Methods 3: Multilevel Statistical Modeling and Machine Learning

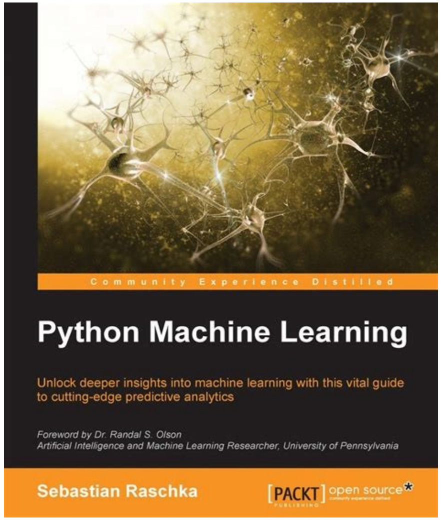
Week 5: *Evaluating and comparing models*
October 12, 2021

by: Lau Møller Andersen

Study Cafe

Python book in Stakbogladen

https://www.stakbogladen.dk/soegning.asp?phrase=9781783555130



I still owe the answers to questions in the CryptPad Poisson and overdispersion, deviance in a logistic model, and the assumption of normality

Code review – pair programming

Learning goals

Evaluating and comparing models

1) Learning tools for comparing models
    1) Variance explained
    2) Likelihood ratio tests
    3) Information criteria
2) Bridging to out-of-sample

Why are we modelling?

Remember Emil's slides from week 03

- To be able to understand the world

dont worry about thia equation, its form physics

$$F = G\frac{m_1 m_2}{r^2}$$

EXPLANATION

newtons formula

- To be able to predict and manipulate the world

NASA/Bill Ingalls

PREDICTION

What constitutes a good model?

Remember Emil’s slides from week 03

- Accurate estimation of the underlying parameters of the population distribution
- Generalisation to new data

EXPLANATION

PREDICTION

Within an **explanatory** framework, how can we assess whether we have done a good job?

1) Variance explained?

Lau vil egentlig kun have os til at extracte det han skal vide e.g coefficients :))

R summary

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## wt           -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

R^2 (coefficient of determination)

estimation

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \mu_y)^2}$$

mean of population

R^2; adjusted

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \mu_y)^2}$$

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

n: number of observations

p: number of predictors beyond the constant term

så altså ikke inkluderende intercept :(

the more predictors we include, the more cariance we explain.. but dont explain it aallll =overfitting

```
lm(formula = mpg ~ wt, data = mtcars)

Coefficients:
(Intercept)           wt
     37.285       -5.344
```

```
"R-squared: 0.753"
"R-squared, adjusted: 0.745"
```

```
lm(formula = mpg ~ I(wt^2) + wt, data = mtcars)

Coefficients:
(Intercept)      I(wt^2)           wt
     49.931        1.171      -13.380
```

```
"R-squared: 0.819"
"R-squared, adjusted: 0.807"
```

```
lm(formula = mpg ~ I(wt^3) + I(wt^2) + wt, data = mtcars)

Coefficients:
(Intercept)      I(wt^3)      I(wt^2)           wt
   48.40370      0.04594      0.68938    -11.82598
```

```
"R-squared: 0.819"
"R-squared, adjusted: 0.8"
```

```
lm(formula = mpg ~ I(wt^4) + I(wt^3) + I(wt^2) + wt, data = mtcars)

Coefficients:
(Intercept)      I(wt^4)      I(wt^3)      I(wt^2)           wt
    14.4558      -0.3863       5.2004     -23.7018      36.6195
```

```
"R-squared: 0.822"
"R-squared, adjusted: 0.796"
```

how do we compare the models? besides r2.. we need more :))

R^2 – mixed models

$$R_{GLMM}(m)^2 = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_\alpha^2 + \sigma_\epsilon^2}$$

$$R_{GLMM}(c)^2 = \frac{\sigma_f^2 + \sigma_\alpha^2}{\sigma_f^2 + \sigma_\alpha^2 + \sigma_\epsilon^2}$$

σ_f^2: variance of the fixed effects

σ_α^2: variance of the random effects

σ_ϵ^2: unexplained variance

Variance explained

- Pros
  - R^2 is intuitive
- Cons
  - More complex models will always explain more variance
  - Hard to interpret in the case of collinearity
  - R^2 doesn't give us what we want

(Yarkoni and Westfall, 2017)

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i. \quad (1)$$

$$\hat{y}_i = 1.6 + 0.35 X_{1i} + 0.62 X_{2i}.$$

$$R^2 = 0.750$$

Tempting interpretation: the parameter estimates are likely to be true, because R^2 is large

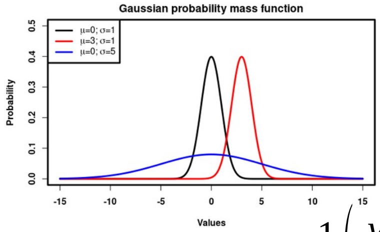
Correct interpretation:

“if one fits a model with the form of equation 1 in each new sample – *each time estimating new values of* b_0, b_1, and b_2 – what will be the average proportional reduction in the sum of squared errors?” (Yarkoni and Westfall 2017)

2) Likelihood ratios

Maximum likelihood estimation

$$\textbf{PDF} = (2\pi\sigma^2)^{-1/2} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



x* B er bare vores y-hat

$$\textbf{Conditional PDF}(y_i|X) = (2\pi\sigma^2)^{-1/2} e^{-\frac{1}{2}\left(\frac{y_i - x_i\beta}{\sigma_0}\right)^2}$$

basically: sum af probabilites, hvad er sandsynligheden for at observere _ting_

$$\textbf{Likelihood function}: L(\sigma^2, \epsilon) = (2\pi\sigma^2)^{-N/2} e^{-\left(\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i\beta)^2\right)}$$

$$\textbf{Log-Likelihood function}: l(\sigma^2, \epsilon) = \log(L)$$

y : dependent variable
$X\beta$: linear predictor
$y - X\beta = \epsilon$: residuals
N : number of observations

we need independence of obsercations, in order to gain probabilites:
cuz when we have independence then we can just take the product of the probabilites, (0.5*0.5 = HEads *heads))
= the product of that is the probabilites for getting two independent heads!

```
model <- lmer(Reaction ~ days_deprived + (days_deprived | Subject),
              data=sleepstudy)
model.ranint <- lmer(Reaction ~ days_deprived + (1 | Subject),
              data=sleepstudy)
```

the more parameters you include, the more variance you explain, just like R^2

LOG LIKELIHOOD

```
print(ll.m <- logLik(model))
```

```
## 'log Lik.' -702.0472 (df=6)
```

LOG LIKELIHOOD

```
print(ll.mr <- logLik(model.ranint))
```

```
## 'log Lik.' -715.01 (df=4)
```

left is the best, because it is highest!

