METHODS 3: MULTILEVEL STATISTICAL MODELLING AND MACHINE LEARNING

COURSE OVERVIEW (FIRST HALF)

W1: Introduction

Setting up R and Python and recollection of the general linear model

W2: Linear Mixed Effects Models

Modelling random effects – and how do they differ from fixed effects?

W3: Generalized Linear Mixed Effects Models

What to do when the response variable is not continuous?

W4: Explanation and prediction

Why are good explanations sometimes bad?

W5: Evaluating and comparing models

How do we assess how models compare to one another?

Fall break:

Machine Learning and Python programming follows

COURSE OVERVIEW (FIRST HALF)

W1: Introduction

Setting up R and Python and recollection of the general linear model

W2: Linear Mixed Effects Models

Modelling random effects – and how do they differ from fixed effects?

W3: Generalized Linear Mixed Effects Models

What to do when the response variable is not continuous?

W4: Explanation and prediction

Why are good explanations sometimes bad?

W5: Evaluating and comparing models

How do we assess how models compare to one another?

Fall break:

Machine Learning and Python programming follows

TODAYS PLAN

- Catch-up
- Output – logistic regression
- QQ-plots
- *(Model selection criteria*
    - *R-squared (revisited)*
    - *AIC/BIC*
    - *Out-of-sample-error)*
- Assignment tips
- Assignment work

TODAYS PLAN

- **Catch-up**
- Output – logistic regression
- QQ-plots
- *(Model selection criteria*
  - *R-squared (revisited)*
  - *AIC/BIC*
  - *Out-of-sample-error)*
- Assignment tips
- Assignment work

CATCH-UP

CATCH-UP

- How are you hanging in? Last assignment? Git/GitHub?
- Extra readings
- Where would you want new messages?
- Bigger visualizations this time around
- Disclaimer

TODAYS PLAN

- Catch-up
- **Output – logistic regression**
- QQ-plots
- *(Model selection criteria*
  - *R-squared (revisited)*
  - *AIC/BIC*
  - *Out-of-sample-error)*
- Assignment tips
- Assignment work

OUTPUT – LOGISTIC REGRESSION

OUTPUT – LOGISTIC REGRESSION

```
logistic_model <- glm(formula = am ~ wt,
                      data = mtcars,
                      family = "binomial")
```

OUTPUT – LOGISTIC REGRESSION

```
logistic_model <- glm(formula = am ~ wt,
                      data = mtcars,
                      family = "binomial")
```

- am = automatic transmission (1 or 0)
- wt = weight (continuous)

OUTPUT – LOGISTIC REGRESSION

```
logistic_model <- glm(formula = am ~ wt,
                      data = mtcars,
                      family = "binomial")
```

- am = automatic transmission (1 or 0)
- wt = weight (continuous)

Uses link function

->

To log-odds, to have linear scale

OUTPUT – LOGISTIC REGRESSION

```
am ~ wt
```

OUTPUT – LOGISTIC REGRESSION

```
am ~ wt

> summary(logistic_model)

Coefficients:
              Estimate Std.
(Intercept)    12.040
wt             -4.024
---
```

OUTPUT – LOGISTIC REGRESSION

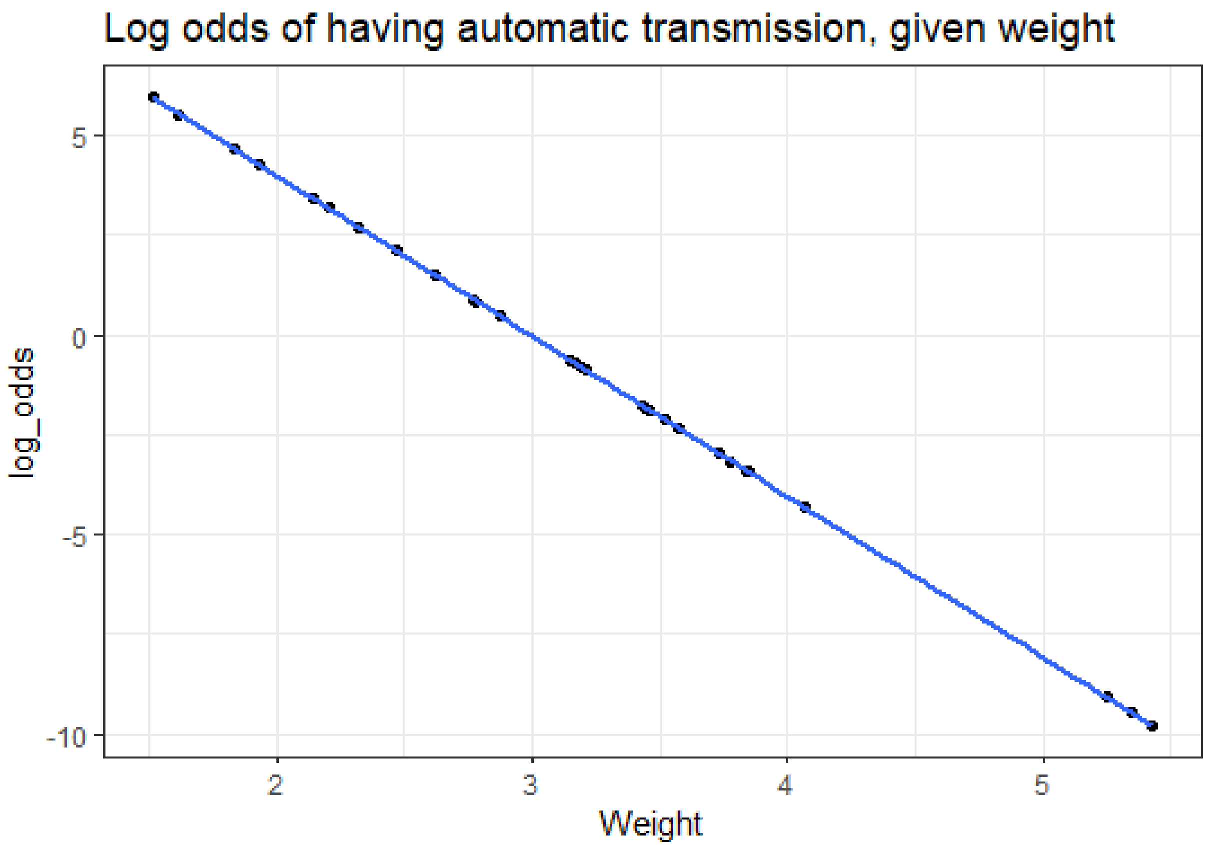
```
am ~ wt

> summary(logistic_model)

Coefficients:
             Estimate Std.
(Intercept)   12.040
wt            -4.024
---
```

- Output in log-odds
  to have linear relationship
  between wt and am

OUTPUT – LOGISTIC REGRESSION

```
Coefficients:
            Estimate Std.
(Intercept)   12.040
wt            -4.024
---
```

OUTPUT – LOGISTIC REGRESSION




```
Coefficients:
            Estimate Std.
(Intercept)   12.040
wt            -4.024
---
```

OUTPUT – LOGISTIC REGRESSION

$y = 12.04 + -4.024 * x$

```
Coefficients:
            Estimate Std.
(Intercept)   12.040
wt            -4.024
---
```

OUTPUT – LOGISTIC REGRESSION

- Going through example of a Mazda

$$y = 12.04 + -4.024 * x$$

```
Coefficients:
            Estimate Std.
(Intercept)   12.040
wt            -4.024
---
```

OUTPUT – LOGISTIC REGRESSION

- Going through example of a Mazda

```
> head(mazda)
          mpg cyl disp  hp drat   wt  qsec vs am gear carb
Mazda RX4  21   6  160 110  3.9 2.62 16.46  0  1    4    4
> y_log_odds = 12.04 + -4.024 * 2.62
```

OUTPUT – LOGISTIC REGRESSION

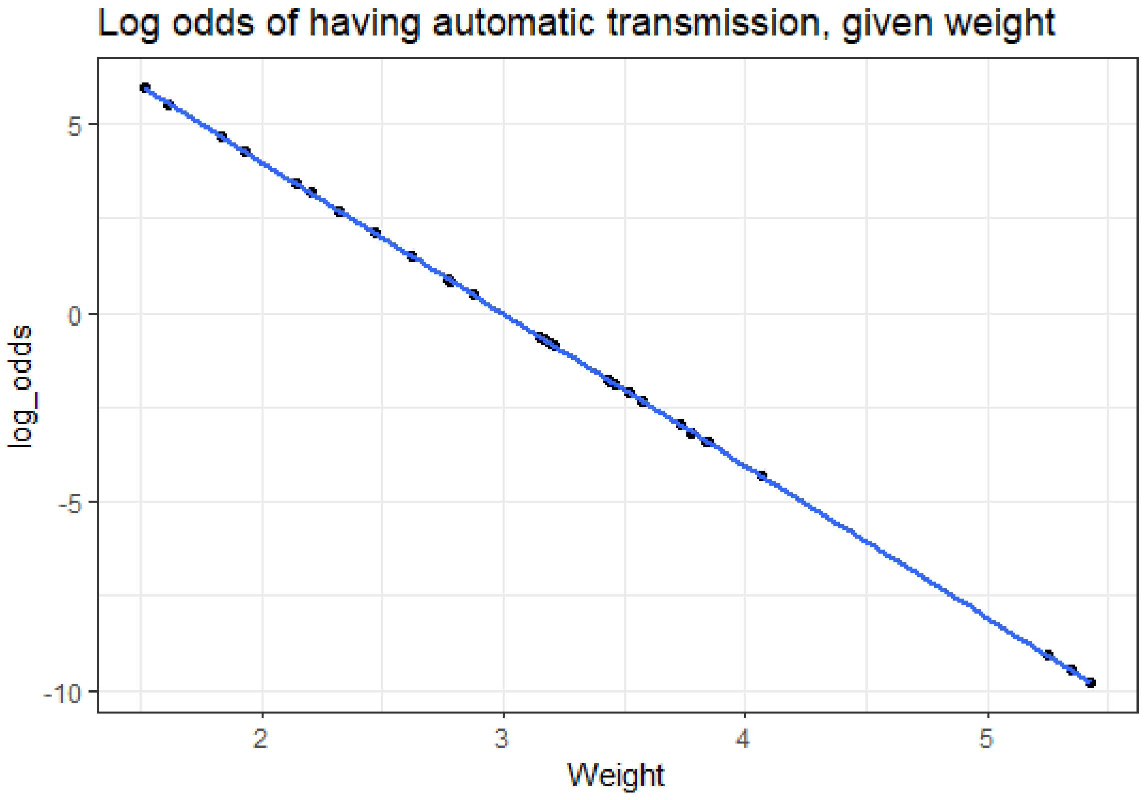
- Going through example of a Mazda

```
> head(mazda)
          mpg cyl disp  hp drat   wt  qsec vs am gear carb
Mazda RX4  21   6  160 110  3.9 2.62 16.46  0  1    4    4
> y_log_odds = 12.04 + -4.024 * 2.62
> y_log_odds
[1] 1.49712
```

OUTPUT – LOGISTIC REGRESSION

- Going through example of a Mazda

```
> head(mazda)
           mpg cyl disp  hp drat   wt  qsec
Mazda RX4   21   6  160 110  3.9 2.62 16.46
> y_log_odds = 12.04 + -4.024 * 2.62
> y_log_odds
[1] 1.49712
```



OUTPUT – LOGISTIC REGRESSION

- Going through example of a Mazda

```
> head(mazda)
          mpg cyl disp  hp drat   wt  qsec vs am gear carb
Mazda RX4  21   6  160 110  3.9 2.62 16.46  0  1    4    4
> y_log_odds = 12.04 + -4.024 * 2.62
> y_log_odds
[1] 1.49712
> y_odds = exp(y_log_odds)
```

OUTPUT – LOGISTIC REGRESSION

- Going through example of a Mazda

```
> head(mazda)
          mpg cyl disp  hp drat   wt  qsec vs am gear carb
Mazda RX4  21   6  160 110  3.9 2.62 16.46  0  1    4    4
> y_log_odds = 12.04 + -4.024 * 2.62
> y_log_odds
[1] 1.49712
> y_odds = exp(y_log_odds)
> y_odds
[1] 4.4688
```

OUTPUT – LOGISTIC REGRESSION

- Going through example of a Mazda

```
> head(mazda)
          mpg cyl disp  hp drat   wt  qsec vs am gear carb
Mazda RX4  21   6  160 110  3.9 2.62 16.46  0  1    4    4
> y_log_odds = 12.04 + -4.024 * 2.62
> y_log_odds
[1] 1.49712
> y_odds = exp(y_log_odds)
> y_odds
[1] 4.4688
> y_probability = y_odds / (1+y_odds)
```

