!\perp \widehat{\sigma^2}_{\mathrm{MLE}}$ and $\widehat{\boldsymbol{\beta}} \perp\!\!\!\perp \widehat{\sigma^2}$

6.2.2 Distributions

Distributions in normal linear model

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
- distributions of point estimators
  - ▷ $\widehat{\boldsymbol{\beta}} \sim \mathrm{N}(\boldsymbol{\beta}, \sigma^2\,(\mathbf{X}^\top\mathbf{X})^{-1})$
  - ▷ $\frac{(n-p)}{\sigma^2}\,\widehat{\sigma^2} \sim \chi^2(n-p)$
  - ▷ $\widehat{\boldsymbol{\beta}} \perp\!\!\!\perp \widehat{\sigma^2}$
- let $\mathbf{a} \in \mathbb{R}^p$ and $\mathbf{A} \in \mathbb{R}^{m \times p}$

- ▷ $\mathbf{a}^\top\widehat{\boldsymbol{\beta}} \sim \mathrm{N}(\mathbf{a}^\top\boldsymbol{\beta}, \sigma^2\,\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{a})$
- ▷ $\mathbf{A}\,\widehat{\boldsymbol{\beta}} \sim \mathrm{N}(\mathbf{A}\,\boldsymbol{\beta}, \sigma^2\,\mathbf{A}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{A}^\top)$
  - ∗ proofs: use MVN 3:
    Let $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and let $\mathbf{A}$ be an $m \times n$ real matrix and $\mathbf{b} \in \mathbb{R}^m$. Then $\mathbf{A}\mathbf{X} + \mathbf{b} \sim \mathrm{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$.

Distributions in normal linear model ctd.

- ∘ model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
- ∘ distributions of statistics
  - ▷ $\mathbf{a}^\top\widehat{\boldsymbol{\beta}} \sim \mathrm{N}(\mathbf{a}^\top\boldsymbol{\beta}, \sigma^2\,\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{a})$ for $\mathbf{a} \in \mathbb{R}^p$
  - ▷ $\mathbf{A}\,\widehat{\boldsymbol{\beta}} \sim \mathrm{N}(\mathbf{A}\,\boldsymbol{\beta}, \sigma^2\,\mathbf{A}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{A}^\top)$ for $\mathbf{A} \in \mathbb{R}^{m\times p}$
  - ▷ $\frac{(n-p)}{\sigma^2}\,\widehat{\sigma^2} \sim \chi^2(n-p)$
  - ▷ $\widehat{\boldsymbol{\beta}} \perp\!\!\!\perp \widehat{\sigma^2}$
- ∘ for $\mathbf{a} \in \mathbb{R}^p$ and $\mathbf{A} \in \mathbb{R}^{m\times p}$, $\mathsf{rank}(\mathbf{A}) = m$
  - ▷ $$\frac{\mathbf{a}^\top\widehat{\boldsymbol{\beta}} - \mathbf{a}^\top\boldsymbol{\beta}}{\sqrt{\widehat{\sigma^2}\,\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{a}}} \sim t(n-p)$$
    - ∗ proof: verify that the definition of $t(n-p)$ is satisfied
  - ▷ $\frac{1}{m\,\widehat{\sigma^2}}\,(\mathbf{A}\widehat{\boldsymbol{\beta}} - \mathbf{A}\boldsymbol{\beta})^\top(\mathbf{A}\,(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{A}^\top)^{-1}(\mathbf{A}\widehat{\boldsymbol{\beta}} - \mathbf{A}\boldsymbol{\beta}) \sim F(m, n-p)$
    - ∗ proof: use QF2:
      $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathsf{rank}(\boldsymbol{\Sigma}) = n \Rightarrow (\mathbf{X}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu}) \sim \chi^2(n)$
      and verify that the definition of $F(m, n-p)$ is satisfied

6.3 Confidence intervals

Confidence intervals

Interval estimation in normal linear model

- ∘ model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
- ∘ let $\mathbf{a} \in \mathbb{R}^p$

- ▷ $\dfrac{\mathbf{a}^\top\widehat{\boldsymbol\beta} - \mathbf{a}^\top\boldsymbol\beta}{\sqrt{\widehat{\sigma^2}\,\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{a}}} \sim t(n-p)$
- ▷ $(1-\alpha)\times 100\,\%$ confidence interval for $\mathbf{a}^\top\boldsymbol\beta$:

$$\Big(\mathbf{a}^\top\widehat{\boldsymbol\beta} - t_{1-\alpha/2}(n-p)\sqrt{\widehat{\sigma^2}\,\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{a}}\,,\quad \mathbf{a}^\top\widehat{\boldsymbol\beta} + t_{1-\alpha/2}(n-p)\sqrt{\widehat{\sigma^2}\,\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{a}}\Big)$$

- ∘ $\frac{(n-p)}{\sigma^2}\,\widehat{\sigma^2} \sim \chi^2(n-p)$
  - ▷ $(1-\alpha)\times 100\,\%$ confidence interval for σ^2:

$$\left(\frac{(n-p)\,\widehat{\sigma^2}}{\chi^2_{1-\alpha/2}(n-p)},\ \frac{(n-p)\,\widehat{\sigma^2}}{\chi^2_{\alpha/2}(n-p)}\right)$$

Confidence intervals for the components of $\boldsymbol\beta$

- ∘ model: $\mathbf{Y} = \mathbf{X}\boldsymbol\beta + \boldsymbol\varepsilon$, $\boldsymbol\varepsilon \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
- ∘ let $\mathbf{a} \in \mathbb{R}^p$ such that $a_i = 1$ and $a_j = 0$ for $j \neq i$
  - ▷ $(1-\alpha)\times 100\,\%$ confidence interval for β_i:

$$\Big(\hat\beta_i - t_{1-\alpha/2}(n-p)\sqrt{\widehat{\sigma^2}\,(\mathbf{X}^\top\mathbf{X})^{-1}_{i,i}}\,,\quad \hat\beta_i + t_{1-\alpha/2}(n-p)\sqrt{\widehat{\sigma^2}\,(\mathbf{X}^\top\mathbf{X})^{-1}_{i,i}}\Big)$$

- ∘ let $\mathbf{a} \in \mathbb{R}^p$ such that $a_1 = 1$, $a_i = 1$ and $a_j = 0$ for $j \neq i$
  - ▷ $(1-\alpha)\times 100\,\%$ confidence interval for $\beta_1 + \beta_i$:

$$\Big(\hat\beta_1 + \hat\beta_i - t_{1-\alpha/2}(n-p)\sqrt{\widehat{\sigma^2}\,\big((\mathbf{X}^\top\mathbf{X})^{-1}_{1,1} + 2(\mathbf{X}^\top\mathbf{X})^{-1}_{1,i} + (\mathbf{X}^\top\mathbf{X})^{-1}_{i,i}\big)}\,,\quad \hat\beta_1 + \hat\beta_i + t_{1-\alpha/2}(n-p)\sqrt{\widehat{\sigma^2}\,\big((\mathbf{X}^\top\mathbf{X})^{-1}_{1,1} + 2(\mathbf{X}^\top\mathbf{X})^{-1}_{1,i} + (\mathbf{X}^\top\mathbf{X})^{-1}_{i,i}\big)}\Big)$$

  - ▷ and analogously for other sums of components of $\boldsymbol\beta$

6.4 Prediction

Prediction

New covariate combinations

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
- what can we say about

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon\ ?$$

- let $\mathbf{x} \in \mathbb{R}^p$ such that $\mathbf{x} = (1, x_1, \ldots, x_k)^\top$
- $Y = \mathbf{x}^\top \boldsymbol{\beta} + \varepsilon$ and $\mathsf{E}\,Y = \mathbf{x}^\top \boldsymbol{\beta}$
- we may estimate $\mathsf{E}\,Y$ by $\widehat{\mathsf{E}\,Y} = \mathbf{x}^\top \widehat{\boldsymbol{\beta}}$
- $(1-\alpha) \times 100\,\%$ confidence interval for $\mathsf{E}\,Y$:

$$\Big(\mathbf{x}^\top \widehat{\boldsymbol{\beta}} - t_{1-\alpha/2}(n-p)\sqrt{\widehat{\sigma^2}\,\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}}\,,\quad \mathbf{x}^\top \widehat{\boldsymbol{\beta}} + t_{1-\alpha/2}(n-p)\sqrt{\widehat{\sigma^2}\,\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}}\Big)$$

Prediction in normal linear model

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
- how can we estimate

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon = \mathbf{x}^\top \boldsymbol{\beta} + \varepsilon\ ?$$

  i.e. how do we predict new Y for new $\mathbf{x}$?
- prediction $\hat{Y} = \mathbf{x}^\top \widehat{\boldsymbol{\beta}}$
- $(1-\alpha) \times 100\,\%$ confidence interval for Y

  (prediction interval):

$$\Big(\mathbf{x}^\top \widehat{\boldsymbol{\beta}} - t_{1-\alpha/2}(n-p)\sqrt{\widehat{\sigma^2}\,(1 + \mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x})}\,,\quad \mathbf{x}^\top \widehat{\boldsymbol{\beta}} + t_{1-\alpha/2}(n-p)\sqrt{\widehat{\sigma^2}\,(1 + \mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x})}\Big)$$

6.5 Confidence bands

Confidence bands

Confidence regions in normal linear model

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
- let $\mathbf{A} \in \mathbb{R}^{m\times p}$, $\mathsf{rank}(\mathbf{A}) = m$

  ▷ $$\frac{1}{m\,\widehat{\sigma^2}}(\mathbf{A}\,\widehat{\boldsymbol{\beta}} - \mathbf{A}\,\boldsymbol{\beta})^\top(\mathbf{A}\,(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{A}^\top)^{-1}(\mathbf{A}\,\widehat{\boldsymbol{\beta}} - \mathbf{A}\,\boldsymbol{\beta}) \sim F(m, n-p)$$
- $(1-\alpha)\times 100\,\%$ confidence bands for $\mathbf{A}\,\boldsymbol{\beta}$:

$$\left\{\mathbf{A}\,\boldsymbol{\beta};\ \frac{1}{m\,\widehat{\sigma^2}}(\mathbf{A}\,\widehat{\boldsymbol{\beta}} - \mathbf{A}\,\boldsymbol{\beta})^\top(\mathbf{A}\,(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{A}^\top)^{-1}(\mathbf{A}\,\widehat{\boldsymbol{\beta}} - \mathbf{A}\,\boldsymbol{\beta}) \le F_{1-\alpha}(m, n-p)\right\}$$

Confidence bands in normal linear model

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$

Lemma 1. *Let* $\mathbf{B} \in \mathbb{R}^{m\times m}$, $\mathbf{B} \succ 0$. *Then for every* $\mathbf{x} \in \mathbb{R}^m$

$$\mathbf{x}^\top\mathbf{B}\,\mathbf{x} \le 1 \Leftrightarrow (\mathbf{b}^\top\mathbf{x})^2 \le \mathbf{b}^\top\mathbf{B}^{-1}\,\mathbf{b}\ \forall\ \mathbf{b} \in \mathbb{R}^m.$$

- a proof can be found in *Jiří Anděl: Základy matematické statistiky (2005). Matfyzpress*; see also multiple comparisons and Scheffé's theorem next semester
- for $\mathbf{A} \in \mathbb{R}^{m\times p}$, $\mathsf{rank}(\mathbf{A}) = m$:

$$\begin{aligned} 1 - \alpha &= \\ &= \mathsf{P}\left(\frac{1}{m\,\widehat{\sigma^2}}(\mathbf{A}\,\widehat{\boldsymbol{\beta}} - \mathbf{A}\,\boldsymbol{\beta})^\top(\mathbf{A}\,(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{A}^\top)^{-1}(\mathbf{A}\,\widehat{\boldsymbol{\beta}} - \mathbf{A}\,\boldsymbol{\beta}) \le F_{1-\alpha}(m, n-p)\right) \\ &= \mathsf{P}\left(\left(\mathbf{b}^\top(\mathbf{A}\,\widehat{\boldsymbol{\beta}} - \mathbf{A}\,\boldsymbol{\beta})\right)^2 \le m\,F_{1-\alpha}(m, n-p)\,\widehat{\sigma^2}\ \mathbf{b}^\top\left(\mathbf{A}\,(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{A}^\top\right)\mathbf{b};\ \forall\mathbf{b} \in \mathbb{R}^m\right)\end{aligned}$$

6.6 Testing hypotheses

6.6.1 Simple hypothesis

Testing $H_0 : \beta_i = 0$

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
- for $\mathbf{a} \in \mathbb{R}^p$
$$\frac{\mathbf{a}^\top\widehat{\boldsymbol{\beta}} - \mathbf{a}^\top\boldsymbol{\beta}}{\sqrt{\widehat{\sigma^2}\,\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{a}}} \sim t(n-p)$$
- let $\mathbf{a} \in \mathbb{R}^p$ such that $a_i = 1$ and $a_j = 0$ for $j \neq i$
- testing
  - ▷ $H_0 : \beta_i = 0$ vs.
  - ▷ $H_1 : \beta_i \neq 0$
- test statistic $T_i = \dfrac{\hat{\beta}_i}{\sqrt{\widehat{\sigma^2}\,(\mathbf{X}^\top\mathbf{X})^{-1}_{i,i}}} \sim t(n-p)$
- reject H_0 in favour of H_1 if $|t_i| > t_{1-\alpha/2}(n-p)$
- analogously for linear combinations of elements of $\boldsymbol{\beta}$
- analogously for testing $H_0 : \beta_i = \beta_{0,i}$

6.6.2 Composite hypothesis

Testing $H_0 : \boldsymbol{\beta}_{i:p} = \mathbf{0}$

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
- for $\mathbf{A} \in \mathbb{R}^{m\times p}$, $\mathsf{rank}(\mathbf{A}) = m$
$$\frac{1}{m\,\widehat{\sigma^2}}\,(\mathbf{A}\widehat{\boldsymbol{\beta}} - \mathbf{A}\boldsymbol{\beta})^\top(\mathbf{A}\,(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{A}^\top)^{-1}(\mathbf{A}\widehat{\boldsymbol{\beta}} - \mathbf{A}\boldsymbol{\beta}) \sim F(m, n-p)$$
- testing
  - ▷ $H_0 : \boldsymbol{\beta}_{i:p} = \mathbf{0}$ vs.
  - ▷ $H_1 : \boldsymbol{\beta}_{i:p} \neq \mathbf{0}$
- test statistic
$$F_{i:p} = \frac{1}{(p-i+1)\,\widehat{\sigma^2}}\,\widehat{\boldsymbol{\beta}}^\top_{i:p}\,(\mathbf{X}^\top\mathbf{X})^{-1}_{i:p,i:p}\,\widehat{\boldsymbol{\beta}}_{i:p} \sim F(p-i+1, n-p)$$

- reject H_0 in favour of H_1 if $f_{i:p} > F_{1-\alpha}(p-i+1, n-p)$

Testing "the model"

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
- for $\mathbf{A} \in \mathbb{R}^{m\times p}$, $\mathsf{rank}(\mathbf{A}) = m$

$$\frac{1}{m\,\widehat{\sigma^2}}\,(\mathbf{A}\widehat{\boldsymbol{\beta}} - \mathbf{A}\boldsymbol{\beta})^\top(\mathbf{A}\,(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{A}^\top)^{-1}(\mathbf{A}\widehat{\boldsymbol{\beta}} - \mathbf{A}\boldsymbol{\beta}) \sim F(m, n-p)$$

- testing
  - $H_0 : \boldsymbol{\beta}_{2:p} = \mathbf{0}$ vs.
  - $H_1 : \boldsymbol{\beta}_{2:p} \neq \mathbf{0}$
- test statistic

$$F = \frac{1}{k\,\widehat{\sigma^2}}\,\widehat{\boldsymbol{\beta}}_{2:p}^\top\,(\mathbf{X}^\top\mathbf{X})^{-1}_{2:p,2:p}\,\widehat{\boldsymbol{\beta}}_{2:p} \sim F(k, n-p)$$

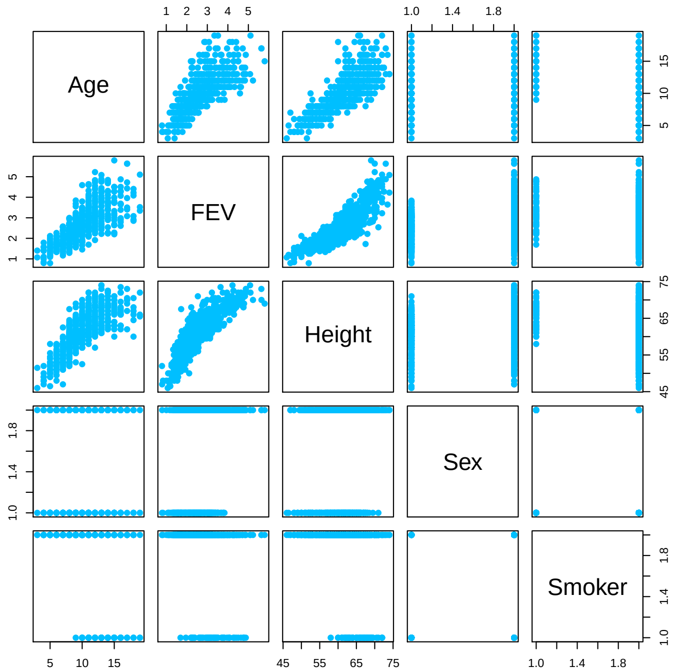
- reject H_0 in favour of H_1 if $f > F_{1-\alpha}(k, n-p)$

6.7 Interpretation

Interpretation of results for normal linear model

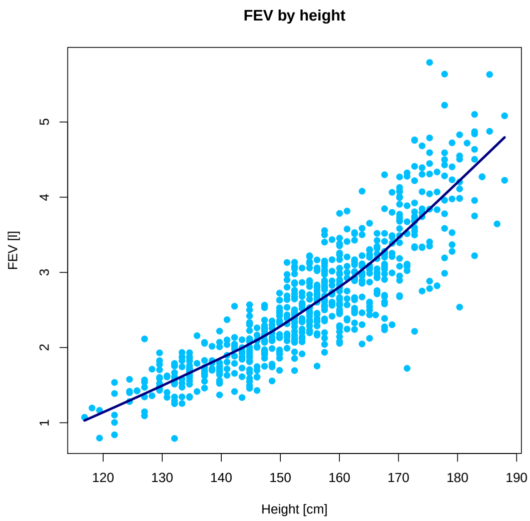
A model for `fev` data

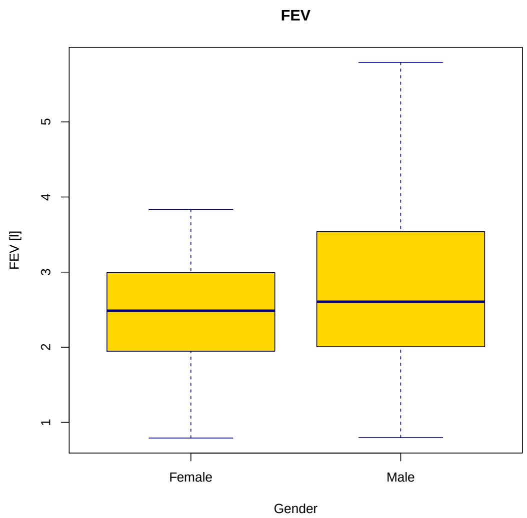
- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
- data: `fev` from `http://www.statsci.org/data/general/fev.html`



A model for `fev` data ctd.