these observations are more probablye under this model than ————————> this model

EXPONENTIAL

```
exp(ll.m)
```

```
## 'log Lik.' 1.272865e-305 (df=6)
```

EXPONENTIAL

```
exp(ll.mr)
```

```
## 'log Lik.' 2.986092e-311 (df=4)
```

Likelihood-ratio test

$$LR = -2(l(\theta_2) - l(\theta_1))$$

ANOVA CASANOVA

```
anova(model.ranint, model.ranslope.and.int)
```

```
## Data: sleepstudy
## Models:
## model.ranint: Reaction ~ days_deprived + (1 | Subject)
## model.ranslope.and.int: Reaction ~ days_deprived + (days_deprived | Subject)
##                         Df    AIC    BIC  logLik deviance  Chisq Chi Df
## model.ranint             4 1446.5 1458.4 -719.25   1438.5
## model.ranslope.and.int   6 1425.2 1443.0 -706.58   1413.2 25.332      2
##                         Pr(>Chisq)
## model.ranint
## model.ranslope.and.int  3.156e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

de her to minus hianden * 2

2 is just the degrees of freedom

Likelihood-ratio test

$$LR = -2\left(l\left(\theta_2\right) - l\left(\theta_1\right)\right)$$

it gives us a principle way of comparing the models better!
Therefore it is better than R^2 because r^2 does not include the “comparing” part of the models.
In r^2 we just look at the highest value, but not COMPARED to other models.

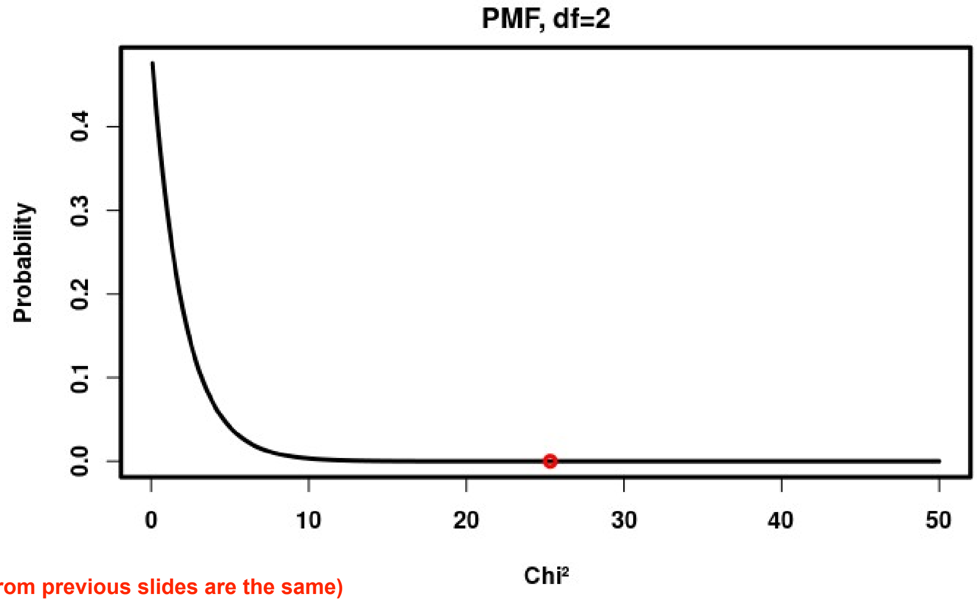
Chisq	Chi Df
25.332	2

Pr(>Chisq)

3.156e-06

the p value

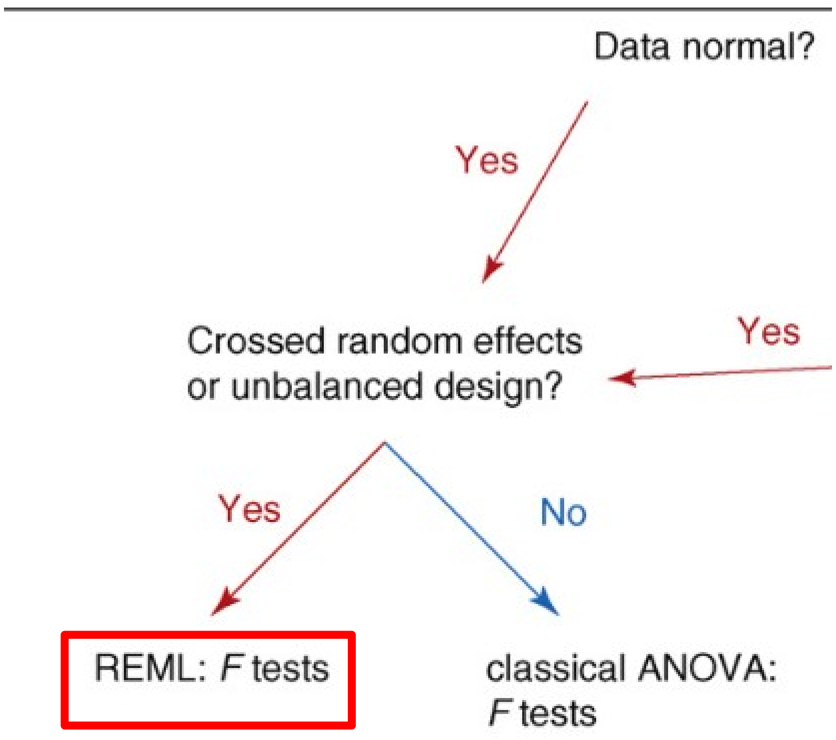
allows us to reject H0 (that the two log likelihoods from previous slides are the same)



Regarding GLMMs

“The LR test is only adequate for testing fixed effects when both the ratio of the total sample size to the number of fixed-effect levels being tested and the number of random-effect levels (blocks) are large. *We have found little guidance and no concrete rules of thumb in the literature* [...]” (my highlights, Bolker et al., 2008)

Requires that degrees of freedom can be defined



(Bolker et al., 2009)

F-values – single level modelling

$$F = \frac{\left(\frac{RSS_1 - RSS_2}{p_2 - p_1}\right)}{\left(\frac{RSS_2}{n - p_2}\right)}$$

$RSS_1 - RSS_2$ → difference in unexplained variance between "null" model and "target" model

$p_2 - p_1$ → numerator degrees of freedom (difference in predictor variables)

RSS_2 → unexplained variance by "target" model

$n - p_2$ → denominator degrees of freedom (degrees of freedom for observations)

RSS: Residual sum of squares
p: number of predictor variables
n: number of observations

F-values – single level modelling

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## wt           -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

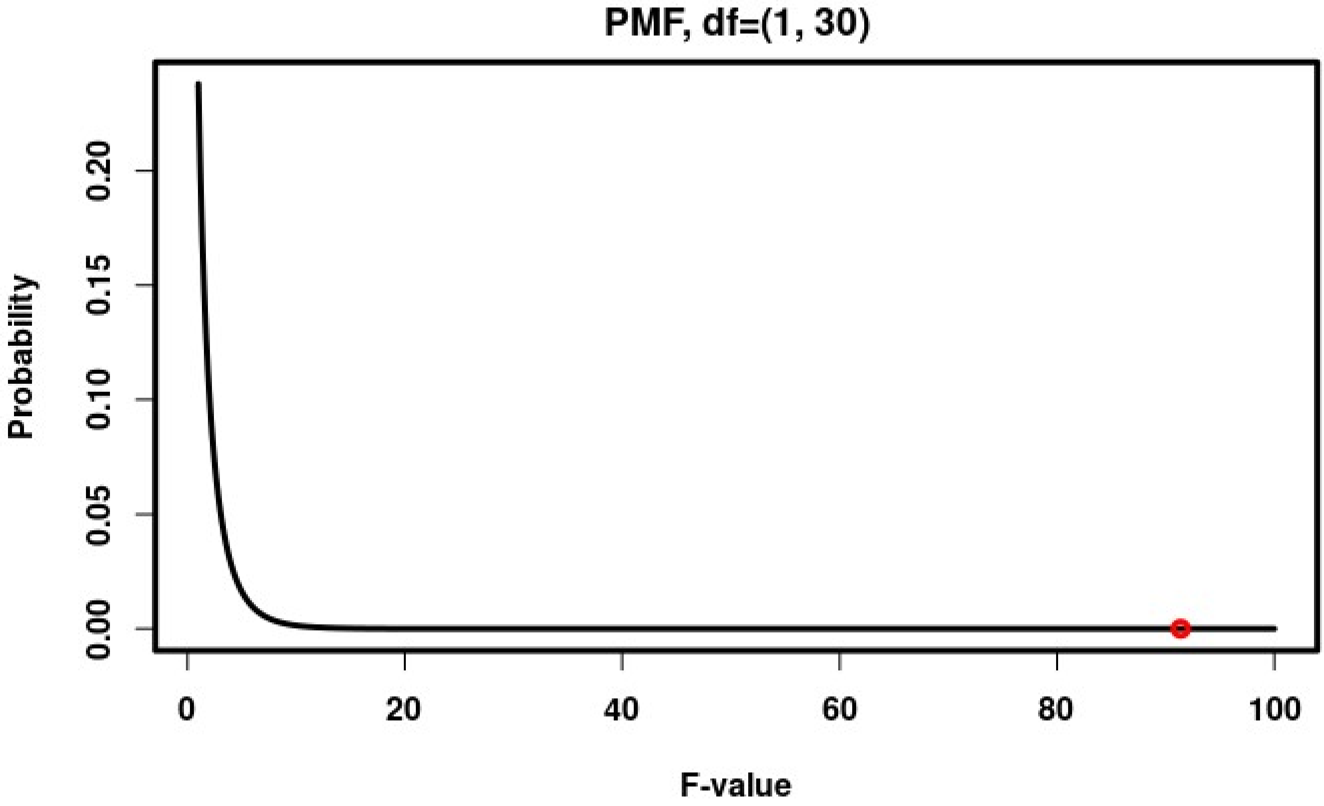
1 = predictor

Target model

```
Call:
lm(formula = mpg ~ 1, data = mtcars)
```

Null model

F-values – single level modelling




```
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

F-values – single level modelling

```
##
## Call:
## lm(formula = mpg ~ I(wt^2) + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.483 -1.998 -0.773  1.462  6.238
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  49.9308     4.2113  11.856 1.21e-12 ***
## I(wt^2)       1.1711     0.3594   3.258  0.00286 **
## wt          -13.3803     2.5140  -5.322 1.04e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.651 on 29 degrees of freedom
## Multiple R-squared:  0.8191, Adjusted R-squared:  0.8066
## F-statistic: 65.64 on 2 and 29 DF,  p-value: 1.715e-11
```

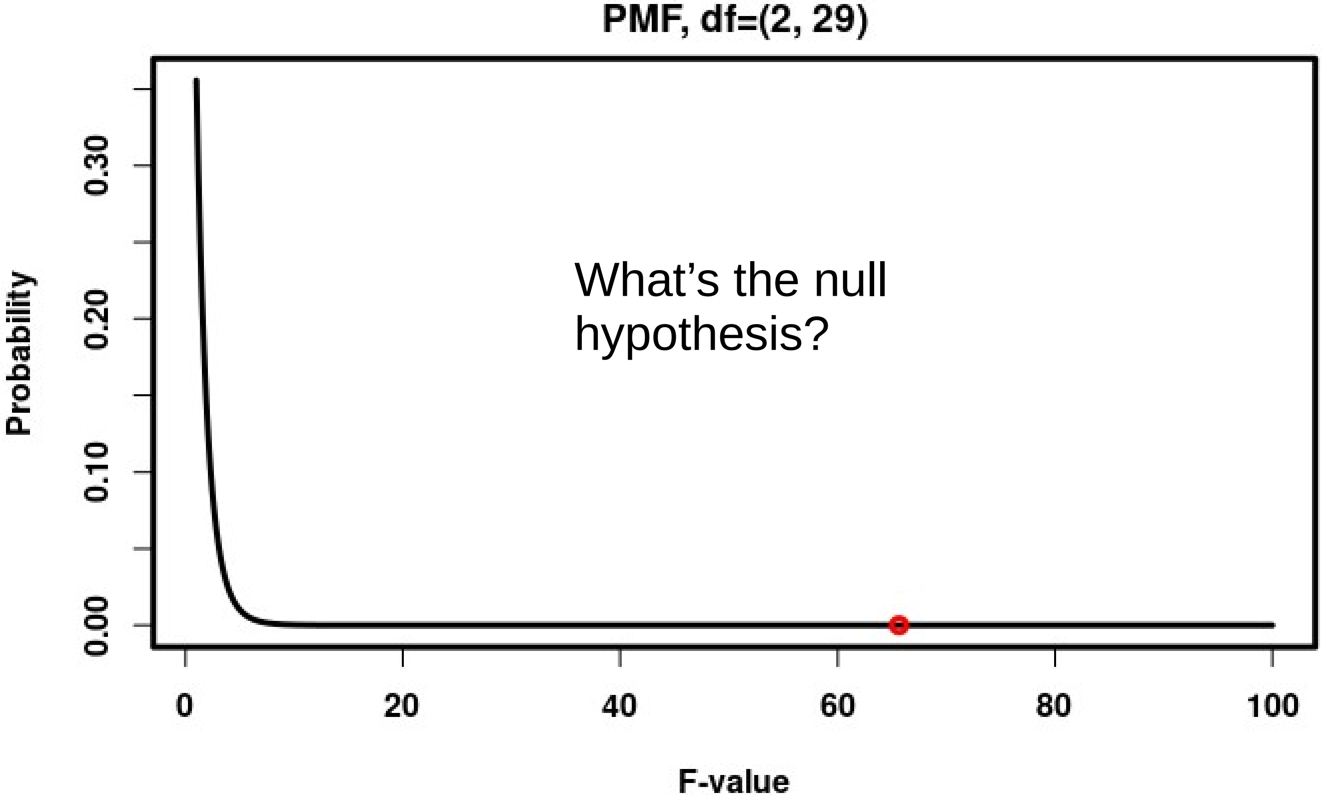
Target model

```
Call:
lm(formula = mpg ~ 1, data = mtcars)
```

Null model,
(still!)