OUTPUT – LOGISTIC REGRESSION

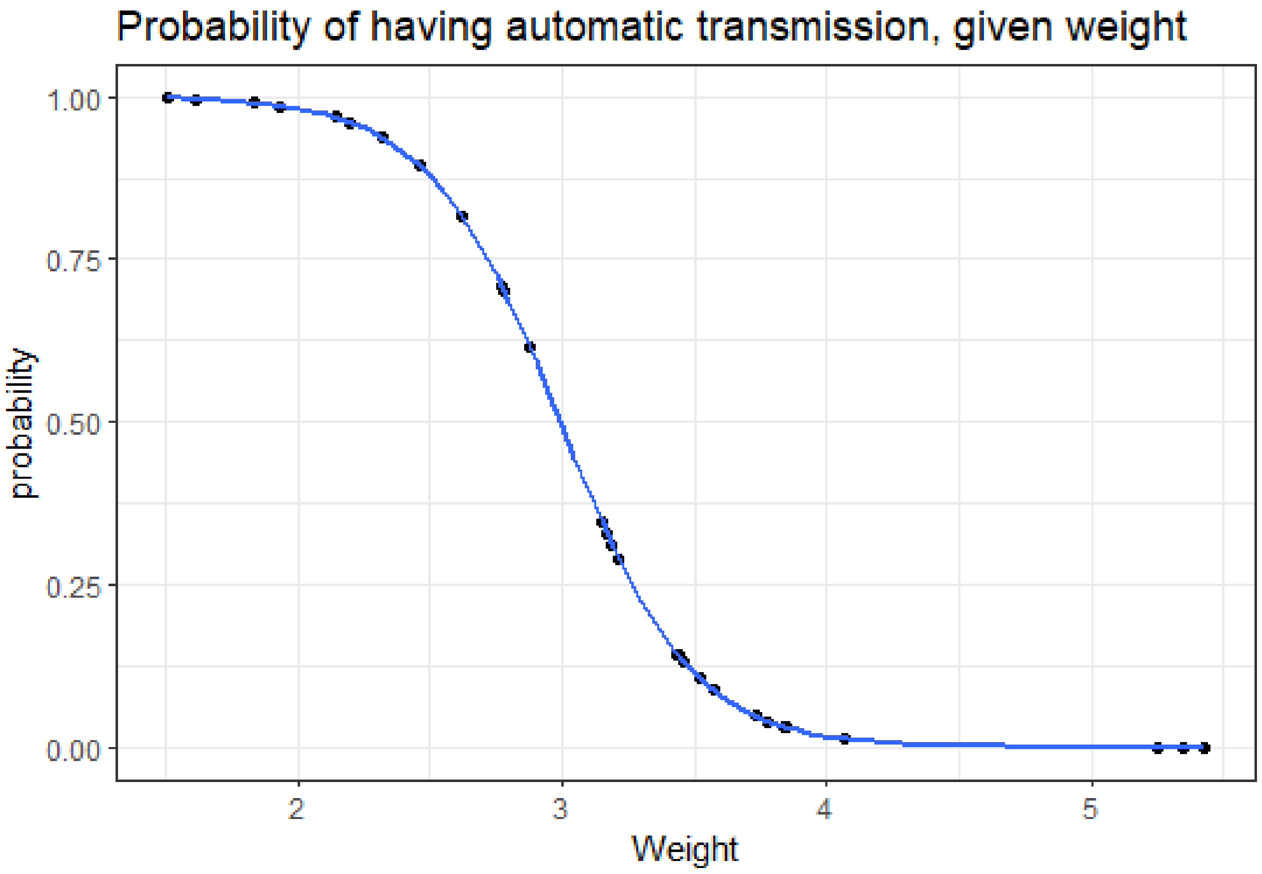
- Going through example of a Mazda

```
> head(mazda)
           mpg cyl disp  hp drat   wt  qsec vs am gear carb
Mazda RX4   21   6  160 110  3.9 2.62 16.46  0  1    4    4
> y_log_odds = 12.04 + -4.024 * 2.62
> y_log_odds
[1] 1.49712
> y_odds = exp(y_log_odds)
> y_odds
[1] 4.4688
> y_probability = y_odds / (1+y_odds)
> y_probability
[1] 0.8171445
```

OUTPUT – LOGISTIC REGRESSION

- Going through example of a Mazda

```
> head(mazda)
           mpg cyl disp  hp drat   wt  q:
Mazda RX4   21   6  160 110  3.9 2.62 16.
> y_log_odds = 12.04 + -4.024 * 2.62
> y_log_odds
[1] 1.49712
> y_odds = exp(y_log_odds)
> y_odds
[1] 4.4688
> y_probability = y_odds / (1+y_odds)
> y_probability
[1] 0.8171445
```



OUTPUT – LOGISTIC REGRESSION

- Going through example of a Mazda

```
> y_log_odds
[1] 1.49712
> y_probability
[1] 0.8171445
```

OUTPUT – LOGISTIC REGRESSION

- Going through example of a Mazda

```
> y_log_odds
[1] 1.49712
> y_probability
[1] 0.8171445
```

"Inverse-link function"

```
inv.logit <- function(x) exp(x) / (1 + exp(x))
```

OUTPUT – LOGISTIC REGRESSION

- Going through example of a Mazda

```
> y_log_odds
[1] 1.49712
> y_probability
[1] 0.8171445

inv.logit <- function(x) exp(x) / (1 + exp(x))

> inv.logit(1.49712)
[1] 0.8171445
```

OUTPUT – LOGISTIC REGRESSION

- Going through example of a Mazda

- logistic_model$fitted.values takes the inv-logit automatically

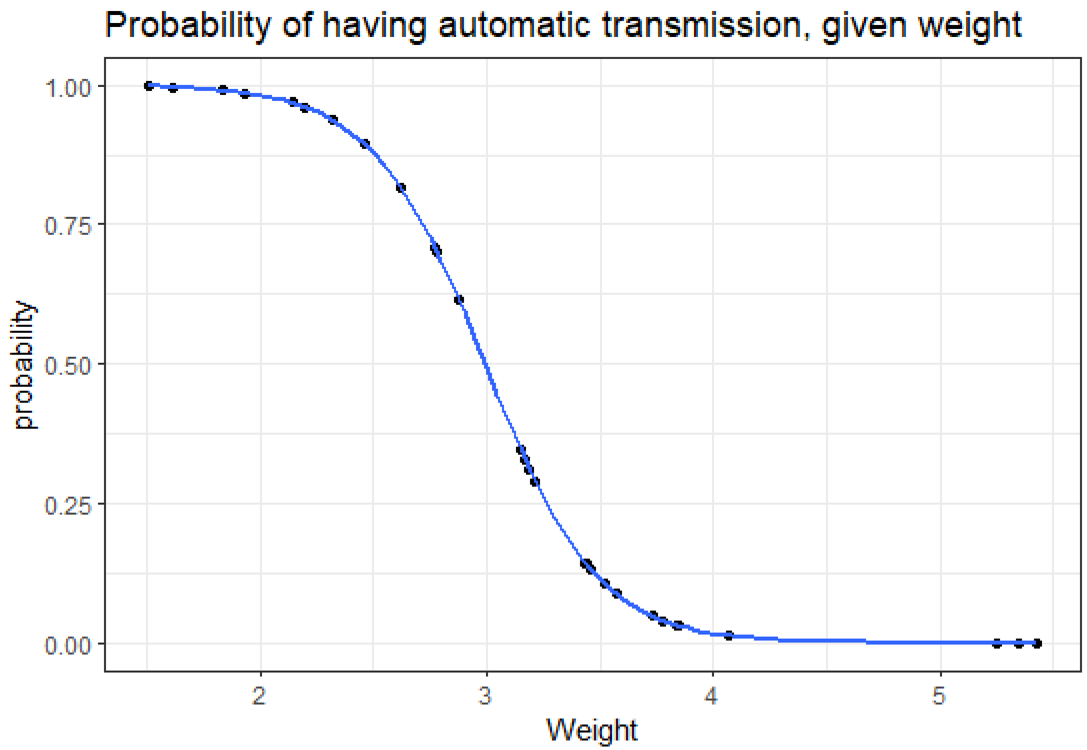
OUTPUT – LOGISTIC REGRESSION

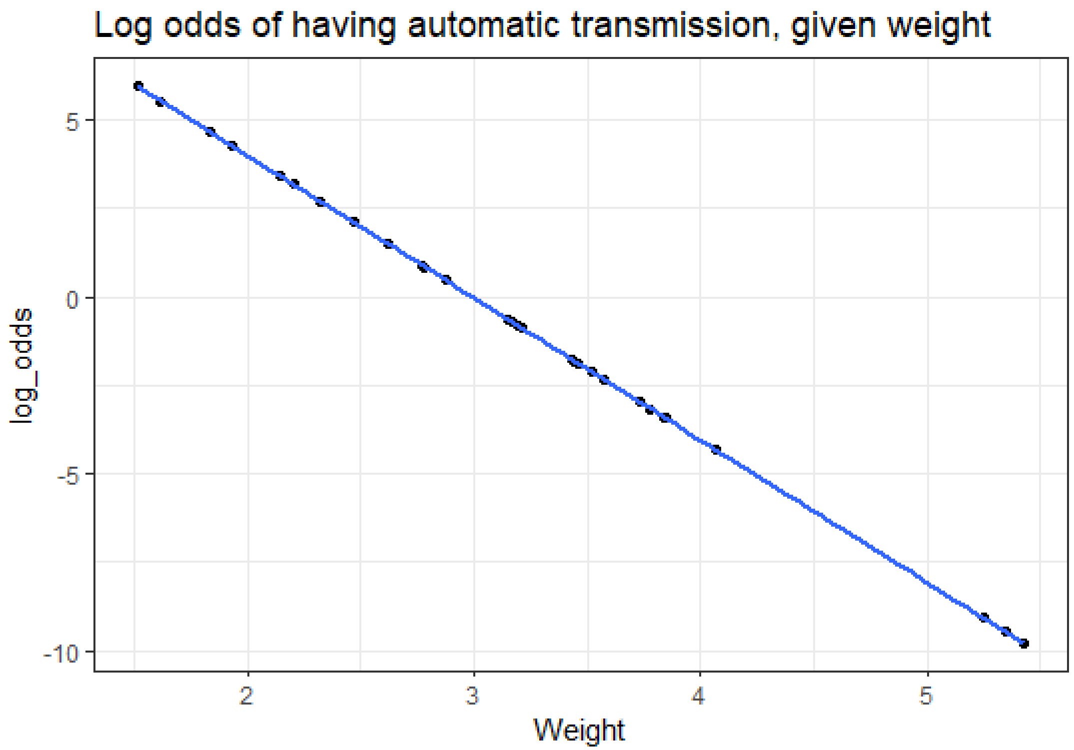
- Going through example of a Mazda

- logistic_model$fitted.values takes the inv-logit automatically

```
> logistic_model$fitted.values
       Mazda RX4    Mazda RX4 Wag
    8.172115e-01     6.157283e-01
```

OUTPUT – LOGISTIC REGRESSION



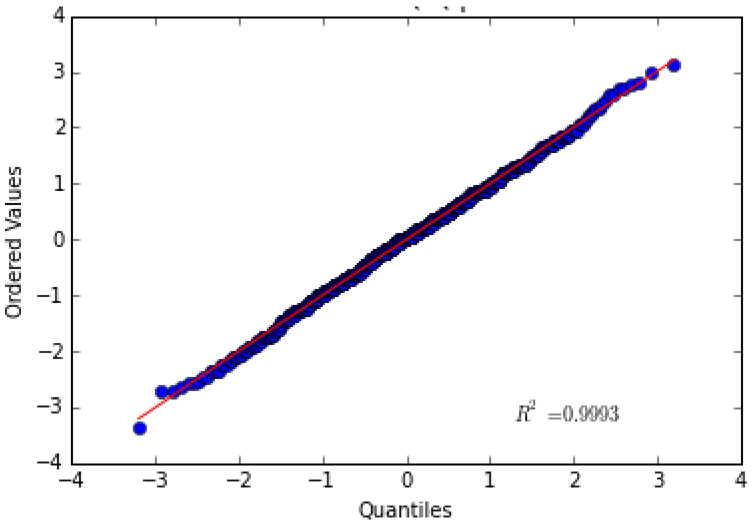


TODAYS PLAN

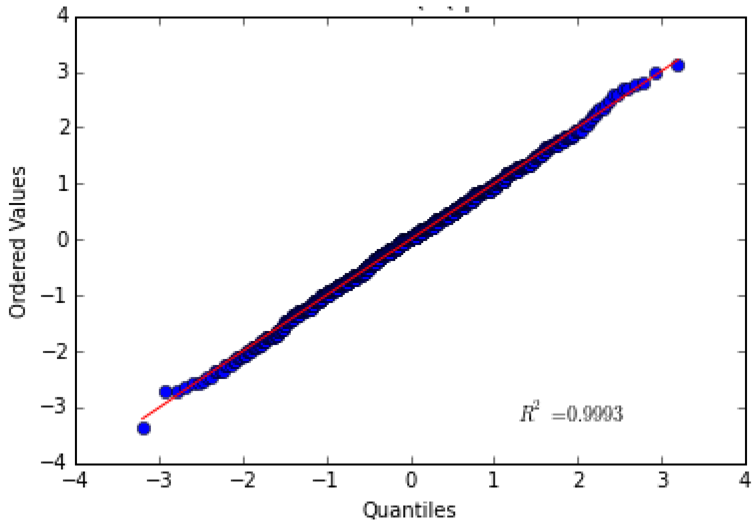
- Catch-up
- Output – logistic regression
- **QQ-plots**
- *(Model selection criteria*
    - *R-squared (revisited)*
    - *AIC/BIC*
    - *Out-of-sample-error)*
- Assignment tips
- Assignment work

