

# M7777 Applied Functional Data Analysis

## 3. From Data to Functions – Smoothing

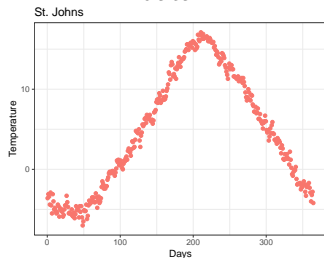
Jan Kolářček (kolacek@math.muni.cz)

Dept. of Mathematics and Statistics, Faculty of Science, Masaryk University, Brno



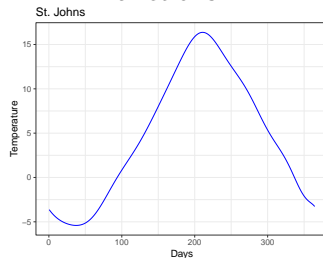
How do we go from

data



to

functions?



# Basis Expansions

We consider

$$y_i = x(t_i) + \varepsilon_i, \quad \varepsilon_i \sim i.i.d$$

and

$$x(t_i) = \sum_{j=1}^K c_j \Phi_j(t_i). \quad (1)$$

Let us denote

- $\mathbf{y} = (y_1, \dots, y_N)'$ ,  $\mathbf{x} = (x(t_1), \dots, x(t_N))'$
- $\Phi$  ... a  $N \times K$  matrix containing values  $\Phi_j(t_i)$
- $\mathbf{c} = (c_1, \dots, c_K)'$  ... **basis coefficients**

We can write (1) as

$$\mathbf{x} = \Phi \mathbf{c}.$$

# Least Squares

**How** to find  $\mathbf{c}$ ?

Minimize the sum of squared errors

$$SSE(\mathbf{c}) = \sum_{i=1}^N (y_i - x(t_i))^2 = (\mathbf{y} - \Phi\mathbf{c})'(\mathbf{y} - \Phi\mathbf{c})$$

This is just linear regression!

The *SSE* is minimized by the **ordinary least squares** estimate

$$\hat{\mathbf{c}} = (\Phi'\Phi)^{-1} \Phi'\mathbf{y}.$$

Thus we have the estimate

$$\hat{\mathbf{x}}(t) = \Phi^*(t) (\Phi'\Phi)^{-1} \Phi'\mathbf{y}.$$

Standard model: suppose i.i.d  $\varepsilon_i \Rightarrow E(\mathbf{y}) = \Phi\mathbf{c}$ ,  $\text{Var}(\mathbf{y}) = \sigma^2\mathbf{I}_N$

# Weighted Least Squares

Practical problems

- heteroscedastic data
- autocorrelated data

Partial solution: **Weighted** Least Squares

$$WSSE(\mathbf{c}) = \sum_{i=1}^N w_i^2 (y_i - x(t_i))^2 = (\mathbf{y} - \Phi\mathbf{c})' \mathbf{W}' \mathbf{W} (\mathbf{y} - \Phi\mathbf{c})$$

with  $\mathbf{W} = \text{diag}\{w_1, \dots, w_N\}$ .

We get an estimate

$$\hat{\mathbf{x}}(t) = \Phi^*(t) (\Phi' \mathbf{W} \Phi)^{-1} \Phi' \mathbf{W} \mathbf{y}.$$

and fitted values

$$\hat{\mathbf{y}} = \Phi (\Phi' \mathbf{W} \Phi)^{-1} \Phi' \mathbf{W} \mathbf{y} = \mathbf{S} \mathbf{y}.$$

**S** ... **smoothing matrix**.

# Choosing the Number of Basis Functions

**How many** basis functions?

- Small numbers of basis functions mean little flexibility.
- Larger numbers of basis functions add flexibility, but may “overfit” .
- For Monomial and Fourier bases, just add functions to the collection.
- Spline bases: adding knots or increasing the order changes the basis.

Trade off:

- Too many basis functions over-fits the data and reflect errors of measurement.
- Too few basis functions fails to capture interesting features of the curves.