different degrees of freedom, cuz you have 1 more predictor than before

F-values – single level modelling




```
## F-statistic: 65.64 on 2 and 29 DF,  p-value: 1.715e-11
```

```
anova(model.1, model.2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ I(wt^2) + wt
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 278.32
## 2     29 203.75  1    74.576 10.615 0.00286 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null model

Target model

What’s the null hypothesis?

Likelihood ratio

use thizzzz

- Pros
  - Models can be compared in a principled way by reference to a theoretical distribution, χ^2. (In the single level case, *F* can be calculated)
- Cons
  - Models have to be nested in one another
  - Maximum likelihood fitting may be biased for complex models
  - Requires large sample sizes
  - Be careful if collinearity is high

3) Information criteria

Information criteria

$$deviance = -2\,l(\hat{\theta})$$

$$AIC = deviance + \boxed{2}\,k$$

$$k : \text{number of predictors}$$

When we add *k* predictors that are pure noise, deviance is reduced by an amount corresponding to a χ^2 distribution with *k* degrees of freedom.

(Gelman and Hill, 2006)

“On average, a predictor needs to reduce the deviance by 2 to in order to improve the fit to new data”

(Gelman and Hill, 2006, p. 525)

```
## Data: sleepstudy
## Models:
## model.ranint: Reaction ~ days_deprived + (1 | Subject)
## model.ranslope.and.int: Reaction ~ days_deprived + (days_deprived | Subject)
##                        Df    AIC    BIC  logLik deviance  Chisq Chi Df
## model.ranint            4 1446.5 1458.4 -719.25   1438.5
## model.ranslope.and.int  6 1425.2 1443.0 -706.58   1413.2 25.332      2
##                        Pr(>Chisq)
## model.ranint
## model.ranslope.and.int  3.156e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How are the degrees of freedom calculated?

Hints: **Likelihood function**: $L(\sigma^2, \epsilon)$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - x_i \hat{\beta}_N)^2$$

How to estimate *k*?

With *J* parameters

- Complete pooling
  - all *J* parameters collapsed into 1 parameter, (*k* = 1)
- No pooling
  - each of the *J* parameters is estimated, (*k* = *J*)
- Partial pooling
  - the number of effective parameters may be estimated by sampling and the *Deviance Information Criterion* can be estimated (beyond this course)
  - the number of parameters, loosely speaking, depend on whether the parameters are estimated mostly by the group average or the individual's own average (*k* = ?)

Information criteria

- Pros
  - Models can be compared even though one is not nested within the other (response data has to be the same though)
- Cons
  - Number of effective parameters not well defined for multilevel models
  - Maximum likelihood fitting may be biased for complex models

Did you learn? (it’s not easy)

Evaluating and comparing models

1) Learning tools for comparing models
    1) Variance explained
    2) Likelihood ratio tests
    3) Information criteria
2) Bridging to out-of-sample

Learning goals

Explanation and prediction

1) Understanding that fitting (explaining) often leads to overfitting

2) Learning methods to prevent overfitting by introducing *bias*

3) Understanding how the error can be decomposed into *bias* and *variance*

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{T} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^{T} \boldsymbol{Y}$$

maximises the likelihood of

$$\boldsymbol{Y} = \boldsymbol{X} \boldsymbol{\beta} + \boldsymbol{\epsilon}$$

for which link function?

To fit is to overfit

(Yarkoni and Westfall, 2017)

Overfitting: fitting sample-specific noise, which is thus not representative of the population

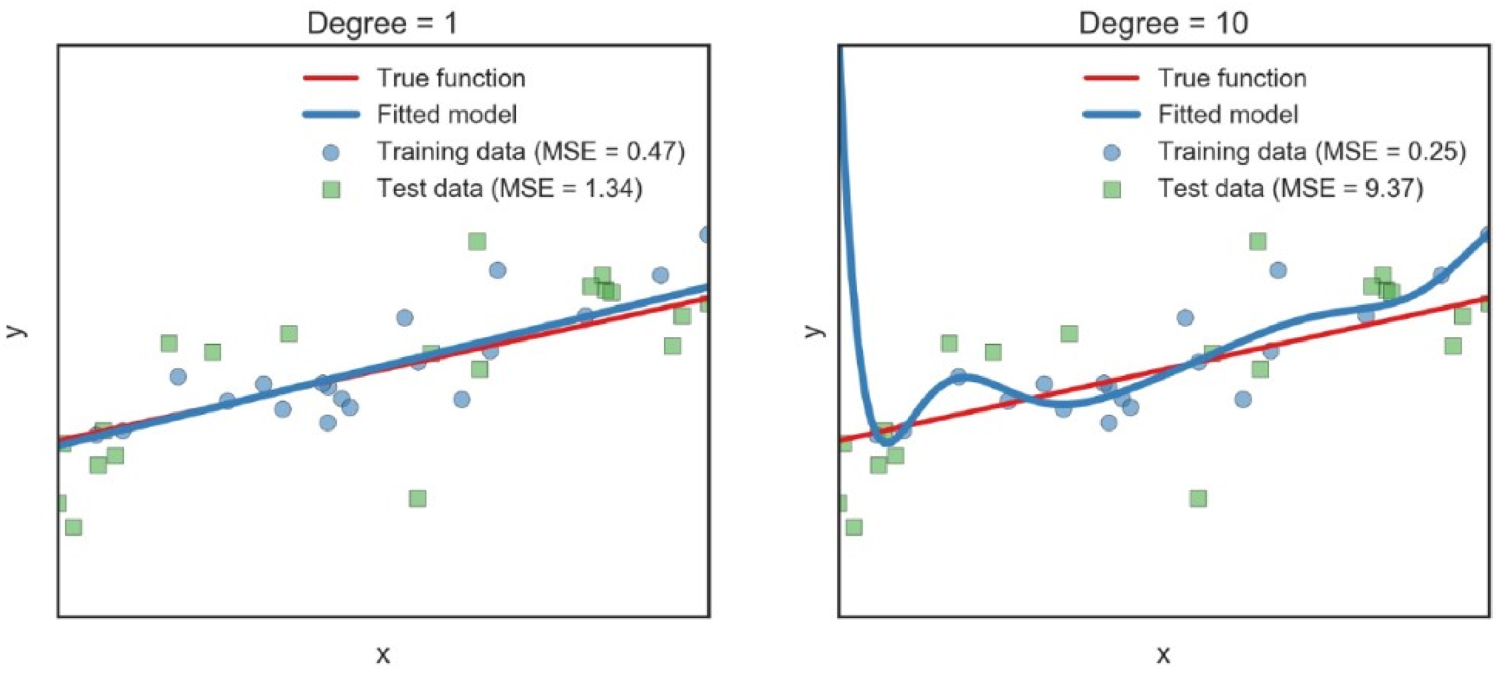
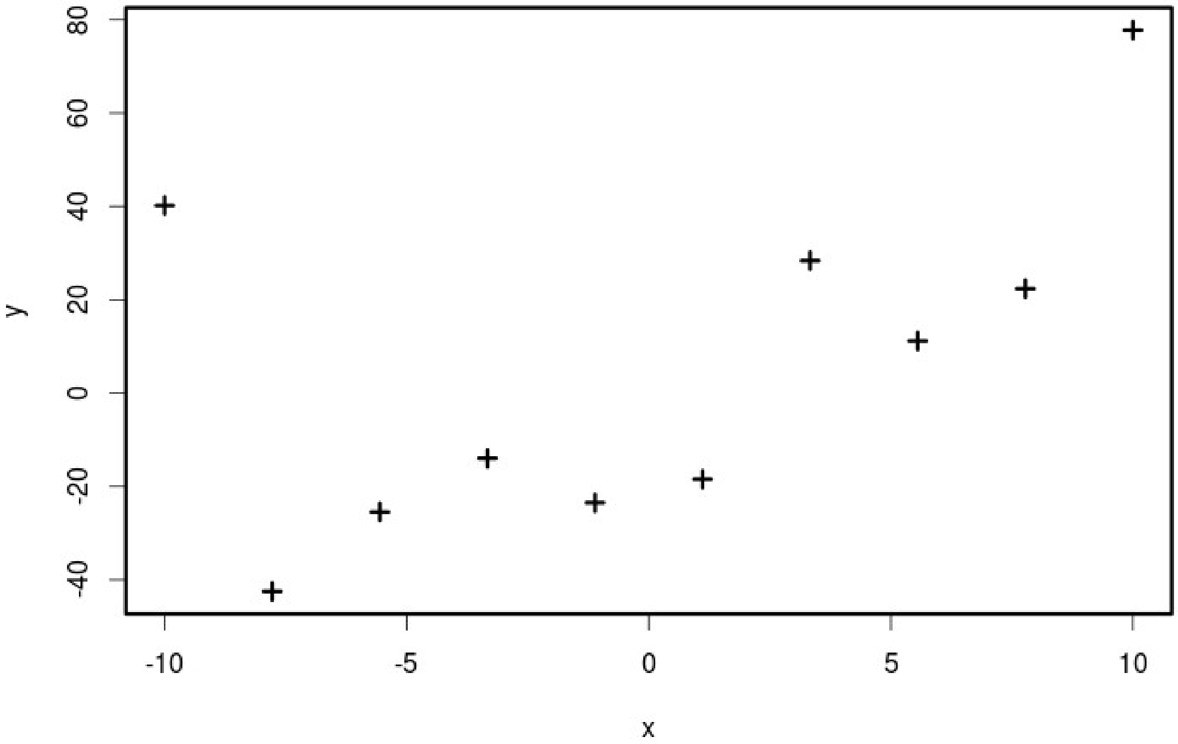