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
- model FEV by Height and Sex





Fitted model for `fev` data

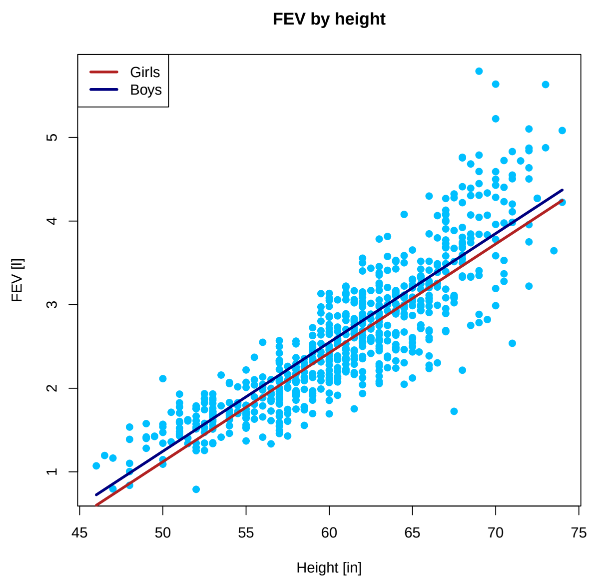
- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
- model FEV by Height and Sex

```
> model.simple <- lm(FEV~Height + Sex, data=fev)
> summary(model.simple)

Call:
lm(formula = FEV ~ Height + Sex, data = fev)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.6763 -0.2505  0.0001  0.2347  2.0722 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5.390263   0.180082 -29.932  < 2e-16 ***
Height       0.130231   0.002964  43.933  < 2e-16 ***
SexMale      0.125123   0.033801   3.702 0.000232 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4265 on 651 degrees of freedom
Multiple R-squared:  0.7587,Adjusted R-squared:  0.758 
F-statistic:  1024 on 2 and 651 DF,  p-value: < 2.2e-16
```

Fitted model for `fev` data ctd.

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
- model FEV by Height and Sex

```
> coefficients(model.simple)
(Intercept)      Height     SexMale 
 -5.3902632   0.1302305   0.1251234 
```



Chapter 7

Model selection

7.1 The problem

7.1.1 Normal linear model

Normal linear model

- $Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k} + \varepsilon_i$, $i \in \{1, \ldots, n\}$
  - Y_i: outcome, response, output, dependent variable
    - random variable, we observe a realization y_i
    - (odezva, závisle proměnná, regresand)
  - $x_{i,1}, \ldots, x_{i,k}$: covariates, predictors, explanatory variables, input, independent variables
    - given, known
    - (nezávisle proměnné, regresory)
  - $\beta_0, \ldots, \beta_k$: coefficients
    - unknown
    - (regresní koeficienty)
  - ε_i: random error
    - random variable, unobserved
- $\varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$, $i \in \{1, \ldots, n\}$
  - $\mathsf{E}\,\varepsilon_i = 0$: no systematic errors
  - $\mathsf{Var}\,\varepsilon_i = \sigma^2$: same precision

Example: `bloodpress` data

- from `sites.stat.psu.edu/~lsimon/stat501wc/sp05/data/`
- association between the mean arterial blood pressure[mmHg] and age[years], weight[kg], body surface area[m^2], duration of hypertension[years], basal pulse[beats/min], stress
- data:

BP	Age	Weight	BSA	DoH	Pulse	Stress
105	47	85.4	1.75	5.1	63	33
115	49	94.2	2.10	3.8	70	14
...	...	...	...	...	...	...
110	48	90.5	1.88	9.0	71	99
122	56	95.7	2.09	7.0	75	99

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

$$\begin{pmatrix} 105 \\ 115 \\ \dots \\ 110 \\ 122 \end{pmatrix} = \begin{pmatrix} 1 & 47 & 85.4 & 1.75 & 5.1 & 63 & 33 \\ 1 & 49 & 94.2 & 2.10 & 3.8 & 70 & 14 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 48 & 90.5 & 1.88 & 9.0 & 71 & 99 \\ 1 & 56 & 95.7 & 2.09 & 7.0 & 75 & 99 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \dots \\ \beta_6 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_{19} \\ \varepsilon_{20} \end{pmatrix}$$

`https://ww2.amstat.org/publications/jse/v13n2/datasets.kahn.html`

Example: `fev` data

- from: `http://www.statsci.org/data/general/fev.html`
- question: association between the FEV[l] and Smoking,

  corrected for Age[years], Height[cm] and Gender

- data:

FEV	Age	Height	Gender	Smoking
1.708	9	144.8	Female	Non
1.724	8	171.5	Female	Non
1.720	7	138.4	Female	Non
1.558	9	134.6	Male	Non
...	...	...	...	...
3.727	15	172.7	Male	Current
2.853	18	152.4	Female	Non
2.795	16	160.0	Female	Current
3.211	15	168.9	Female	Non

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

$$\begin{pmatrix} 1.708 \\ 1.724 \\ 1.720 \\ 1.558 \\ \dots \\ 3.727 \\ 2.853 \\ 2.795 \\ 3.211 \end{pmatrix} = \begin{pmatrix} 1 & 9 & 144.8 & 0 & 0 \\ 1 & 8 & 171.5 & 0 & 0 \\ 1 & 7 & 138.4 & 0 & 0 \\ 1 & 9 & 134.6 & 1 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 15 & 172.7 & 1 & 1 \\ 1 & 18 & 152.4 & 0 & 0 \\ 1 & 16 & 160.0 & 0 & 1 \\ 1 & 15 & 168.9 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \dots \\ \beta_5 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \dots \\ \varepsilon_{651} \\ \varepsilon_{652} \\ \varepsilon_{653} \\ \varepsilon_{654} \end{pmatrix}$$

7.1.2 Task for this chapter

Model building/selection

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

- ▷ outcome $\mathbf{Y}$
  - $*$ random vector, we observe a realization $\mathbf{y}$
- ▷ predictors $\mathbf{x}_{,1}, \ldots, \mathbf{x}_{,k}$
  - $*$ vector of given (known) constants
- ▷ coefficients $\boldsymbol{\beta}$
  - $*$ vector of unknown constants
- ▷ error $\boldsymbol{\varepsilon}$
  - $*$ unknown random vector, we do not observe its realization
- ▷ assumptions: $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
  - $*$ $\mathsf{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$: the expected value of $\mathbf{Y}$ is a linear function of $\boldsymbol{\beta}$
  - $*$ $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$: no systematic errors
  - $*$ $\mathsf{Var}\,\boldsymbol{\varepsilon} = \sigma^2\,\mathbf{I}$: independence and same precision

- ∘ task: given the observed data $\mathbf{y}$ and values of potential covariates, construct $\mathbf{X}$
- ∘ Note: $\mathbf{X}$ should ideally be known a priori based on background knowledge and various optimality considerations but …

7.2 Why consider various models?

7.2.1 Should we leave out covariates that appear unnecessary?

Testing hypotheses about null coefficients

- ∘ model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
- ∘ testing
  - ▷ $H_0 : \beta_i = 0$ vs.
  - ▷ $H_1 : \beta_i \neq 0$
- ∘ test statistic $T_i = \dfrac{\hat{\beta}_i}{\sqrt{\widehat{\sigma^2}\,(\mathbf{X}^\top\mathbf{X})^{-1}_{i,i}}} \sim t(n-p)$
- ∘ reject H_0 in favour of H_1 if $|t_i| > t_{1-\alpha/2}(n-p)$
- ∘ testing
  - ▷ $H_0 : \boldsymbol{\beta}_{i:p} = \mathbf{0}$ vs.

- ▷ $H_1 : \boldsymbol{\beta}_{i:p} \neq \mathbf{0}$

- test statistic

$$F_{i:p} = \frac{1}{(p-i+1)\,\widehat{\sigma^2}}\,\widehat{\boldsymbol{\beta}}_{i:p}^{\top}\,(\mathbf{X}^{\top}\mathbf{X})^{-1}_{i:p,i:p}\,\widehat{\boldsymbol{\beta}}_{i:p} \sim F(p-i+1, n-p)$$

- reject H_0 in favour of H_1 if $f_{i:p} > F_{1-\alpha}(p-i+1, n-p)$

What if we do not reject?

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
- testing
  - ▷ $H_0 : \beta_i = 0$ vs. $H_1 : \beta_i \neq 0$
- if $|t_i| < t_{1-\alpha/2}(n-p)$
  - ▷ at $\alpha\,\%$ level, we do not reject that $\beta_i = 0$ in favour of $\beta_i \neq 0$
- testing
  - ▷ $H_0 : \boldsymbol{\beta}_{i:p} = \mathbf{0}$ vs. $H_1 : \boldsymbol{\beta}_{i:p} \neq \mathbf{0}$
- if $f_{i:p} < F_{1-\alpha}(p-i+1, n-p)$
  - ▷ at $\alpha\,\%$ level, we do not reject $\boldsymbol{\beta}_{i:p} = \mathbf{0}$ in favour of $\boldsymbol{\beta}_{i:p} \neq \mathbf{0}$
- if we do not reject that some components of $\boldsymbol{\beta}$ are 0, should we change the model?
  - ▷ original model
    $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
  - ▷ new model
    $\mathbf{Y} = \mathbf{X}_{,1:(i-1)}\boldsymbol{\beta}_{1:(i-1)} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$

Example: `bloodpress` data

- original model

$$Y_i = \beta_0 + \beta_1 \times \mathsf{Age}_i + \beta_2 \times \mathsf{Weight}_i + \beta_3 \times \mathsf{BSA}_i + \\ + \beta_4 \times \mathsf{Dur}_i + \beta_5 \times \mathsf{Pulse}_i + \beta_6 \times \mathsf{Stress}_i + \varepsilon_i,\ 1 \le i \le 20$$

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -12.870476   2.556650  -5.034 0.000229 ***
Age           0.703259   0.049606  14.177 2.76e-09 ***
Weight        0.969920   0.063108  15.369 1.02e-09 ***
BSA           3.776491   1.580151   2.390 0.032694 *
Dur           0.068383   0.048441   1.412 0.181534
Pulse        -0.084485   0.051609  -1.637 0.125594
Stress        0.005572   0.003412   1.633 0.126491

Residual standard error: 0.4072 on 13 degrees of freedom

> coef.table <- summary(model.full)$coefficients
> V <- vcov(model.full)
> A <- diag(rep(1, 7))[5:7, ]
> F.stat <- t(A%*%coef.table[, 1])%*%solve(A%*%V%*%t(A))%*%(A%*%coef.table[, 1])/3
> 1-pf(F.stat, df1=3, df2=13)
          [,1]
[1,] 0.1950807
```

- should we rather use the new model?

$$Y_i = \beta_0 + \beta_1 \times \mathsf{Age}_i + \beta_2 \times \mathsf{Weight}_i + \beta_3 \times \mathsf{BSA}_i + \varepsilon_i,\ 1 \le i \le 20$$

Example: `bloodpress` data

- original model
  - ▷ $Y_i = \beta_0 + \beta_1 \times \mathsf{Age}_i + \beta_2 \times \mathsf{Weight}_i + \beta_3 \times \mathsf{BSA}_i + {}$ $+ \beta_4 \times \mathsf{Dur}_i + \beta_5 \times \mathsf{Pulse}_i + \beta_6 \times \mathsf{Stress}_i + \varepsilon_i,\ 1 \le i \le 20$
  - ▷ $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

$$\begin{pmatrix} 105 \\ 115 \\ \dots \\ 110 \\ 122 \end{pmatrix} = \begin{pmatrix} 1 & 47 & 85.4 & 1.75 & 5.1 & 63 & 33 \\ 1 & 49 & 94.2 & 2.10 & 3.8 & 70 & 14 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 48 & 90.5 & 1.88 & 9.0 & 71 & 99 \\ 1 & 56 & 95.7 & 2.09 & 7.0 & 75 & 99 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \dots \\ \beta_6 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_{19} \\ \varepsilon_{20} \end{pmatrix}$$

- new model
  - ▷ $Y_i = \beta_0 + \beta_1 \times \mathsf{Age}_i + \beta_2 \times \mathsf{Weight}_i + \beta_3 \times \mathsf{BSA}_i + \varepsilon_i,\ 1 \le i \le 20$
  - ▷ $\mathbf{Y} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

$$\begin{pmatrix} 105 \\ 115 \\ \dots \\ 110 \\ 122 \end{pmatrix} = \begin{pmatrix} 1 & 47 & 85.4 & 1.75 \\ 1 & 49 & 94.2 & 2.10 \\ \dots & \dots & \dots & \dots \\ 1 & 48 & 90.5 & 1.88 \\ 1 & 56 & 95.7 & 2.09 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_{19} \\ \varepsilon_{20} \end{pmatrix}$$

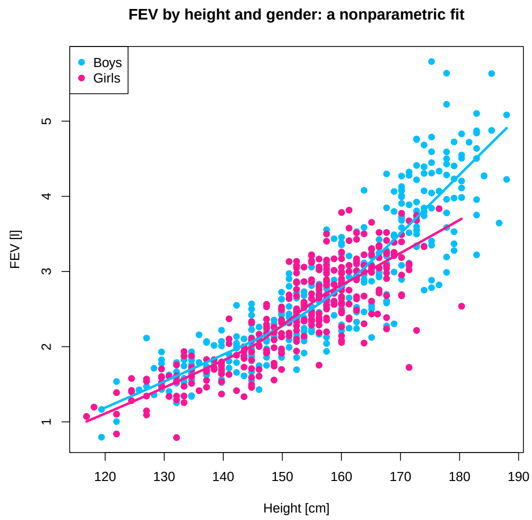
7.2.2 What is the right form of the dependence on covariates?

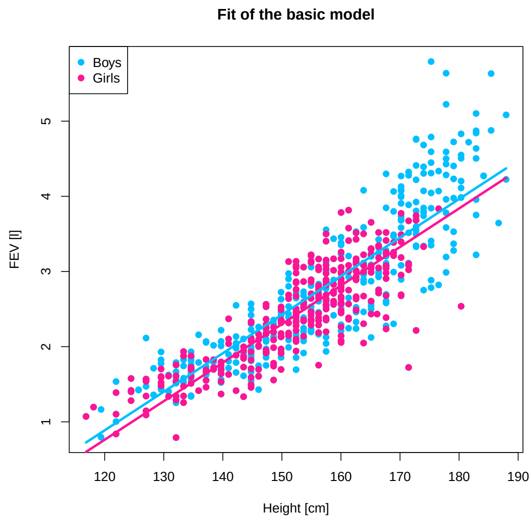
Specifying the form of dependence in the `fev` data

- basic model for the dependence of FEV on Height and Sex:

$$Y_i = \beta_0 + \beta_1 \times \mathsf{Height}_i + \beta_2 \times \mathbb{I}\{\text{the } i^{\text{th}} \text{ person is male}\},\ 1 \le i \le 654$$

- does the basic model fit the data well enough?





Example: `fev` data

- original model
  - ▷ $Y_i = \beta_0 + \beta_1 \times \mathsf{Height}_i + \beta_2 \times \mathbb{I}\{\text{the } i^{\text{th}} \text{ child is male}\} + \varepsilon_i,\ 1 \le i \le 654$

$$\begin{pmatrix} 1.708 \\ 1.724 \\ 1.720 \\ 1.558 \\ \ldots \\ 3.727 \\ 2.853 \\ 2.795 \\ 3.211 \end{pmatrix} = \begin{pmatrix} 1 & 144.8 & 0 \\ 1 & 171.5 & 0 \\ 1 & 138.4 & 0 \\ 1 & 134.6 & 1 \\ \ldots & \ldots & \ldots \\ 1 & 172.7 & 1 \\ 1 & 152.4 & 0 \\ 1 & 160.0 & 0 \\ 1 & 168.9 & 0 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \ldots \\ \varepsilon_{651} \\ \varepsilon_{652} \\ \varepsilon_{653} \\ \varepsilon_{654} \end{pmatrix}$$

- new model
  - ▷ $Y_i = \beta_0 + \beta_1 \times \mathsf{Height}_i + \beta_2 \times \mathsf{Height}_i^2 +$

$$+ \beta_3 \times \mathbb{I}\{\text{the } i^{\text{th}} \text{ child is male}\} + \beta_4 \times \mathsf{Height}_i \mathbb{I}\{\text{the } i^{\text{th}} \text{ child is male}\} +$$

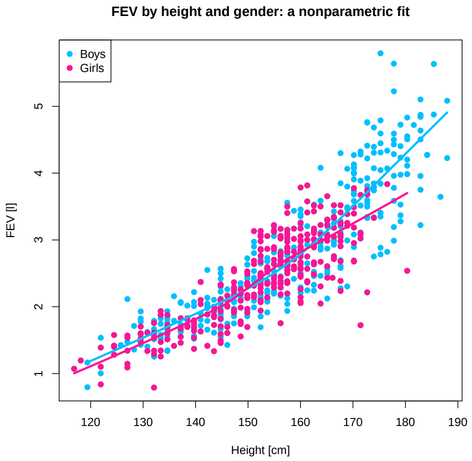
$$+ \beta_5 \times \mathsf{Height}_i^2 \mathbb{I}\{\text{the } i^{\text{th}} \text{ child is male}\} + \varepsilon_i,\ 1 \le i \le 654$$

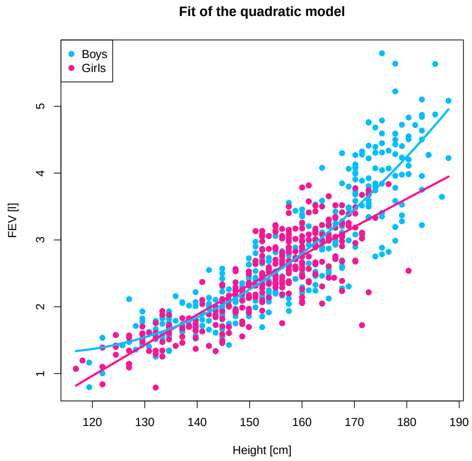
$$\begin{pmatrix} 1.708 \\ 1.724 \\ 1.720 \\ 1.558 \\ \ldots \\ 3.727 \\ 2.853 \\ 2.795 \\ 3.211 \end{pmatrix} = \begin{pmatrix} 1 & 144.8 & 20961.3 & 0 & 0 & 0 \\ 1 & 171.5 & 29395.1 & 0 & 0 & 0 \\ 1 & 138.4 & 19162.9 & 0 & 0 & 0 \\ 1 & 134.6 & 18122.5 & 1 & 134.6 & 18122.5 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 172.7 & 29832.2 & 1 & 172.7 & 29832.2 \\ 1 & 152.4 & 23225.8 & 0 & 0 & 0 \\ 1 & 160.0 & 25606.4 & 0 & 0 & 0 \\ 1 & 168.9 & 28530.6 & 0 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \ldots \\ \beta_5 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \ldots \\ \varepsilon_{651} \\ \varepsilon_{652} \\ \varepsilon_{653} \\ \varepsilon_{654} \end{pmatrix}$$

Example: `fev` data

- fit of the new model

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -5.194e+00  2.740e+00  -1.895   0.0585 .
Height              5.611e-02  3.692e-02   1.520   0.1291
I(Height^2)        -3.977e-05  1.238e-04  -0.321   0.7482
SexMale             1.392e+01  3.423e+00   4.067 5.34e-05 ***
Height:SexMale     -1.903e-01  4.545e-02  -4.188 3.20e-05 ***
I(Height^2):SexMale 6.471e-04  1.501e-04   4.310 1.89e-05 ***
```





