Introductory Econometrics
Multicollinearity and Heteroskedasticity
Suggested Solution

by Hieu Nguyen

Fall 2024

1.

We estimate a linear regression model for the years 1972 to 1991:

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \epsilon_t,$$

where ϵ_t are normally and independently distributed, but we suspect that the variance of the error term is heteroskedastic and depends on x_{t1}. We estimate the following regression where e_t are residuals from regression (1):

$$e_t^2 = \delta_0 + \delta_1 x_{t1} + u_t.$$

We find that R^2 for regression (2) is 0.201. Use these results to test for the presence of heteroskedasticity.

Extract from statistical table of χ^2 distribution (area under right-hand tail):

d.f.	0.05	0.025	0.01
1	3.841	5.324	6.635
2	5.991	7.378	9.210
3	7.815	9.348	11.345
4	9.488	11.143	13.277

Solution:
Information in the setup suggests the Breusch-Pagan test:

$$\text{LM test statistic (in this case n = T):} \quad nR^2 = 20 \cdot 0.201 = 4.02 \sim \chi_1^2,$$

$$\text{critical value} \quad \chi^2_{1,0.95} = 3.84 < 4.02,$$

$$\Rightarrow \text{we reject the } H_0 : \delta_1 = 0 \quad \text{(meaning: no heteroskedasticity)} \quad \text{at the 5\% significance level in favor of the } H_A : \delta_1 \neq 0,$$

i.e., we conclude there is a problem with heteroskedasticity in the model.

2.

Use data `htv_selected.gdt` to estimate the returns to education in the 'wage equation.'

(a) Estimate the baseline model of the impact of education and experience on wages:

$$\ln(\text{wage}_i) = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \epsilon_i.$$

Interpret the estimated coefficient $\hat{\beta}_1$.

(b) Re-estimate the model using robust standard errors, comment on the differences.

(c) Test for heteroskedasticity in the model in part (a). Is it necessary to use robust standard errors in this case?

(d) Perform RESET (specification test) and discuss the results.

(e) Generate variable exper^2. Why we include this variable in the model and what is the expected sign of its coefficient?

(f) Estimate the model with quadratic specification (polynomial functional form) of experience:

$$\ln(\text{wage}_i) = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \beta_3 \text{exper}_i^2 + u_i.$$

Comment on how and why the estimated coefficient $\hat{\beta}_2$ changed with respect to part (a). Did the estimated coefficient $\hat{\beta}_1$ change as well? Why or why not? Compare R^2 and R^2_{adj} with the previous specification. Perform RESET again.

(g) Find $\frac{\partial \ln(\text{wage})}{\partial \text{exper}}$, which describes the marginal effect of a 1 year increase in work experience on wage. Compare the result with the marginal effect from the estimated model without exper^2.

(h) Do you believe that the coefficient β_1 is correctly estimated? Is there any issue that could create a bias in this equation? If yes, how would you solve for this problem? What is the expected sign of this bias?

(i) In the dataset, there are two proxies for inherent abilities and skills of the observed individuals, `abil1` and `abil2`. Estimate the model with just one of those. Is there an impact on the coefficient $\hat{\beta}_1$? Does this signalize there likely was a problem with bias in the model from part (f)?

Estimate the model with both proxies and discuss the differences and potential multicollinearity. Which Classical Assumption might be violated in this case? How do we check for this assumption?

(j) Include in the model from part (f) the education of the mother and of the father of the observed individuals:

$$\ln(\text{wage}_i) = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \beta_3 \text{exper}_i^2 + \beta_4 \text{motheduc}_i + \beta_5 \text{fatheduc}_i + v_i.$$

   i. What is the idea beyond including these variables in the model?
   ii. Is there an impact on the estimated coefficient $\hat{\beta}_1$? Does this signalize there likely was a problem with bias in the model from part (f)? Comment on the sign of this bias.
   iii. Are both motheduc and fatheduc individually significant? Are they jointly significant? Check potential multicollinearity.
   iv. What happens if you exclude one these variables from the regression? Which one would you keep?