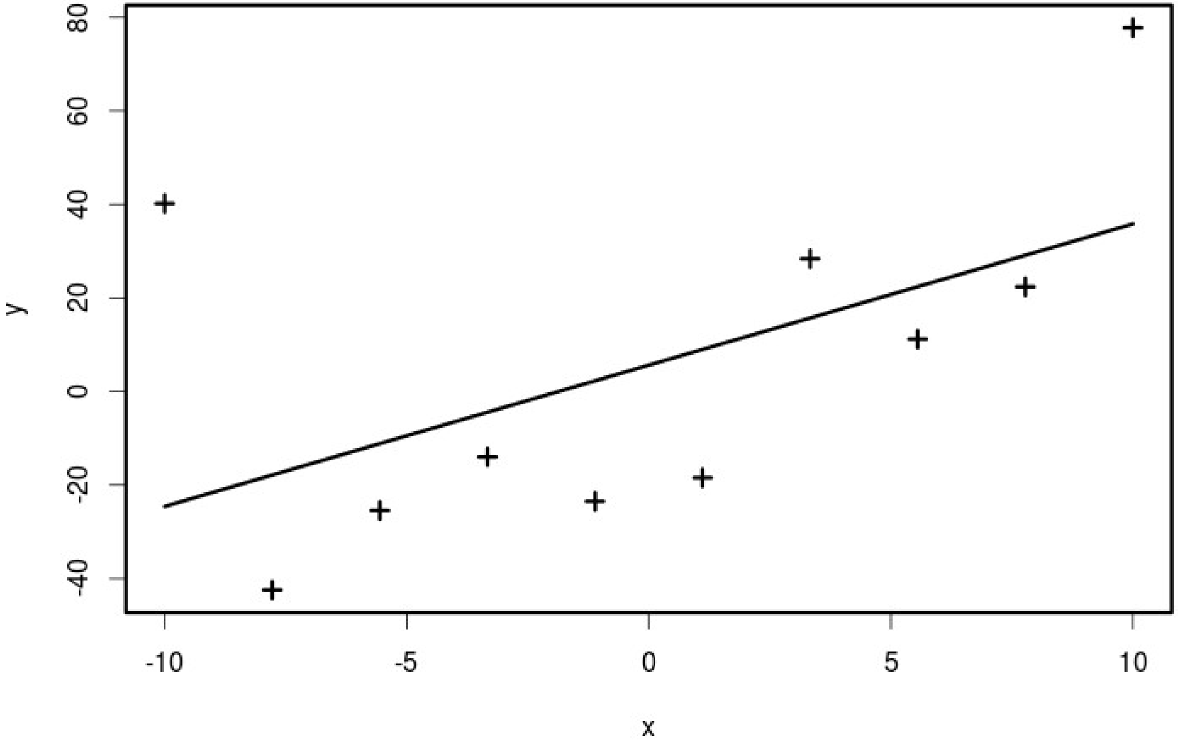
Fig. 1. Training and test error produced by fitting either a linear regression (left) or a 10th-order polynomial regression (right) when the true relationship in the population (red line) is linear. In both cases, the test data (green) deviate more from the model's predictions (blue line) than the training data (blue). However, the flexibility of the 10th-order polynomial model facilitates much greater overfitting, resulting in lower training error but much higher test error than the linear model. MSE = mean squared error.

(Yarkoni and Westfall, 2017)

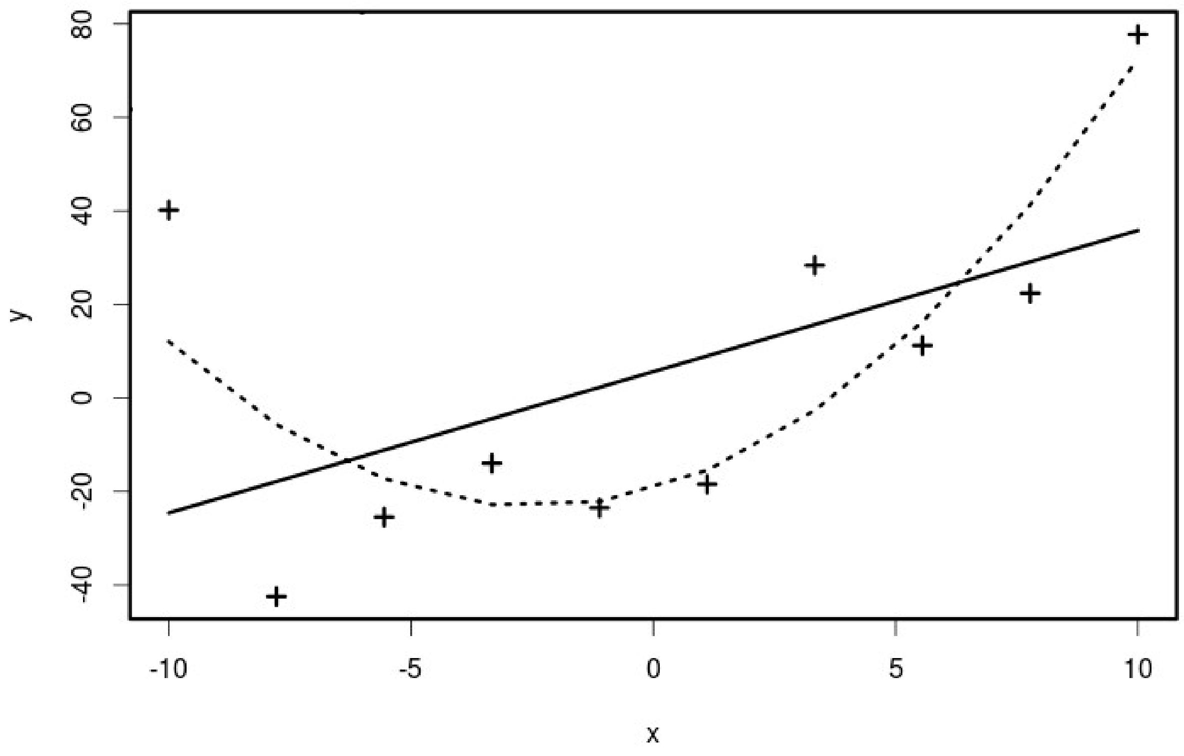
A sample of 10
linear or quadratic?



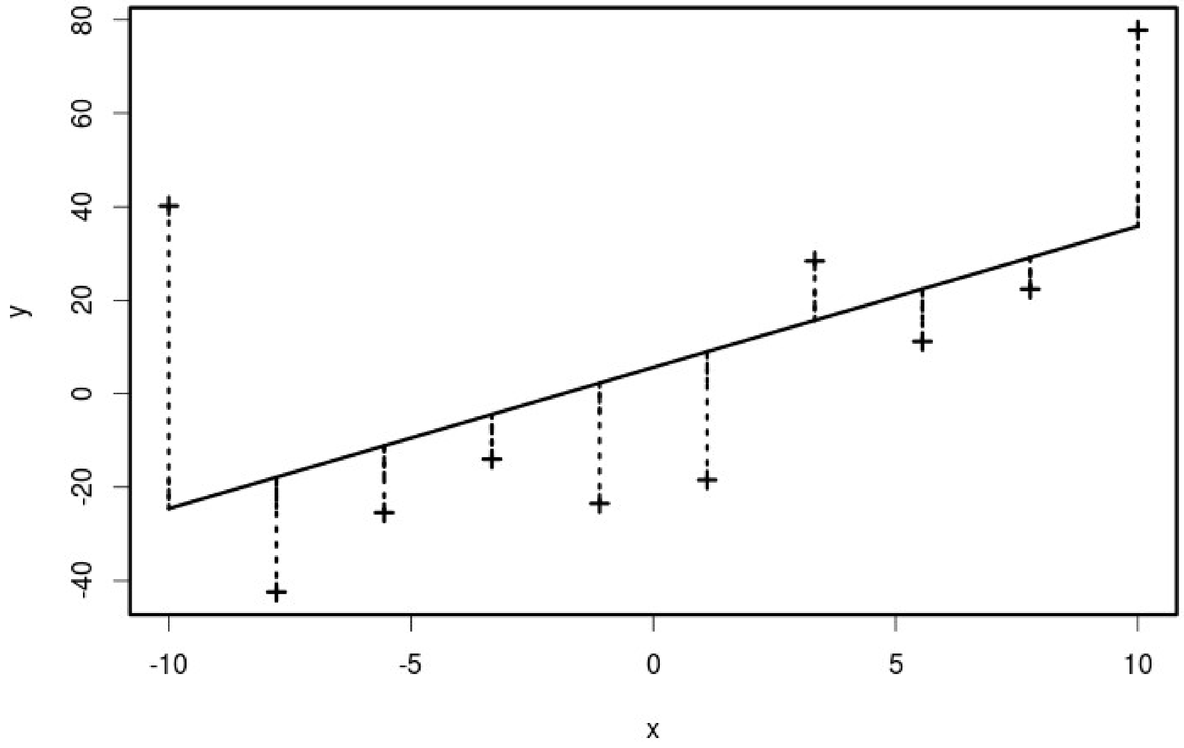
A sample of 10
linear or quadratic?



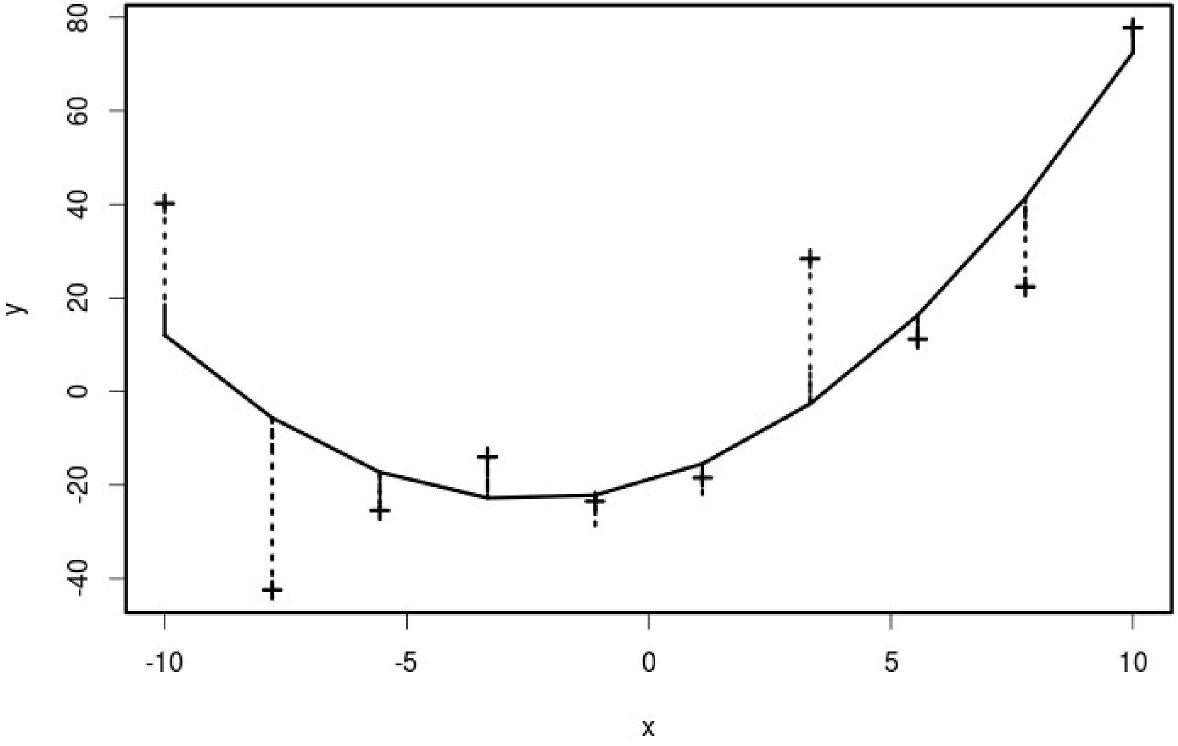
A sample of 10
linear or quadratic?