7.3 Nested models

Submodel

Nested models

- Bigger model: $\mathbf{Y} = \mathbf{X}_\mathrm{b}\,\boldsymbol{\beta}_\mathrm{b} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\,\mathbf{I})$,

$$\mathbf{X}_\mathrm{b} = \big(\mathbf{1} \mid \mathbf{x}_{,1} \mid \mathbf{x}_{,2} \mid \ \ldots\ \mid \mathbf{x}_{,k-1} \mid \mathbf{x}_{,k}\big)$$

  - ▷ $\widehat{\boldsymbol{\beta}}_\mathrm{b} = (\mathbf{X}_\mathrm{b}^\top\mathbf{X}_\mathrm{b})^{-1}\mathbf{X}_\mathrm{b}^\top\mathbf{Y}$
  - ▷ $\hat{\mathbf{Y}}_\mathrm{b} = \mathbf{X}_\mathrm{b}\widehat{\boldsymbol{\beta}}_\mathrm{b} = \mathbf{H}_\mathrm{b}\mathbf{Y}$
  - ▷ $\mathbf{e}_\mathrm{b} = \mathbf{Y} - \widehat{\mathbf{Y}}_\mathrm{b} = (\mathbf{I} - \mathbf{H}_\mathrm{b})\mathbf{Y}$
  - ▷ $\widehat{\sigma^2}_\mathrm{b} = \frac{1}{n-p}\,||\mathbf{e}_\mathrm{b}||^2$

- Smaller model: $\mathbf{Y} = \mathbf{X}_\mathrm{s}\,\boldsymbol{\beta}_\mathrm{s} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\,\mathbf{I})$,

$$\mathbf{X}_\mathrm{s} = \big(\mathbf{1} \mid \mathbf{x}_{,1} \mid \mathbf{x}_{,2} \mid \ \ldots\ \mid \mathbf{x}_{,k-r}\big)$$

  - ▷ $\widehat{\boldsymbol{\beta}}_\mathrm{s} = (\mathbf{X}_\mathrm{s}^\top\mathbf{X}_\mathrm{s})^{-1}\mathbf{X}_\mathrm{s}^\top\mathbf{Y}$
  - ▷ $\hat{\mathbf{Y}}_\mathrm{s} = \mathbf{X}_\mathrm{s}\widehat{\boldsymbol{\beta}}_\mathrm{s} = \mathbf{H}_\mathrm{s}\mathbf{Y}$
  - ▷ $\mathbf{e}_\mathrm{s} = \mathbf{Y} - \widehat{\mathbf{Y}}_\mathrm{s} = (\mathbf{I} - \mathbf{H}_\mathrm{s})\mathbf{Y}$

  - ▷ $\widehat{\sigma^2}_{\mathrm{s}} = \frac{1}{n-p+r}\,||\mathbf{e}_{\mathrm{s}}||^2$

- ∘ more generally, any $\mathbf{X}_{\mathrm{s}}$ such that $\mathsf{im}(\mathbf{X}_{\mathrm{s}}) \le \mathsf{im}(\mathbf{X}_{\mathrm{b}})$
  - ▷ $\exists\ \mathbf{A} \in \mathbb{R}^{p\times(p-r)}$ such that $\mathbf{X}_{\mathrm{s}} = \mathbf{X}_{\mathrm{b}}\,\mathbf{A}$
  - ▷ $\mathbf{X}_{\mathrm{s}} = \left(\sum_{i=1}^{p} a_{i,1}\,\mathbf{x}_{,i} \mid \ldots \mid \sum_{i=1}^{p} a_{i,p-r}\,\mathbf{x}_{,i}\right)$

Relationship between the two models

- ∘ if the smaller model holds, so does the bigger one
- ∘ $\exists\ \mathbf{A} \in \mathbb{R}^{p\times(p-r)}$ such that $\mathbf{X}_{\mathrm{s}} = \mathbf{X}_{\mathrm{b}}\,\mathbf{A}$
  - ▷ $\mathbf{X}_{\mathrm{s}} = \left(\sum_{i=1}^{p} a_{i,1}\,\mathbf{x}_{,i} \mid \ldots \mid \sum_{i=1}^{p} a_{i,p-r}\,\mathbf{x}_{,i}\right)$
- ∘ bigger model: $\mathbf{Y} = \mathbf{X}_{\mathrm{b}}\,\boldsymbol{\beta}_{\mathrm{b}} + \boldsymbol{\varepsilon}$
- ∘ smaller model: $\mathbf{Y} = \mathbf{X}_{\mathrm{s}}\,\boldsymbol{\beta}_{\mathrm{s}} + \boldsymbol{\varepsilon} = \mathbf{X}_{\mathrm{b}}\,\underbrace{\mathbf{A}\,\boldsymbol{\beta}_{\mathrm{s}}}_{\boldsymbol{\beta}_{\mathrm{b}}} + \boldsymbol{\varepsilon}$
- ∘ smaller model is the bigger model with a condition on $\boldsymbol{\beta}_{\mathrm{b}}$
  - ▷ $\underbrace{\boldsymbol{\beta}_{\mathrm{b}}}_{p\times 1} = \underbrace{\mathbf{A}}_{p\times(p-r)}\ \underbrace{\boldsymbol{\beta}_{\mathrm{s}}}_{(p-r)\times 1} = \begin{pmatrix} \sum_{j=1}^{n-p} a_{1,j}\,\beta_{\mathrm{s},j} \\ \ldots \\ \sum_{j=1}^{n-p} a_{p,j}\,\beta_{\mathrm{s},j} \end{pmatrix}$
  - ▷ $\exists\ \mathbf{B} \in \mathbb{R}^{r\times p}$ such that $\mathbf{B}\,\boldsymbol{\beta}_{\mathrm{b}} = \mathbf{0}$
- ∘ in the bigger normal linear model we may test for the validity of the smaller model by testing whether $\mathbf{B}\,\boldsymbol{\beta}_{\mathrm{b}} = \mathbf{0}$ (see Week 7)

Relationship between the fits of the two models

- ∘ difference between the fits
  - ▷ $\hat{\mathbf{Y}}_{\mathrm{b}} - \hat{\mathbf{Y}}_{\mathrm{s}} = (\mathbf{H}_{\mathrm{b}} - \mathbf{H}_{\mathrm{s}})\,\mathbf{Y}$
- ∘ difference between the residuals
  - ▷ $\mathbf{e}_{\mathrm{s}} - \mathbf{e}_{\mathrm{b}} = (\mathbf{I} - \mathbf{H}_{\mathrm{s}})\,\mathbf{Y} - (\mathbf{I} - \mathbf{H}_{\mathrm{b}})\,\mathbf{Y} = (\mathbf{H}_{\mathrm{b}} - \mathbf{H}_{\mathrm{s}})\,\mathbf{Y}$
- ∘ comparison of the nested models' fits
  - ▷ $||\mathbf{e}_{\mathrm{s}}||^2 = ||\mathbf{e}_{\mathrm{b}}||^2 + ||\mathbf{e}_{\mathrm{s}} - \mathbf{e}_{\mathrm{b}}||^2$
    - ∗ proof: realize that $< \mathbf{e}_{\mathrm{b}}, \mathbf{e}_{\mathrm{s}} - \mathbf{e}_{\mathrm{b}} >= 0$
    - ∗ note: $||\mathbf{e}_{\mathrm{s}}||^2 \ge ||\mathbf{e}_{\mathrm{b}}||^2 \Rightarrow$ the fit of the bigger model is closer to the observed data

- note: $||\mathbf{e}_\mathrm{s} - \mathbf{e}_\mathrm{b}||^2 = ||\mathbf{e}_\mathrm{s}||^2 - ||\mathbf{e}_\mathrm{b}||^2$

◦ in the normal linear model ($\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$)

▷ $\mathbf{e}_\mathrm{b} \perp\!\!\!\perp (\mathbf{e}_\mathrm{s} - \mathbf{e}_\mathrm{b})$

- proof: Corollary of MVN 7:

  Let $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then $\mathbf{AX} \perp\!\!\!\perp \mathbf{BX}$ iff $\mathbf{A\Sigma B}^\top = \mathbf{0}$.

Does the bigger model fit significantly better?

◦ assume that both models hold (i.e. the smaller model holds) and that $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$ (normal linear model)

◦ $\frac{1}{\sigma^2}\,||\mathbf{e}_\mathrm{b}||^2 \sim \chi^2_{n-p}$

▷ proof: see Week 6

◦ $\frac{1}{\sigma^2}\,||\mathbf{e}_\mathrm{s} - \mathbf{e}_\mathrm{b}||^2 \sim \chi^2_r$

▷ proof: MVN 3:

Let $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and let $\mathbf{A}$ be an $m \times n$ real matrix and $\mathbf{b} \in \mathbb{R}^m$. Then $\mathbf{AX} + \mathbf{b} \sim \mathrm{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A\Sigma A}^\top)$.

▷ and QF 4:

Let $\mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$ and let $\mathbf{P}$ be an $n \times n$ projection matrix of rank r. Then $\mathbf{Z}^\top\mathbf{P}\,\mathbf{Z} \sim \chi^2(r)$.

◦ $||\mathbf{e}_\mathrm{b}||^2 \perp\!\!\!\perp ||\mathbf{e}_\mathrm{s} - \mathbf{e}_\mathrm{b}||^2$

▷ proof: see the previous slide

◦ $\dfrac{||\mathbf{e}_\mathrm{s} - \mathbf{e}_\mathrm{b}||^2/r}{||\mathbf{e}_\mathrm{b}||^2/(n-p)} = \dfrac{(||\mathbf{e}_\mathrm{s}||^2 - ||\mathbf{e}_\mathrm{b}||^2)/r}{||\mathbf{e}_\mathrm{b}||^2/(n-p)} \sim F_{r,n-p}$

▷ proof: verify that the definition of $F_{r,n-p}$ is satisfied

More submodels

Several models nested within one another

◦ Big model: $\mathbf{Y} = \mathbf{X}_\mathrm{b}\,\boldsymbol{\beta}_\mathrm{b} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\,\mathbf{I})$,

▷ $\hat{\mathbf{Y}}_\mathrm{b} = \mathbf{X}_\mathrm{b}\widehat{\boldsymbol{\beta}}_\mathrm{b} = \mathbf{H}_\mathrm{b}\mathbf{Y}$

▷ $\mathbf{e}_\mathrm{b} = \mathbf{Y} - \widehat{\mathbf{Y}}_\mathrm{b} = (\mathbf{I} - \mathbf{H}_\mathrm{b})\mathbf{Y}$

◦ Small model: $\mathbf{Y} = \mathbf{X}_\mathrm{s}\,\boldsymbol{\beta}_\mathrm{s} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\,\mathbf{I})$,

- ▷ $\hat{\mathbf{Y}}_{\mathrm{s}} = \mathbf{X}_{\mathrm{s}}\widehat{\boldsymbol{\beta}}_{\mathrm{s}} = \mathbf{H}_{\mathrm{s}}\mathbf{Y}$
- ▷ $\mathbf{e}_{\mathrm{s}} = \mathbf{Y} - \widehat{\mathbf{Y}}_{\mathrm{s}} = (\mathbf{I} - \mathbf{H}_{\mathrm{s}})\mathbf{Y}$
- ▷ $\widehat{\sigma^2}_{\mathrm{s}} = \frac{1}{n-p+r}\,||\mathbf{e}_{\mathrm{s}}||^2$

∘ Super-small model: $\mathbf{Y} = \mathbf{X}_{\mathrm{ss}}\,\boldsymbol{\beta}_{\mathrm{ss}} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\,\mathbf{I})$,

- ▷ $\hat{\mathbf{Y}}_{\mathrm{ss}} = \mathbf{X}_{\mathrm{ss}}\widehat{\boldsymbol{\beta}}_{\mathrm{ss}} = \mathbf{H}_{\mathrm{ss}}\mathbf{Y}$
- ▷ $\mathbf{e}_{\mathrm{ss}} = \mathbf{Y} - \widehat{\mathbf{Y}}_{\mathrm{ss}} = (\mathbf{I} - \mathbf{H}_{\mathrm{ss}})\mathbf{Y}$
- ▷ $\widehat{\sigma^2}_{\mathrm{ss}} = \frac{1}{n-p+s}\,||\mathbf{e}_{\mathrm{ss}}||^2$

∘ $\mathsf{im}(\mathbf{X}_{\mathrm{ss}}) \leq \mathsf{im}(\mathbf{X}_{\mathrm{s}}) \leq \mathsf{im}(\mathbf{X}_{\mathrm{b}})$

Relationship between the fits of the models

∘ difference between the fits

- ▷ $\hat{\mathbf{Y}}_{\mathrm{b}} - \hat{\mathbf{Y}}_{\mathrm{s}} = (\mathbf{H}_{\mathrm{b}} - \mathbf{H}_{\mathrm{s}})\,\mathbf{Y}$
- ▷ $\hat{\mathbf{Y}}_{\mathrm{b}} - \hat{\mathbf{Y}}_{\mathrm{ss}} = (\mathbf{H}_{\mathrm{b}} - \mathbf{H}_{\mathrm{ss}})\,\mathbf{Y}$
- ▷ $\hat{\mathbf{Y}}_{\mathrm{s}} - \hat{\mathbf{Y}}_{\mathrm{ss}} = (\mathbf{H}_{\mathrm{s}} - \mathbf{H}_{\mathrm{ss}})\,\mathbf{Y}$

∘ difference between the residuals

- ▷ $\mathbf{e}_{\mathrm{s}} - \mathbf{e}_{\mathrm{b}} = (\mathbf{I} - \mathbf{H}_{\mathrm{s}})\,\mathbf{Y} - (\mathbf{I} - \mathbf{H}_{\mathrm{b}})\,\mathbf{Y} = (\mathbf{H}_{\mathrm{b}} - \mathbf{H}_{\mathrm{s}})\,\mathbf{Y}$
- ▷ $\mathbf{e}_{\mathrm{ss}} - \mathbf{e}_{\mathrm{b}} = (\mathbf{I} - \mathbf{H}_{\mathrm{ss}})\,\mathbf{Y} - (\mathbf{I} - \mathbf{H}_{\mathrm{b}})\,\mathbf{Y} = (\mathbf{H}_{\mathrm{b}} - \mathbf{H}_{\mathrm{ss}})\,\mathbf{Y}$
- ▷ $\mathbf{e}_{\mathrm{ss}} - \mathbf{e}_{\mathrm{s}} = (\mathbf{I} - \mathbf{H}_{\mathrm{ss}})\,\mathbf{Y} - (\mathbf{I} - \mathbf{H}_{\mathrm{s}})\,\mathbf{Y} = (\mathbf{H}_{\mathrm{s}} - \mathbf{H}_{\mathrm{ss}})\,\mathbf{Y}$

∘ in the normal linear model ($\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$)

- ▷ $\mathbf{e}_{\mathrm{b}} \perp\!\!\!\perp (\mathbf{e}_{\mathrm{s}} - \mathbf{e}_{\mathrm{b}})$
- ▷ $\mathbf{e}_{\mathrm{b}} \perp\!\!\!\perp (\mathbf{e}_{\mathrm{ss}} - \mathbf{e}_{\mathrm{b}})$
- ▷ $\mathbf{e}_{\mathrm{b}} \perp\!\!\!\perp (\mathbf{e}_{\mathrm{ss}} - \mathbf{e}_{\mathrm{s}})$
  - ∗ proof: Corollary of MVN 7:
    Let $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then $\mathbf{AX} \perp\!\!\!\perp \mathbf{BX}$ iff $\mathbf{A\Sigma B}^\top = \mathbf{0}$.

How about the super-small model's fit?

∘ assume that all three models hold (i.e. the super-small model holds) and that $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$ (normal linear model)

∘ $\frac{1}{\sigma^2}\,||\mathbf{e}_{\mathrm{b}}||^2 \sim \chi^2_{n-p}$

∘ $\frac{1}{\sigma^2}\,||\mathbf{e}_{\mathrm{ss}} - \mathbf{e}_{\mathrm{s}}||^2 \sim \chi^2_{s-r}$

- ▷ proof: MVN 3:

  Let $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and let $\mathbf{A}$ be an $m \times n$ real matrix and $\mathbf{b} \in \mathbb{R}^m$. Then $\mathbf{A}\mathbf{X} + \mathbf{b} \sim \mathrm{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$.

- ▷ and QF 4:

  Let $\mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$ and let $\mathbf{P}$ be an $n \times n$ projection matrix of rank r. Then $\mathbf{Z}^\top \mathbf{P}\,\mathbf{Z} \sim \chi^2(r)$.