# Bias and Variance Tradeoff

Measure of quality of  $\hat{x}(t)$

- **Bias**

$$\text{Bias}[\hat{x}(t)] = E[\hat{x}(t)] - x(t)$$

- **Variance**

$$\text{Var}[\hat{x}(t)] = E \{ \hat{x}(t) - E[\hat{x}(t)] \}^2$$

- Too many basis functions  $\Rightarrow$  small bias, large sampling variance.
- Too few basis functions  $\Rightarrow$  small sampling variance, large bias.
- **M**ean **S**quared **E**rror

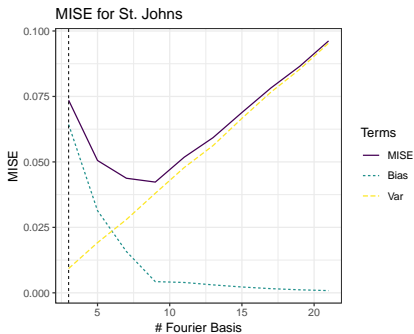
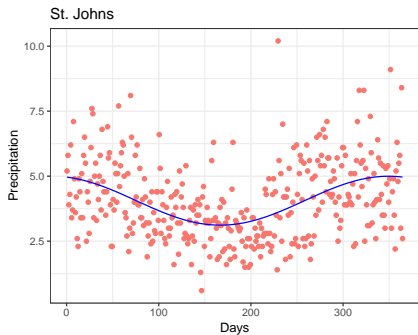
$$\text{MSE}[\hat{x}(t)] = E [ \{ \hat{x}(t) - x(t) \}^2 ] = \text{Bias}^2[\hat{x}(t)] + \text{Var}[\hat{x}(t)]$$

- **I**ntegrated **M**ean **S**quared **E**rror

$$\text{IMSE}[\hat{x}(t)] = \int \text{MSE}[\hat{x}(t)] dt$$

# Choosing the Number of Basis Functions

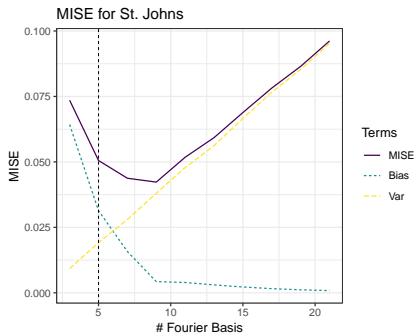
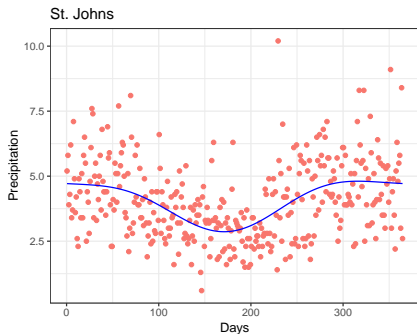
## St. Johns Precipitation: 3 Fourier Bases





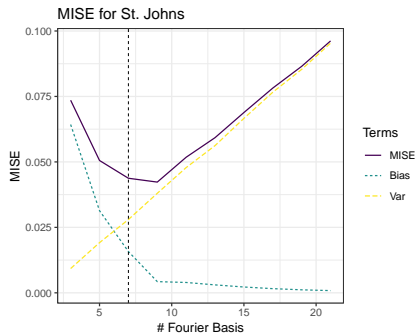
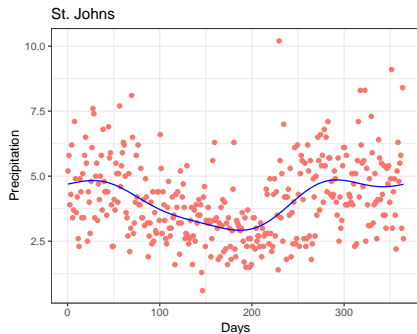
# Choosing the Number of Basis Functions

## St. Johns Precipitation: 5 Fourier Bases



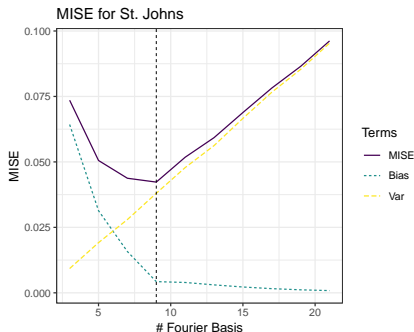
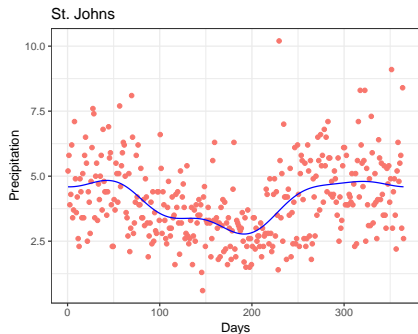
# Choosing the Number of Basis Functions