(k) Compare the final models from parts (i) and (j). Which is a better model (based on the dataset in hand)? Try RESET again to potentially support your answer.

Solution:

(a) Baseline model:

```
Model 1: OLS, using observations 1{1230
Dependent variable: ln(wage)

Coefficient     Std. Error    t-ratio    p-value
------------------------------------------------
const           0.372975      0.175655   2.1233    0.0339
educ            0.130245      0.00896398 14.5299   0.0000
exper           0.0319110     0.00678060 4.7062    0.0000

Mean dependent var  2.413807    S.D. dependent var  0.593715
Sum squared resid   356.7888    S.E. of regression  0.539242
R^2                 0.176424    Adjusted R^2        0.175082
F (2, 1227)         131.4226    P-value(F )         1.92e{52
Log-likelihood      984.1547   Akaike criterion    1974.309
Schwarz criterion 1989.654    Hannan{Quinn         1980.083
```

$\hat{\beta}_1 = 0.13$: an increase in education by 1 year is associated with a 13% increase in wage, ceteris paribus.

(b) Robust standard errors:

```
Model 2: OLS, using observations 1{1230
Dependent variable: ln(wage)

Heteroskedasticity-robust standard errors, variant HC1

Coefficient     Std. Error    t-ratio    p-value
------------------------------------------------
const           0.372975      0.186249   2.0026    0.0454
educ            0.130245      0.00996372 13.0720   0.0000
exper           0.0319110     0.00688478 4.6350    0.0000

Mean dependent var  2.413807    S.D. dependent var  0.593715
Sum squared resid   356.7888    S.E. of regression  0.539242
R^2                  0.176424   Adjusted R^2        0.175082
F (2, 1227)        100.4045    P-value(F )         4.13e{41
Log-likelihood     984.1547   Akaike criterion    1974.309
Schwarz criterion 1989.654    Hannan{Quinn         1980.083
```

We observe an increase in standard errors primarily for educ and a related decrease of t-statistics. We also observe a reduction of the F -statistic related to the test of the overall significance of the regression. Many other results keep identical, bcs. not influenced by heteroskedasticity, namely: coefficient estimates, RSS, R^2, R^2_{adj}.

(c) White test:

```
White's test for heteroskedasticity

OLS, using observations 1-1230

Dependent variable: uhat^2

Coefficient     Std. Error    t-ratio    p-value
------------------------------------------------
const           2.77166       1.83933    1.507     0.1321
educ            -0.335010     0.169778   -1.973    0.0487 **
exper           -0.0968352    0.146825   -0.6595   0.5097
sq_educ         0.0107375     0.00403572 2.661     0.0079 ***
X2_X3           0.00600539    0.00652069 0.9210    0.3572
sq_exper        0.00184526    0.00304466 0.6061    0.5446

Unadjusted R-squared = 0.021218

Test statistic: TR^2 = 26.097743,

with p-value = P(Chi-square(5) > 26.097743) = 0.000085
```

Critical value $\chi^2_{5,0.95} = 11.07 < 26.1$.

We reject the H_0 of the overall insignificance (meaning: 'no heteroskedasticity') at the 5% significance level, i.e., we conclude there is a problem with heteroskedasticity in the model $\Rightarrow$ using robust standard errors in part (b) is well justified.

We might confirm this result with the Breusch-Pagan test.

We might also inspect residuals graphically. Can you see any pattern w.r.t educ or exper?