- ○ $||\mathbf{e}_{\mathrm{b}}||^2 \perp\!\!\!\perp ||\mathbf{e}_{\mathrm{ss}} - \mathbf{e}_{\mathrm{s}}||^2$

- ○ $\dfrac{||\mathbf{e}_{\mathrm{ss}} - \mathbf{e}_{\mathrm{s}}||^2/(s-r)}{||\mathbf{e}_{\mathrm{b}}||^2/(n-p)} = \dfrac{(||\mathbf{e}_{\mathrm{ss}}||^2 - ||\mathbf{e}_{\mathrm{s}}||^2)/(s-r)}{||\mathbf{e}_{\mathrm{b}}||^2/(n-p)} \sim F_{s-r,n-p}$

  - ▷ proof: verify that the definition of $F_{s-r,n-p}$ is satisfied

7.4 Selecting the model

7.4.1 Model selection tools

Model selection based on sequential testing

- ○ statistical tests
  - ▷ t test for testing $\beta_i = 0$ vs. $\beta_i \neq 0$
  - ▷ F test for testing $\mathbf{A}\,\boldsymbol{\beta} = \mathbf{0}$ vs. $\mathbf{A}\,\boldsymbol{\beta} \neq \mathbf{0}$
  - ▷ likelihood ratio test
    - ∗ $2(\max_{\boldsymbol{\theta}_{\mathrm{b}}} \ell\,(\text{Big model}) - \max_{\boldsymbol{\theta}_{\mathrm{s}}} \ell\,(\text{Small model})) \overset{as.}{\sim} \chi^2_{|\boldsymbol{\theta}_{\mathrm{b}}|-|\boldsymbol{\theta}_{\mathrm{s}}|}$
    - ∗ details next semester
- ○ we may start with a big model and sequentially leave out terms that do not appear significant
  - ▷ multiple testing $\Rightarrow$ we do not keep the overall α
    - ∗ often $\alpha > 0.05$ is used at this stage (even $\alpha \approx 0.2$)
    - ∗ the procedure is an ad-hoc one (rather than valid testing)
    - ∗ "clean" ways exist (e.g. error-spending function)
  - ▷ an approach of this kind is often applied when the interest is in $\boldsymbol{\beta}$ and the model is there to explain the phenomenon
- ○ words of caution
  - ▷ $p > 0.05$ does not guarantee the absence of the relationship

- ▷ significance of the terms in the final model may be amplified

Model selection based on "criteria"

- ∘ model selection criterion
  - ▷ a number that describes the overall fit of the model
- ∘ often applied when the interest is in prediction
- ∘ focus is on $||\mathbf{e}||^2 = ||\mathbf{Y} - \hat{\mathbf{Y}}||^2$
- ∘ already seen
  - ▷ coefficient of determination

    $$R^2 = 1 - \frac{||\mathbf{e}||^2}{||\mathbf{Y} - \bar{Y}\mathbf{1}||^2}$$

    - ∗ always bigger for a bigger model
    - ∗ bigger model is not necessarily better, so is the difference big enough to justify the use of the bigger model?
  - ▷ adjusted coefficient of determination

    $$R^2_{adj} = 1 - \frac{||\mathbf{e}||^2/(n-p)}{||\mathbf{Y} - \bar{Y}\mathbf{1}||^2/(n-1)}$$

    - ∗ penalizes for the model complexity

Likelihood-based information criteria

- ∘ model fit versus model complexity trade-off
- ∘ Akaike information criterion
  - ▷ $AIC = -2\ \max_{\boldsymbol{\theta}} \ell\,(\text{model}) + 2 \times |\boldsymbol{\theta}|$
  - ▷ motivation
    - ∗ information theory
    - ∗ prediction
  - ▷ favours bigger models
- ∘ Bayesian information criterion
  - ▷ $BIC = -2\ \max_{\boldsymbol{\theta}} \ell\,(\text{model}) + \log(n) \times |\boldsymbol{\theta}|$
  - ▷ motivation
    - ∗ Bayesian model comparison

      * selection of covariates
    - ▷ favours smaller models

- ∘ smaller is better
- ∘ can be used to compare non-nested models
- ∘ can be used for more general models (cf. next semester)

Mallows's C_P

- ∘ criterion specific for linear regression:
  - ▷ suppose that the full model has $\boldsymbol{\beta}$ of length p
  - ▷ describe the fit (focus on prediction) of its submodel with $\widetilde{\boldsymbol{\beta}}$ of length P
  - ▷ estimate the average mean square error of prediction $\frac{1}{\sigma^2}\sum_{i=1}^{n} \mathsf{E}(\hat{Y}_i - \mathsf{E}\, Y_i)^2$ by

  $$\frac{1}{\widehat{\sigma^2}_{\mathrm{b}}}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 = \frac{||\mathbf{e}_{\mathrm{s}}||^2}{||\mathbf{e}_{\mathrm{b}}||^2/(n-p)}$$

- ∘ $C_P = \dfrac{||\mathbf{e}_{\mathrm{s}}||^2}{||\mathbf{e}_{\mathrm{b}}||^2/(n-p)} - n + 2P$
  - ▷ for the full model: $C_p = p$
  - ▷ models with $C_P \approx P$ are considered good
  - ▷ we may plot C_P against P and choose a small model that has $C_P \approx P$ (if small is preferred)
  - ▷ related to the AIC

7.4.2 Model selection strategies

To leave out or not to leave out?

- ∘ setting $\beta_i = 0$ if the true $\beta_i \neq 0$

  i.e. leaving out a covariate that should have been kept
  - ▷ possible bias in the estimators of β_j for $i \neq j$
  - ▷ possible invalidity of the resulting model (cf. Week 10)

- ∘ allowing $\beta_i \neq 0$ if the true $\beta_i \approx 0$

  i.e. keeping unnecessary covariates in the model

- ▷ possibly worse estimation of β_j for $i \neq j$ and larger confidence intervals (cf. Week 11)
- ▷ possibility of overfitting
- ▷ sometimes/often simple explanations are preferable

- ∘ conclusion
  - ▷ avoid blind automatic model selection procedures if possible

Model selection strategies

- ∘ step-wise procedures based on p-values of the t/F test
  - ▷ backward selection
    - ∗ start with a biggest model, leave out the covariate with the largest p-value, end when p-values for all included covariates are smaller than α_{crit}
  - ▷ forward selection
    - ∗ start with a smallest model, add the covariate with the smallest p-value, end when p-values of all non-included covariates are larger than α_{crit}
  - ▷ step-wise selection
    - ∗ a combination of forward and backward selection
  - ▷ issues
    - ∗ non-exhaustive search
    - ∗ multiple testing; tests invalid unless the smaller model is true
    - ∗ not recommended for prediction
- ∘ step-wise procedures with a model selection criterion
- ∘ exhaustive search with a model selection criterion
  - ▷ e.g. plot C_P or R^2 against the number of predictors

Notes on model selection

- ∘ hierarchical modelling
  - ▷ powers of lower order should be kept in the model if powers of higher order are present
  - ▷ main terms and interactions of lower order should be kept in the model if interactions of higher order are present
  - ▷ there may be a good reason for a non-hierarchical model but such a model is not invariant to affine transformations and rotations of covariates

- several models may fit equally well
  - ▷ if they give qualitatively different answers, reconsider the use of the data to answer the question
- avoid blind automatic model selection procedures if possible
  - ▷ if impossible, choose a selection procedure to fit the purpose of the modelling and carefully examine the final model
- make sure that the models you considered were fitted to the same data

Concluding notes

- there is no best/foolproof way to do the model selection except for common sense and sound understanding of the phenomenon

- *A model should be as simple as possible but no simpler.*

  Albert Einstein

- *All models are wrong but some are useful.*

  George Box

Chapter 8

Model diagnostics

8.1 The problem

8.1.1 Normal linear model

Normal linear model

- $Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k} + \varepsilon_i$, $i \in \{1, \ldots, n\}$
  - ▷ Y_i: outcome, response, output, dependent variable
    - ∗ random variable, we observe a realization y_i
    - ∗ (odezva, závisle proměnná, regresand)
  - ▷ $x_{i,1}, \ldots, x_{i,k}$: covariates, predictors, explanatory variables, input, independent variables
    - ∗ given, known
    - ∗ (nezávisle proměnné, regresory)
  - ▷ $\beta_0, \ldots, \beta_k$: coefficients
    - ∗ unknown
    - ∗ (regresní koeficienty)
  - ▷ ε_i: random error
    - ∗ random variable, unobserved
- $\varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$, $i \in \{1, \ldots, n\}$
  - ▷ $\mathsf{E}\,\varepsilon_i = 0$: no systematic errors
  - ▷ $\mathsf{Var}\,\varepsilon_i = \sigma^2$: same precision

Example: `bloodpress` data

- from `sites.stat.psu.edu/~lsimon/stat501wc/sp05/data/`
- association between the mean arterial blood pressure[mmHg] and age[years], weight[kg], body surface area[m^2], duration of hypertension[years], basal pulse[beats/min], stress
- data:

BP	Age	Weight	BSA	DoH	Pulse	Stress
105	47	85.4	1.75	5.1	63	33
115	49	94.2	2.10	3.8	70	14
...	...	...	...	...	...	...
110	48	90.5	1.88	9.0	71	99
122	56	95.7	2.09	7.0	75	99

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

$$\begin{pmatrix} 105 \\ 115 \\ \dots \\ 110 \\ 122 \end{pmatrix} = \begin{pmatrix} 1 & 47 & 85.4 & 1.75 & 5.1 & 63 & 33 \\ 1 & 49 & 94.2 & 2.10 & 3.8 & 70 & 14 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 48 & 90.5 & 1.88 & 9.0 & 71 & 99 \\ 1 & 56 & 95.7 & 2.09 & 7.0 & 75 & 99 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \dots \\ \beta_6 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_{19} \\ \varepsilon_{20} \end{pmatrix}$$

Example: `fev` data

- from: `http://www.statsci.org/data/general/fev.html`
- question: association between the FEV[l] and Smoking,

  corrected for Age[years], Height[cm] and Gender

- data:

FEV	Age	Height	Gender	Smoking
1.708	9	144.8	Female	Non
1.724	8	171.5	Female	Non
1.720	7	138.4	Female	Non
1.558	9	134.6	Male	Non
...	...	...	...	...
3.727	15	172.7	Male	Current
2.853	18	152.4	Female	Non
2.795	16	160.0	Female	Current
3.211	15	168.9	Female	Non

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

$$\begin{pmatrix} 1.708 \\ 1.724 \\ 1.720 \\ 1.558 \\ \dots \\ 3.727 \\ 2.853 \\ 2.795 \\ 3.211 \end{pmatrix} = \begin{pmatrix} 1 & 9 & 144.8 & 0 & 0 \\ 1 & 8 & 171.5 & 0 & 0 \\ 1 & 7 & 138.4 & 0 & 0 \\ 1 & 9 & 134.6 & 1 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 15 & 172.7 & 1 & 1 \\ 1 & 18 & 152.4 & 0 & 0 \\ 1 & 16 & 160.0 & 0 & 1 \\ 1 & 15 & 168.9 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \dots \\ \beta_5 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \dots \\ \varepsilon_{651} \\ \varepsilon_{652} \\ \varepsilon_{653} \\ \varepsilon_{654} \end{pmatrix}$$

8.1.2 Task for this chapter

Checking the model assumptions

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
  - ▷ outcome $\mathbf{Y}$

    * random vector, we observe a realization $\mathbf{y}$
  - predictors $\mathbf{x}_{,1}, \dots, \mathbf{x}_{,k}$
    * vector of given (known) constants
  - coefficients $\boldsymbol{\beta}$
    * vector of unknown constants
  - error $\boldsymbol{\varepsilon}$
    * unknown random vector, we do not observe its realization
  - assumptions: $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I})$
    * $\mathsf{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$: the expected value of $\mathbf{Y}$ is a linear function of $\boldsymbol{\beta}$
    * $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$: no systematic errors
    * $\mathsf{Var}\,\boldsymbol{\varepsilon} = \sigma^2 \mathbf{I}$: independence and same precision

- task: do the assumptions appear to be satisfied?
- Note: if they are not, inference is not valid ...

8.2 Random errors and residuals

Random errors

Random errors in the normal linear model

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
- assumptions
  - $\mathsf{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$: the expected value of $\mathbf{Y}$ is a linear function of $\boldsymbol{\beta}$
  - $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I})$
    * $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$: no systematic errors
    * $\mathsf{Var}\,\boldsymbol{\varepsilon} = \sigma^2 \mathbf{I}$: independence and the same precision
- we need to verify the assumptions on
  - expectation: $\mathsf{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$, i.e. $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$
  - variance: $\mathsf{Var}\,\boldsymbol{\varepsilon} = \sigma^2 \mathbf{I}$
  - distribution: $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I})$
- all assumptions are made on <u>unobserved</u> random errors $\boldsymbol{\varepsilon}$
- fitted model: $\mathbf{Y} = \hat{\mathbf{Y}} + (\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{e}$

- residuals $\mathbf{e}$ sometimes seen as "estimates" of $\boldsymbol{\varepsilon}$
  - ▷ $\boldsymbol{\varepsilon}$ is an unobserved random vector, not a parameter (constant)
  - ▷ $\mathbf{e}$ are not estimates in the usual sense

Residuals

Residuals in the normal linear model

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
- fitted model: $\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{e} = \mathbf{H}\,\mathbf{Y} + (\mathbf{I} - \mathbf{H})\,\mathbf{Y}$
- $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,(\mathbf{I} - \mathbf{H}))$
  - ▷ proof: cf. Week 6 or use MVN 3
  - ▷ $\mathsf{rank}(\mathbf{I} - \mathbf{H}) = n - p$ if $\mathsf{rank}(\mathbf{X}) = p$
    - ∗ $\mathbf{e} \overset{d}{=} \mathbf{A}\,\mathbf{Z}$ for an $(n-p)$-dimensional $\mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$
      $\mathbf{I} - \mathbf{H} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$ (spec. dec.) $\Rightarrow \mathbf{A} = \mathbf{U}_{n\times(n-p)}\,\boldsymbol{\Lambda}^{1/2}_{(n-p)\times(n-p)}$
      (cf. Week 4 or use MVN 3)
- if the assumptions are satisfied, residuals are
  - ▷ zero-mean
  - ▷ with unequal variances: $\mathsf{Var}\,e_i = \sigma^2\,(1 - h_{i,i})$
  - ▷ with a degenerate normal distribution
  - ▷ correlated: $\mathsf{Cor}(e_i, e_j) = -\frac{h_{i,j}}{\sqrt{(1-h_{i,i})\,(1-h_{j,j})}}$
- compare to $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$ ...

Standardized residuals in the normal linear model

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
  - ▷ $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
- fitted model: $\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{e} = \mathbf{H}\,\mathbf{Y} + (\mathbf{I} - \mathbf{H})\,\mathbf{Y}$
  - ▷ $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,(\mathbf{I} - \mathbf{H}))$
- to check the assumptions, we often use

  standardized residuals $r_i = \dfrac{e_i}{\sqrt{\widehat{\sigma^2}\,(1 - h_{i,i})}}$, $1 \le i \le n$

- if the assumptions are satisfied
  - ▷ we expect that $r_i \approx \mathrm{N}(0,1)$
    - ∗ it can be shown that $\mathsf{E}\, r_i = 0$ and $\mathsf{Var}\, r_i = 1$ (some technical work needed to prove this)
    - ∗ we did not derive the distribution of r_i's
    - ∗ we did not try to get rid of the correlation
- compare to $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$ ...

8.3 Model diagnostics I: checking the assumptions

8.3.1 General principles

Checking the assumptions

- Specifying the possible departures
  - ▷ need to specify in what sense the assumption might be violated
  - ▷ if the assumption is H_0, need to specify H_1

1. Graphical checking
   - plots that allows us to "see" departures from the assumptions
   - based on residuals ($\mathbf{e}$ or $\mathbf{r}$)
2. Testing the validity of assumptions
   - usually by fitting a more general model that allows them not to be satisfied and testing whether the generalization is needed
   - useful as numerical indications BUT
   - we cannot "prove the null hypothesis"
   - problems with the validity of inference:
     - ▷ chains of tests and multiple testing
     - ▷ assumptions on assumptions
     - ▷ we should *know* in advance they are satisfied

Overall check: residuals versus fitted values

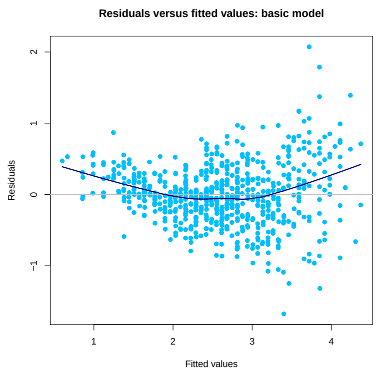
- $\mathbf{e} \perp \hat{\mathbf{Y}}$ by definition

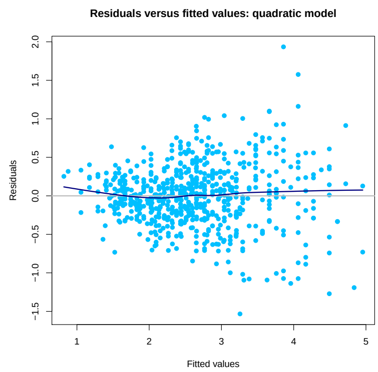
- no systematic patterns should appear between $\mathbf{e}$ and $\hat{\mathbf{Y}}$
- example: `fev` data
  - ▷ basic model:

$$Y_i = \beta_0 + \beta_1 \times \mathsf{Height}_i + \beta_2 \times \mathbb{I}\{\text{the } i^{\text{th}} \text{ child is male}\} + \varepsilon_i, \ 1 \le i \le 654$$

  - ▷ quadratic model:

$$\begin{aligned} Y_i = \beta_0 + \beta_1 \times \mathsf{Height}_i + \beta_2 \times \mathsf{Height}_i^2 + \\ + \beta_3 \times \mathbb{I}\{\text{the } i^{\text{th}} \text{ child is male}\} + \beta_4 \times \mathsf{Height}_i \mathbb{I}\{\text{the } i^{\text{th}} \text{ child is male}\} + \\ + \beta_5 \times \mathsf{Height}_i^2 \mathbb{I}\{\text{the } i^{\text{th}} \text{ child is male}\} + \varepsilon_i, \ 1 \le i \le 654 \end{aligned}$$





8.3.2 Assumptions on the expectation

Checking $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$, i.e. $\mathsf{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$

- suspected departures from the assumption
  - ▷ incorrectly specified form of dependence
    - ∗ plot $\mathbf{e}$ against the included covariates
    - ∗ $\mathbf{e} \perp \mathbf{x}_{,i}$, $1 \le i \le p$, by definition
    - ∗ no systematic patterns should appear between $\mathbf{e}$ and $\mathbf{x}_{,i}$
    - ∗ a trend indicates a dependence not captured by the model
    - ∗ a formal test: fit a more complicated dependence and test against the original model
  - ▷ missing covariates
    - ∗ plot $\mathbf{e}$ against covariates that are not included in the model
    - ∗ no systematic patterns should appear
    - ∗ a trend indicates a dependence not captured by the model
    - ∗ a formal test: fit a larger model and test the effect of the additional covariate

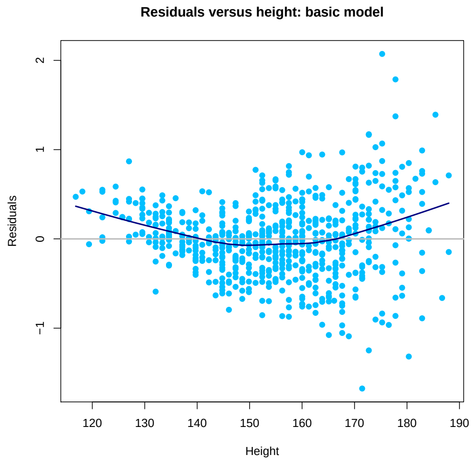
Incorrectly specified form of dependence

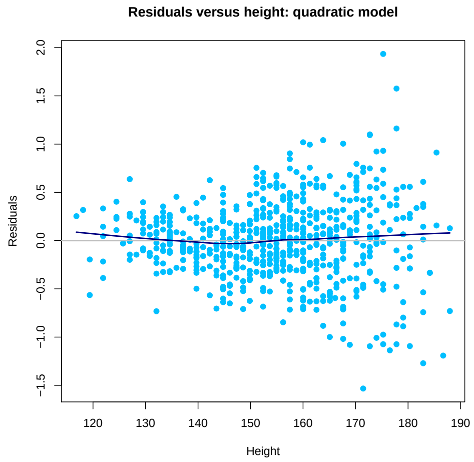
- example: `fev` data
  - basic model:

$$Y_i = \beta_0 + \beta_1 \times \mathsf{Height}_i + \beta_2 \times \mathbb{I}\{\text{the } i^{\text{th}} \text{ child is male}\} + \varepsilon_i, \ 1 \le i \le 654$$