Residuals (linear)



Residuals (quadratic)



Quadratic:

$ax^2 + bx + c$

Linear:

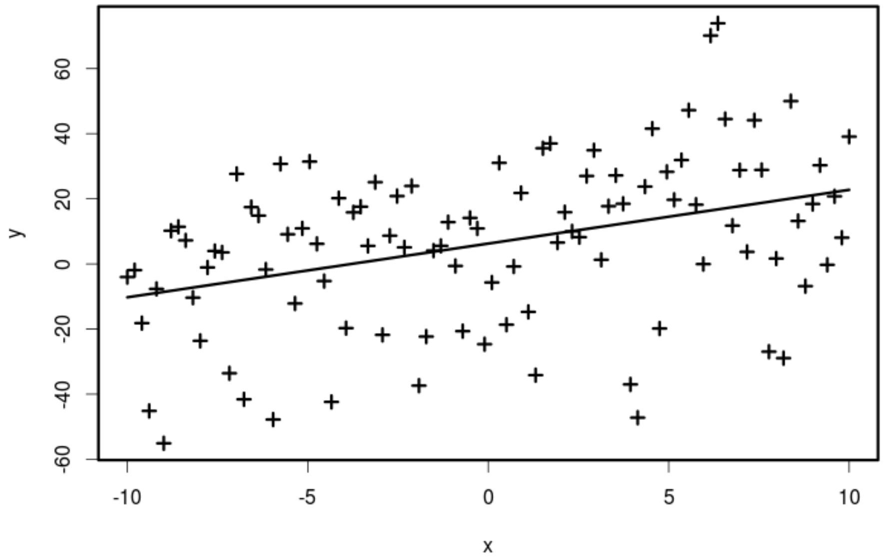
$bx + c$

Estimates

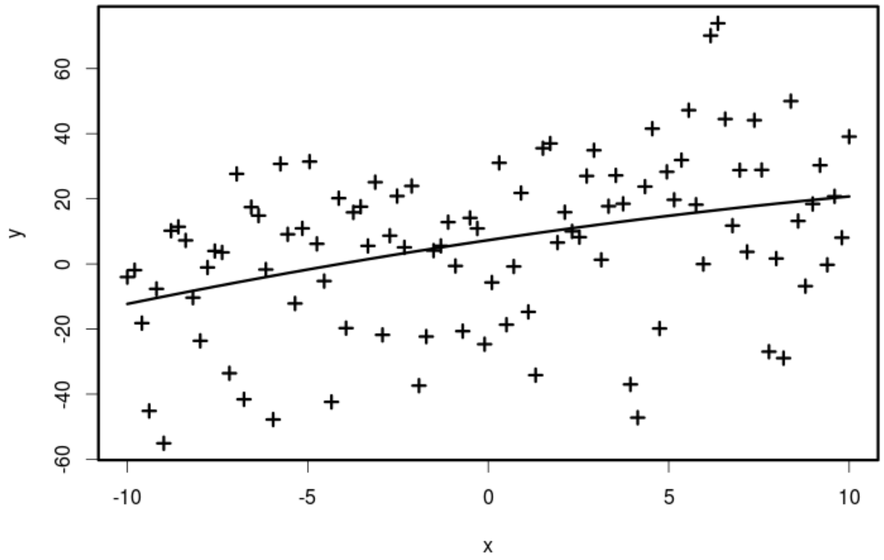
a = 0,6184; b = 3,0201

b = 3,020

Now a sample of 100



b = 1,650



a = -0,03074 ≈ 0; b = 1,650

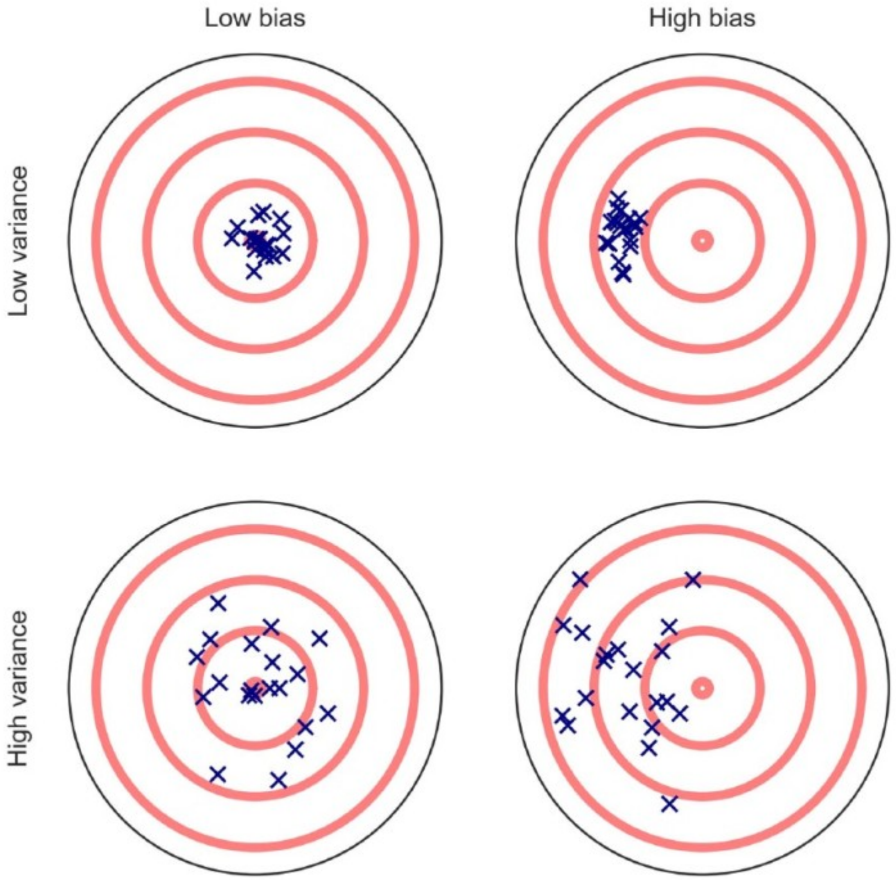


Fig. 2. An estimator's predictions can deviate from the desired outcome (or true scores) in two ways. First, the predictions may display a systematic tendency (or *bias*) to deviate from the central tendency of the true scores (compare right panels with left panels). Second, the predictions may show a high degree of *variance*, or imprecision (compare bottom panels with top panels).

(Yarkoni and Westfall, 2017)

Which two of these does our least-squares solution prefer?

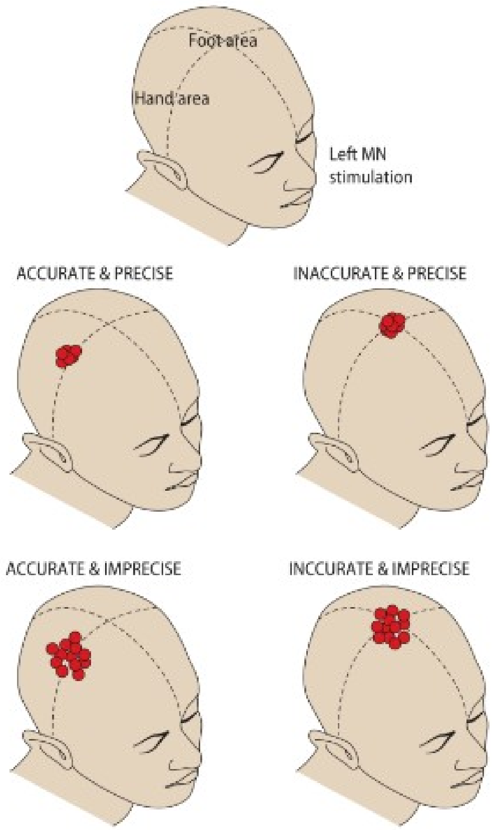


FIGURE 3.7. **Accuracy versus precision.** A schematic illustration of the differences between accuracy and precision of source localization. After left median-nerve stimulation, activations is expected in the right-hemisphere hand region of the primary somatosensory cortex. The foot area is shown at the top of the head. See text for further explanation.

(Hari and Puce, 2017)

(Yarkoni and Westfall, 2017)