QQ-PLOTS

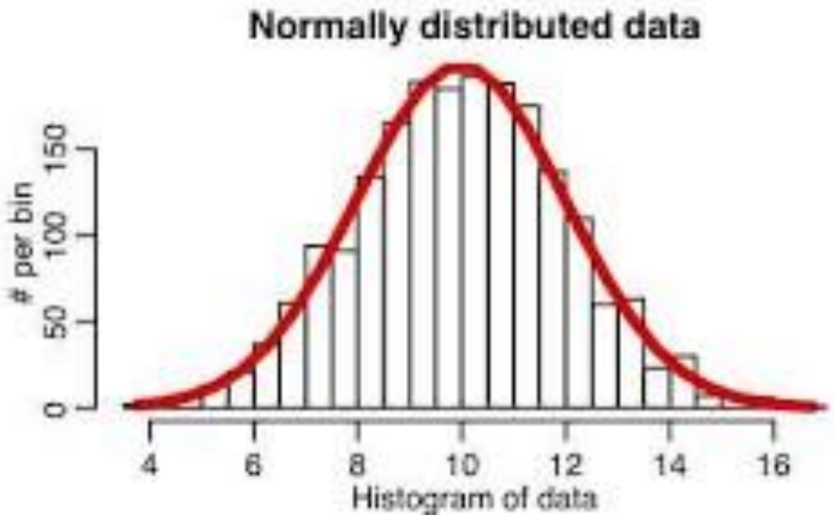
QQ-PLOTS



QQ-PLOTS

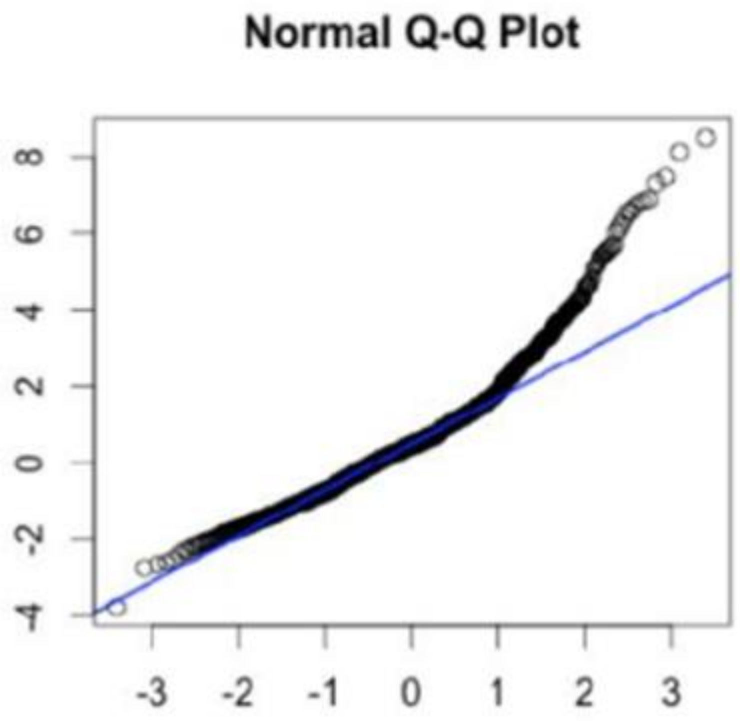


Ordered Values
Quantiles
R^2 =0.9993
4
3
2
1
0
−1
−2
−3
−4
−4
−3
−2
−1
0
1
2
3
4

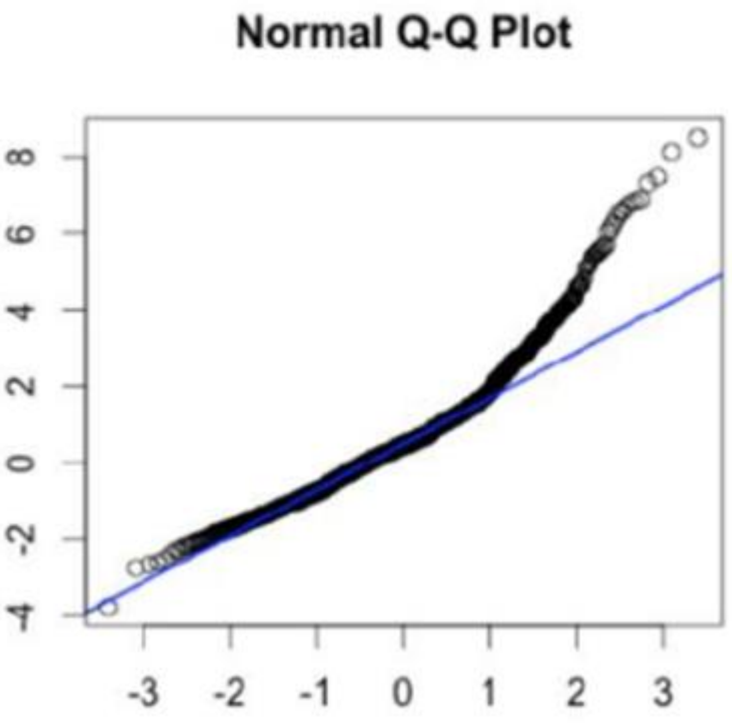


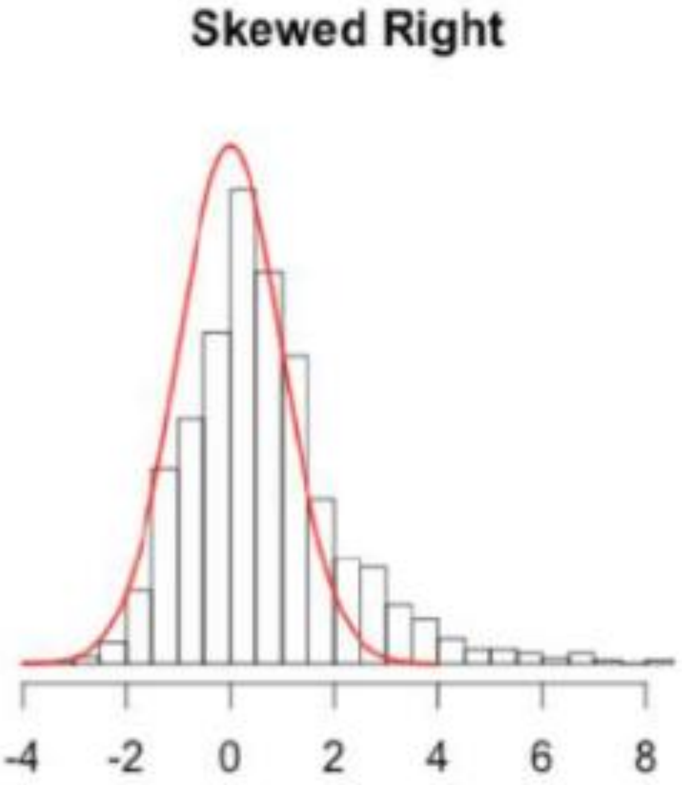
Normally distributed data
per bin
0
50
100
150
4
6
8
10
12
14
16
Histogram of data

QQ-PLOTS



QQ-PLOTS





QQ-PLOTS

- Quantiles-quantile plots

QQ-PLOTS

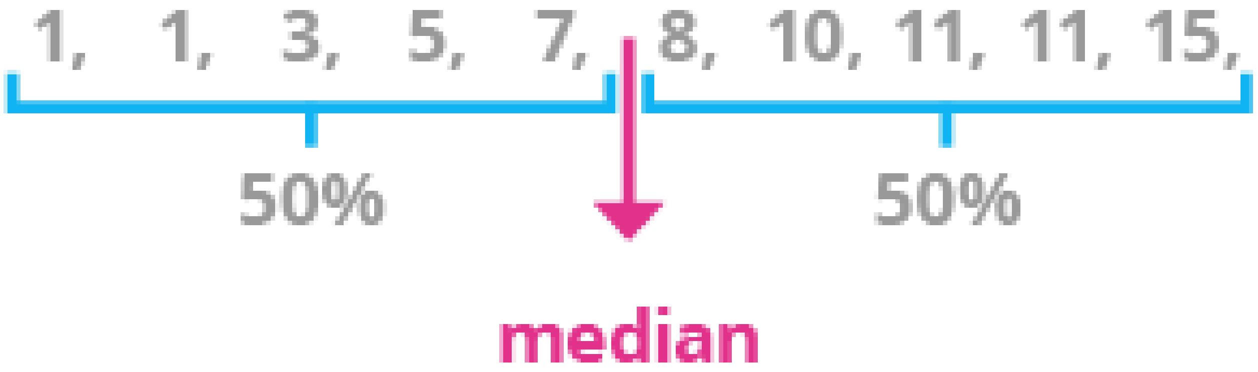
- Quantiles-quantile plots
- What is a quantile?

QQ-PLOTS

- Quantiles-quantile plots
- What is a quantile?
- Generalization of median, quartile, percentile, etc.

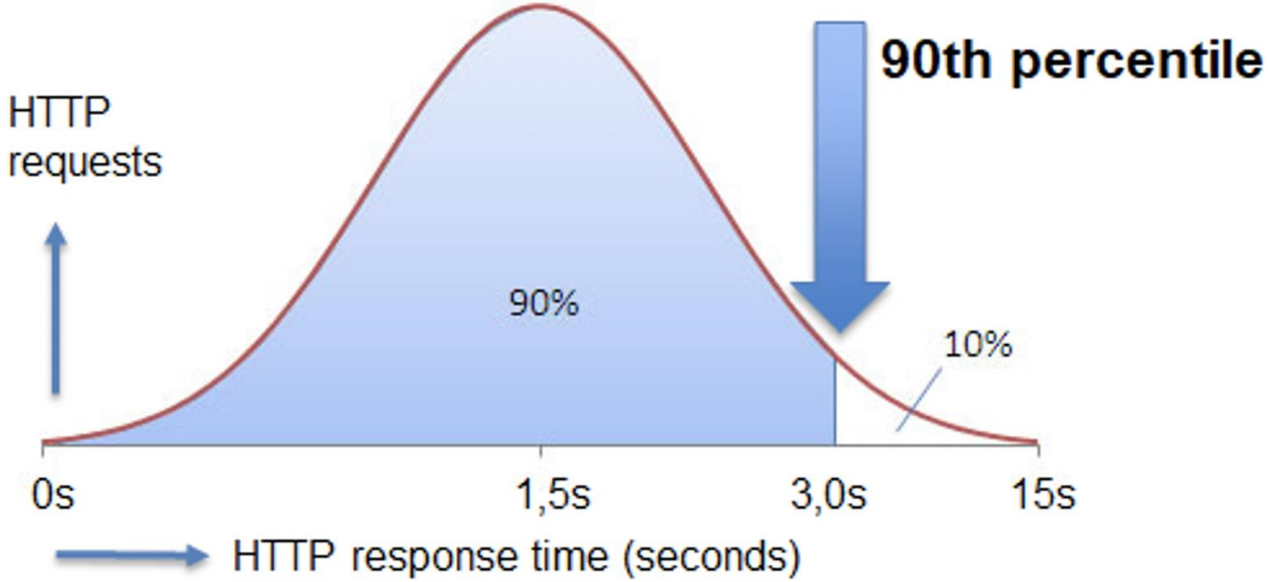
QQ-PLOTS

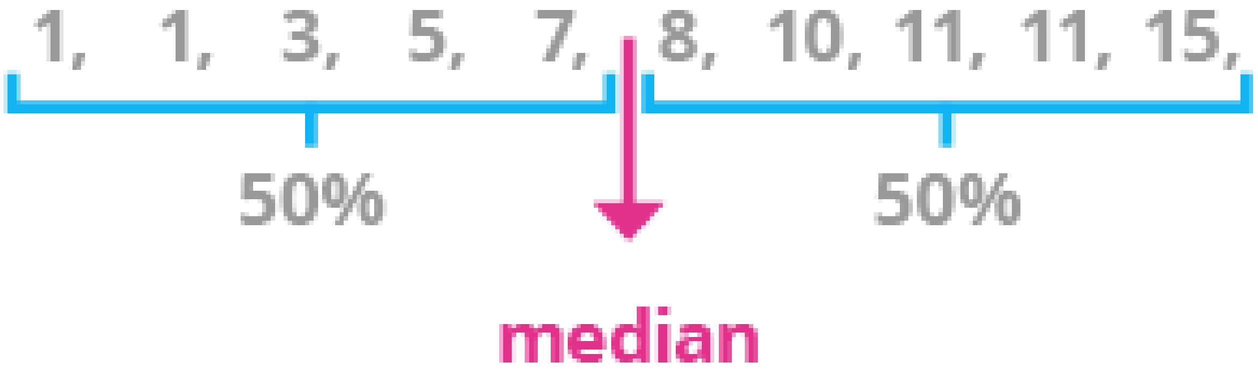
- Quantiles-quantile plots
- What is a quantile?
- Generalization of median, quartile, percentile, etc.



QQ-PLOTS

- Quantiles-quantile plots
- What is a quantile?
- Generalization of median, quartile, percentile, etc.