  - quadratic model:

$$\begin{aligned} Y_i = {} & \beta_0 + \beta_1 \times \mathsf{Height}_i + \beta_2 \times \mathsf{Height}_i^2 + \\ & + \beta_3 \times \mathbb{I}\{\text{the } i^{\text{th}} \text{ child is male}\} + \beta_4 \times \mathsf{Height}_i \mathbb{I}\{\text{the } i^{\text{th}} \text{ child is male}\} + \\ & + \beta_5 \times \mathsf{Height}_i^2 \mathbb{I}\{\text{the } i^{\text{th}} \text{ child is male}\} + \varepsilon_i, \ 1 \le i \le 654 \end{aligned}$$

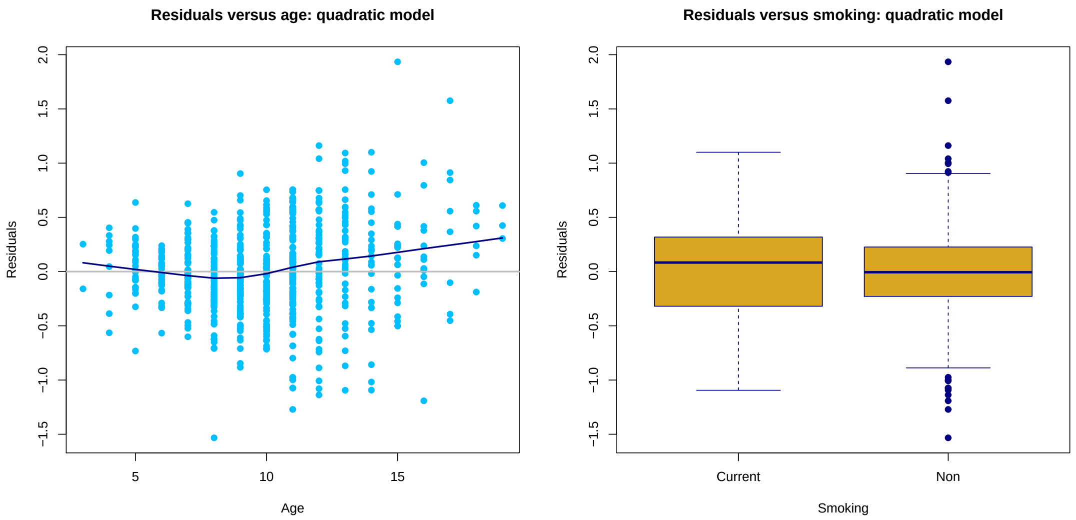




Missing covariates

- example: `fev` data
- quadratic model:

$$\begin{aligned} Y_i = {} & \beta_0 + \beta_1 \times \mathsf{Height}_i + \beta_2 \times \mathsf{Height}_i^2 + \\ & + \beta_3 \times \mathbb{I}\{\text{the } i^{\text{th}} \text{ child is male}\} + \beta_4 \times \mathsf{Height}_i \mathbb{I}\{\text{the } i^{\text{th}} \text{ child is male}\} + \\ & + \beta_5 \times \mathsf{Height}_i^2 \mathbb{I}\{\text{the } i^{\text{th}} \text{ child is male}\} + \varepsilon_i, \ 1 \le i \le 654 \end{aligned}$$



8.3.3 Assumptions on the variance

Checking $\mathsf{Var}\,\varepsilon = \sigma^2\mathbf{I}$: homoskedasticity

- suspected departures from the assumption
  - ▷ variance changing with fitted values (usually increasing)
    - ∗ plot standardized residuals (usually square root of the absolute value) against fitted values
    - ∗ no pattern should appear
  - ▷ variance changing with covariates
    - ∗ plot standardized residuals (usually square root of the absolute value) against covariates
    - ∗ no pattern should appear
    - ∗ a formal test: studentized Breusch–Pagan test
  - ▷ subgroups with the same within-group variance
    - ∗ plot boxplots of standardized residuals by groups
    - ∗ boxes should be of approximately equal sizes
    - ∗ a formal test: fit a more general model and test against the original model

Breusch–Pagan test

- original model
  - ▷ $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

  ▷ $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I})$

- more general model

  ▷ $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

  ▷ $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2))$

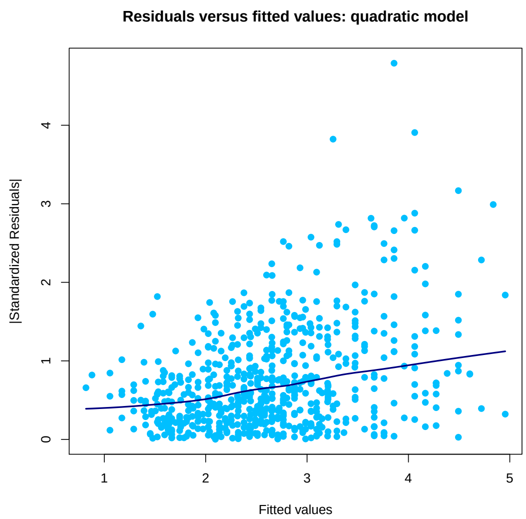
  ▷ $\sigma^2 = \mathbf{X}\boldsymbol{\alpha}$

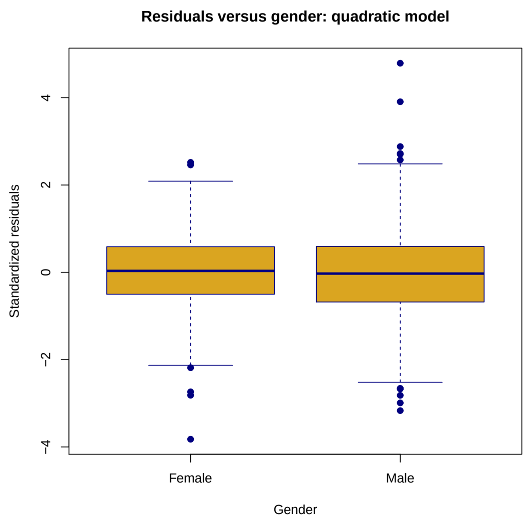
- Breusch–Pagan test: test $\boldsymbol{\alpha}_{2:p} = \mathbf{0}$ in the more general model
- studentized Breusch–Pagan test less sensitive to the assumption of normality
- more general versions of the Breusch–Pagan test and more general tests exist

Checking $\mathsf{Var}\,\boldsymbol{\varepsilon} = \sigma^2 \mathbf{I}$: homoskedasticity

- example: `fev` data
- quadratic model:

$$Y_i = \beta_0 + \beta_1 \times \mathsf{Height}_i + \beta_2 \times \mathsf{Height}_i^2 +$$
$$+ \beta_3 \times \mathbb{I}\{\text{the } i^{\text{th}} \text{ child is male}\} + \beta_4 \times \mathsf{Height}_i \mathbb{I}\{\text{the } i^{\text{th}} \text{ child is male}\} +$$
$$+ \beta_5 \times \mathsf{Height}_i^2 \mathbb{I}\{\text{the } i^{\text{th}} \text{ child is male}\} + \varepsilon_i, \ 1 \le i \le 654$$





Checking $\mathsf{Var}\,\boldsymbol{\varepsilon} = \sigma^2 \mathbf{I}$: independence

- suspected departures from the assumption

  ▷ clustering

- suspected e.g. when several data points collected from one individual (e.g. same individuals followed over time)
- plot boxplots of residuals by the suspected groups
- no pattern should appear
- a formal test: fit a more general model allowing for the within-group dependence and test against the original model

▷ serial correlation

- suspected when data collected over time or space
- plot e_i against e_{i-1}
- no pattern should appear
- plot the (partial) autocorrelation function
- a formal test: fit a more general model and test against the original model
- a formal test: Durbin–Watson test

Durbin–Watson test

- original model
  - ▷ $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
  - ▷ $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$
- more general model
  - ▷ $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
  - ▷ $\varepsilon_i = \rho\,\varepsilon_{i-1} + w_i$, $w_i \overset{\text{iid}}{\sim} (0, \sigma^2)$, $|\rho| < 1$

    (autoregression of the first order on the error terms)
- Durbin–Watson test: test $\rho = 0$ against $\rho > 0$ in the more general model
- also possible to test $\rho = 0$ against $\rho < 0$ and $\rho = 0$ against $\rho \neq 0$
- more general tests available

Time series models

- time series is a random sequence $\{X_t, t \in \mathbb{Z}\}$
  - ▷ stationary if $\mathsf{E}\, X_t = \mu$, $\mathsf{Var}\, X_t = \sigma^2$, $\mathsf{Cov}(X_t, X_{t+s}) = \gamma(s)$
- The autocovariance function of a stationary random sequence $\{X_t, t \in \mathbb{Z}\}$ is defined as $\gamma(h) = \mathsf{Cov}(X_t, X_{t+h})$, $h \in \mathbb{Z}$.

- The autocorrelation function (ACF) is defined as $\rho(h) = \mathsf{Cor}(X_t, X_{t+h}) = \gamma(h)/\gamma(0)$, $h \in \mathbb{Z}$.
- The partial autocorrelation function (PACF) is defined as $\alpha(1) = \mathsf{Cor}(X_t, X_{t+1}) = \rho(1)$ and $\alpha(h) = \mathsf{Cor}(X_t - \hat{X}_t, X_{t+h} - \hat{X}_{t+h})$, $h = 2, 3, \ldots$, where $\hat{X}_t$ and $\hat{X}_{t+h}$ are the fitted values from the linear regressions $X_t \sim X_{t+1}, \ldots, X_{t+h-1}$ and $X_{t+h} \sim X_{t+1}, \ldots, X_{t+h-1}$.

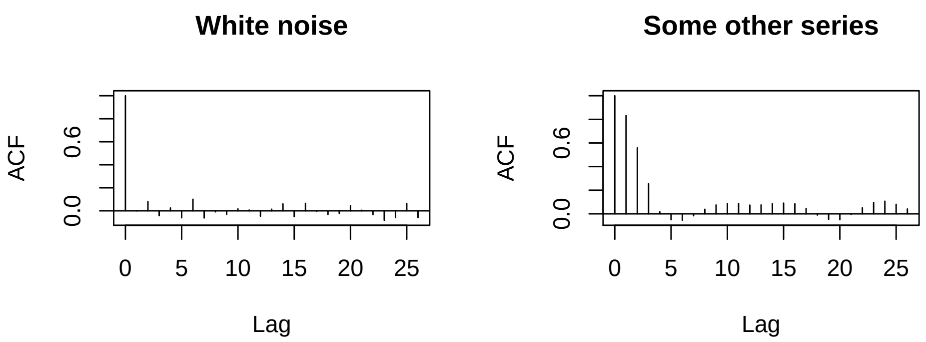
ACF and PACF for ARMA models

- special time series models
- Let $\{\epsilon_t\} \overset{iid}{\sim} (0, \sigma^2)$. Then $\{X_t, t \in \mathbb{Z}\}$ is
  - ▷ AR(p) if
    - ∗ $X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \epsilon_t$;
  - ▷ MA(q) if
    - ∗ $X_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$;
  - ▷ ARMA(p, q) if
    - ∗ $X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$.
- ACF and PACF for AR/MA

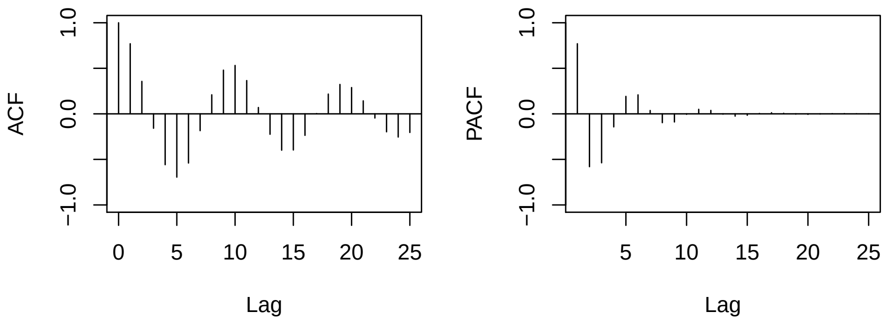
	ACF	PACF
AR(p)	Exponential decay	Cuts off after lag p
MA(q)	Cuts off after lag q	Exponential decay
ARMA(p, q)	Exponential decay	Exponential decay

ACFs and PACFs

- simulated ACFs

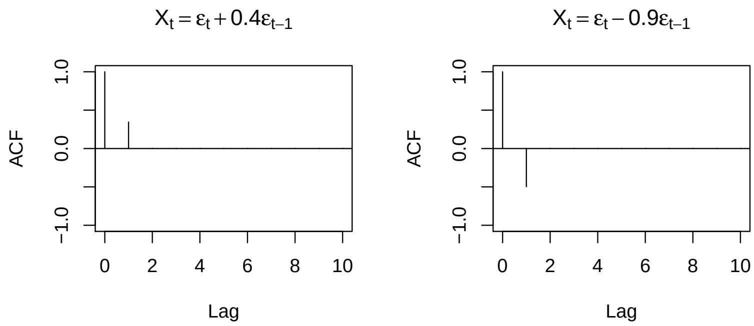


- theoretical ACF and PACF for an ARMA

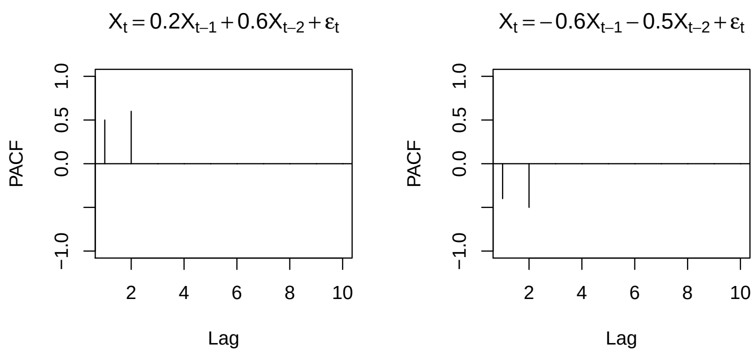


ACFs and PACFs

- theoretical ACF for MA(1)



- theoretical PACF for AR(2)



Other types of dependence

- spatial correlation diagnosed via semivariogram
  - ▷ for a stationary isotropic random field $\{Z(\mathbf{x}); \mathbf{x} \in \mathbb{R}^2\}$, semivariogram is
    - ∗ $\gamma(\mathbf{x}, \mathbf{y}) = \frac{1}{2} \mathsf{Var}(Z(\mathbf{x}) - Z(\mathbf{y})) = \frac{1}{2} \mathsf{E}(Z(\mathbf{x}) - Z(\mathbf{y}))^2 = \gamma(h)$, where $h = ||\mathbf{x} - \mathbf{y}||^2$
- clustering (boxplots of residuals by group)

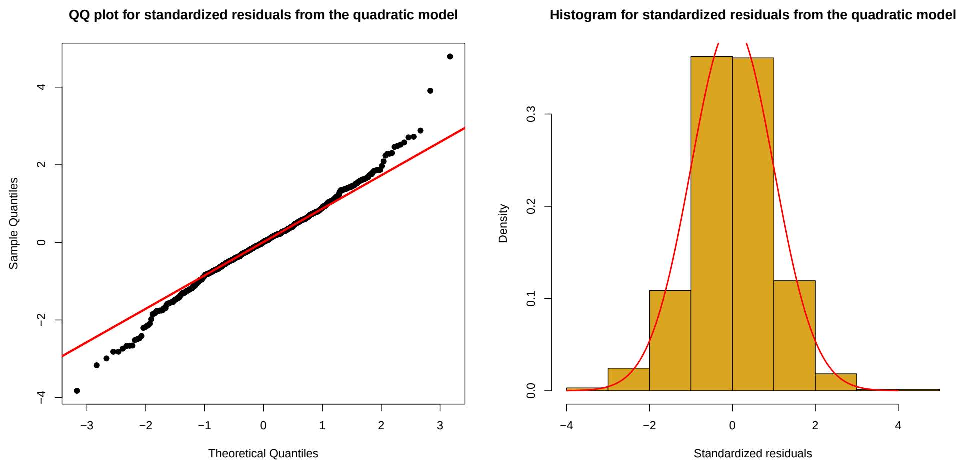
8.3.4 Assumptions on the distribution

Checking $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$

- suspected departures from the assumption
  - ▷ non-normal distribution
    - ∗ skewed distribution
    - ∗ heavy-tailed distribution
- plot a QQ plot for (standardized) residuals
- plot a histogram for (standardized) residuals
- formal tests: Shapiro–Wilk test, Kolmogorov–Smirnov test
  - ▷ warning: valid for iid's (and residuals are not iid's)

QQ plot and histogram

- QQ plot (preferred)
  - ▷ quantiles of $\mathrm{N}(0, 1)$ against empirical quantiles
  - ▷ should be near a straight line
  - ▷ problems to look for
    - ∗ S shape (heavy tails)
    - ∗ an arc (skewness)



Shapiro–Wilk test and Kolmogorov–Smirnov test

- valid for iid's (and residuals are not iid's)

- ▷ Shapiro–Wilk test
  - ∗ can be seen as a numerical summary of the QQ plot
  - ∗ rather a strong one

```
> shapiro.test(rstandard(model.basic.quad))

Shapiro-Wilk normality test

data:  rstandard(model.basic.quad)
W = 0.9865, p-value = 9.713e-06

> shapiro.test(rstandard(model.basic.quad)[sample(1:654, 50)])

Shapiro-Wilk normality test

data:  rstandard(model.basic.quad)[sample(1:654, 50)]
W = 0.97011, p-value = 0.2338
```

- ▷ Kolmogorov–Smirnov test
  - ∗ rather a weak one

Importance of the assumption

- ∘ large-sample distribution of $\widehat{\boldsymbol{\beta}}$
  - ▷ Let $\mathbf{X}_n$; $n \in \mathbb{N}$ be a sequence of $n \times p$ design matrices of full rank defining a sequence of linear models $\mathbf{Y}_n = \mathbf{X}_n \boldsymbol{\beta} + \boldsymbol{\varepsilon}_n$ with $\boldsymbol{\varepsilon}_n \sim (\mathbf{0}, \sigma^2 \mathbf{I}_n)$. If $\max_{1 \le i \le n} \mathbf{x}_{i,}^\top (\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \mathbf{x}_{i,} \xrightarrow[n\to\infty]{} 0$ then

$$(\mathbf{X}_n^\top \mathbf{X}_n)^{1/2} \, (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow[n\to\infty]{d} \mathrm{N}(\mathbf{0}, \sigma^2 \, \mathbf{I}),$$

  where $\widehat{\boldsymbol{\beta}}_n = (\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \mathbf{X}_n^\top \mathbf{Y}_n$.