## St. Johns Precipitation: 7 Fourier Bases



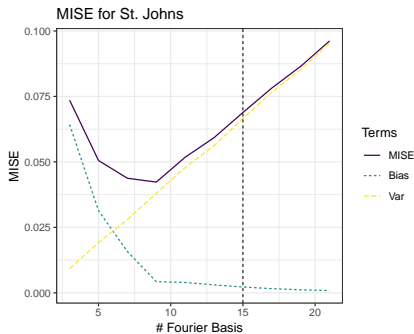
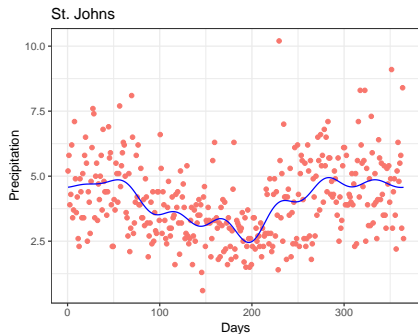
# Choosing the Number of Basis Functions

## St. Johns Precipitation: 9 Fourier Bases



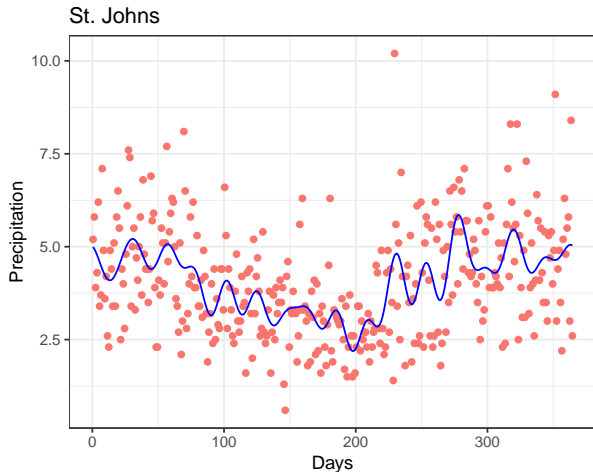
# Choosing the Number of Basis Functions

St. Johns Precipitation: **15** Fourier Bases



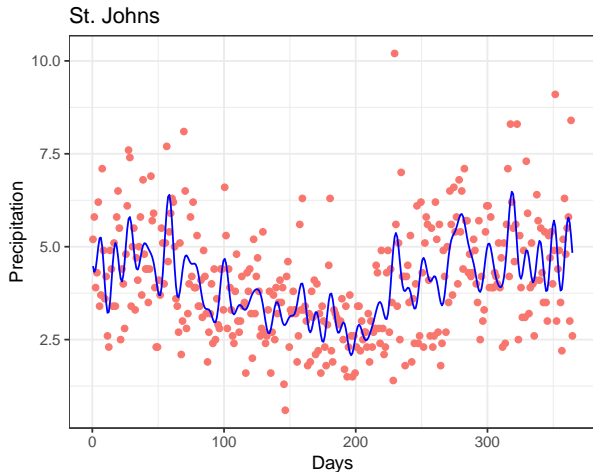
# Choosing the Number of Basis Functions

St. Johns Precipitation: **35** Fourier Bases



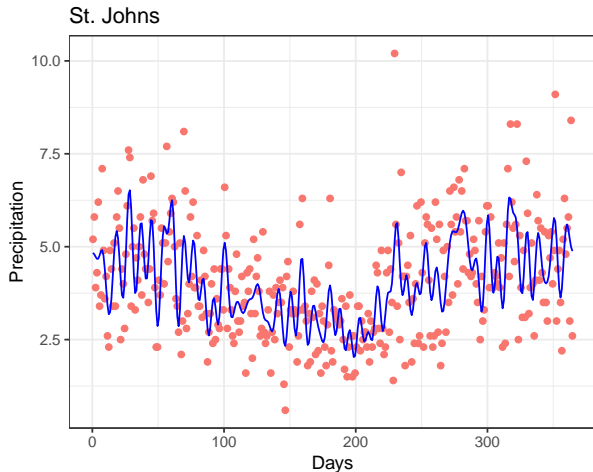
# Choosing the Number of Basis Functions