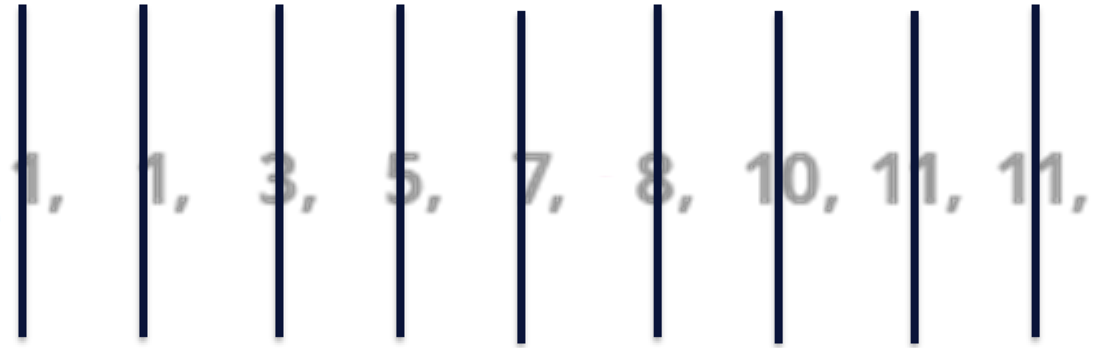
QQ-PLOTS

- Quantiles

1, 1, 3, 5, 7, 8, 10, 11, 11,

QQ-PLOTS

- Quantiles



QQ-PLOTS

- Quantiles

1, 1, 3, 5, 7, 8, 10, 11, 11,

QQ-PLOTS

- Quantiles

1, 1, 3, 5, 7, 8, 10, 11, 11,

TODAYS PLAN

- Catch-up
- Output – logistic regression
- QQ-plots
- *(Model selection criteria*
    - *R-squared (revisited)*
    - *AIC/BIC*
    - *Out-of-sample-error)*
- **Assignment tips**
- Assignment work

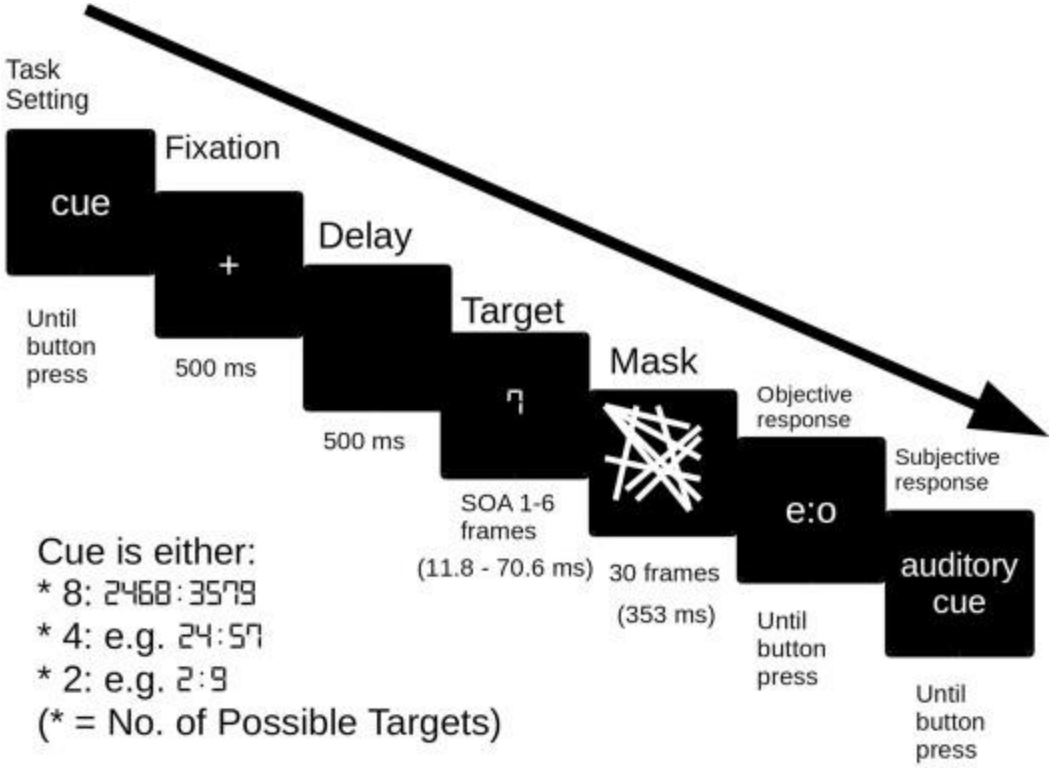
ASSIGNMENT TIPS

ASSIGNMENT TIPS

- Understanding assignment data
    - Experiment paradigm

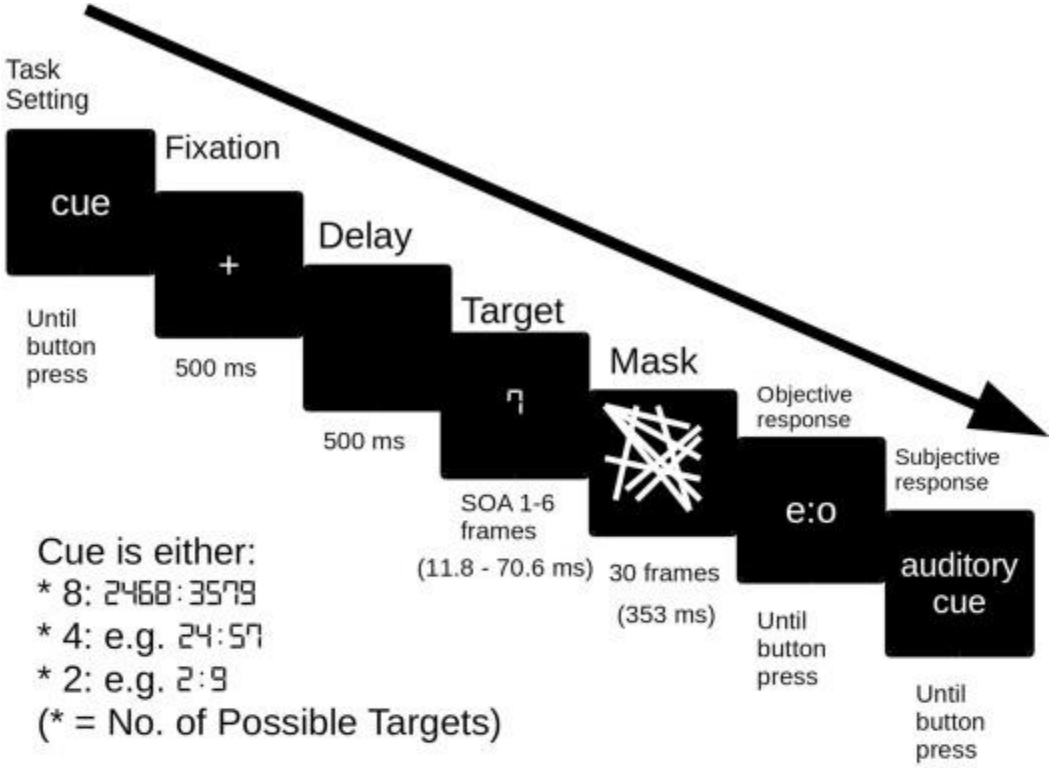
ASSIGNMENT TIPS

- Understanding the paradigm



ASSIGNMENT TIPS

- Understanding the paradigm
  - Task (8, quadrant, pairs)
  - PAS (how clearly was target seen?
  - rt.obj (Reaction time -> Even/Odd)
  - rt.sub (Reaction time -> PAS)

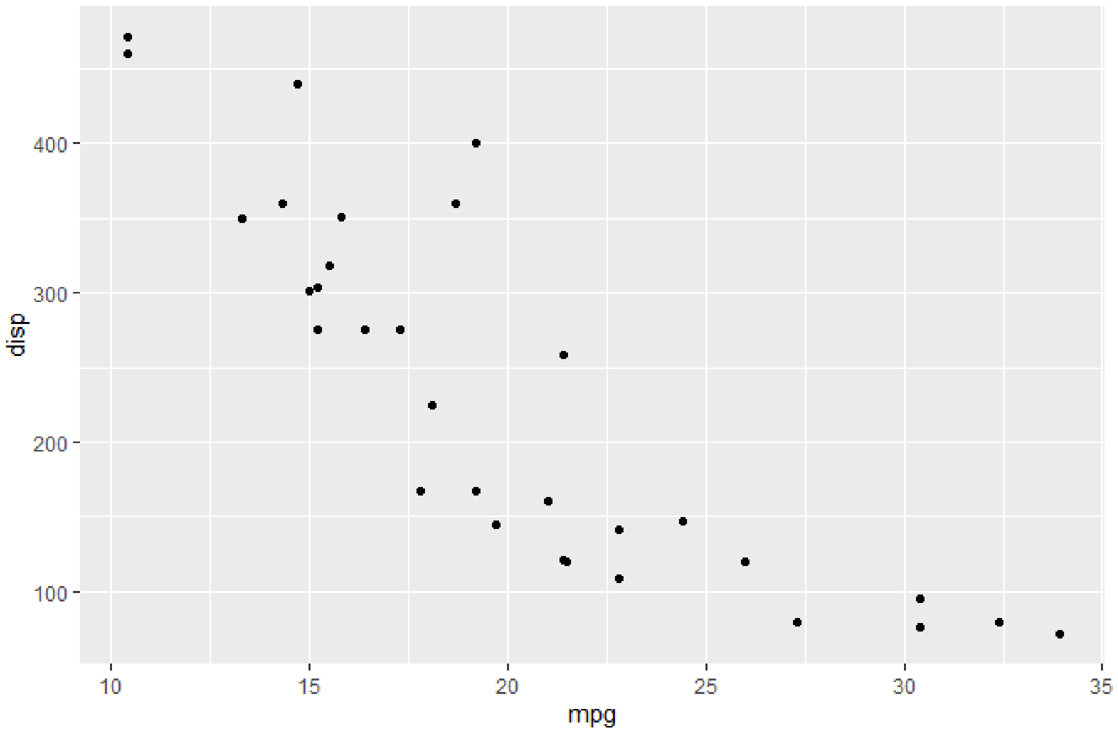


ASSIGNMENT TIPS

- ggplot(mtcars, aes(mpg, disp)) +
  geom_point()

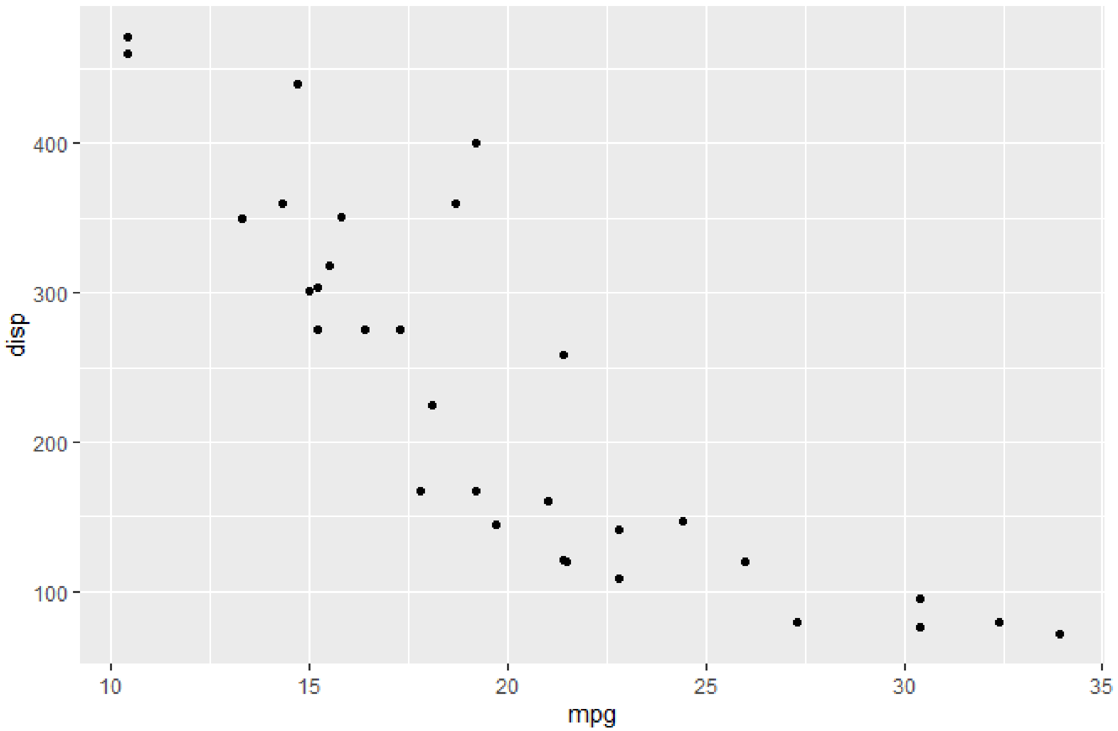
ASSIGNMENT TIPS

- ggplot(mtcars, aes(mpg, disp)) + geom_point()