- ∘ normality not crucial in large samples unless there are special observations

8.4 Model diagnostics II: influential and unusual observations

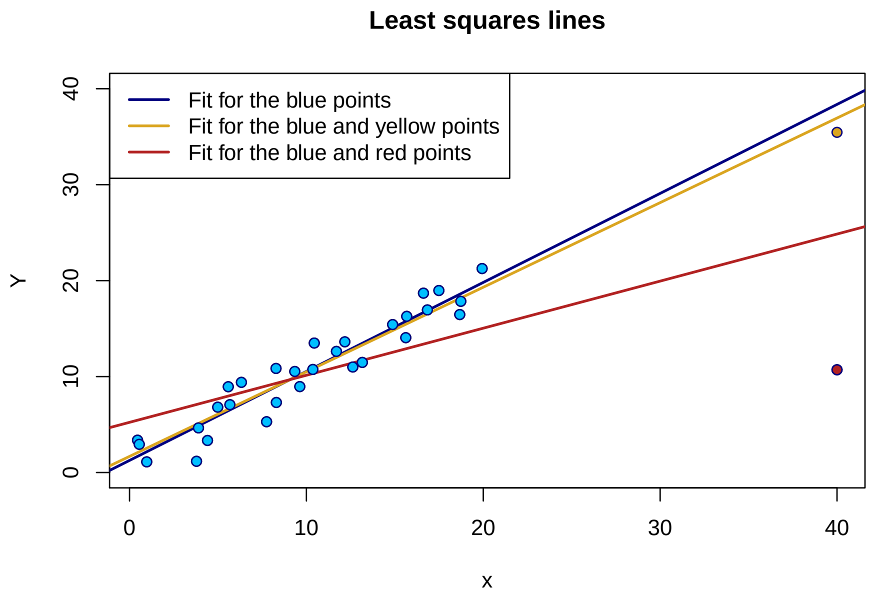
8.4.1 Observations to look at

Leverage

- ∘ $\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{e} = \mathbf{H}\,\mathbf{Y} + (\mathbf{I} - \mathbf{H})\,\mathbf{Y}$
- ∘ $\mathsf{Var}\,\hat{Y}_i = h_{i,i}$ . . . leverage
- ∘ $\mathbf{H} = \mathbf{X}\,(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ and $\mathsf{rank}(\mathbf{H}) = \mathrm{tr}(\mathbf{H}) = p$
- ∘ $h_{i,i} = \mathbf{x}_{i,}^\top\,(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_{i,}$ and $\sum_{i=1}^n h_{i,i} = p$

- the variance of $\hat{Y}_i$ determined by the corresponding covariates
- we want all observations to contribute $\approx$ equally to the fit
  - ▷ we want that $h_{i,i} \approx \frac{p}{n}$
- if $h_{i,i}$ much larger for some i, the fit may be influenced by $(Y_i, \mathbf{x}_{i,})$ much more than by the other observations
- observations with $h_{i,i} > \frac{2p}{n}$ should be checked

Potentially influential and influential observations



- both points have a high leverage, but only one is influential

Model with an excluded observation

- consider a model $\mathbf{Y}_{[-i]} = \mathbf{X}_{[-i]}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{[-i]}$ without the i^{th} observation
- fit the model
  - ▷ compute $\widehat{\boldsymbol{\beta}}_{[-i]}$ and $\widehat{\sigma^2}_{[-i]}$
- compute $\hat{y}_{[-i]} = \mathbf{x}_{i,}^\top \widehat{\boldsymbol{\beta}}_{[-i]}$
  - ▷ prediction of y_i based on the model without the i^{th} observation
- if $y_i - \hat{y}_{[-i]}$ is large, the i^{th} observation is an outlier
  - ▷ how large is “too large?”

- ▷ $\mathsf{Var}(y_i - \hat{y}_{[-i]}) = \sigma^2(1 + \mathbf{x}_{i,}^\top(\mathbf{X}_{[-i]}^\top \mathbf{X}_{[-i]})^{-1}\mathbf{x}_{i,})$
- ▷ define jackknife residuals $t_i = \dfrac{y_i - \hat{y}_{[-i]}}{\sqrt{\widehat{\sigma^2}_{[-i]}(1 + \mathbf{x}_{i,}^\top(\mathbf{X}_{[-i]}^\top \mathbf{X}_{[-i]})^{-1}\mathbf{x}_{i,})}}$
- ▷ there is a simpler equivalent formula that does not require fitting n models with excluded observations

Influential and unusual observations

- ∘ in the normal linear model:
  - ▷ $t_i \sim t_{n-p-1}$
- ∘ we can test whether and observation is an outlier
  - ▷ heavy multiple testing ⇝ Bonferroni correction
    - ∗ use $t_{n-p-1}(1 - \alpha/(2n))$ instead of $t_{n-p-1}(1 - \alpha/2)$
- ∘ to evaluate whether the observation is influential
  - ▷ Cook's distance: $d_i = \dfrac{1}{p\,\hat{\sigma}^2}||\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{[-i]}||^2 = \dfrac{1}{p}\, r_i^2 \dfrac{h_{i,i}}{1 - h_{i,i}}$
  - ▷ how large is "too large"?
    - ∗ rule of thumb: $d_i \geq 0.5$ deserve some attention
      $d_i \geq 1$ ⇝ highly influential observation

Chapter 9

Reduced-rank design matrix and multicolllinearity

9.1 The problem

9.1.1 Normal linear model

Normal linear model

- $Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k} + \varepsilon_i$, $i \in \{1, \ldots, n\}$
  - ▷ Y_i: outcome, response, output, dependent variable
    - * random variable, we observe a realization y_i
    - * (odezva, závisle proměnná, regresand)
  - ▷ $x_{i,1}, \ldots, x_{i,k}$: covariates, predictors, explanatory variables, input, independent variables
    - * given, known
    - * (nezávisle proměnné, regresory)
  - ▷ $\beta_0, \ldots, \beta_k$: coefficients
    - * unknown
    - * (regresní koeficienty)
  - ▷ ε_i: random error
    - * random variable, unobserved
- $\varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$, $i \in \{1, \ldots, n\}$
  - ▷ $\mathsf{E}\,\varepsilon_i = 0$: no systematic errors
  - ▷ $\mathsf{Var}\,\varepsilon_i = \sigma^2$: same precision

Example: `bloodpress` data

- from `sites.stat.psu.edu/~lsimon/stat501wc/sp05/data/`
- association between the mean arterial blood pressure[mmHg] and age[years], weight[kg], body surface area[m^2], duration of hypertension[years], basal pulse[beats/min], stress
- data:

BP	Age	Weight	BSA	DoH	Pulse	Stress
105	47	85.4	1.75	5.1	63	33
115	49	94.2	2.10	3.8	70	14
...	...	...	...	...	...	...
110	48	90.5	1.88	9.0	71	99
122	56	95.7	2.09	7.0	75	99

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

$$\begin{pmatrix} 105 \\ 115 \\ \dots \\ 110 \\ 122 \end{pmatrix} = \begin{pmatrix} 1 & 47 & 85.4 & 1.75 & 5.1 & 63 & 33 \\ 1 & 49 & 94.2 & 2.10 & 3.8 & 70 & 14 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 48 & 90.5 & 1.88 & 9.0 & 71 & 99 \\ 1 & 56 & 95.7 & 2.09 & 7.0 & 75 & 99 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \dots \\ \beta_6 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_{19} \\ \varepsilon_{20} \end{pmatrix}$$

`https://ww2.amstat.org/publications/jse/v13n2/datasets.kahn.html`

Example: `fev` data

- from: `http://www.statsci.org/data/general/fev.html`
- question: association between the FEV[l] and Smoking,

  corrected for Age[years], Height[cm] and Gender
- data:

FEV	Age	Height	Gender	Smoking
1.708	9	144.8	Female	Non
1.724	8	171.5	Female	Non
1.720	7	138.4	Female	Non
1.558	9	134.6	Male	Non
...	...	...	...	...
3.727	15	172.7	Male	Current
2.853	18	152.4	Female	Non
2.795	16	160.0	Female	Current
3.211	15	168.9	Female	Non

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

$$\begin{pmatrix} 1.708 \\ 1.724 \\ 1.720 \\ 1.558 \\ \dots \\ 3.727 \\ 2.853 \\ 2.795 \\ 3.211 \end{pmatrix} = \begin{pmatrix} 1 & 9 & 144.8 & 0 & 0 \\ 1 & 8 & 171.5 & 0 & 0 \\ 1 & 7 & 138.4 & 0 & 0 \\ 1 & 9 & 134.6 & 1 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 15 & 172.7 & 1 & 1 \\ 1 & 18 & 152.4 & 0 & 0 \\ 1 & 16 & 160.0 & 0 & 1 \\ 1 & 15 & 168.9 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \dots \\ \beta_5 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \dots \\ \varepsilon_{651} \\ \varepsilon_{652} \\ \varepsilon_{653} \\ \varepsilon_{654} \end{pmatrix}$$

9.1.2 Task for this chapter

Rank-deficiency/near-rank deficiency of X

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
  - ▷ outcome $\mathbf{Y}$
    - ∗ random vector, we observe a realization $\mathbf{y}$
  - ▷ predictors $\mathbf{x}_{,1}, \dots, \mathbf{x}_{,k}$
    - ∗ vector of given (known) constants
  - ▷ coefficients $\boldsymbol{\beta}$
    - ∗ vector of unknown constants
  - ▷ error $\boldsymbol{\varepsilon}$
    - ∗ unknown random vector, we do not observe its realization
  - ▷ assumptions: $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I})$
    - ∗ $\mathsf{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$: the expected value of $\mathbf{Y}$ is a linear function of $\boldsymbol{\beta}$
    - ∗ $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$: no systematic errors
    - ∗ $\mathsf{Var}\,\boldsymbol{\varepsilon} = \sigma^2 \mathbf{I}$: independence and same precision
- task: so far we have assumed that $\mathsf{rank}(\mathbf{X}) = p$

  What happens if $\mathsf{rank}(\mathbf{X}) < p$ or “nearly so”?

9.2 Rank-deficient design matrix

9.2.1 Rank-deficient design matrix

Full-rank design matrix X

- design matrix $\mathbf{X}$ is $n \times p$, $n > p$
- SVD: $\mathbf{X} = \underbrace{\mathbf{U}}_{n\times n}\underbrace{\boldsymbol{\Sigma}}_{n\times p}\underbrace{\mathbf{V}^\top}_{p\times p}$
- if all covariates are linearly independent
  - ▷ $\mathsf{rank}(\mathbf{X}) = p$
  - ▷ $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p > 0$
  - ▷ thin SVD: $\mathbf{X} = \underbrace{\mathbf{U}_1}_{n\times p}\underbrace{\boldsymbol{\Sigma}_1}_{p\times p}\underbrace{\mathbf{V}^\top}_{p\times p}$
  - ▷ the columns generate a p-dimensional space $\mathsf{im}(\mathbf{X})$

  * $\{\mathbf{x}_{,1}, \ldots, \mathbf{x}_{,p}\}$ is a basis of $\mathsf{im}(\mathbf{X})$
  * $\{\mathbf{u}_{,1}, \ldots, \mathbf{u}_{,p}\}$ is an orthonormal basis of $\mathsf{im}(\mathbf{X})$
  * $\mathbf{H} = \mathbf{U}_1\mathbf{U}_1^\top = \mathbf{X}\,(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is a projection matrix on $\mathsf{im}(\mathbf{X})$

Rank-deficient design matrix $\mathbf{X}$

- design matrix $\mathbf{X}$ is $n \times p$, $n > p$
- SVD: $\mathbf{X} = \underbrace{\mathbf{U}}_{n\times n}\underbrace{\mathbf{\Sigma}}_{n\times p}\underbrace{\mathbf{V}^\top}_{p\times p}$
- if covariates are not linearly independent
  - ▷ $\mathsf{rank}(\mathbf{X}) = r < p$
  - ▷ $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0 = \sigma_{r+1} = \ldots = \sigma_p$
  - ▷ compact SVD: $\mathbf{X} = \underbrace{\mathbf{U}_1}_{n\times r}\underbrace{\mathbf{\Sigma}_1}_{r\times r}\underbrace{\mathbf{V}^\top}_{r\times r}$
  - ▷ the columns generate an r-dimensional space $\mathsf{im}(\mathbf{X})$
    * $\{\mathbf{u}_{,1}, \ldots, \mathbf{u}_{,r}\}$ is an orthonormal basis of $\mathsf{im}(\mathbf{X})$
    * $\mathbf{H} = \mathbf{U}_1\mathbf{U}_1^\top = \mathbf{X}\,(\mathbf{X}^\top\mathbf{X})^{+}\mathbf{X}^\top$ is a projection matrix on $\mathsf{im}(\mathbf{X})$

$\widehat{\beta}$ motivated by orthogonal projection (reminder)

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon}$ unknown, $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$
- idea: set $\boldsymbol{\varepsilon} \stackrel{!}{=} \mathbf{0}$ and solve $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$ w.r.t. $\boldsymbol{\beta}$
  - ▷ then $\underbrace{\mathbf{Y}}_{n\times 1} \stackrel{!}{=} \underbrace{\mathbf{X}}_{n\times p}\underbrace{\boldsymbol{\beta}}_{p\times 1}$
  - ▷ n linear equations with p unknowns and $n > p$

    $\Rightarrow$ a solution exists only if $\mathbf{Y} \in \mathsf{im}(\mathbf{X})$
- modified idea: find $\hat{\mathbf{Y}} \in \mathsf{im}(\mathbf{X})$ such that $||\mathbf{Y} - \hat{\mathbf{Y}}||^2$ is the smallest possible and solve $\hat{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta}$ w.r.t. $\boldsymbol{\beta}$
  - ▷ then $\hat{\mathbf{Y}}$ is the orthogonal projection of $\mathbf{Y}$ onto $\mathsf{im}(\mathbf{X})$
  - ▷ projection matrix onto $\mathsf{im}(\mathbf{X})$ is $\underbrace{\mathbf{H}}_{\text{hat matrix}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{+}\mathbf{X}^\top$
  - ▷ solving $\hat{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta}$ is solving $\mathbf{X}\,(\mathbf{X}^\top\mathbf{X})^{+}\mathbf{X}^\top\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$
  - ▷ estimate $\boldsymbol{\beta}$ by $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{+}\mathbf{X}^\top\mathbf{Y}$
  - ▷ but $\widehat{\boldsymbol{\beta}}$ is the unique solution of $\hat{\mathbf{Y}} = \mathbf{X}\,\boldsymbol{\beta}$ iff $\mathsf{rank}(\mathbf{X}) = p$

- and then $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$

$\widehat{\beta}$ as least squares estimator (reminder)

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon}$ unknown, $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$
- idea: make the residuals as small as possible
  - minimize $||\boldsymbol{\varepsilon}||^2 = \sum_{i=1}^n \varepsilon_i^2$ w.r.t. $\boldsymbol{\beta}$
    $\rightsquigarrow$ Least Squares Estimator (LSE) $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n \varepsilon_i^2$
  - also called the OLS (Ordinary Least Squares) solution
- computation:
  - $\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$
  - $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 = \arg\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$
- look for the minimum by differentiating:
  - $\frac{\partial}{\partial\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \overset{!}{=} 0$
  - $-2\,\mathbf{X}^\top\mathbf{Y} + 2\,\mathbf{X}^\top\mathbf{X}\,\boldsymbol{\beta} \overset{!}{=} 0$
  - $\mathbf{X}^\top\mathbf{X}\,\boldsymbol{\beta} \overset{!}{=} \mathbf{X}^\top\mathbf{Y}$: normal equations
- normal equations have unique solution iff $\mathsf{rank}(\mathbf{X}) = p$: then
  - the solution is $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$
  - $\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = 2\,\mathbf{X}^\top\mathbf{X} \succ 0$ for all $\boldsymbol{\beta}$
    $\Rightarrow$ the solution is the minimum $\Rightarrow \widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$

If $\mathsf{rank}(\mathbf{X}) = r < p$

- orthogonal projection approach
  - $\hat{\mathbf{Y}}$ exists and is unique
  - $\hat{\boldsymbol{\beta}}$ such that $\hat{\mathbf{Y}} = \mathbf{X}\,\hat{\boldsymbol{\beta}}$ is a vector of coordinates of $\hat{\mathbf{Y}} \in \mathsf{im}(\mathbf{X})$ w.r.t. $\{\mathbf{x}_{,1}, \ldots, \mathbf{x}_{,p}\}$
    - if $\{\mathbf{x}_{,1}, \ldots, \mathbf{x}_{,p}\}$ is not a basis of $\mathsf{im}(\mathbf{X})$, $\hat{\boldsymbol{\beta}}$ is not unique
  - $\{\hat{\boldsymbol{\beta}};\ \hat{\mathbf{Y}} = \mathbf{X}\,\hat{\boldsymbol{\beta}}\}$ is a linear subspace of $\mathbb{R}^p$ of dimension $p - r$
  - neither $\hat{\mathbf{Y}}$ nor $||\mathbf{Y} - \hat{\mathbf{Y}}||^2$ depend on the choice of $\hat{\boldsymbol{\beta}}$
- ordinary least squares approach
  - normal equations $\mathbf{X}^\top\mathbf{X}\,\boldsymbol{\beta} = \mathbf{X}^\top\mathbf{Y}$ are consistent

    * $\mathsf{rank}(\mathbf{X}^\top\mathbf{X}) = \mathsf{rank}((\mathbf{X}^\top\mathbf{Y}, \mathbf{X}^\top\mathbf{X}))$
- ▷ normal equations have infinitely many solutions
    * the linear subspace of $\mathbb{R}^p$ of dimension $p - r$
- ▷ the minimum $\min_{\boldsymbol{\beta}} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2$ is attained for each of the solutions and its value is the same for all the solutions
    * the $||\mathbf{Y} - \hat{\mathbf{Y}}||^2$
- ▷ proofs can be found in *Anděl: Základy matematické statistiky*

9.2.2 Identifiability

Identifiable parameters

- $\hat{\mathbf{Y}}$ and $||\mathbf{Y} - \hat{\mathbf{Y}}||^2 = \min_{\boldsymbol{\beta}} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2$ does not depend on $\hat{\boldsymbol{\beta}}$
- any other quantities with such properties?

Theorem. *Let $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\mathbf{X}$ is an $n \times p$ matrix, $\mathsf{rank}(\mathbf{X}) = r < p$, $\boldsymbol{\beta} \in \mathbb{R}^p$, and $\boldsymbol{\varepsilon}$ is an n-dimensional random vector with $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$ and $\mathsf{Var}\,\boldsymbol{\varepsilon} = \sigma^2\mathbf{I}$. Let $\mathbf{c} \in \mathbb{R}^p$ and $\theta = \mathbf{c}^\top\boldsymbol{\beta}$.*
If $\theta \in \mathsf{im}((\mathbf{X}\,\boldsymbol{\beta})^\top)$, equivalently if $\mathbf{c} \in \mathsf{im}(\mathbf{X}^\top)$, then

(i) the value of $\hat{\theta} = \mathbf{c}^\top\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}}$ is a solution to the normal equations does not depend on the choice of the solution;

(ii) $\exists$ a linear unbiased estimator of θ;

(iii) $\hat{\theta} = \mathbf{c}^\top\hat{\boldsymbol{\beta}}$ is BLUE for θ.