(d) Heteroskedasticity might also be caused by an incorrect specification of the model (incorrect functional specification or possibly an omitted variable with a heteroskedastic element, often represented by a nonlinear relationship in variables). We thus perform RESET (specification test):

```
Auxiliary regression for RESET specification test

OLS, using observations 1-1230

Dependent variable: ln(wage)

Coefficient     Std. Error    t-ratio    p-value
-----------------------------------------------
const           4.77602       2.95771    1.615     0.1066
educ            -1.45174      0.880991   -1.648    0.0996 *
exper           -0.356826     0.215819   -1.653    0.0985 *
yhat^2          5.40170       2.76440    1.954     0.0509 *
yhat^3          -0.785906     0.373362   -2.105    0.0355 **

Test statistic: F = 5.532007,

with p-value = P(F(2,1225) > 5.53201) = 0.00406
```

Critical value $F_{2,1225,0.95} = 3 < 5.53$

We reject the $H_0 : \gamma_1 = \gamma_2 = 0$ at the 5% significance level in favor of the $H_A : \gamma_1 \neq 0$ or $\gamma_2 \neq 0$, i.e., we conclude there is a misspecification problem in the model.

(e) Select exper and follow the path in the Gretl menu: Add—Squares of selected variables. Variable exper^2 is included to capture an expected nonlinear (decreasing) marginal effect of the variable exper via estimating a quadratic relationship (polynomial functional form). We expect a negative sign and a small absolute magnitude of the coefficient of exper^2 (compared to the coefficient of exper, discussed in detail during lecture #6 and the seminar).

(f) Model with exper^2:

If we expect exper^2 to be omitted, we should first perform the expected bias analysis: $\gamma < 0, \alpha_1 > 0$ $\Rightarrow$ expected bias negative, exper in part (a) is likely underestimated.

```
Model 3: OLS, using observations 1{1230
Dependent variable: ln(wage)

Coefficient     Std. Error    t-ratio    p-value
-----------------------------------------------
const           -0.0594569    0.226862   -0.2621   0.7933
educ            0.133631      0.00900606 14.8379   0.0000
exper           0.110042      0.0269265  4.0867    0.0000
sq_exper        -0.00360586   0.00120292 -2.9976   0.0028

Mean dependent var  2.413807    S.D. dependent var  0.593715
Sum squared resid   354.1929    S.E. of regression  0.537495
R^2                0.182417    Adjusted R^2        0.180416
F (3, 1226)        91.18041    P-value(F )         2.89e{53
Log-likelihood     979.6638   Akaike criterion    1967.328
Schwarz criterion 1987.787    Hannan{Quinn         1975.025
```

The regression result supports our suspicion from the EBA; we observe a considerable increase of the estimated coefficient of exper and the expected negative sign of the coefficient of exper^2. On the other hand, $\hat{\beta}_1$ has almost not changed (it is not as strongly correlated to exper^2, and it is not related to the functional relationship between wage and exper).

R^2 and R^2_{adj} both naturally increase.

Nonetheless, the RESET specification test still suggests a misspecification problem in the model.

As both heteroskedasticity tests still suggest heteroskedasticity, we should use heteroskedasticity-robust standard errors, but the impact is minimal.

(g) We need to differentiate the RHS of the model w.r.t. exper and plug in estimated coefficients to obtain the estimated marginal effect:

$$\frac{\partial \ln(\text{wage})}{\partial \text{exper}} = \hat{\beta}_2 + \hat{\beta}_3 \cdot 2 \cdot \text{exper} = 0.11 - 0.0072\text{exper}.$$

The estimated marginal effect of a 1-year increase in work experience on wage is thus decreasing, nonconstant, and considerably different from the model's constant marginal effect without exper^2 (suggesting its incorrect specification). The interpretation still follows the log-level functional form. However, the effect is dependent on the actual value of exper: for exper = 10, the 1-year increase in work experience is associated with a 3.8% increase in wage, ceteris paribus. If we put the estimated marginal effect equal to zero (FOC) to find the maximum, we can compute the saturation/turnaround point: circa 15.3 years (discussed in detail during the seminar).