St. Johns Precipitation: **75** Fourier Bases



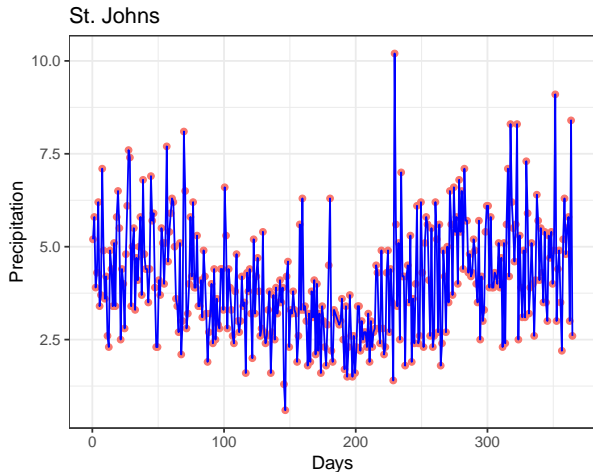
# Choosing the Number of Basis Functions

St. Johns Precipitation: **105** Fourier Bases



# Choosing the Number of Basis Functions

St. Johns Precipitation: **365** Fourier Bases





## Cross-Validation

- Leave out one observation  $(t_i, y_i)$  and construct an estimate  $\hat{x}_{-i}(t)$  from remaining data.
- Choose  $K$  to minimize the **ordinary cross-validation** score

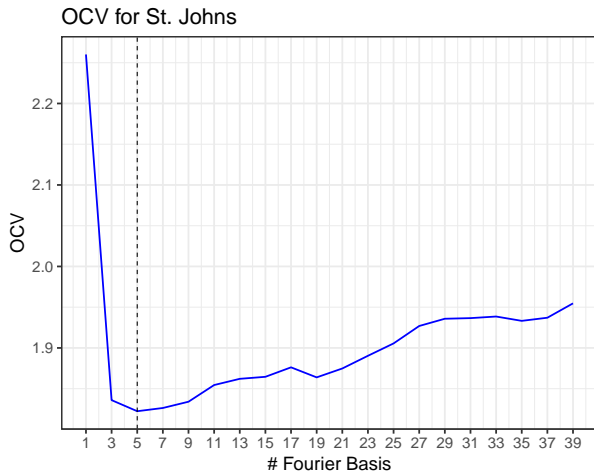
$$OCV(\hat{x}) = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{x}_{-i}(t_i))^2.$$

- For a linear smooth  $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$

$$OCV(\hat{x}) = \frac{1}{N} \sum_{i=1}^N \frac{(y_i - \hat{x}(t_i))^2}{(1 - s_{ii})^2}.$$

# Choosing the Number of Basis Functions

St. Johns Precipitation: Cross-Validated Error



# Pointwise Confidence Bands

## Estimating the Variance

- Standard model assumes

$$\text{Var}[\mathbf{y}] = \sigma^2 \mathbf{I}_N$$

- An unbiased estimate (can be more sophisticated for correlated residuals)

$$\hat{\sigma}^2 = \frac{1}{N - K} \text{SSE}(\hat{\mathbf{c}})$$

- For a linear smooth  $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$

$$\text{Var}[\hat{\mathbf{y}}] = \sigma^2 \mathbf{S}\mathbf{S}'.$$

- More generally, for  $\text{Var}[\mathbf{y}] = \mathbf{\Sigma}$

$$\text{Var}[\hat{\mathbf{y}}] = \mathbf{S}\mathbf{\Sigma}\mathbf{S}'.$$

## Pointwise Confidence Intervals

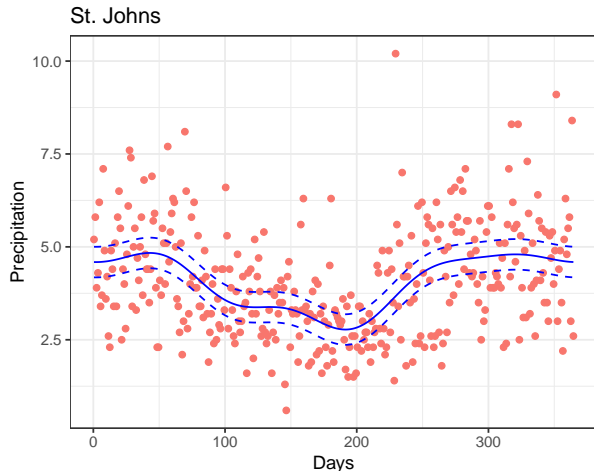
- For each point we calculate lower and upper bands for  $\hat{\mathbf{y}}$  by

$$\hat{\mathbf{y}} \pm u_{0.975} \sqrt{\text{Var}[\hat{\mathbf{y}}]}.$$

- These bands are not confidence bands for the entire curve, but only for the value of the curve at a fixed point.
- Ignores bias in the estimated curve.
- Provide an impression of how well the curve is estimated.