- parameter θ that is a linear combination of $\mathsf{E}\,\mathbf{Y}$ is identifiable
- a proof can be found in *Jiří Anděl: Základy matematické statistiky*

Inference for identifiable parameters

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} = \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$, $\mathsf{rank}(\mathbf{X}) = r < p$
- $\mathsf{E}\,\mathbf{Y}$ is identifiable
- $\hat{\mathbf{Y}} = \mathbf{X}\,\hat{\boldsymbol{\beta}}$ is BLUE for $\mathsf{E}\,\mathbf{Y}$ for any $\hat{\boldsymbol{\beta}}$ that solves the normal equations
- it can be shown that
    - ▷ $\frac{n-r}{\sigma^2}\hat{\sigma^2} \sim \chi^2_{n-r}$
    - ▷ $\hat{\sigma^2} = \frac{1}{n-r}||\mathbf{Y} - \hat{\mathbf{Y}}||^2$ is an unbiased estimator of σ^2

- ▷ $\hat{\sigma^2} \perp\!\!\!\perp \hat{\boldsymbol\beta}$ for any $\hat{\boldsymbol\beta}$ that solves the normal equations
- ▷ proofs are similar to the full-rank case
  - ∗ can be found in *Jiří Anděl: Základy matematické statistiky (2005). Matfyzpress.*

- ◦ inference for identifiable parameters and vectors is as in the full-rank model but we need to adjust the degrees of freedom
  - ▷ $n - r$ instead of $n - p$

9.2.3 Choice of the solution

Choice of $\hat{\beta}$

- ◦ $\hat{\mathbf{Y}}$ and $\hat{\sigma^2}$ do not depend on the choice of $\hat{\boldsymbol\beta}$
- ◦ $\{\hat{\boldsymbol\beta};\, \hat{\mathbf{Y}} = \mathbf{X}\,\hat{\boldsymbol\beta}\}$ is a linear subspace of $\mathbb{R}^p$ of dimension $p - r$
  - ▷ we can choose $\hat{\boldsymbol\beta}$ by specifying $p - r$ linear constraints
    - ∗ choose an $(p - r) \times p$ matrix $\mathbf{D}$, $\mathsf{rank}(\mathbf{D}) = p - r$
    - ∗ require that $\mathbf{D}\,\boldsymbol\beta = \mathbf{0}$
- ◦ for a given $\mathbf{D}$
  - ▷ QR decompose $\mathbf{D}^\top = (\mathbf{Q}_1 \,|\, \mathbf{Q}_2)\begin{pmatrix}\mathbf{R}_1\\ \mathbf{0}\end{pmatrix} = \mathbf{Q}_1\,\mathbf{R}_1$
  - ▷ $\mathbf{C_D} = \mathbf{Q}_2$ is a $p \times r$ matrix, $\mathsf{rank}(\mathbf{C_D}) = r$
  - ▷ $\mathbf{X_D} = \mathbf{X}\,\mathbf{C_D}$ is an $n \times r$ matrix, $\mathsf{rank}(\mathbf{X_D}) = r$
  - ▷ fit the (full-rank) model $\mathbf{Y} = \mathbf{X_D}\,\boldsymbol\beta_{\mathbf{D}} + \boldsymbol\varepsilon$
  - ▷ $\hat{\boldsymbol\beta} = \mathbf{C_D}\,\widehat{\boldsymbol\beta}_{\mathbf{D}}$ is the solution to the original normal equations satisfying the constraints given by $\mathbf{D}$

Common example: factor variables (`fev` data)

- ◦ basic model: $\mathsf{FEV} \sim \mathsf{Height} + \mathsf{Gender}$
  - ▷ naïve parametrization

$$
\begin{aligned}
Y_i = \beta_0 + \beta_{\mathrm{H}} \times \mathsf{Height}_i + \\
+ \beta_{\mathrm{M}} \times \mathbb{I}\{\text{the } i^{\text{th}} \text{ child is male}\} + \beta_{\mathrm{F}} \times \mathbb{I}\{\text{the } i^{\text{th}} \text{ child is female}\} + \\
+ \varepsilon_i,\ 1 \le i \le 654
\end{aligned}
$$

$$
\begin{pmatrix} 1.708\\ 1.724\\ 1.720\\ 1.558\\ \dots\\ 3.211 \end{pmatrix} = \begin{pmatrix} 1 & 144.8 & 0 & 1\\ 1 & 171.5 & 0 & 1\\ 1 & 138.4 & 0 & 1\\ 1 & 134.6 & 1 & 0\\ \dots & \dots & \dots & \dots\\ 1 & 168.9 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} \beta_0\\ \beta_{\mathrm{H}}\\ \beta_{\mathrm{M}}\\ \beta_{\mathrm{F}} \end{pmatrix} + \begin{pmatrix} \varepsilon_1\\ \varepsilon_2\\ \varepsilon_3\\ \varepsilon_4\\ \dots\\ \varepsilon_{654} \end{pmatrix}
$$

- ▷ standard parametrization

$$Y_i = \beta_0 + \beta_{\mathrm{H}} \times \mathsf{Height}_i + \\ + \beta_{\mathrm{M}} \times \mathbb{I}\{\text{the } i^{\text{th}} \text{ child is male}\} + \\ + \varepsilon_i, \; 1 \le i \le 654$$

- ○ basic model with interaction: FEV ∼ Height * Gender
  - ▷ standard parametrization

$$Y_i = \beta_0 + \beta_{\mathrm{H}} \times \mathsf{Height}_i + \\ + \beta_{\mathrm{M}} \times \mathbb{I}\{\text{the } i^{\text{th}} \text{ child is male}\} + \\ + \beta_{\mathrm{H:M}} \times \mathbb{I}\{\text{the } i^{\text{th}} \text{ child is male}\} \times \mathsf{Height}_i + \; \varepsilon_i, \; 1 \le i \le 654$$

One-way ANOVA

- ○ $Y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j}$, $\varepsilon_{i,j} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$
  $i \in \{1, \dots, I\}$, $j \in \{1, \dots n_i\}$
- ○ matrix form $\mathbf{Y} = \mathbf{X}\,(\mu, \boldsymbol{\alpha})^\top + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$

$$\begin{pmatrix} Y_{1,1} \\ \dots \\ Y_{1,n_1} \\ Y_{2,1} \\ \dots \\ Y_{2,n_2} \\ \dots \\ \dots \\ \dots \\ Y_{I,1} \\ \dots \\ Y_{I,n_I} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & 0 & 0 & \dots & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \dots \\ \alpha_I \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,1} \\ \dots \\ \varepsilon_{1,n_1} \\ \varepsilon_{2,1} \\ \dots \\ \varepsilon_{2,n_2} \\ \dots \\ \dots \\ \dots \\ \varepsilon_{I,1} \\ \dots \\ \varepsilon_{I,n_I} \end{pmatrix}$$

  - ▷ $\mathbf{X}$ is an $n \times (I+1)$ matrix with $\mathsf{rank}(\mathbf{X}) = I$

ANOVA

- ○ one-way ANOVA
  - ▷ $Y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j}$, $\varepsilon_{i,j} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$
    $i \in \{1, \dots, I\}$, $j \in \{1, \dots n_i\}$
  - ▷ matrix form $\mathbf{Y} = (\mathbf{1}\,|\,\mathbf{X}_{\boldsymbol{\alpha}})\,(\mu, \boldsymbol{\alpha})^\top + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
    - ∗ $\mathbf{X}$ is an $n \times (I+1)$ matrix with $\mathsf{rank}(\mathbf{X}) = I$
- ○ two-way ANOVA

- ▷ $Y_{i,j,k} = \mu + \alpha_i + \beta_j + \varepsilon_{i,j,k}$, $\varepsilon_{i,j,k} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$
  $i \in \{1, \dots, I\}$, $j \in \{1, \dots, J\}$, $k \in \{1, \dots n_{i,j}\}$
- ▷ matrix form $\mathbf{Y} = (\mathbf{1} \,|\, \mathbf{X}_{\boldsymbol{\alpha}} \,|\, \mathbf{X}_{\boldsymbol{\beta}})\, (\mu, \boldsymbol{\alpha}, \boldsymbol{\beta})^\top + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\, \mathbf{I})$
  - ∗ $\mathbf{X}$ is an $n \times (I + J + 1)$ matrix with $\mathsf{rank}(\mathbf{X}) = I + J - 1$

○ two-way ANOVA with interactions

- ▷ $Y_{i,j,k} = \mu + \alpha_i + \beta_j + \gamma_{i,j} + \varepsilon_{i,j,k}$, $\varepsilon_{i,j,k} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$
  $i \in \{1, \dots, I\}$, $j \in \{1, \dots, J\}$, $k \in \{1, \dots n_{i,j}\}$
- ▷ $\mathbf{Y} = (\mathbf{1} \,|\, \mathbf{X}_{\boldsymbol{\alpha}} \,|\, \mathbf{X}_{\boldsymbol{\beta}} \,|\, \mathbf{X}_{\boldsymbol{\alpha}} \,\boldsymbol{.}\, \mathbf{X}_{\boldsymbol{\beta}})\, (\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})^\top + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\, \mathbf{I})$
  ($\boldsymbol{.}$ denotes component-wise multiplication in the $n \times (I \times J)$ matrix)
  - ∗ $\mathbf{X}$ is an $n \times (I + J + (I \times J) + 1)$ matrix, $\mathsf{rank}(\mathbf{X}) = I \times J$

ANOVA parametrizations

○ one-way ANOVA

- ▷ $Y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j}$, $\varepsilon_{i,j} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$
  $i \in \{1, \dots, I\}$, $j \in \{1, \dots n_i\}$
  - ∗ R parametrization: $\alpha_1 = 0$
  - ∗ other parametrizations: e.g. $\sum_{i=1}^{I} n_i \alpha_i = 0$

○ two-way ANOVA

- ▷ $Y_{i,j,k} = \mu + \alpha_i + \beta_j + \varepsilon_{i,j,k}$, $\varepsilon_{i,j,k} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$
  $i \in \{1, \dots, I\}$, $j \in \{1, \dots, J\}$, $k \in \{1, \dots n_{i,j}\}$
  - ∗ R parametrization: $\alpha_1 = 0$, $\beta_1 = 0$
  - ∗ other: e.g. $\sum_{i=1}^{I} \alpha_i \sum_{j=1}^{J} n_{i,j} = 0$, $\sum_{j=1}^{J} \beta_j \sum_{i=1}^{I} n_{i,j} = 0$

○ two-way ANOVA with interactions

- ▷ $Y_{i,j,k} = \mu + \alpha_i + \beta_j + \gamma_{i,j} + \varepsilon_{i,j,k}$, $\varepsilon_{i,j,k} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$
  $i \in \{1, \dots, I\}$, $j \in \{1, \dots, J\}$, $k \in \{1, \dots n_{i,j}\}$
  - ∗ R parametrization: $\alpha_1 = 0$, $\beta_1 = 0$, $\gamma_{1,j} = 0\ \forall\ j$, $\gamma_{i,1} = 0\ \forall\ i$
  - ∗ other: e.g. $\sum_{i=1}^{I} \alpha_i \sum_{j=1}^{J} n_{i,j} = 0$, $\sum_{j=1}^{J} \beta_j \sum_{i=1}^{I} n_{i,j} = 0$, $\sum_{i=1}^{I} n_{i,j} \gamma_{i,j} = 0\ \forall\ j$, $\sum_{j=1}^{J} n_{i,j} \gamma_{i,j} = 0\ \forall\ i$

ANOVA parametrizations via matrices of contrasts

- one-way ANOVA
  - $\mathbf{Y} = (\mathbf{1}\,|\,\mathbf{X}_{\boldsymbol{\alpha}})\,(\mu, \boldsymbol{\alpha})^\top + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
  - replace $(\mathbf{1}\,|\,\mathbf{X}_{\boldsymbol{\alpha}})$ by $(\mathbf{1}\,|\,\mathbf{X}_{\boldsymbol{\alpha}}\,\mathbf{C}_{\boldsymbol{\alpha}})$
    - $\mathbf{C}_{\boldsymbol{\alpha}} \in \mathbb{R}^{I\times(I-1)}$, $\mathsf{rank}((\mathbf{1}\,|\,\mathbf{X}_{\boldsymbol{\alpha}}\,\mathbf{C}_{\boldsymbol{\alpha}})) = I$
  - estimate $\boldsymbol{\alpha}$ by $\mathbf{C}_{\boldsymbol{\alpha}}\,\hat{\boldsymbol{\alpha}}$ from the fitted model
- two-way ANOVA
  - $\mathbf{Y} = (\mathbf{1}\,|\,\mathbf{X}_{\boldsymbol{\alpha}}\,|\,\mathbf{X}_{\boldsymbol{\beta}})\,(\mu, \boldsymbol{\alpha}, \boldsymbol{\beta})^\top + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
  - replace $(\mathbf{1}\,|\,\mathbf{X}_{\boldsymbol{\alpha}}\,|\,\mathbf{X}_{\boldsymbol{\beta}})$ by $(\mathbf{1}\,|\,\mathbf{X}_{\boldsymbol{\alpha}}\,\mathbf{C}_{\boldsymbol{\alpha}}\,|\,\mathbf{X}_{\boldsymbol{\beta}}\,\mathbf{C}_{\boldsymbol{\beta}})$
    - $\mathbf{C}_{\boldsymbol{\alpha}} \in \mathbb{R}^{I\times(I-1)}$, $\mathbf{C}_{\boldsymbol{\beta}} \in \mathbb{R}^{J\times(J-1)}$
    - $\mathsf{rank}((\mathbf{1}\,|\,\mathbf{X}_{\boldsymbol{\alpha}}\,\mathbf{C}_{\boldsymbol{\alpha}}\,|\,\mathbf{X}_{\boldsymbol{\beta}}\,\mathbf{C}_{\boldsymbol{\beta}})) = I + J - 1$
  - estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by $\mathbf{C}_{\boldsymbol{\alpha}}\,\hat{\boldsymbol{\alpha}}$ and $\mathbf{C}_{\boldsymbol{\beta}}\,\hat{\boldsymbol{\beta}}$ from the fitted model
- two-way ANOVA with interactions
  - $\mathbf{Y} = (\mathbf{1}\,|\,\mathbf{X}_{\boldsymbol{\alpha}}\,|\,\mathbf{X}_{\boldsymbol{\beta}}\,|\,\mathbf{X}_{\boldsymbol{\alpha}} \,\boldsymbol{.}\, \mathbf{X}_{\boldsymbol{\beta}})\,(\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma)^\top + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
  - replace $(\mathbf{1}\,|\,\mathbf{X}_{\boldsymbol{\alpha}}\,|\,\mathbf{X}_{\boldsymbol{\beta}}\,|\,\mathbf{X}_{\boldsymbol{\alpha}} \,\boldsymbol{.}\, \mathbf{X}_{\boldsymbol{\beta}})$ by $(\mathbf{1}\,|\,\mathbf{X}_{\boldsymbol{\alpha}}\,\mathbf{C}_{\boldsymbol{\alpha}}\,|\,\mathbf{X}_{\boldsymbol{\beta}}\,\mathbf{C}_{\boldsymbol{\beta}}\,|\,\mathbf{X}_{\boldsymbol{\alpha}}\,\mathbf{C}_{\boldsymbol{\alpha}} \,\boldsymbol{.}\, \mathbf{X}_{\boldsymbol{\beta}}\,\mathbf{C}_{\boldsymbol{\beta}})$
    - $\mathbf{C}_{\boldsymbol{\alpha}} \in \mathbb{R}^{I\times(I-1)}$, $\mathbf{C}_{\boldsymbol{\beta}} \in \mathbb{R}^{J\times(J-1)}$
    - $\mathsf{rank}((\mathbf{1}\,|\,\mathbf{X}_{\boldsymbol{\alpha}}\,\mathbf{C}_{\boldsymbol{\alpha}}\,|\,\mathbf{X}_{\boldsymbol{\beta}}\,\mathbf{C}_{\boldsymbol{\beta}}\,|\,\mathbf{X}_{\boldsymbol{\alpha}}\,\mathbf{C}_{\boldsymbol{\alpha}} \,\boldsymbol{.}\, \mathbf{X}_{\boldsymbol{\beta}}\,\mathbf{C}_{\boldsymbol{\beta}})) = I\,J$
  - estimate $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ by $\mathbf{C}_{\boldsymbol{\alpha}}\,\hat{\boldsymbol{\alpha}}$, $\mathbf{C}_{\boldsymbol{\beta}}\,\hat{\boldsymbol{\beta}}$ and $(\mathbf{C}_{\boldsymbol{\alpha}} \otimes \mathbf{C}_{\boldsymbol{\beta}})\,\hat{\boldsymbol{\gamma}}$

9.3 Multicollinearity

Multicollinearity

Multicollinearity

- we have seen that if $\mathsf{rank}(\mathbf{X}) = r < p$, we do not lose

  anything by leaving out $p - r$ columns
- but what if $\mathsf{rank}(\mathbf{X}) = p$ but "only nearly so"?
  - the columns of $\mathbf{X}$ linearly independent BUT
  - $\frac{<\mathbf{x}_{,i},\mathbf{x}_{,j}>}{||\mathbf{x}_{,i}||\,||\mathbf{x}_{,j}||} \approx \pm 1$ for some (i, j)

    and/or for some linear combinations of the columns
- we would lose information by leaving out columns but keeping them all is a problem as well

- ▷ $\mathbf{X}^\top\mathbf{X}$ is ill-conditioned
  - ∗ $\widehat{\boldsymbol{\beta}}$ solves $(\mathbf{X}^\top\mathbf{X})\,\boldsymbol{\beta} = \mathbf{X}^\top\mathbf{Y}$
  - ∗ small change in $\mathbf{Y} \Rightarrow$ large change in $\widehat{\boldsymbol{\beta}}$
  - ∗ fit extremely sensitive to errors $\boldsymbol{\varepsilon}$
- ▷ large $\mathsf{Var}\,\widehat{\boldsymbol{\beta}}$
  - ∗ imprecise estimation of $\boldsymbol{\beta}$
  - ∗ wide confidence intervals for β's
  - ∗ large p-values of the t-tests
    (not necessarily of the overall F-test)

Detecting multicollinearity

- ∘ pairwise relationships
  - ▷ graphically: plot pairs of covariates one against another
  - ▷ numerically: compute pairwise correlations
- ∘ pairwise and/or higher-order relationships
  - ▷ regressing each covariate in turn on all the others
    - ∗ large values of the corresponding R^2 problematic
  - ▷ compute eigenvalues of $\mathbf{X}^\top\mathbf{X}$
    - ∗ large values of $\sqrt{\lambda_1/\lambda_j}$ problematic
- ∘ other indications
  - ▷ large p-values of the individual t-tests but a small p-value of the overall F-test
  - ▷ estimates of $\boldsymbol{\beta}$ and $\mathsf{Var}(\widehat{\boldsymbol{\beta}})$ very sensitive to adding/leaving out covariates and/or perturbing $\mathbf{Y}$

