Module 3: Linear Regression Part 2

Daesoo Lee,

Department of Mathematical Sciences, NTNU

NTNU
Norwegian University of Science and Technology

26/01/2023

Recap

Recap

Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

where X_j is the j-th predictor and β_j is the respective regression coefficient.

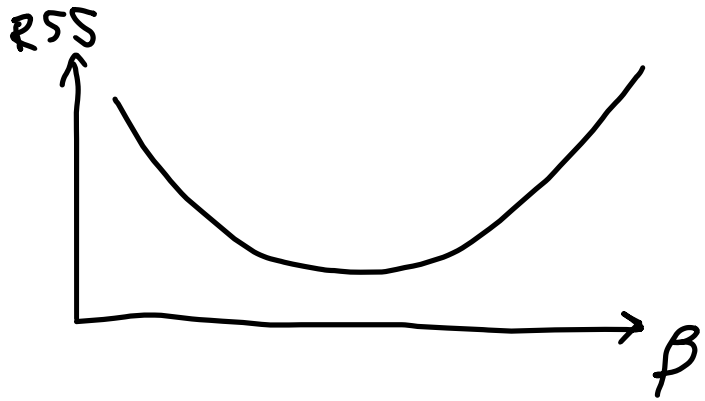
$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

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

Recap

Parameter estimation for $\boldsymbol{\beta}$

- In multiple linear regression, $\boldsymbol{\beta}$ is estimated with *maximum likelihood* and *least squares*.
- Aim: we want to estimate $\boldsymbol{\beta}$ where RSS is at its minimum



$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

then, $\boldsymbol{\beta}$ can be found by solving

$$\frac{\partial \text{RSS}}{\partial \boldsymbol{\beta}} = \mathbf{0}$$

- We derived:

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

Recap

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

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

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

```r
# Linear regression
```{r}
reg.price = lm(medv ~ rm, data=df)

summary(reg.price)
```
```

```
Call:
lm(formula = medv ~ rm, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-23.346  -2.547   0.090   2.986  39.433

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -34.671      2.650  -13.08   <2e-16 ***
rm             9.102      0.419   21.72   <2e-16 ***
```

```r
```{r}
# estimating the betas with the analytical solution
X = model.matrix(reg.price)
Y = df$medv
betahat = solve(t(X) %*% X) %*% t(X) %*% Y
```
```

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

```
> head(Y)
     [,1]
[1,] 24.0
[2,] 21.6
[3,] 34.7
[4,] 33.4
[5,] 36.2
[6,] 28.7
```

```
> head(X)
  (Intercept)    rm
1           1 6.575
2           1 6.421
3           1 7.185
4           1 6.998
5           1 7.147
6           1 6.430
```

```
> betahat
                   [,1]
(Intercept) -34.670621
rm            9.102109
```

Multiple Linear Regression (continued)

Multiple Linear Regression (continued)

Distribution of the regression parameter estimator

- Given

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

- What are
  - The mean $\mathrm{E}[\widehat{\boldsymbol{\beta}}]$?
  - The mean $\mathrm{Cov}(\widehat{\boldsymbol{\beta}})$?
  - The distribution of $\widehat{\boldsymbol{\beta}}$?
- Hint:
  - $\widehat{\boldsymbol{\beta}} = \boldsymbol{C}\boldsymbol{Y}$ where $\boldsymbol{C} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$
  - $\mathrm{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$
  - $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) = \mathrm{Cov}(\boldsymbol{C}\boldsymbol{Y}) = \boldsymbol{C}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\boldsymbol{C}^{\boldsymbol{T}}$
  - $\boldsymbol{Y}$ follows $\boldsymbol{N}(\dots)$. Therefore, its component $\widehat{\boldsymbol{\beta}}$ also follows $\boldsymbol{N}(\dots)$
- You'll work on it on the recommended exercise 3 (w/ the solutions attached)
- $\widehat{\boldsymbol{\beta}} \sim \boldsymbol{N}\left(\boldsymbol{\beta}, \mathrm{Cov}(\widehat{\boldsymbol{\beta}})\right)$