(h) We might still suspect and omitted variable bias from omitting a variable measuring observed individuals' inherent skills and abilities. This variable should have a direct impact on wage, but it is also likely correlated with educ, and thus we expect a bias of the estimated coefficient of educ. If possible, we should add a variable measuring inherent skills and abilities (a proxy) to the model. EBA: $\gamma > 0, \alpha_1 > 0 \Rightarrow$ expected bias positive, educ in part (f) is likely overestimated.

(i) Model with `abil1`:

```
Model 4: OLS, using observations 1{1230
Dependent variable: ln(wage)

Coefficient     Std. Error    t-ratio    p-value
-------------------------------------------------
const           0.247283      0.228520   1.0821     0.2794
educ            0.105432      0.00992112 10.6270    0.0000
exper           0.0992598     0.0265617  3.7370     0.0002
sq_exper        -0.00297871   0.00118832 -2.5067    0.0123
abil1           0.0547057     0.00863816 6.3330     0.0000

Mean dependent var  2.413807    S.D. dependent var  0.593715
Sum squared resid   342.9640    S.E. of regression  0.529123
R^2                  0.208336   Adjusted R^2        0.205751
F (4, 1225)         80.59346    P-value(F )        9.24e{61
Log-likelihood      959.8509   Akaike criterion    1929.702
Schwarz criterion 1955.276     Hannan{Quinn         1939.324
```

We indeed observe a considerable decrease of $\hat{\beta}_1$ supporting our suspicion of a positive bias.

Model with `abil1` and `abil2`:

```
Model 5: OLS, using observations 1{1230
Dependent variable: ln(wage)

Coefficient     Std. Error    t-ratio    p-value
-------------------------------------------------
const           0.242456      0.228728   1.0600     0.2893
educ            0.105465      0.00992392 10.6274    0.0000
```

```
exper           0.100148      0.0266116  3.7633    0.0002
sq_exper        -0.00301812   0.00119052 -2.5351   0.0114
abil1           0.0423309     0.0227208  1.8631    0.0627
abil2           0.0124320     0.0211108  0.5889    0.5560

Mean dependent var   2.413807    S.D. dependent var  0.593715
Sum squared resid    342.8669    S.E. of regression  0.529264
R^2                 0.208560    Adjusted R^2        0.205327
F (5, 1224)        64.50975    P-value(F )        7.45e{60
Log-likelihood     959.6767   Akaike criterion    1931.353
Schwarz criterion 1962.042    Hannan{Quinn        1942.900
```

We observe a loss of the statistical significance of `abil1` (standard errors almost tripled) and a decreased estimated coefficient. Both impacts are perhaps caused by strong multicollinearity between `abil1` and `abil2`. Correlation between these variables is very large: 95%, and e.g. $VIF_{\hat{\beta}_4} = 1/(1 - 0.907) = 10.75$, both suggesting strong multicollinearity. Since this is not perfect multicollinearity, CA 6. is not violated, but the variance of the OLS estimator of related coefficients markedly increases and estimated standard errors. Solution: keep only one of 'abil' variables.

(j) Model with `motheduc` and `fatheduc`, but without `abil1`:

```
Model 6: OLS, using observations 1{1230
Dependent variable: ln(wage)

Coefficient     Std. Error    t-ratio    p-value
-------------------------------------------------
const           -0.242051     0.231347   -1.0463   0.2956
educ            0.119298      0.00955563 12.4845   0.0000
exper           0.109470      0.0267203  4.0969    0.0000
sq_exper        -0.00347614   0.00119414 -2.9110   0.0037
motheduc        0.00864702    0.00864412 1.0003    0.3173
fatheduc        0.0204120     0.00600822 3.3973    0.0007

Mean dependent var   2.413807    S.D. dependent var  0.593715
Sum squared resid    348.1603    S.E. of regression  0.533334
R^2                 0.196342    Adjusted R^2        0.193059
F (5, 1224)        59.80700    P-value(F )        8.04e{56
Log-likelihood     969.0990   Akaike criterion    1950.198
Schwarz criterion 1980.887    Hannan{Quinn        1961.744
```

i. The 'first-glance' idea might be that the education of one's parents might also be used as a proxy for inherent abilities and skills.