ASSIGNMENT TIPS

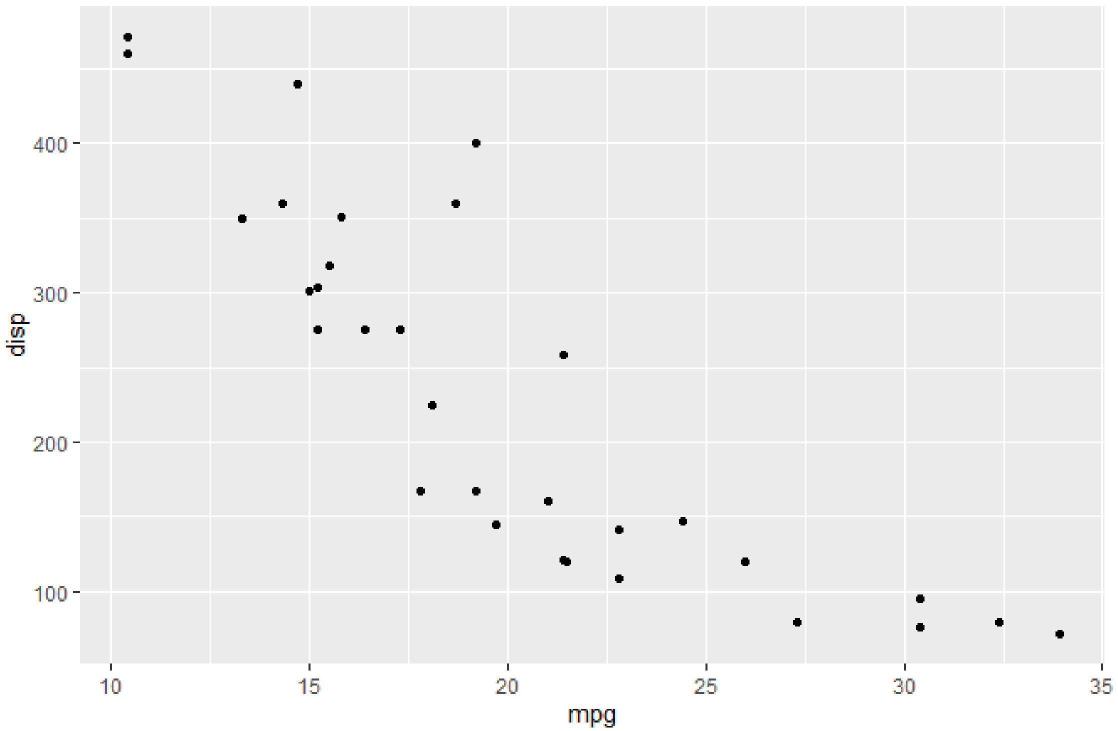
- ggplot(mtcars, aes(mpg, disp)) + geom_point()



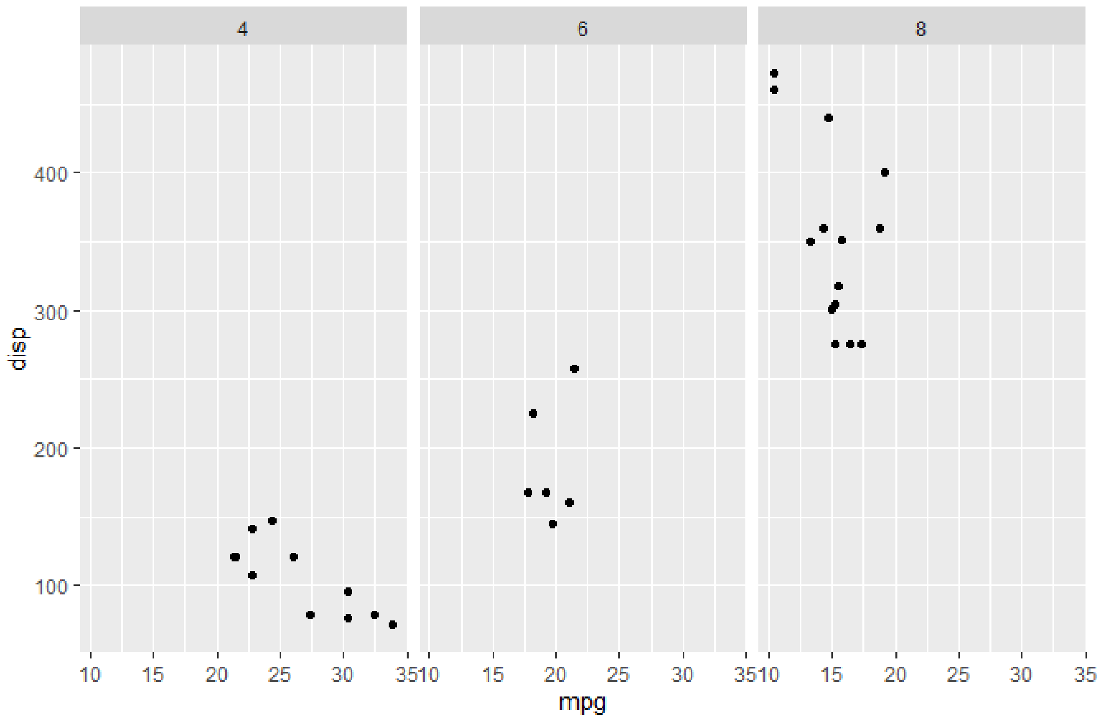
- ggplot(mtcars, aes(mpg, disp)) + geom_point() + **facet_wrap(~ cyl)**

ASSIGNMENT TIPS

- ggplot(mtcars, aes(mpg, disp)) + geom_point()



- ggplot(mtcars, aes(mpg, disp)) + geom_point() + **facet_wrap(~ cyl)**



ASSIGNMENT TIPS

- Loading in the files
  - pwd in R = getwd()
  - ls in R = list.files()

ASSIGNMENT TIPS

- Loading in the files
  - pwd in R = getwd()
  - ls in R = list.files()

```
> getwd()
[1] "C:/Users/Lenovo/Desktop/Diverse dokumenter/C
nce/github_methods_3/week_03/non_student"
```

```
list.files("/experiment_2/")
[1] "001.csv" "002.csv" "003.csv" "004.csv" "005.csv"
[6] "006.csv" "007.csv" "008.csv" "009.csv" "010.csv"
11] "011.csv" "012.csv" "013.csv" "014.csv" "015.csv"
16] "016.csv" "017.csv" "018.csv" "019.csv" "020.csv"
21] "021.csv" "022.csv" "023.csv" "024.csv" "025.csv"
26] "026.csv" "027.csv" "028.csv" "029.csv"
```

BREAK TIME


THIS IS GIT. IT TRACKS COLLABORATIVE WORK ON PROJECTS THROUGH A BEAUTIFUL DISTRIBUTED GRAPH THEORY TREE MODEL.
COOL. HOW DO WE USE IT?
NO IDEA. JUST MEMORIZE THESE SHELL COMMANDS AND TYPE THEM TO SYNC UP. IF YOU GET ERRORS, SAVE YOUR WORK ELSEWHERE, DELETE THE PROJECT, AND DOWNLOAD A FRESH COPY.

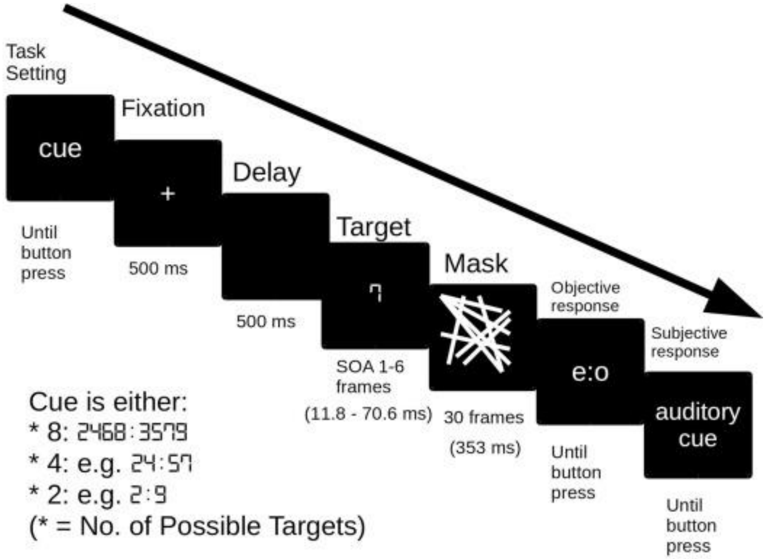
TODAYS PLAN

- Catch-up
- Output – logistic regression
- QQ-plots
- *(Model selection criteria*
  - *R-squared (revisited)*
  - *AIC/BIC*
  - *Out-of-sample-error)*
- Assignment tips
- **Assignment work**

ASSIGNMENT

ASSIGNMENT

1. *Pull new assignment from upstream*
2. **Download files from BrightSpace, put into new folder "github_methods_3/week_03/"**
3. Create copy of "practical_exercise_3.Rmd", give unique filename ending
4. Work on new copied version of assignment



- list.files()
- getwd()
- + facet_wrap(~ subject)

AARHUS
UNIVERSITY

MODEL SELECTION CRITERIA

MODEL SELECTION CRITERIA

MODEL SELECTION CRITERIA

- Why are we modeling?

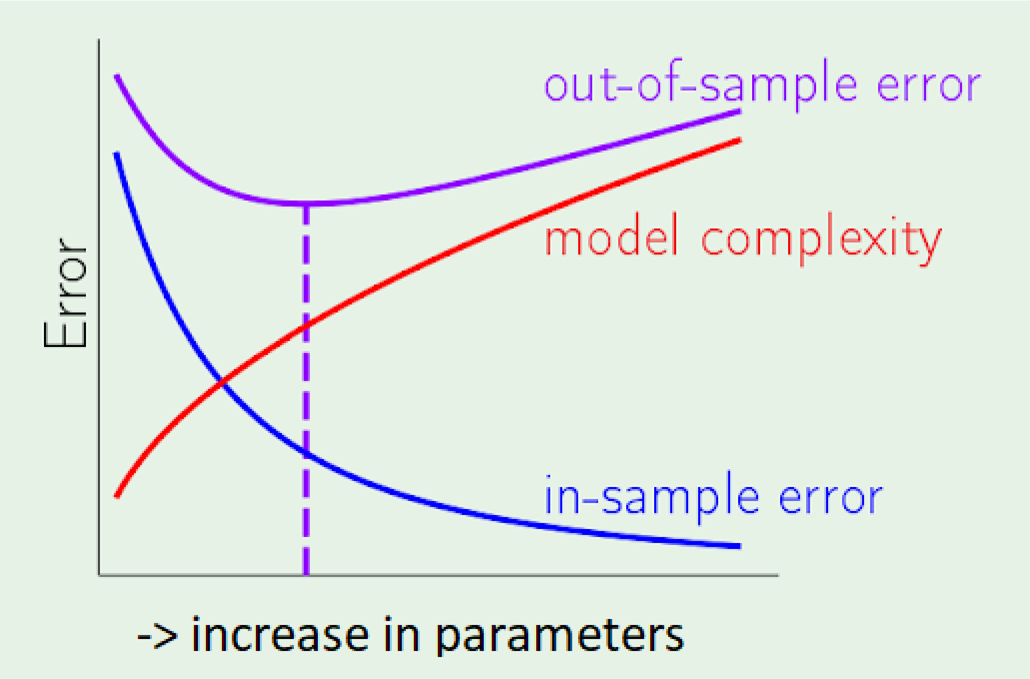
- What constitutes a good model?

MODEL SELECTION CRITERIA

- Why are we modeling?
  - *To be able to understand the world*
  - *Why understand? To be able to predict/manipulate*

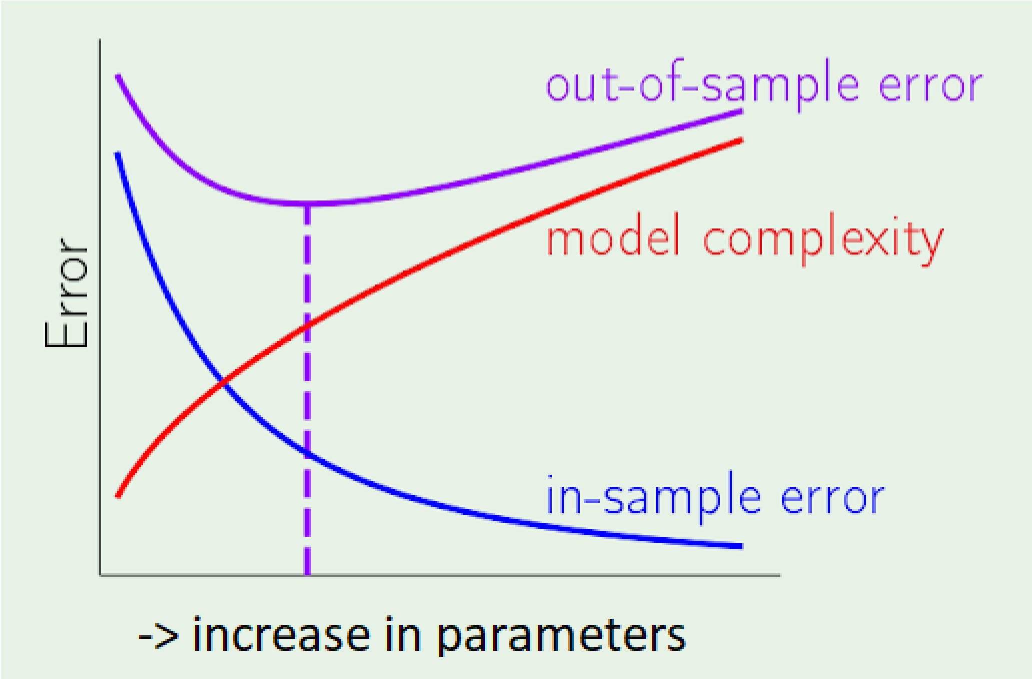
- What constitutes a good model?
  - *A model is good if it accurately models the world*
  - *… is good if it generalizes to new data*
  - *… is good if we accurately have estimated the underlying parameters of the population distribution*