Random Vector

Linear combinations

$\boldsymbol{Z} = \boldsymbol{C}\boldsymbol{X}$

- scalar matrix $\boldsymbol{C} \in \mathbb{R}^{k\times p}$
- random matrix $\boldsymbol{X} \in \mathbb{R}^{p\times n}$

Then,

- $\mathrm{E}(\boldsymbol{Z}) = \mathrm{E}(\boldsymbol{C}\boldsymbol{X}) = \boldsymbol{C}\mathrm{E}(\boldsymbol{X}) = \boldsymbol{C}\boldsymbol{\mu}_{\boldsymbol{X}}$
- $\mathrm{Cov}(\boldsymbol{Z}) = \mathrm{Cov}(\boldsymbol{C}\boldsymbol{X}) = \boldsymbol{C}\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{C}^T$
  - $\boldsymbol{\mu}_{\boldsymbol{X}}$: $\boldsymbol{\mu}$ of $\boldsymbol{X}$
  - $\boldsymbol{\Sigma}_{\boldsymbol{X}}$: $\boldsymbol{\Sigma}$ of $\boldsymbol{X}$

Multiple Linear Regression (continued)

The variance-covariance matrix of $\widehat{\beta}$ in R

```r
# variance-covariance matrix of \hat{beta}
reg.price = lm(medv ~ rm + age, data=df)
vcov(reg.price)
```

```
             (Intercept)           rm          age
(Intercept)  8.16105965 -1.137029517 -0.0136552875
rm          -1.13702952  0.169810025  0.0010183841
age         -0.01365529  0.001018384  0.0001057984
```

Four important questions

Four important questions

1. Is at least one of the predictors $X_1, \ldots, X_p$ useful in predicting the response? (statistical test)
2. Is a particular subset of X useful in predicting the response? (statistical test)
3. How well does the model fit the data? (model assessment)
4. How much uncertainty is in our estimated model and prediction? (uncertainty measure)

Four important questions

1. Is at least one of the predictors $X_1, \ldots, X_p$ useful in predicting the response?

- To answer the question, we can compare

$$Y = \mathrm{E}[Y] = \beta_0$$

vs

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- How do we do that?
  → some sort of statistical test?
- What are our hypotheses?
  - H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$
  - H_1: at least one β_j is non-zero.
- What test can we use? → F-test!
- In case we're a bit fuzzy about it, let's refresh our memory.
  (Even if it's new to you, I'll make sure you get on board!)

Four important questions

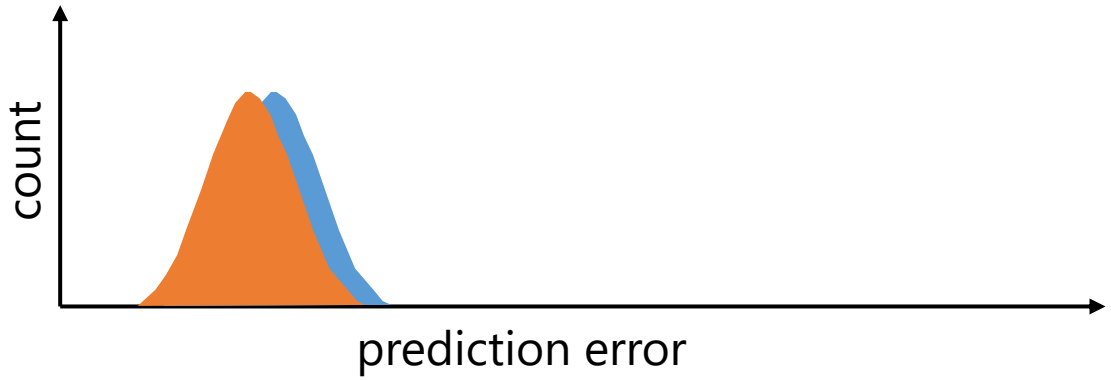
F-test

- Are the two distributions statistically different from each other?