Variance inflation factors

- ∘ fit $\mathrm{lm}(\mathbf{X}_{,j} \sim \mathbf{X}_{,1} + \ldots + \mathbf{X}_{,j-1} + \mathbf{X}_{,j+1} + \ldots + \mathbf{X}_{,p})$
  - ▷ R_j^2 . . . the corresponding coefficient of determination
- ∘ it can be shown that $\mathsf{Var}(\hat{\beta}_j) = \dfrac{s^2}{(n-1)s^2_{X,j}} \times \dfrac{1}{1-R_j^2}$

$$\text{in } \mathrm{lm}(\mathbf{Y} \sim \mathbf{X}_{,1} + \ldots + \mathbf{X}_{,p})$$

- ∘ variance inflation factor $\mathrm{VIF}_j = \frac{1}{1-R_j^2}$

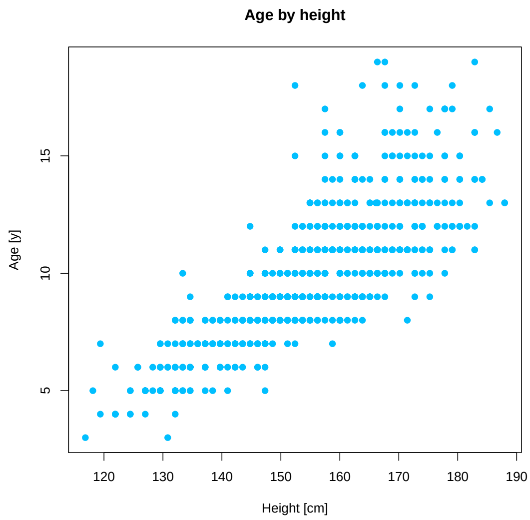
- ▷ measures linear dependence of the j^{th} covariate

  on the other covariates
- ▷ interpretation
  - ∗ standard error of $\hat{\beta}_j$ is $\approx \sqrt{\text{VIF}_j} \times$ larger than it would be were the j^{th} covariate independent of the other covariates
- ▷ $=$ 1 for orthogonal covariates, large values indicate problems
- ▷ how big is "too big"?
  - ∗ some consider VIF > 5 problematic
  - ∗ VIF > 10 is definitely considered problematic

- ∘ a generalization gVIF exists for categorical variables

Example: `fev` data

- ∘ $\mathsf{Cor}(\mathsf{Age}, \mathsf{Height}) = 0.79$



- ∘ $R^2_{\mathsf{Age}} = 0.69$
- ∘ $\text{VIF}_{\mathsf{Age}} = 3.24$

Ill-conditioned $\mathbf{X}^\top\mathbf{X}$

- ∘ linear model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
- ∘ model fitting: $\underbrace{(\mathbf{X}^\top\mathbf{X})}_{(p\times p)} \underbrace{\hat{\boldsymbol{\beta}}}_{(p\times 1)} = \underbrace{\mathbf{X}^\top\mathbf{Y}}_{(p\times 1)} \ \ldots \ \underbrace{\mathbf{A}}_{(p\times p)} \underbrace{\mathbf{x}}_{(p\times 1)} = \underbrace{\mathbf{b}}_{(p\times 1)}$

- solving for $\hat{\boldsymbol{\beta}}$ with machine precision
  - ▷ if the error in $\mathbf{b}$ is $\boldsymbol{\epsilon}$, the error in the solution $\mathbf{A}^{-1}\mathbf{b}$ is $\mathbf{A}^{-1}\boldsymbol{\epsilon}$
  - ▷ relative error in the solution divided by the relative error in $\mathbf{b}$:
    - $\frac{||\mathbf{A}^{-1}\boldsymbol{\epsilon}||/||\mathbf{A}^{-1}\mathbf{b}||}{||\boldsymbol{\epsilon}||/||\mathbf{b}||}$ for some norm $|| \bullet ||^2$
    - maximal value: $\frac{||\mathbf{A}^{-1}||}{||\mathbf{A}||}$
  - ▷ for Euclidean/spectral norm: $\frac{||\mathbf{A}^{-1}||}{||\mathbf{A}||} = \sqrt{\frac{\lambda_1}{\lambda_p}}$: $\surd$ of the ratio of the smallest and largest eigenvalue: condition number
    - some consider ≥ 30 problematic
    - the condition number depends also on the scales of covariates (not only on their relationships)
    - can improve a lot if all covariates are on similar scales

Tackling multicollinearity

- having independent covariates helps a lot but inherent relationships cannot be circumvented
- with collinear covariates, information does not increase as we would expect with the number of covariates
- "solutions"
  - ▷ excluding covariates
    - we avoid "repeating the same thing" but lose information
    - keep covariates that are of interest and/or are easy to measure
    - do not misinterpret leaving out a covariate as implying that it has no significant influence on the outcome
  - ▷ orthogonalizing and/or standardizing the predictors
    - more complicated interpretation
    - not a problem for prediction (but then multicollinearity might not have been a big issue unless extrapolation was planned)
  - ▷ a different method for estimation (e.g. ridge regression)
    - we loose some nice properties of the estimators

Chapter 10

Miscellanea and recap

10.1 The problem

10.1.1 Normal linear model

Normal linear model

- $Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k} + \varepsilon_i$, $i \in \{1, \ldots, n\}$
  - Y_i: outcome, response, output, dependent variable
    - random variable, we observe a realization y_i
    - (odezva, závisle proměnná, regresand)
  - $x_{i,1}, \ldots, x_{i,k}$: covariates, predictors, explanatory variables, input, independent variables
    - given, known
    - (nezávisle proměnné, regresory)
  - $\beta_0, \ldots, \beta_k$: coefficients
    - unknown
    - (regresní koeficienty)
  - ε_i: random error
    - random variable, unobserved
- $\varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$, $i \in \{1, \ldots, n\}$
  - $\mathsf{E}\, \varepsilon_i = 0$: no systematic errors
  - $\mathsf{Var}\, \varepsilon_i = \sigma^2$: same precision

Example: `bloodpress` data

- from `sites.stat.psu.edu/~lsimon/stat501wc/sp05/data/`
- association between the mean arterial blood pressure[mmHg] and age[years], weight[kg], body surface area[m^2], duration of hypertension[years], basal pulse[beats/min], stress
- data:

BP	Age	Weight	BSA	DoH	Pulse	Stress
105	47	85.4	1.75	5.1	63	33
115	49	94.2	2.10	3.8	70	14
...	...	...	...	...	...	...
110	48	90.5	1.88	9.0	71	99
122	56	95.7	2.09	7.0	75	99

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

$$\begin{pmatrix} 105 \\ 115 \\ \dots \\ 110 \\ 122 \end{pmatrix} = \begin{pmatrix} 1 & 47 & 85.4 & 1.75 & 5.1 & 63 & 33 \\ 1 & 49 & 94.2 & 2.10 & 3.8 & 70 & 14 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 48 & 90.5 & 1.88 & 9.0 & 71 & 99 \\ 1 & 56 & 95.7 & 2.09 & 7.0 & 75 & 99 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \dots \\ \beta_6 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_{19} \\ \varepsilon_{20} \end{pmatrix}$$

`https://ww2.amstat.org/publications/jse/v13n2/datasets.kahn.html`

Example: `fev` data

- from: `http://www.statsci.org/data/general/fev.html`
- question: association between the FEV[l] and Smoking,

  corrected for Age[years], Height[cm] and Gender
- data:

FEV	Age	Height	Gender	Smoking
1.708	9	144.8	Female	Non
1.724	8	171.5	Female	Non
1.720	7	138.4	Female	Non
1.558	9	134.6	Male	Non
...	...	...	...	...
3.727	15	172.7	Male	Current
2.853	18	152.4	Female	Non
2.795	16	160.0	Female	Current
3.211	15	168.9	Female	Non

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

$$\begin{pmatrix} 1.708 \\ 1.724 \\ 1.720 \\ 1.558 \\ \dots \\ 3.727 \\ 2.853 \\ 2.795 \\ 3.211 \end{pmatrix} = \begin{pmatrix} 1 & 9 & 144.8 & 0 & 0 \\ 1 & 8 & 171.5 & 0 & 0 \\ 1 & 7 & 138.4 & 0 & 0 \\ 1 & 9 & 134.6 & 1 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 15 & 172.7 & 1 & 1 \\ 1 & 18 & 152.4 & 0 & 0 \\ 1 & 16 & 160.0 & 0 & 1 \\ 1 & 15 & 168.9 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \dots \\ \beta_5 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \dots \\ \varepsilon_{651} \\ \varepsilon_{652} \\ \varepsilon_{653} \\ \varepsilon_{654} \end{pmatrix}$$

10.1.2 Task for this chapter

Miscellanea & recap

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
  - ▷ outcome $\mathbf{Y}$
    - ∗ random vector, we observe a realization $\mathbf{y}$
  - ▷ predictors $\mathbf{x}_{,1}, \dots, \mathbf{x}_{,k}$
    - ∗ vector of given (known) constants
  - ▷ coefficients $\boldsymbol{\beta}$
    - ∗ vector of unknown constants
  - ▷ error $\boldsymbol{\varepsilon}$
    - ∗ unknown random vector, we do not observe its realization
  - ▷ assumptions: $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
    - ∗ $\mathsf{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$: the expected value of $\mathbf{Y}$ is a linear function of $\boldsymbol{\beta}$
    - ∗ $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$: no systematic errors
    - ∗ $\mathsf{Var}\,\boldsymbol{\varepsilon} = \sigma^2\,\mathbf{I}$: independence and same precision
- task: miscellanea & recap

10.2 Linear regression in practice

10.2.1 Linear regression in practice

Statistical analysis with linear regression

1. build a mathematical model, i.e. define
   - what is known
   - what is uncertain

   linear regression example: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
2. build a probabilistic model for what is uncertain

   linear regression example: $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$
3. use probability calculus to draw conclusions

   linear regression example:

- $\widehat{\boldsymbol{\beta}} \sim \mathrm{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$
- $\frac{n-p}{\sigma^2}\,\widehat{\sigma^2} \sim \chi^2_{n-p}$
- $\widehat{\boldsymbol{\beta}} \perp\!\!\!\perp \widehat{\sigma^2}$

$\rightsquigarrow$ confidence intervals & hypotheses testing

4. "translate back" to the original problem (interpret the results)

   linear regression example:
   - $\widehat{\boldsymbol{\beta}}$, $\mathbf{a}^\top\widehat{\boldsymbol{\beta}}$, $\mathbf{A}\widehat{\boldsymbol{\beta}}$,
   - $\widehat{\mathsf{E}\,\mathbf{Y}}$, $\mathbf{a}^\top(\widehat{\mathsf{E}\,\mathbf{Y}})$, $\mathbf{A}(\widehat{\mathsf{E}\,\mathbf{Y}})$

   - confidence intervals
   - hypotheses testing

Usual additions to the basic analysis

1. find a suitable mathematical model
   - propose a suitable functional dependence of $\mathbf{Y}$ on $\mathbf{X}$
   - propose a suitable model for the error

   linear regression example: model selection

2. build a probabilistic model for what is uncertain

   linear regression example: check the normality, potentially propose a different error distribution

3. use probability calculus to draw conclusions
   - might need to adjust for multiple testing, post-hoc testing, poor design, ...

4. "translate back" to the original problem (interpret the results)

   linear regression example:
   - explanation
   - prediction

10.3 Notes on interpretation

10.3.1 Notes on the explanation

Explanation using linear regression

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

- estimate $\boldsymbol{\beta}$ by $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$
- estimate $\mathbf{a}^\top \boldsymbol{\beta}$ by $\mathbf{a}^\top \widehat{\boldsymbol{\beta}}$
- $(1-\alpha) \times 100\,\%$ confidence interval for $\mathbf{a}^\top \boldsymbol{\beta}$:

$$\left(\mathbf{a}^\top \widehat{\boldsymbol{\beta}} - t_{1-\alpha/2}(n-p) \sqrt{\widehat{\sigma^2}\, \mathbf{a}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{a}}\,, \quad \mathbf{a}^\top \widehat{\boldsymbol{\beta}} + t_{1-\alpha/2}(n-p) \sqrt{\widehat{\sigma^2}\, \mathbf{a}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{a}}\right)$$

- estimate $\mathbf{A}\,\boldsymbol{\beta}$ by $\mathbf{A}\,\widehat{\boldsymbol{\beta}}$
- $(1-\alpha) \times 100\,\%$ confidence bands for $\mathbf{A}\,\boldsymbol{\beta}$:

$$\left\{\mathbf{A}\,\boldsymbol{\beta};\ \frac{1}{m\,\widehat{\sigma^2}} (\mathbf{A}\,\widehat{\boldsymbol{\beta}} - \mathbf{A}\,\boldsymbol{\beta})^\top (\mathbf{A}\,(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{A}^\top)^{-1} (\mathbf{A}\,\widehat{\boldsymbol{\beta}} - \mathbf{A}\,\boldsymbol{\beta}) \leq F_{1-\alpha}(m, n-p)\right\}$$

Interpretation

- “keeping the values of all the other covariates fixed, a unit increase in x_i is associated with a $\hat{\beta}_i$ increase in $\mathsf{E}\,Y$”
  - suitably adapted for categorical predictors and potentially interactions, and depends on the choice of the identifiability conditions
  - polynomials need a more complex interpretation
- is it meaningful to imagine that a covariate changes while all the other remain fixed?

Be careful with

- confounding: suppose that
  - the truth is $Y_i = \beta_0 + \beta_E\, E_i + \beta_C\, C_i + \varepsilon_i$
  - we do not know about C and use $Y_i = \beta_0 + \beta_E\, E_i + \varepsilon_i$ instead
  - C and E are connected, e.g. $E_i = \gamma_0 + \gamma_C\, C_i + \tilde{\varepsilon}_i$
  - then if C has an effect on Y, we will (erroneously) attribute an effect on Y to E
  - may be solved by multiple regression model, provided the confounders and the form of their association to the outcome are known
- causality
  - very hard to be confident about a causal relationship rather that the “association”
- both can be helped by a sound design

10.3.2 Notes on the prediction

Prediction from linear regression

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
- what can we say about $Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon$

  for a new $\mathbf{x} = (1, x_1, \ldots, x_k)^\top$?
- $Y = \mathbf{x}^\top\boldsymbol{\beta} + \varepsilon$ and $\mathsf{E}\,Y = \mathbf{x}^\top\boldsymbol{\beta}$
- estimate $\mathsf{E}\,Y$ and Y by $\mathbf{x}^\top\widehat{\boldsymbol{\beta}}$
- $(1-\alpha)\times 100\,\%$ confidence interval for $\mathsf{E}\,Y$:

$$\Big(\mathbf{x}^\top\widehat{\boldsymbol{\beta}} - t_{1-\alpha/2}(n-p)\sqrt{\widehat{\sigma^2}\,\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}}\,,\quad \mathbf{x}^\top\widehat{\boldsymbol{\beta}} + t_{1-\alpha/2}(n-p)\sqrt{\widehat{\sigma^2}\,\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}}\Big)$$

- $(1-\alpha)\times 100\,\%$ confidence interval for Y

$$\Big(\mathbf{x}^\top\widehat{\boldsymbol{\beta}} - t_{1-\alpha/2}(n-p)\sqrt{\widehat{\sigma^2}\,(1+\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x})}\,,\quad \mathbf{x}^\top\widehat{\boldsymbol{\beta}} + t_{1-\alpha/2}(n-p)\sqrt{\widehat{\sigma^2}\,(1+\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x})}\Big)$$

Be careful with

- extrapolation
  - predicting Y for $\mathbf{x}$ that is far from the $\mathbf{x}_{i,}$'s in $\mathbf{X}$
  - predicting for different situations/populations than the one satisfying $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
- overfitting
  - fitting a model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ that is "too close to the data"
  - estimated σ^2 is small
- having seen enough data

10.4 Transformations

10.4.1 Transformations

Transformations of variables

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
- we have seen transformations of predictors to find a suitable functional dependence of Y on x
- how about transforming Y?
  - ▷ done in practice to improve the functional dependence or fix heteroskedasticity
  - ▷ most common are $\log(Y)$, $\sqrt{Y}$, some use other powers of Y
  - ▷ this is a fundamental change to the model
    - ∗ leaving the simple linear regression framework . . .

Log-transformation of the response

- original model: $Y_i = \beta_0 + \beta_1\, x_i + \varepsilon_i$
- model after the log transform: $\log(Y_i) = \beta_0 + \beta_1\, x_i + \varepsilon_i$
  - ▷ on the original scale: $Y_i = \exp\{\beta_0\} \times \exp\{\beta_1\, x_i\} \times \exp\{\varepsilon_i\}$
  - ▷ the effects of covariates are on the multiplicative scale
  - ▷ the error enters multiplicatively and the multiplicative error has log-normal distribution
    - ∗ $\exp\{x\} \approx 1 + x$ for small x
      $\Rightarrow Y_i = \exp\{\beta_0\} \times \exp\{\beta_1\, x_i\} \times (1 + \varepsilon_i)$ for small ε_i
      non-linear regression model with non-constant variance
      $Y_i = \exp\{\beta_0\} \times \exp\{\beta_1\}\, \exp\{x_i\} + \sigma_i^2\, \varepsilon_i$ for small ε_i . . .
  - ▷ prediction on the original scale
    - ∗ predict by $\exp\{\hat{Y}\}$ with CI $(\exp\{L\}, \exp\{U\})$
  - ▷ interpretation of $\boldsymbol{\beta}$ on the log-scale
  - ▷ problems with interpretation on the original scale
    - ∗ $\log(\mathsf{E}\, Y) \neq \mathsf{E} \log(Y)$ but the median is preserved
    - ∗ $\log(1 + x) \approx x$ for small x . . .
    - ∗ e.g. $\hat{\beta}_1 = 0.09$ can be interpreted as a 9 % increase in med$Y$ associated with a unit increase in x

Box–Cox transformation of the response

- original model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- looking for a more general transform...
- suppose that $Y > 0$
- Box–Cox transformation:
  - $g_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases}$ (a continuous function of λ)
  - λ can be viewed as a parameter and $\hat{\lambda}$ found by MLE
    - also gives a CI
  - for prediction, you may use $y^{\hat{\lambda}}$
  - for interpretation, you had better round $\hat{\lambda}$ to the nearest interpretable value (check the CI)
  - use CI to see if you need a transform at all

10.5 Concluding remarks

10.5.1 Reflection

It's an uncertain world ... use statistics to decide

- How much of
  - chocolate and other goodies is good for our health?
  - levels of bacteria, fertilizers, chemicals, ... is safe?
- What is the right size for
  - the height of a dam?
  - insurance premium?
  - mortgage interest?
- What is
  - the average salary?
  - public opinion on ...?
  - results in upcoming elections?

- uncertainty at the beginning ⇢ imperfect answers at the end
- statistics is used for quantifying uncertainty,

  not for getting rid of it

Statistics is collaboration

- *The best thing about being a statistician is that you get to play at everyone's backyard.*

  John Tukey

Statistics does not guarantee the right answers

- if there is no uncertainty, there is no need for statistics

$\hookrightarrow$ statistics might give a wrong answer

!!!but we should not abuse this!!!

- only incompetent statisticians do not know how to lie with statistics
- good statisticians know the pitfalls and know they must be cautious

Ingredients of a statistical analysis

- mathematics, programming, communication ...
- but above all: **COMMON SENSE**