MODEL SELECTION CRITERIA



MODEL SELECTION CRITERIA

- R-squared
- AIC, BIC
- Out-of-sample performance



R SQUARED

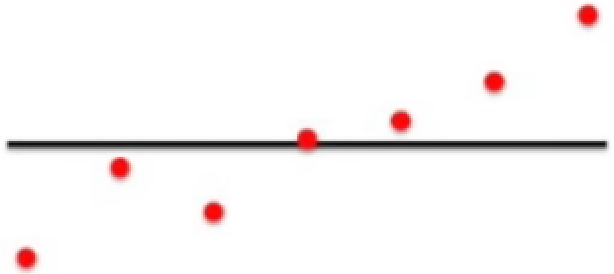
$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \mu)^2}$$

R SQUARED

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \mu)^2}$$

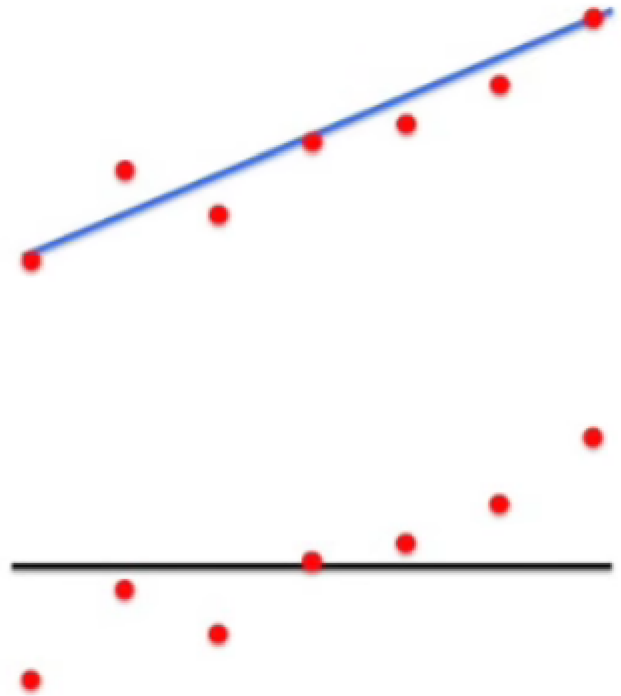
R SQUARED

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \mu)^2}$$



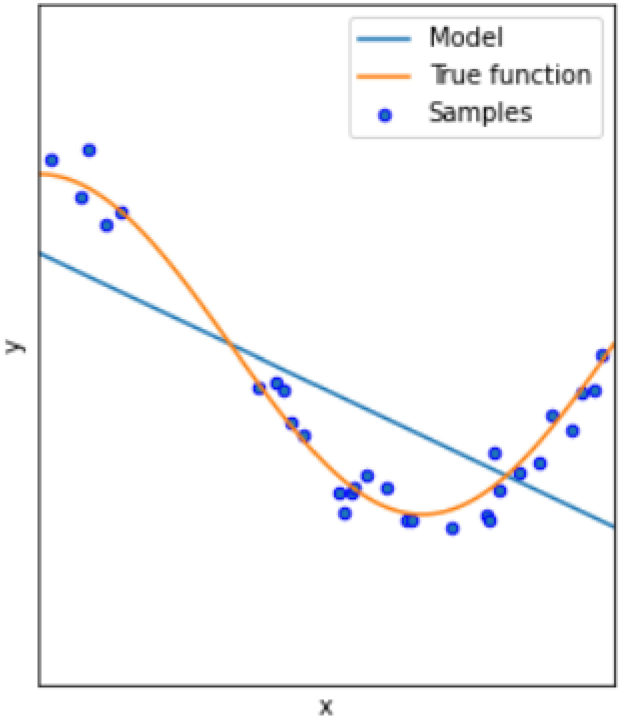
R SQUARED

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \mu)^2}$$



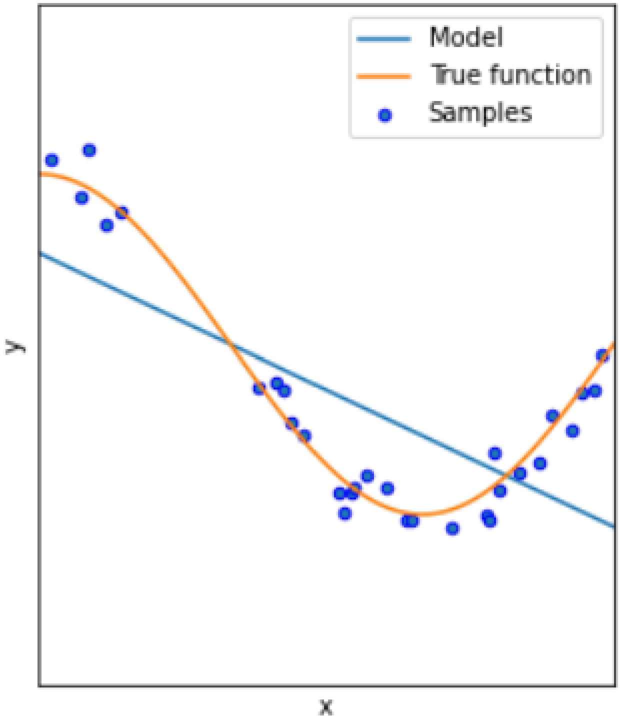
R SQUARED

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \mu)^2}$$

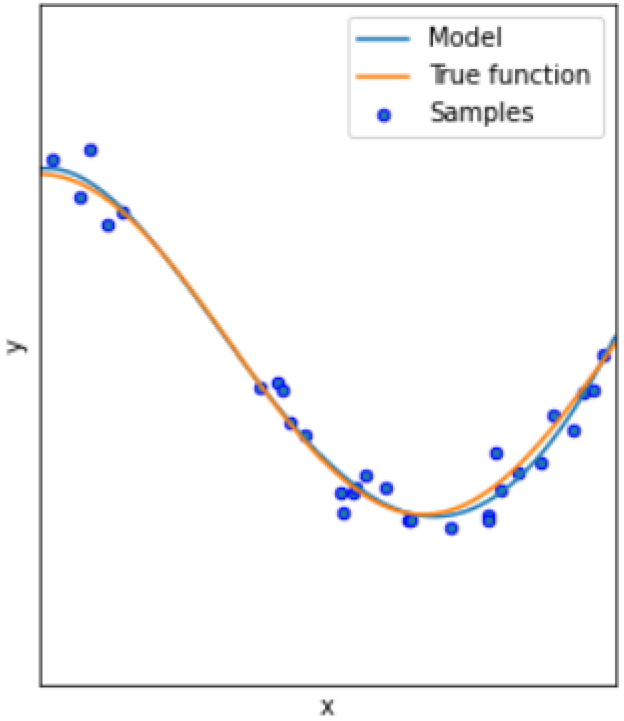


R SQUARED

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \mu)^2}$$

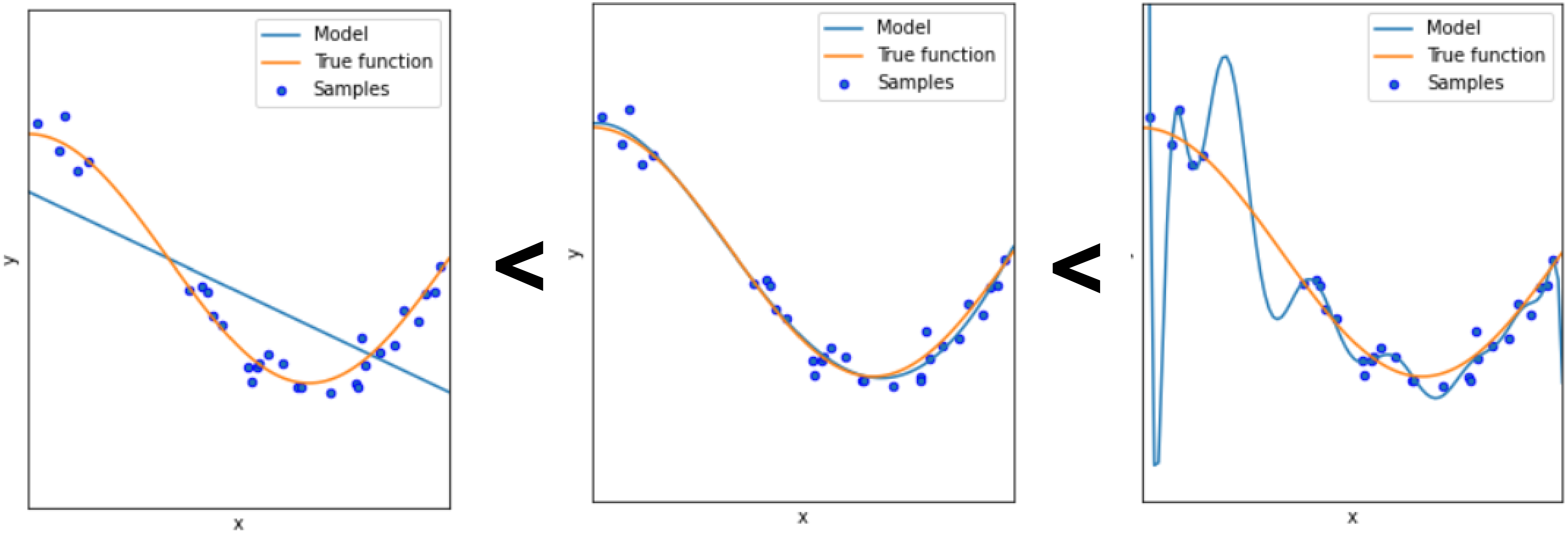


<



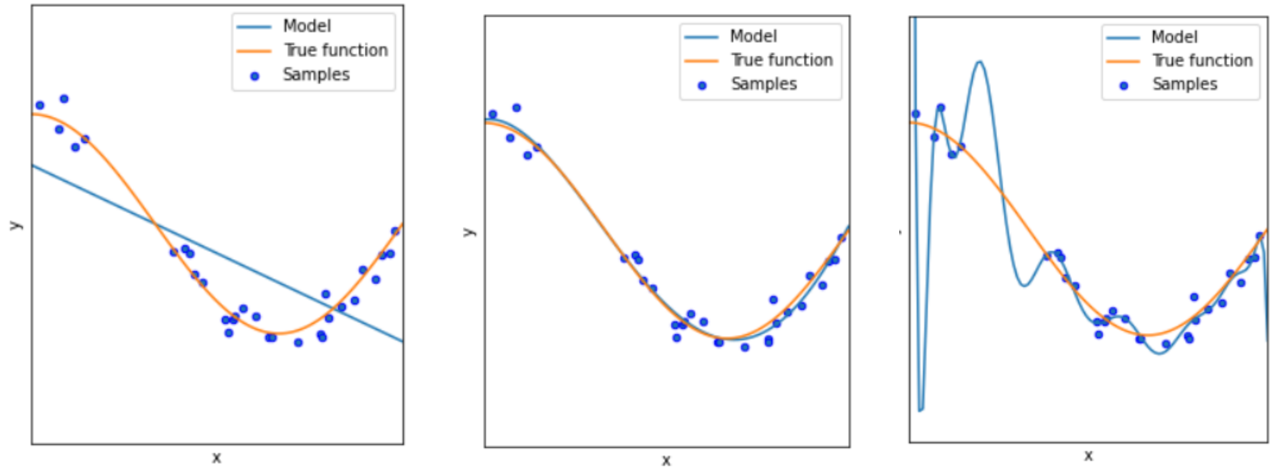
R SQUARED

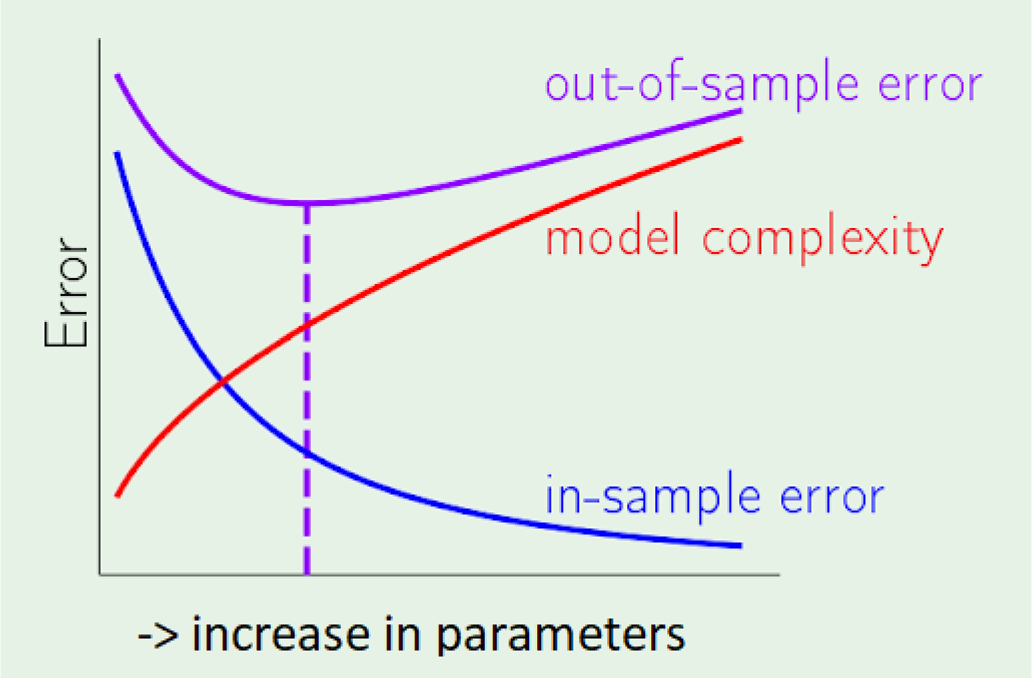
$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \mu)^2}$$



R SQUARED

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \mu)^2}$$





R SQUARED

- Can we use a criteria that accounts for overfitting?