$$Y_S = \mathrm{E}[Y] = \beta_0$$
vs
$$Y_L = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

S: small
L: large

"statistically different"



"statistically not different"

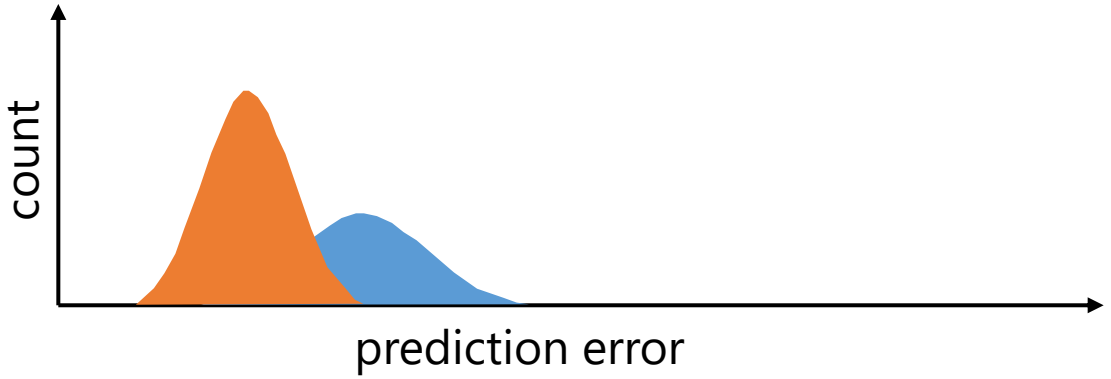
Four important questions

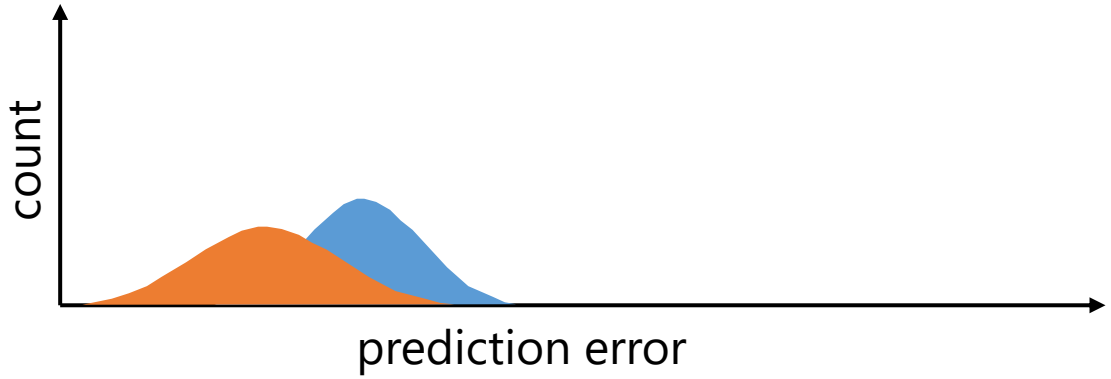
F-test

- It gets tricky in the real world

$$Y_S = \mathrm{E}[Y] = \beta_0$$
vs
$$Y_L = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

S: small
L: large





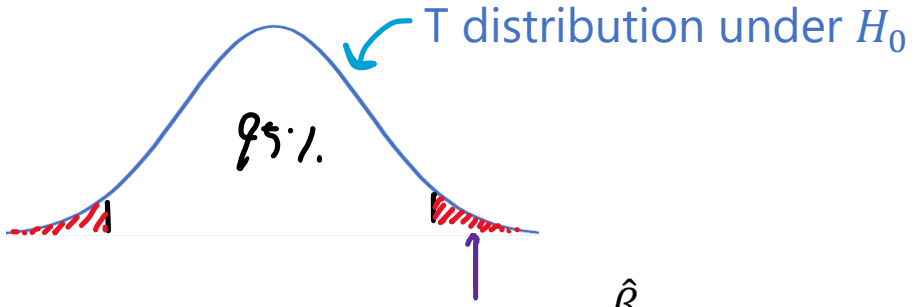
In short,
we want to measure if two distributions are different.
(not by my eyes but some systematic way)