# Choosing the Number of Basis Functions

Fitted St. Johns Precipitation Data with 9 Fourier Bases and Confident Bands



## ① Melanoma Data

- Load the variable `melanoma` from the `fda` package and plot it.
- Fit these data with a Fourier basis, choosing the number of basis functions by minimizing the `gcv` value returned by `smooth.basis`. Plot the Cross-Validation function and the final fit (see Figures 1,2).
- Try removing a linear trend for these data first, by looking at the residuals after a call to `lm`. Repeat the steps above; does the optimal number of basis functions change?
- Re-fit the data using an `gcv`-optimal B-spline basis. Plot the CV function for this basis and the final fit (see Figures 3,4).
- Plot the previous fit with its pointwise 95% confidence bands (see Figure 5). What's the observed value of incidences in 1950? What's the estimated mean confidence band for this year?

[2.9, (2.49, 2.96)]

## ② Canadian Weather Data

- Load the variable `CanadianWeather` from the `fda` package and select the precipitation in St. Johns.
- Fit these data using a B-spline basis with 5 basis functions.
- What's the observed value of precipitation in St. Johns on January 23?  
What's the estimated mean confidence band for this day?

[4.4, (4.37, 5.07)]

# Problems to solve

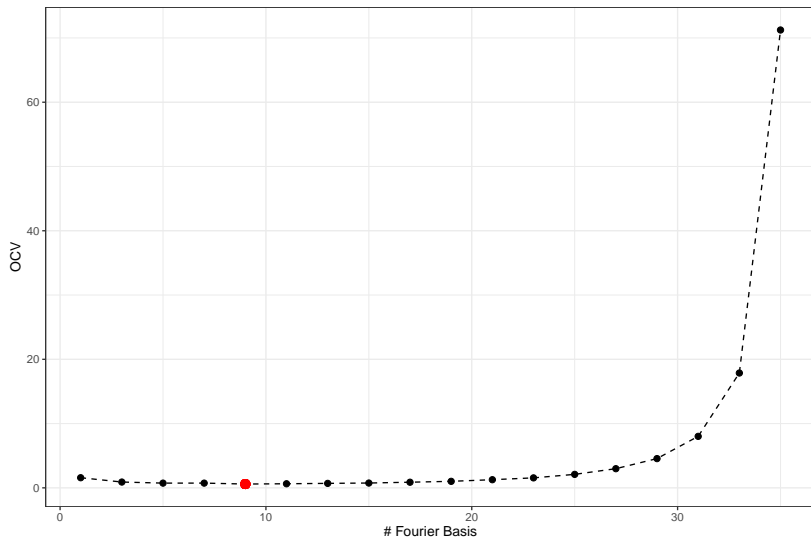


Figure 1.



# Problems to solve

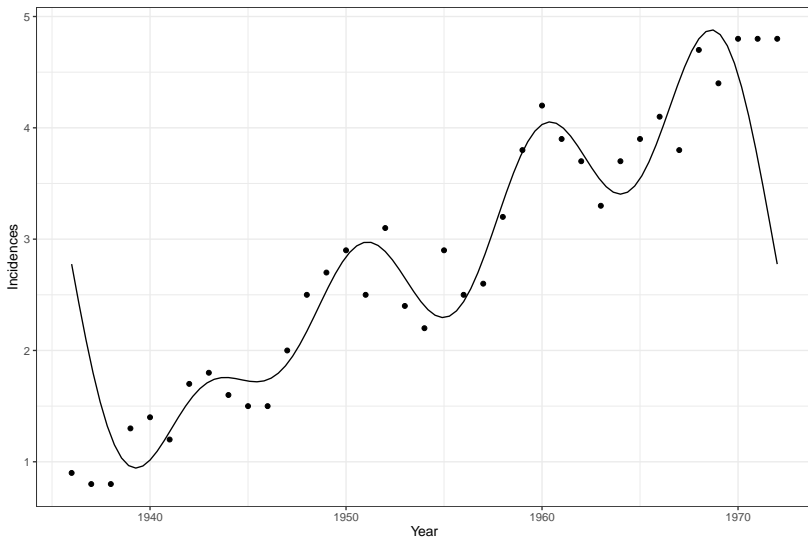


Figure 2.