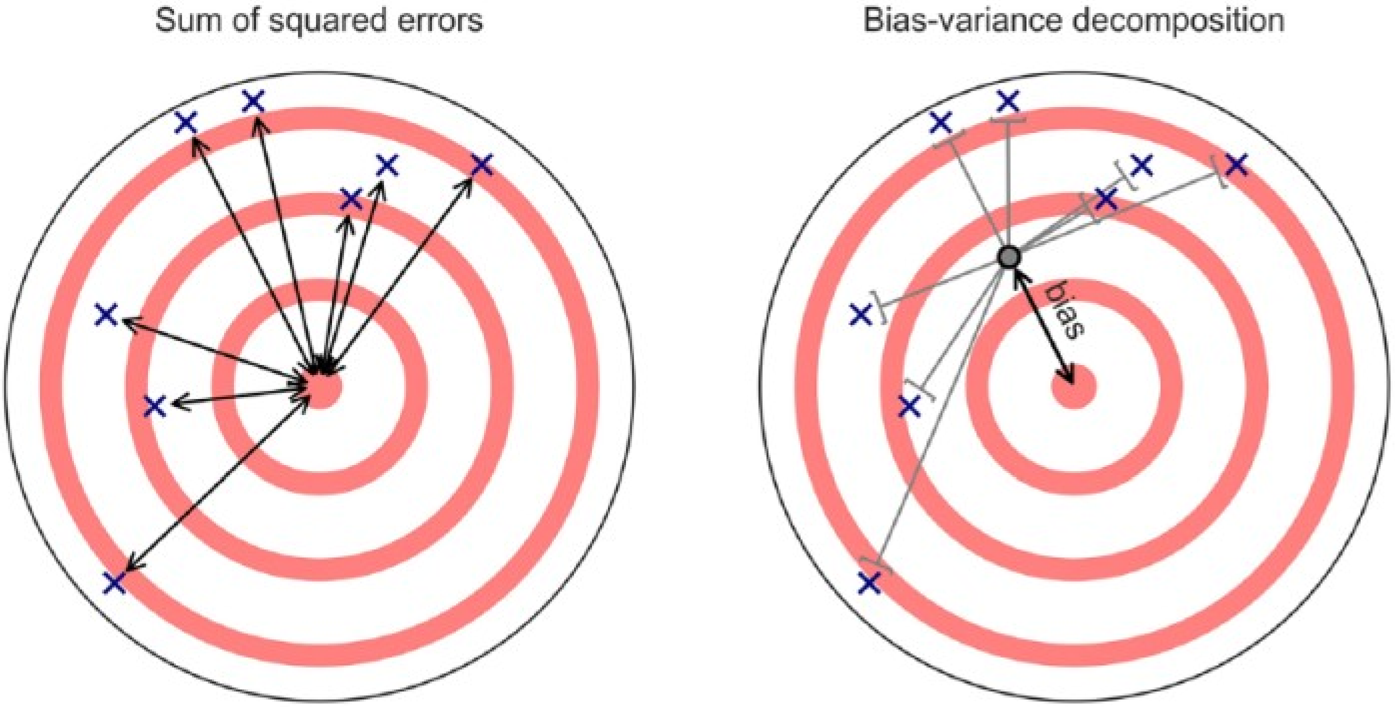


Fig. 3. Schematic illustration of the bias-variance decomposition. (Left) Under the classical error model, prediction error is defined as the sum of squared differences between true scores and observed scores (black lines). (Right) The bias-variance decomposition partitions the total sum of squared errors into two separate components: a bias term that captures a model's systematic tendency to deviate from the true scores in a predictable way (black line) and a variance term that represents the deviations of the individual observations from the model's expected prediction (gray lines).

Multilevel modelling as a *bias* introducer

“*For example, some readers may be surprised to learn that multilevel modeling approaches to analyzing clustered data—which have recently seen a dramatic increase in adoption in psychology—improve on ordinary least squares (OLS) approaches to estimating individual cluster effects by deliberately biasing (through “shrinking” or “pooling”) the cluster estimates toward the estimated population average*”

(Yarkoni and Westfall, 2017)

Introducing *bias*

“In a widely used form of penalized regression called lasso regression (Tibshirani, 1996, 2011), this leastsquares criterion is retained, but the overall cost function that the estimation seeks to minimize now includes an additional penalty term that is proportional to the sum of the absolute values of the coefficients.”

(Yarkoni and Westfall, 2017)

Penalised regression

$$\text{RSS} = \sum (y_i - \hat{y}_i)^2 (\text{minimise to obtain least squares solution })$$

$$\text{lasso regression}: \text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j| (\text{minimise this sum})$$

$$\text{ridge regression}: \text{RSS} + \lambda \sum_{j=1}^{p} (\beta_j^2) (\text{minimise this sum})$$

i : observations
p : predictor variables
λ : a constant

Penalised regression

$$\text{RSS} = \sum (y_i - \hat{y}_i)^2 (\text{minimise to obtain least squares solution})$$

$$\text{lasso regression}: \text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j| (\text{minimise this sum})$$

$$\text{ridge regression}: \text{RSS} + \lambda \sum_{j=1}^{p} (\beta_j^2) (\text{minimise this sum})$$

i: observations
p: predictor variables
λ: a constant

Group discussion

In each case: what happens when?

1. λ increases?
2. λ decreases?
3. λ is 0?
4. λ goes towards infinity?

Penalised regression

$$\mathrm{RSS} = \sum (y_i - \hat{y}_i)^2 (\text{minimise to obtain least squares solution })$$

$$\underset{\lambda}{\operatorname{argmin}} = \mathrm{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$

$$\underset{\lambda}{\operatorname{argmin}} = \mathrm{RSS} + \lambda \sum_{j=1}^{p} (\beta_j^2)$$

$i:$ observations

$p:$ predictor variables

$\lambda:$ a constant

How to choose λ?

```
##
## Call:
## lm(formula = hp ~ mpg + wt + drat + qsec, data = mtcars)
##
## Coefficients:
## (Intercept)          mpg           wt         drat         qsec
##     473.779       -2.877       26.037        4.819      -20.751
```

What is λ equal to here?

```
sum.of.squares.total <- sum((y - mean(y))^2)
sum.of.squared.errors.lm <- sum(residuals(linear_model)^2)
print(r.squared.lm <- 1 - sum.of.squared.errors.lm/sum.of.squares.total)
```

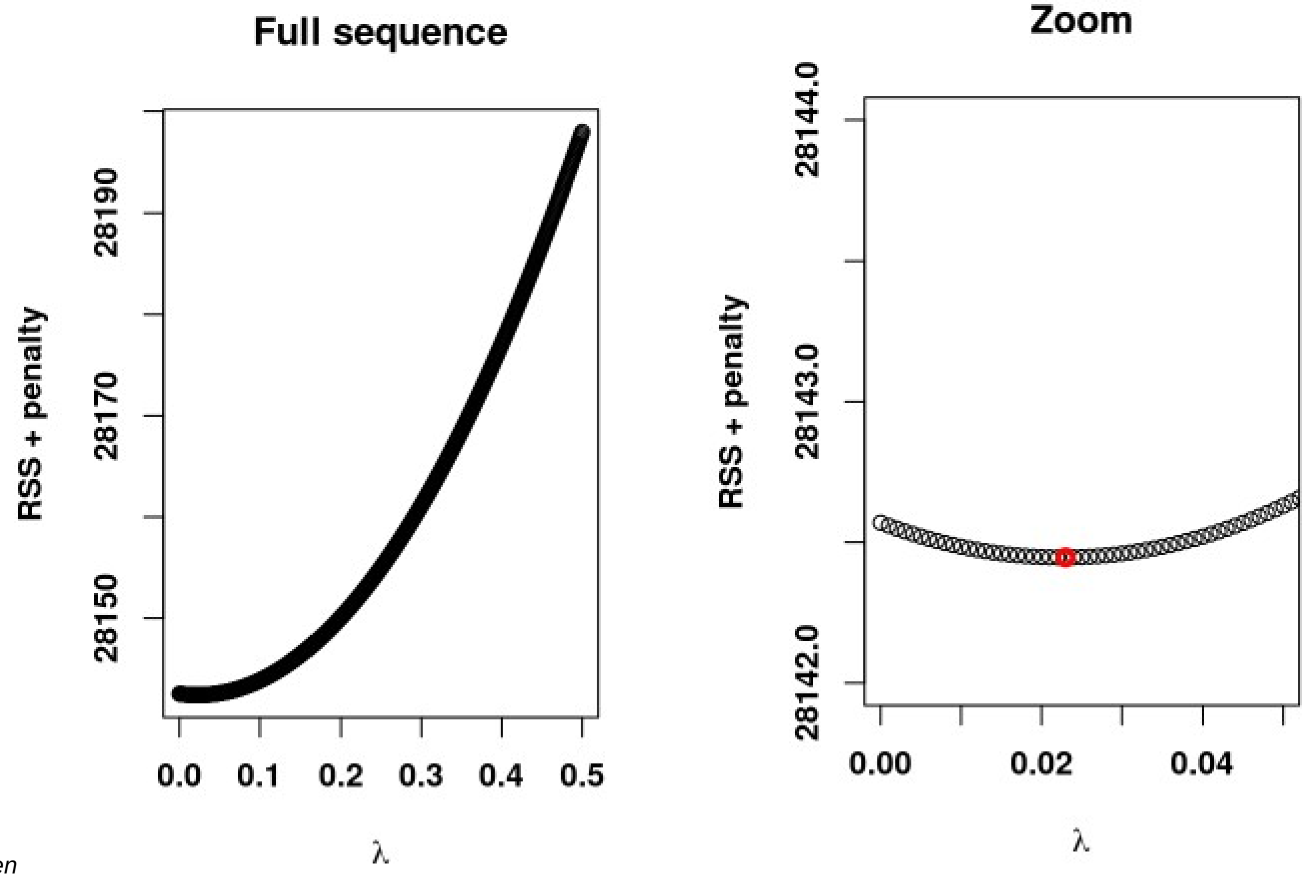
```
## [1] 0.8072553
```

How to choose λ (lasso)?

```
##
## Call:  glmnet(x = x, y = y, alpha = 1, lambda = c(0, 0.2, 2, 4, 20,
100))
##
##   Df   %Dev Lambda
## 1  0 0.0000  100.0
## 2  2 0.6567   20.0
## 3  3 0.8004    4.0
## 4  3 0.8051    2.0
## 5  4 0.8072    0.2
## 6  4 0.8073    0.0
```

lambda	RSS	penalty	sum
100.0	145726.9	0.0	145726.9
20.00000	50025.64218	14.05496	50039.69714
4.00000	29082.92003	41.34363	29124.26366
2.00000	28408.50683	44.99057	28453.49740
0.20000	28097.60764	52.47741	28150.08505
0.00000	28088.09951	54.46997	28142.56948

How to choose λ?