ii. $\hat{\beta}_1$ slightly decreases. This impact seems comparable to adding omitted `abil1` to the model in part (g), i.e., reducing the omitted variable bias. Still, the crucial question is whether the education of one's mother and father should belong to the equation (should influence one's wage directly). This problem of potential exogeneity of parents' education was discussed in detail during the seminar. The seeming reduction of the bias might only be an effect of multicollinearity between all three 'educ' variables.

iii. `motheduc` is not (t-test). Based on the test of linear restrictions (F -test of the joint significance) in Gretl, they are jointly significant:

```
Restriction set

1: b[motheduc] = 0

2: b[fatheduc] = 0
```

```
Test statistic: F(2, 1224) = 10.6041, with p-value = 2.71745e-05

Restricted estimates:

Coefficient Std. Error t-ratio p-value
---------------------------------------------------------
const -0.0594569 0.226862 -0.2621 0.7933
educ 0.133631 0.00900606 14.84 6.25e-46 ***
exper 0.110042 0.0269265 4.087 4.66e-05 ***
sq_exper -0.00360586 0.00120292 -2.998 0.0028 ***
motheduc 0.00000 0.00000 NA NA
fatheduc 0.00000 0.00000 NA NA

Standard error of the regression = 0.537495
```

Correlation between the new variables `motheduc` and `fatheduc` is 60%, and, e.g., $VIF_{\hat{\beta}_4} = 1/(1-0.4) = 1.67$ ($VIF_{\hat{\beta}_5}$ is almost similar), both suggesting some but not completely serious level of multicollinearity.

iv. Both variables are statistically significant if put to the regression alone. Empirically, there are only small differences, but intuitively, the father's education might be more determinative for somebody. Also, the t-statistic and R^2_{adj} is higher for `fatheduc`, so I would personally keep this one:

```
Model 7: OLS, using observations 1{1230
Dependent variable: ln(wage)

Coefficient     Std. Error    t-ratio    p-value
-------------------------------------------------
const           -0.198545     0.227222   -0.8738   0.3824
educ            0.121271      0.00934975 12.9705   0.0000
exper           0.109179      0.0267188  4.0862    0.0000
sq_exper        -0.00346108   0.00119405 -2.8986   0.0038
fatheduc        0.0234096     0.00520761 4.4953    0.0000

Mean dependent var  2.413807    S.D. dependent var  0.593715
Sum squared resid   348.4450    S.E. of regression  0.533334
R^2                  0.195684    Adjusted R^2        0.193058
F (4, 1225)        74.50854    P-value(F )         1.43e{56
Log-likelihood     969.6016   Akaike criterion    1949.203
Schwarz criterion 1974.777    Hannan{Quinn        1958.825
```

(k) Models 4 and 7 seem relatively comparable, but `abil` appears to be (intuitively as well as theoretically due to the potential exogeneity of `fatheduc`) a dominant proxy for inherent abilities and skills. Also, based on other specification criteria, Model 4 seems better: `abil1` has a higher t-statistic (although both variables are very statistically significant), R^2_{adj} is higher for Model 4, the expected positive bias seems to be reduced in both models but more in Model 4.

RESET run for Model 4 finally suggests no other specification problem in the model.

However, this is a rare case when RESET is suitable for detecting an omitted variable (a proxy for inherent abilities and skills). However, its performance is generally poor in this respect and very sample dependent $\Rightarrow$ do not rely on RESET when thinking of omitted variables, it is primarily a test for general functional form misspecification:

```
Test statistic: F = 0.586120,

with p-value = P(F(2,1223) > 0.58612) = 0.557
```

We do not reject the $H0 : \gamma_1 = \gamma_2 = 0$ at the 5% significance level, i.e., we conclude the model is correctly specified.