# Problems to solve

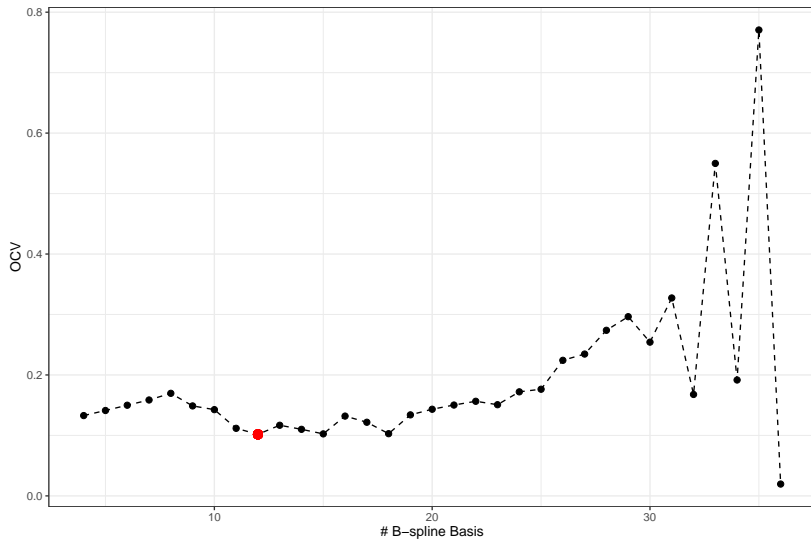


Figure 3.

# Problems to solve

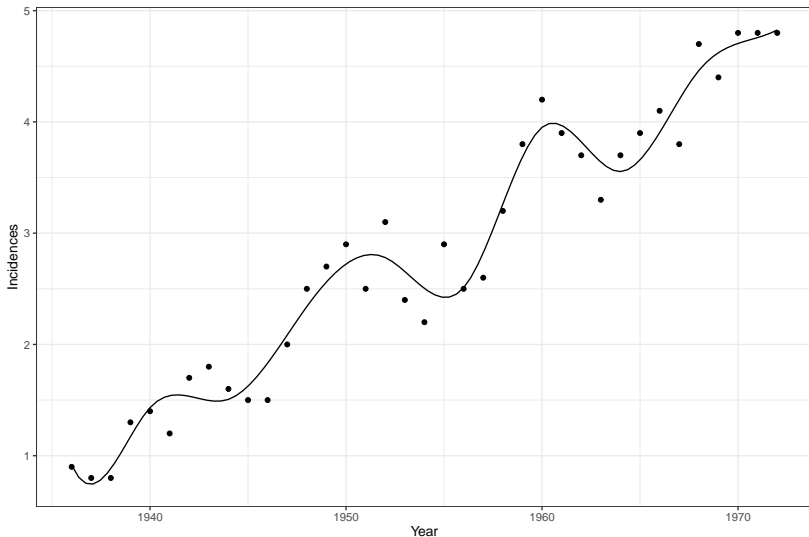


Figure 4.

# Problems to solve

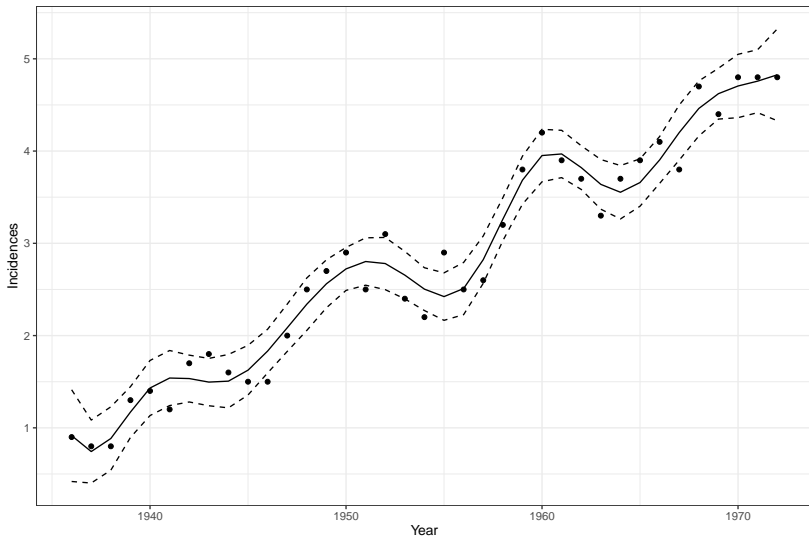


Figure 5.