What does λ do?

$$\hat{\beta}_{ridge} = \left(X^T X + \lambda I \right)^{-1} X^T y$$

I : an identity matrix with p predictor variables

λ : a constant

What is $X^T X$?

```r
head(X)
```

```
##                        (Intercept)  mpg    wt drat  qsec
## Mazda RX4                        1 21.0 2.620 3.90 16.46
## Mazda RX4 Wag                    1 21.0 2.875 3.90 17.02
## Datsun 710                       1 22.8 2.320 3.85 18.61
## Hornet 4 Drive                   1 21.4 3.215 3.08 19.44
## Hornet Sportabout                1 18.7 3.440 3.15 17.02
## Valiant                          1 18.1 3.460 2.76 20.22
```

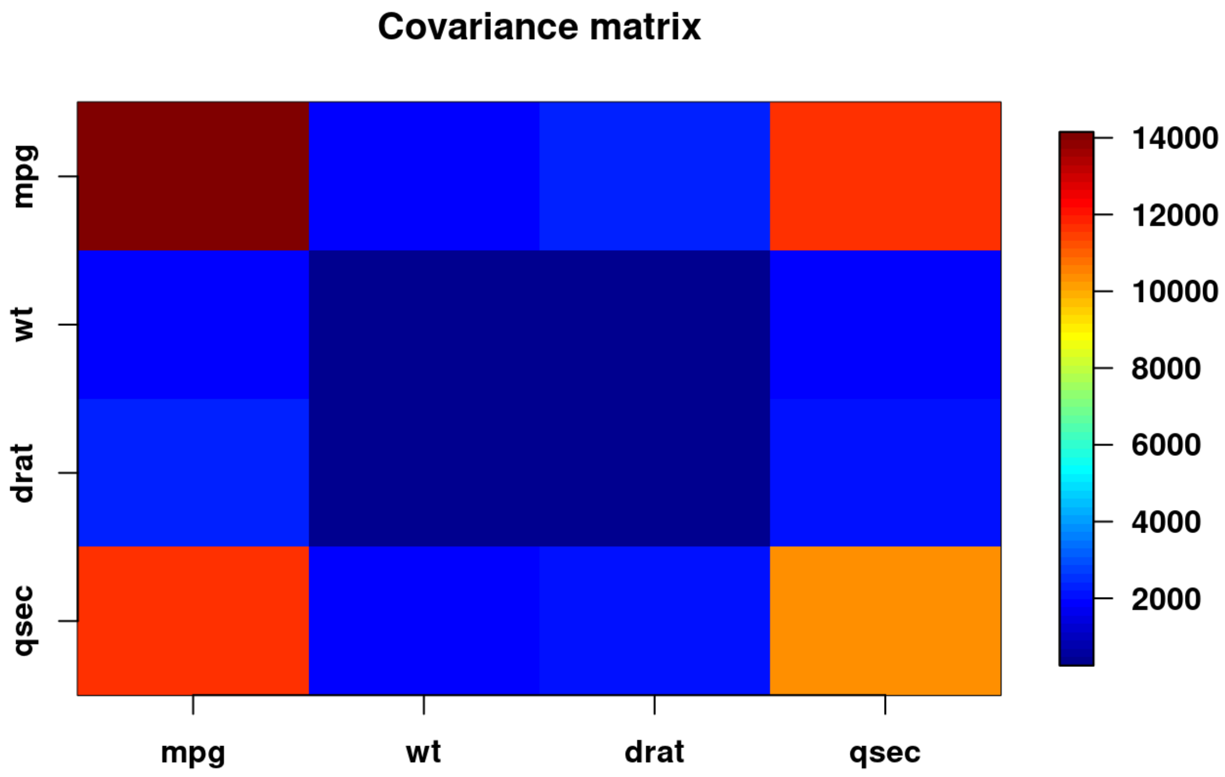
$$X_{COV} = X^T X$$

```r
print(cov.X)
```

```
##            mpg        wt      drat      qsec
## mpg  14042.310 1909.7528 2380.2770 11614.745
## wt    1909.753  360.9011  358.7190  1828.095
## drat  2380.277  358.7190  422.7907  2056.914
## qsec 11614.745 1828.0946 2056.9140 10293.480
```

$$X_{COV} = X^T X$$

The fact that the off-diagonal > 0, indicates that there is **collinearity**



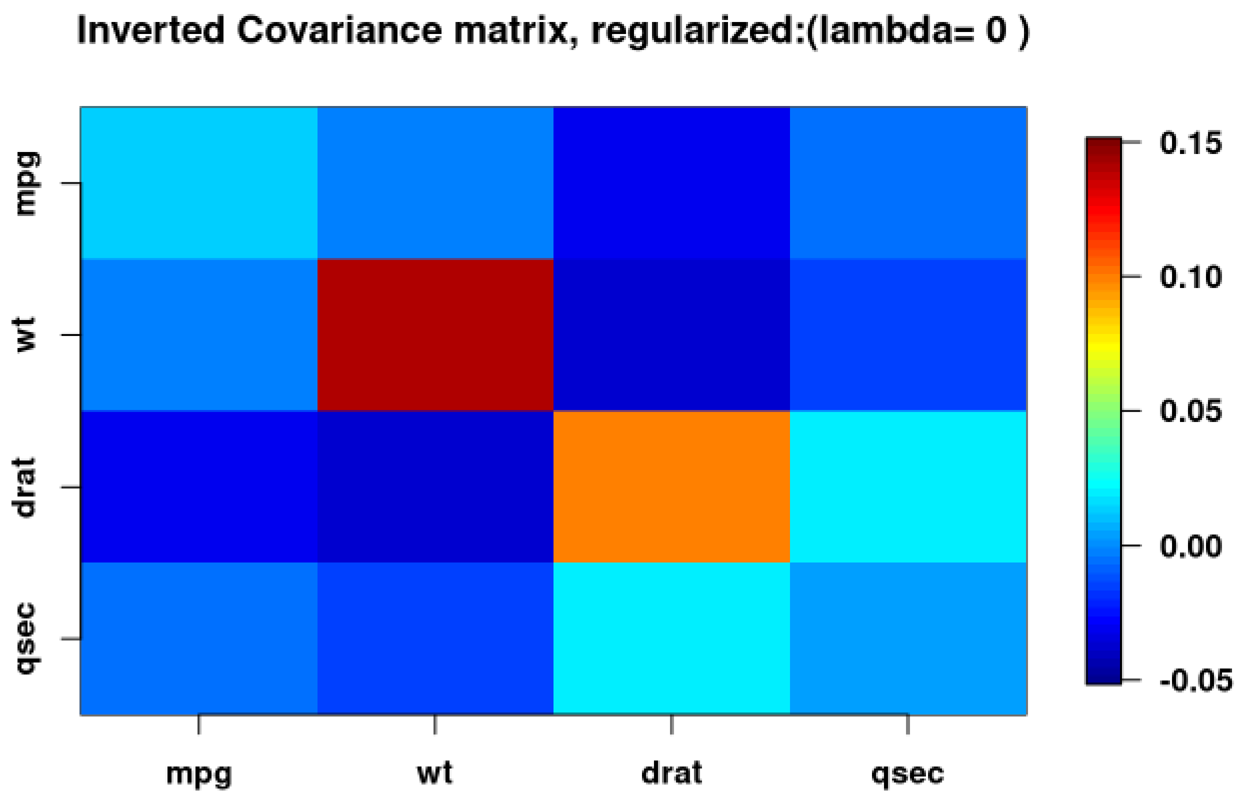
Collinearity can be bad

```
##
## Call:
## lm(formula = hp ~ mpg + wt + drat + qsec, data = mtcars)
##
## Coefficients:
## (Intercept)          mpg           wt         drat         qsec
##     473.779       -2.877       26.037        4.819      -20.751
```

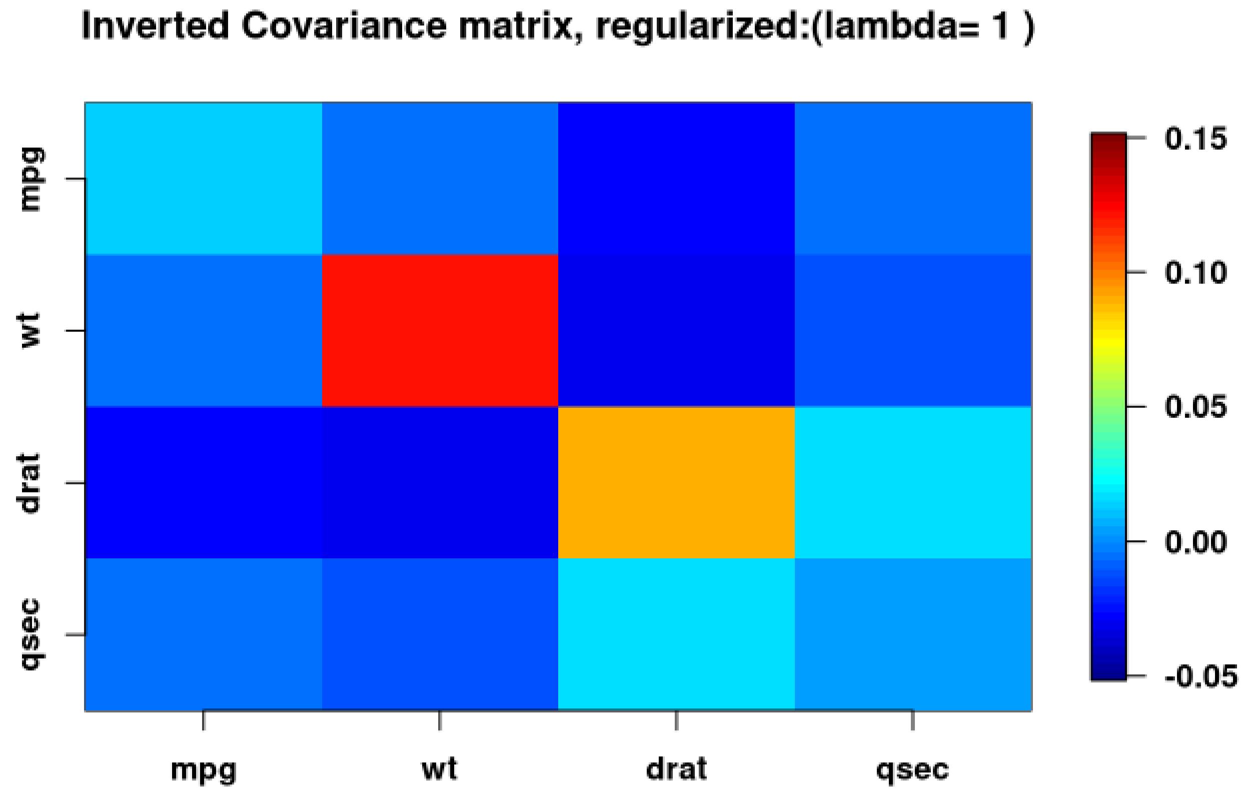
Assuming no collinearity, what is the interpretation of the coefficients?

With collinearity, is that interpretation possible?