AIC

AIC

$$AIC = -2\ln(L) + 2k$$

AIC

$$AIC = -2\ln(L) + 2k$$

$$=$$

$$\mathrm{AIC} \;=\; 2k - 2\ln(\hat{L})$$

AIC

$$AIC = -2\ln(L) + 2k$$

AIC - BIC

$$AIC = -2\ln(L) + 2k$$

$$\mathrm{BIC} = k\ln(n) - 2\ln(\widehat{L})$$

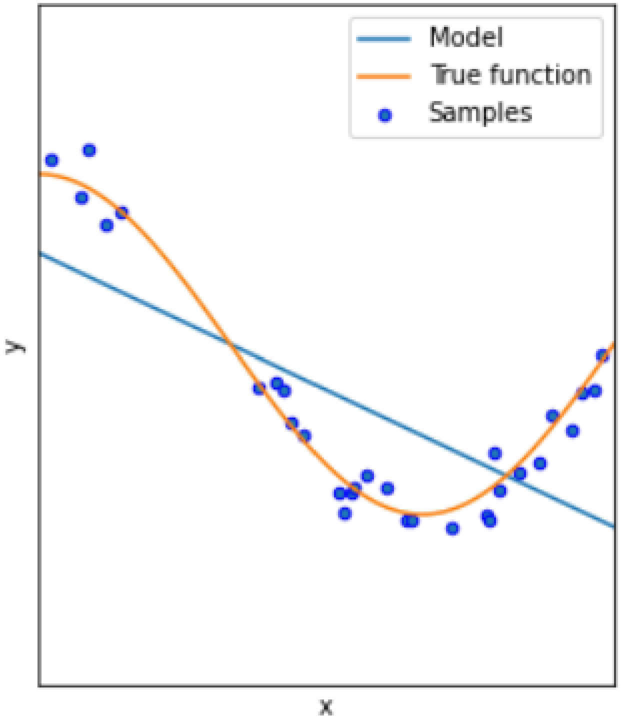
AIC - BIC

$$AIC = -2\ln(L) + 2k$$

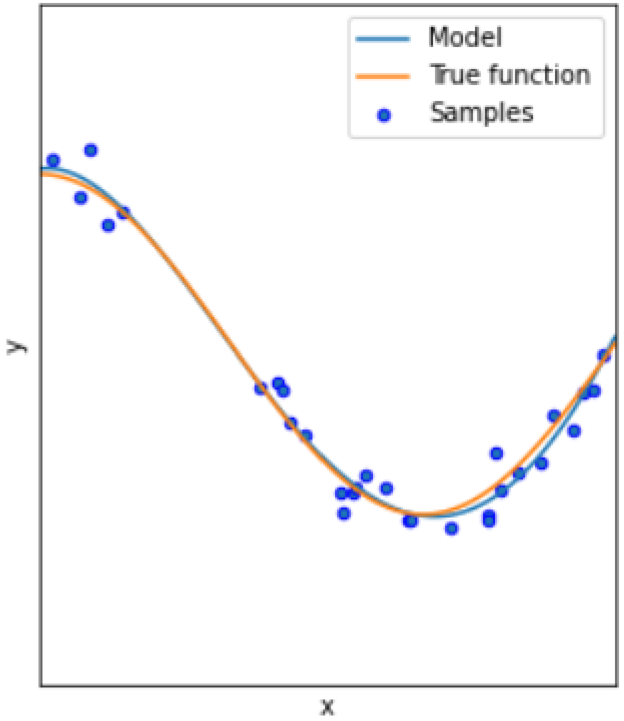
$$\mathrm{BIC} = k\ln(n) - 2\ln(\widehat{L})$$

AIC - BIC

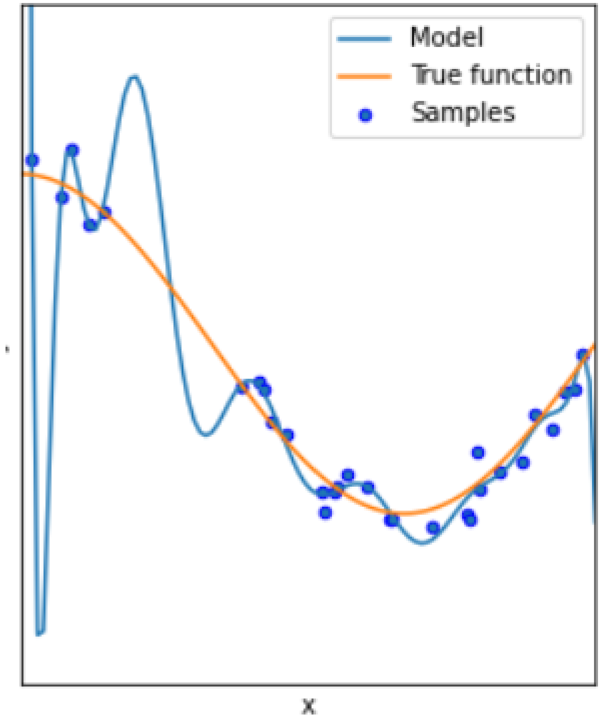
R-squared model rating:



<

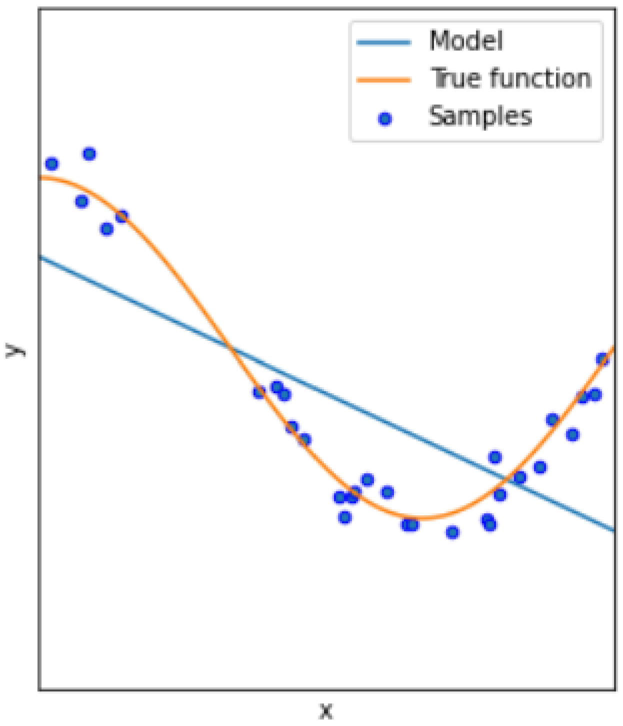


<

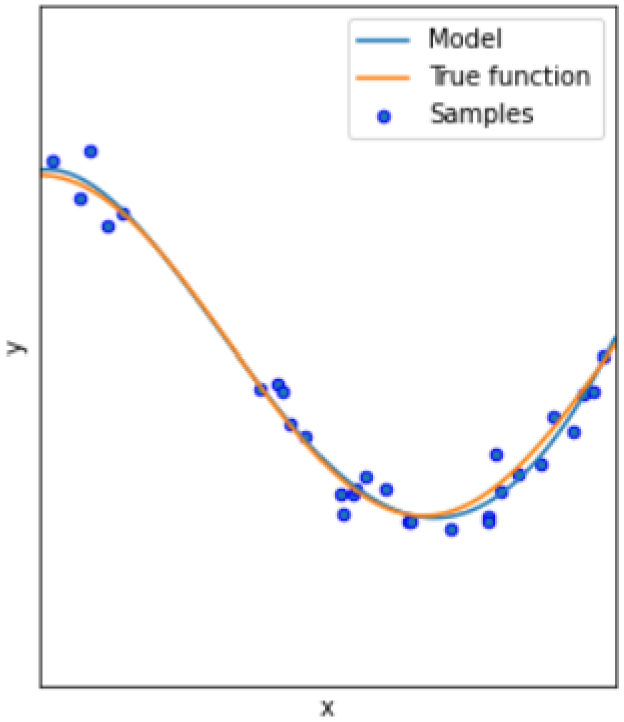


AIC - BIC

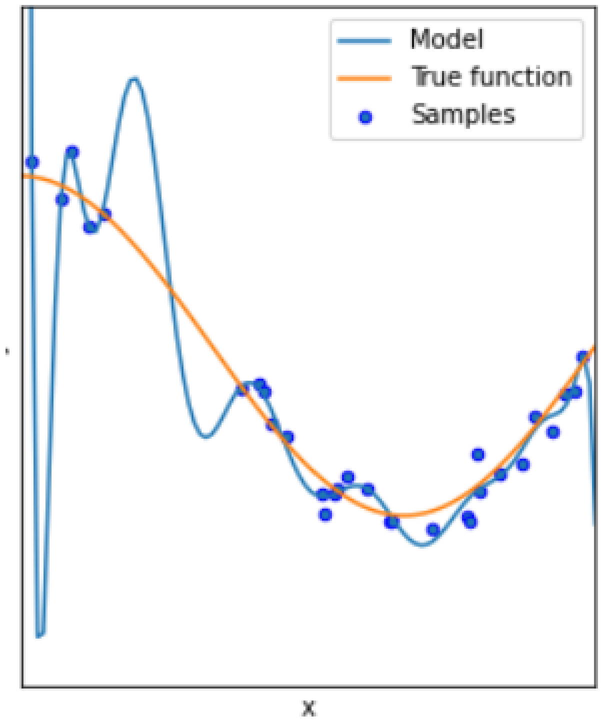
R-squared model rating:



<

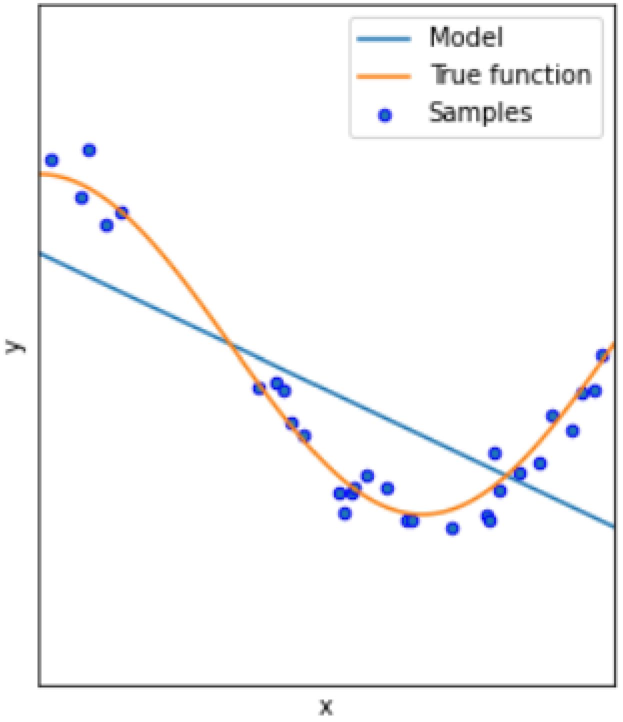


<

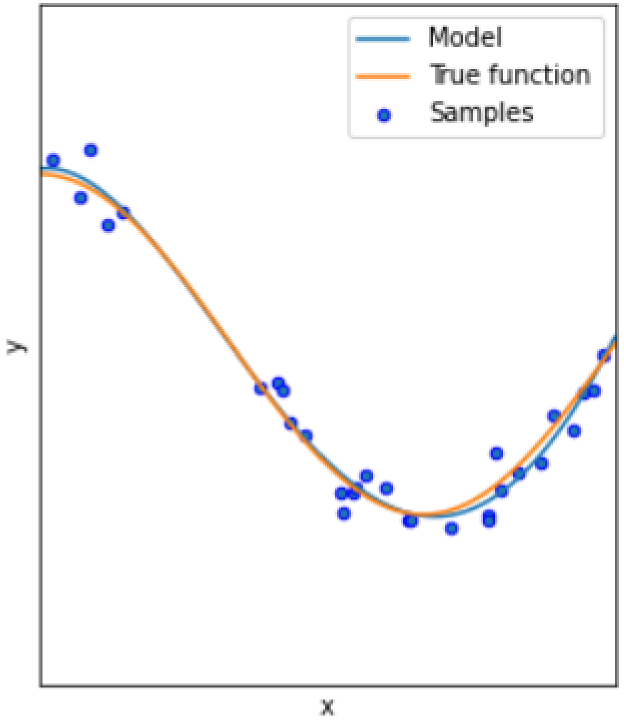


AIC - BIC

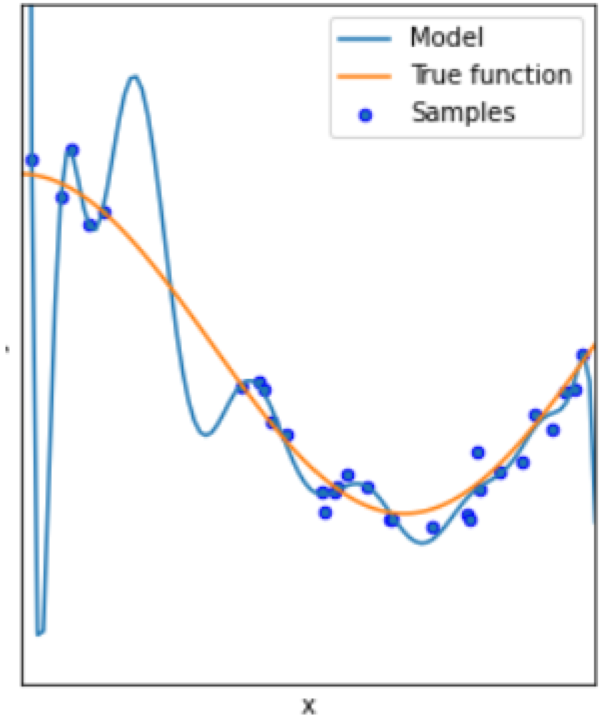
AIC / BIC model rating:



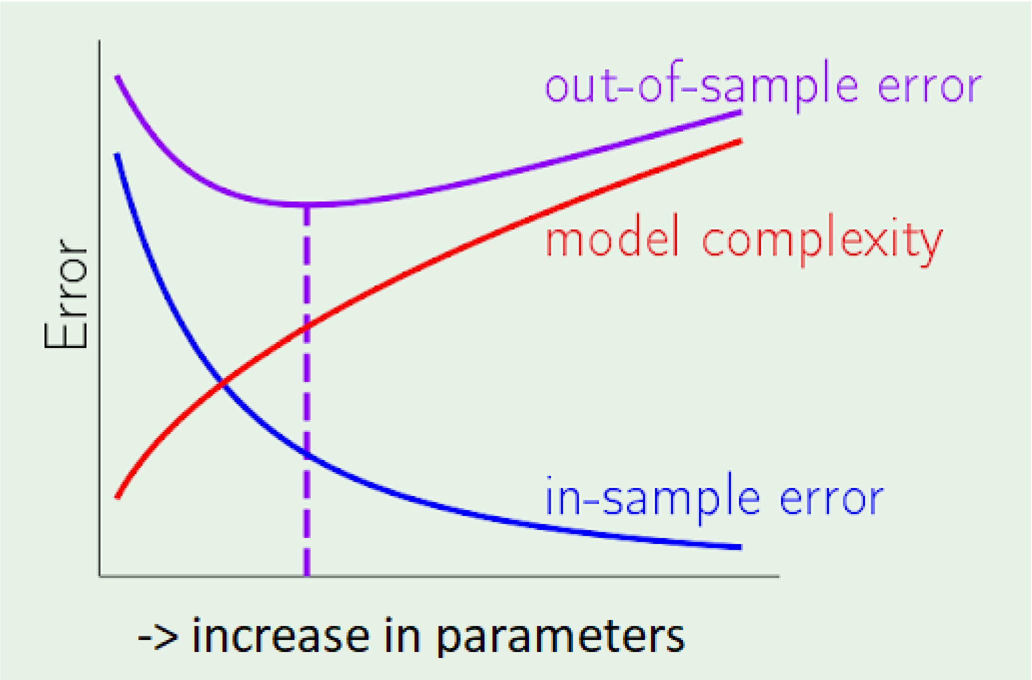
<



>

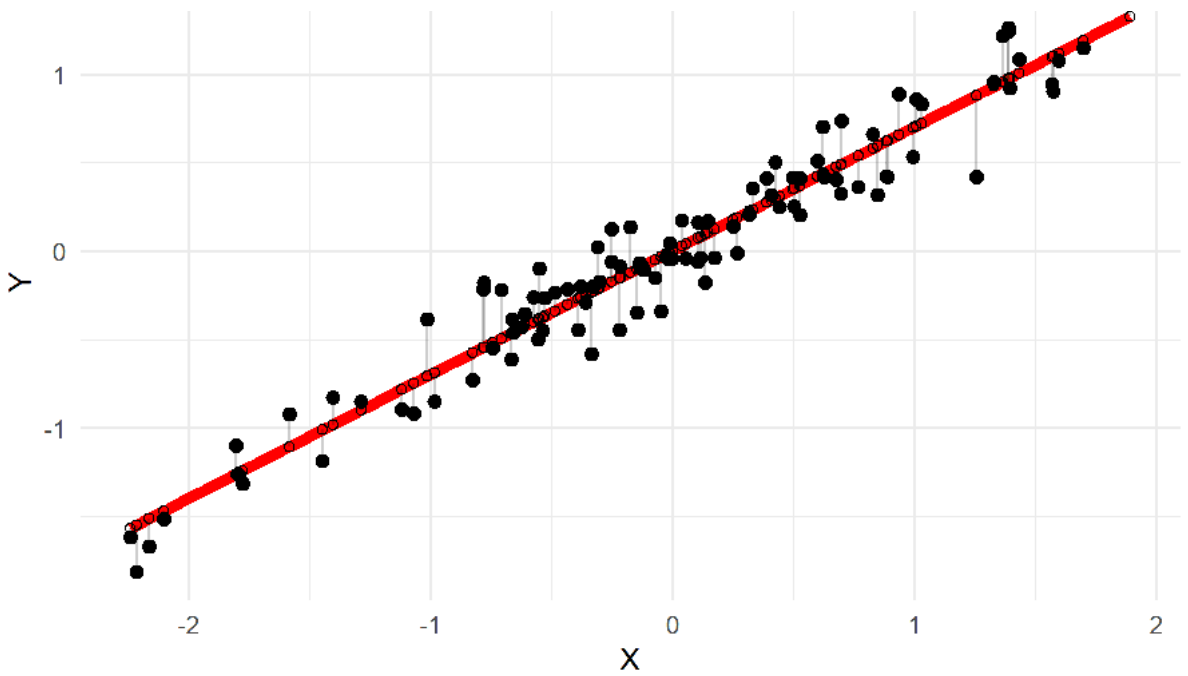


AIC - BIC

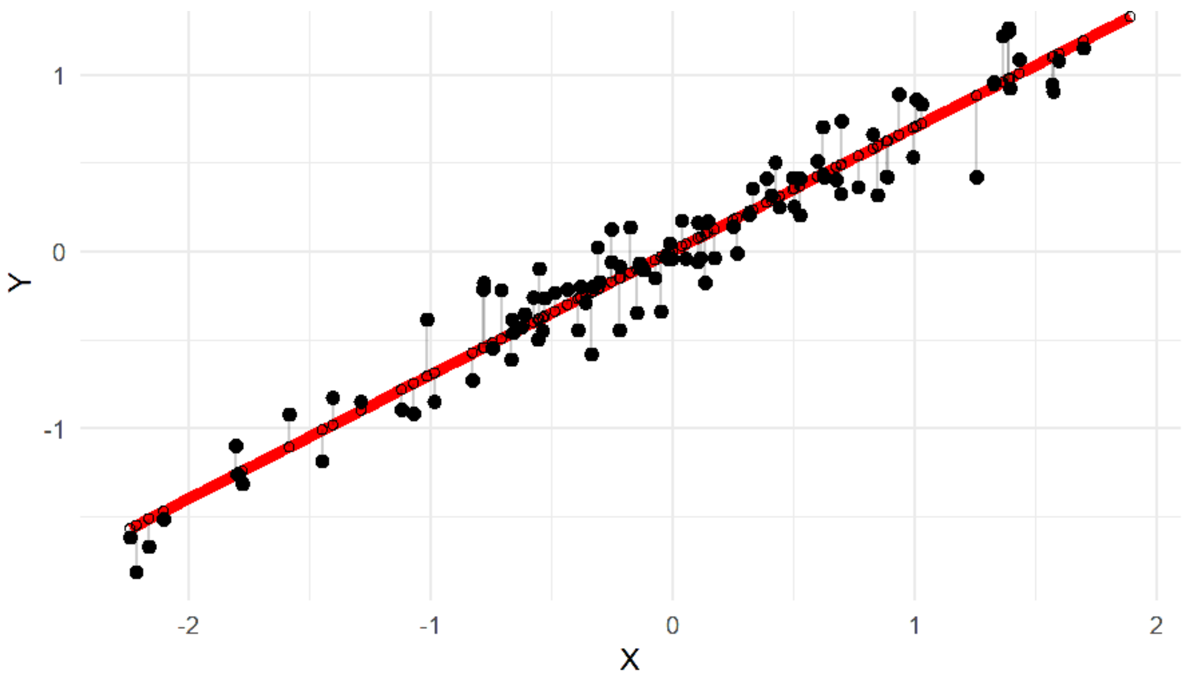

out-of-sample error
model complexity
Error
in-sample error
-> increase in parameters

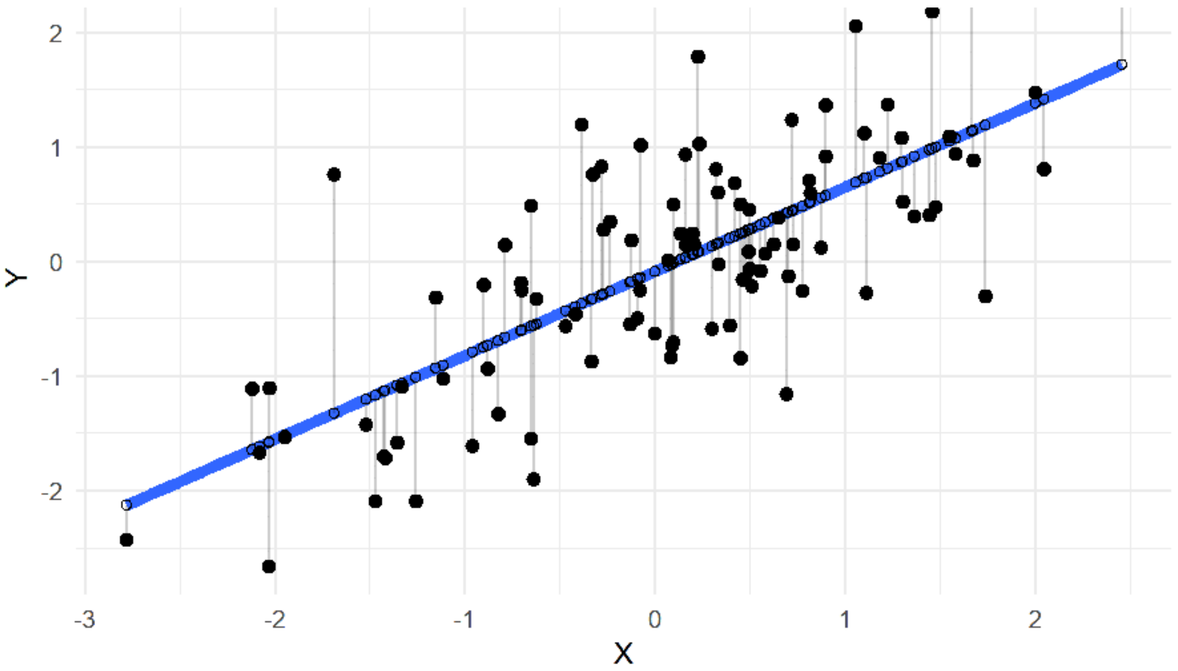
OUT-OF-SAMPLE ERROR

OUT-OF-SAMPLE ERROR

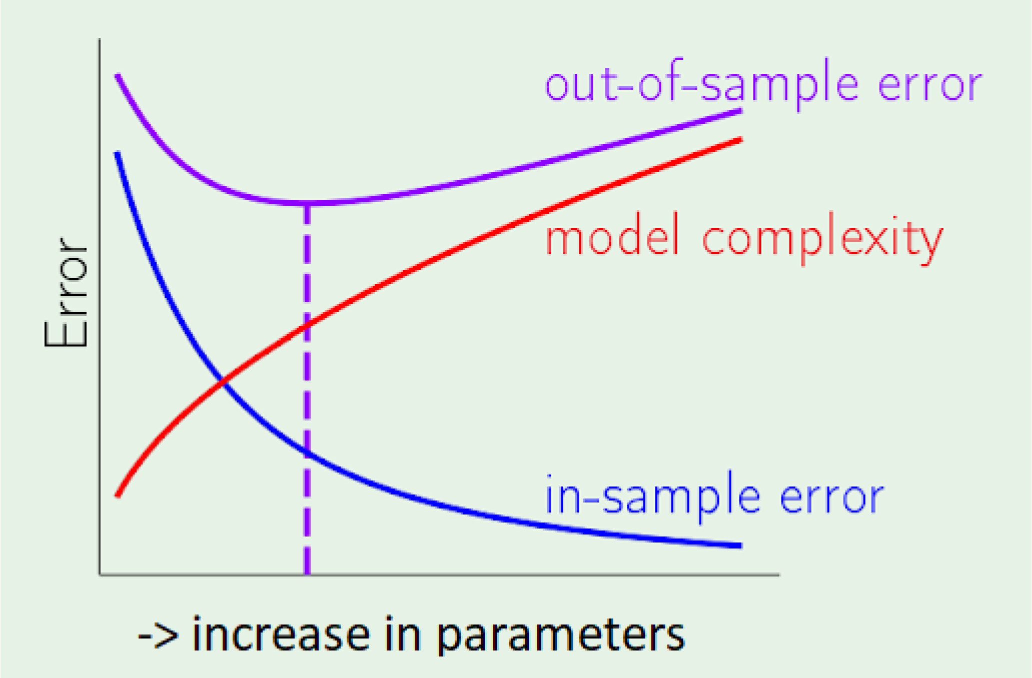


OUT-OF-SAMPLE ERROR





OUT-OF-SAMPLE ERROR



AARHUS
UNIVERSITY