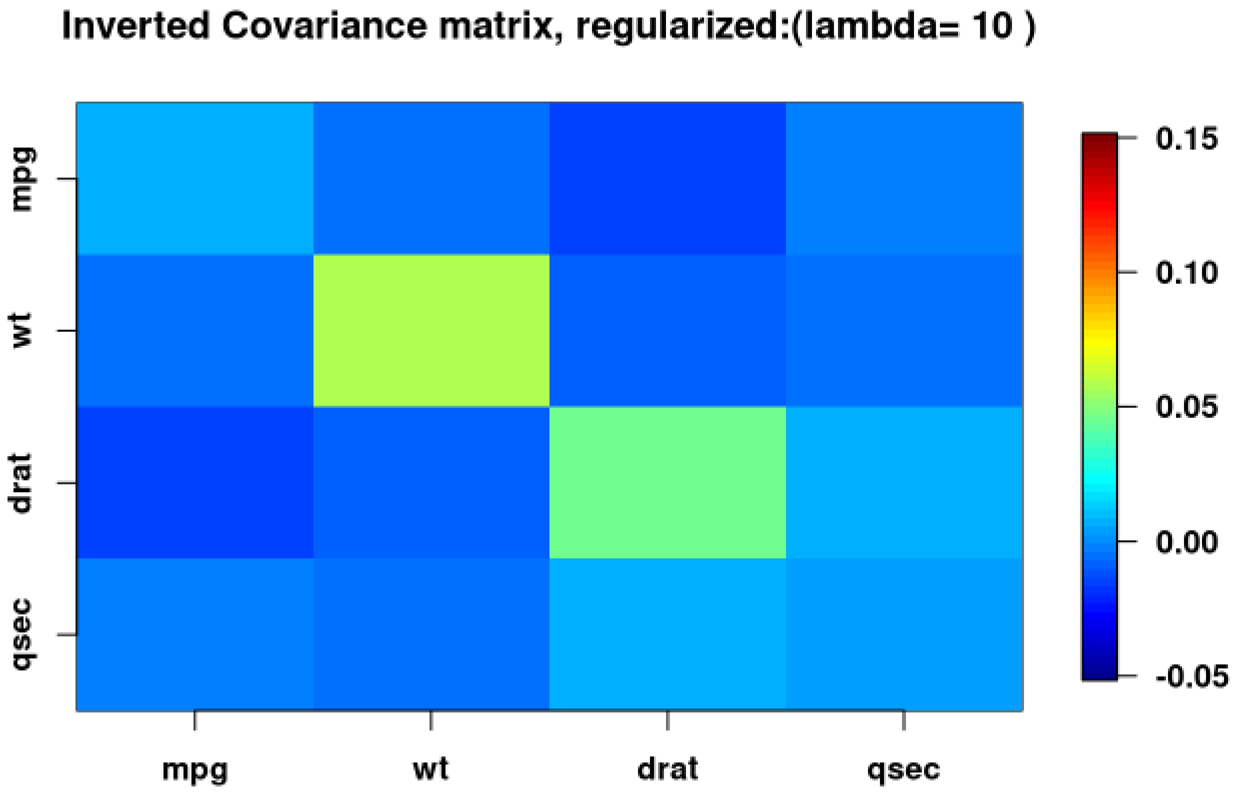
$$\hat{\beta}_{ridge} = \boxed{\left(X^T X + \lambda I \right)^{-1}} X^T y$$



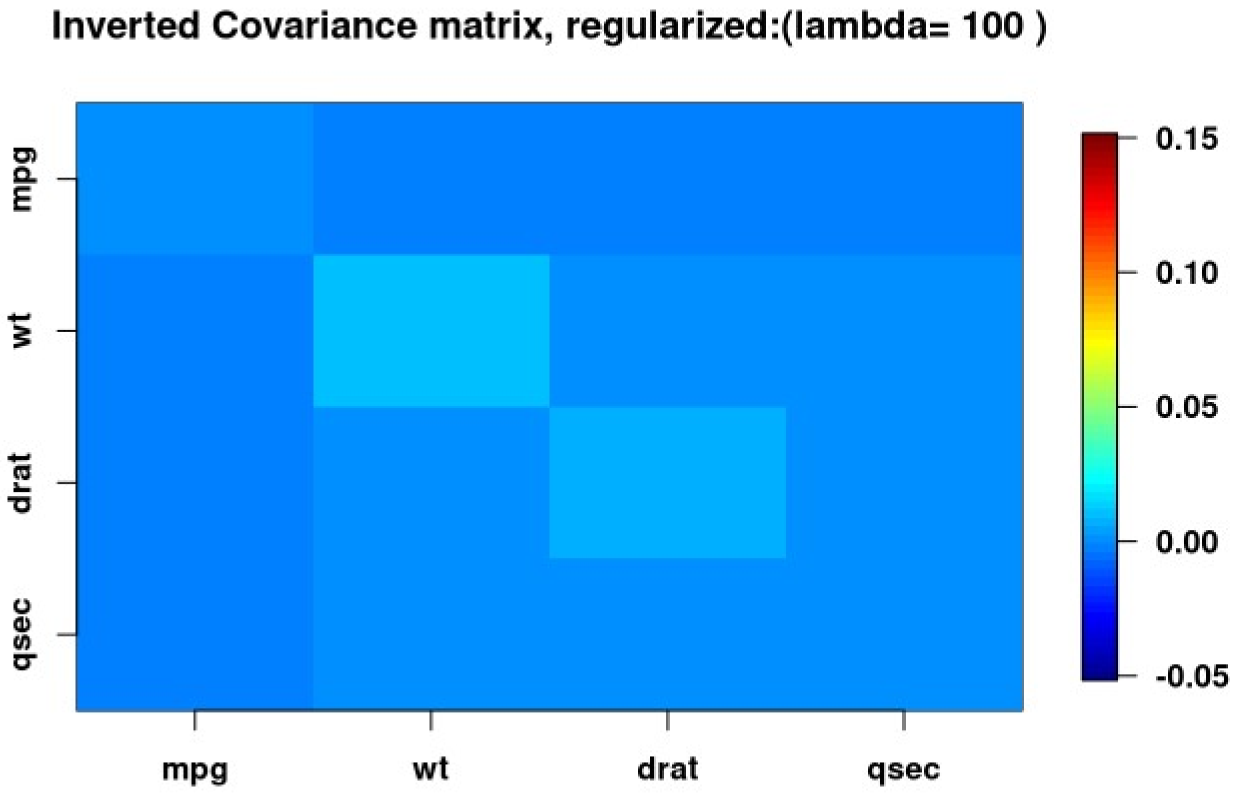
$$\hat{\beta}_{ridge} = \boxed{(X^T X + \lambda I)^{-1}} X^T y$$



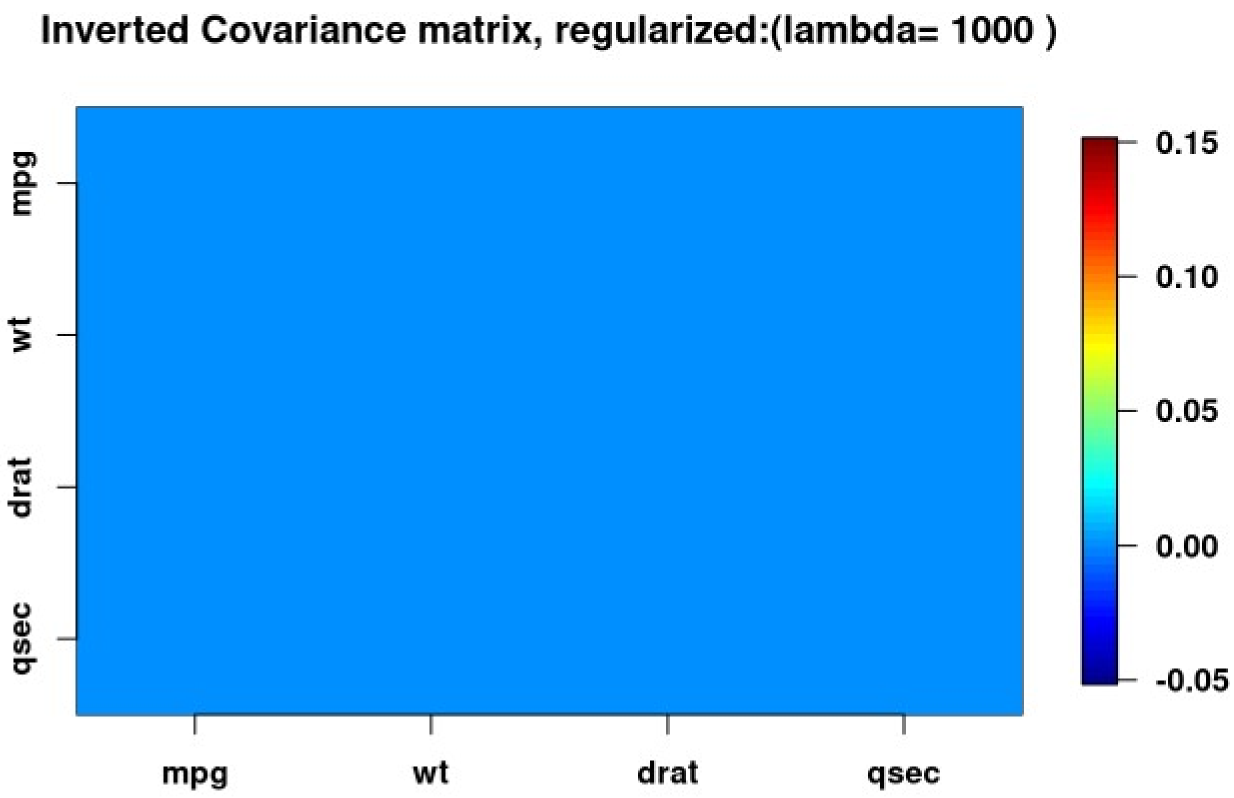
$$\hat{\beta}_{ridge} = \boxed{\left(X^T X + \lambda I\right)^{-1}} X^T y$$



$$\hat{\beta}_{ridge} = \boxed{\left(X^T X + \lambda I \right)^{-1}} X^T y$$



$$\hat{\beta}_{ridge} = \boxed{\left(X^T X + \lambda I\right)^{-1}} X^T y$$



So why is it called regularisation?

Two notes about the inverted matrix:

with increase of *λ*

1. Diagonal shrinks (bias is added)

2. Off-diagonal shrinks (collinearity is reduced, which improves the stability of the model)

Stability of a model:

Feeding new data or adding new predictor variables will not change the parameter estimates a lot

We have succeeded in finding a λ making our model more stable (improved **in-sample** validity), but we haven’t found a λ that optimises predictive power – (**out-of-sample**) – week 6)

Did you learn?

Explanation and prediction

1) Understanding that fitting (explaining) often leads to overfitting

2) Learning methods to prevent overfitting by introducing *bias*

3) Understanding how the error can be decomposed into *bias* and *variance*

References

- Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H., White, J.-S.S., 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology & Evolution 24, 127–135. https://doi.org/10.1016/j.tree.2008.10.008
- Gelman, A., Hill, J., 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.
- Hari, R., Puce, A., 2017. MEG-EEG Primer. Oxford University Press, New York, NY, US.
- Yarkoni, T., Westfall, J., 2017. Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. Perspect Psychol Sci 12, 1100–1122. https://doi.org/10.1177/1745691617693393