"F-test"

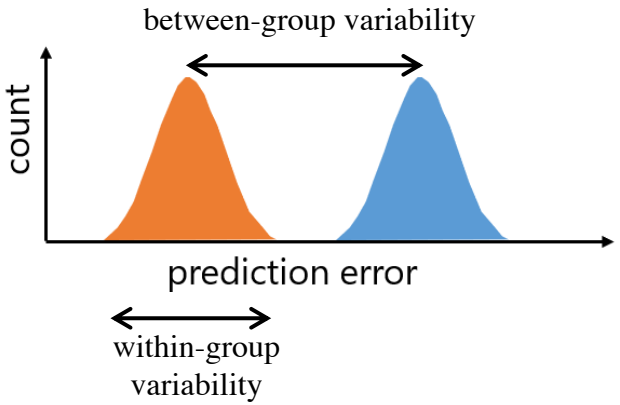
Four important questions

F-test

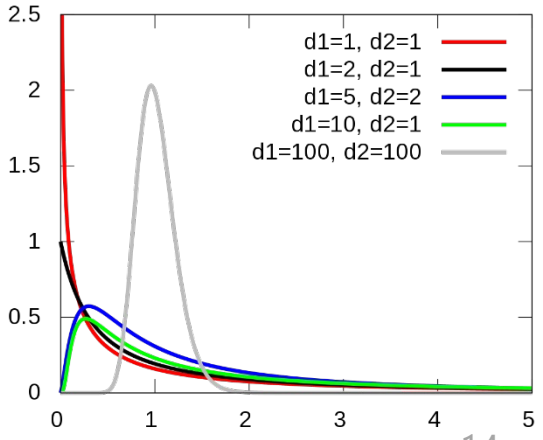
- In the t-test,
  - We got t-value and could calculate p-value.
- In the F-test, we use F-statistics,
  - We get F-value and do the same thing as that.



- In our case, ***"Is at least one of the predictors $X_1, \ldots, X_p$ useful in predicting the response?"***
  - H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$
  - H_1: at least one β_j is non-zero.
  - $F = \frac{(\mathrm{RSS}_S - \mathrm{RSS}_L)/p}{\mathrm{RSS}_L/(n-p-1)} = \frac{\text{between-group variability}}{\text{within-group varability}} \sim F_{df_1, df_2}$
    where $\mathrm{TSS} = \mathrm{RSS}_S = \sum_i (y_i - \bar{y})^2$ and $\mathrm{RSS}_L = \sum_i (y_i - \hat{y}_i)^2$
  - F distribution F_{df_1, df_2} has the following shapes:
    - df_1: p
    - df_2: $(n - p - 1)$



F_{df_1, df_2}



Four important questions

1. **Is at least one of the predictors $X_1, \ldots, X_p$ useful in predicting the response?**

```r
# checking the F-value in the R output
reg.price = lm(medv ~ rm + age, data=df)
summary(reg.price)
```

```
Call:
lm(formula = medv ~ rm + age, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-20.555  -2.882  -0.274   2.293  40.799

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -25.27740    2.85676  -8.848  < 2e-16 ***
rm            8.40158    0.41208  20.388  < 2e-16 ***
age          -0.07278    0.01029  -7.075 5.02e-12 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.316 on 503 degrees of freedom
Multiple R-squared:  0.5303,    Adjusted R-squared:  0.5284
F-statistic: 283.9 on 2 and 503 DF,  p-value: < 2.2e-16
```

F df_1 df_2

Four important questions

2. Is a particular subset of X useful in predicting the response? (statistical test)

- A subset $X = \{X_1, X_2, \dots, X_q\}$ where $q < p$
- we are now comparing

$$Y_S = \beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q$$

vs

$$Y_L = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

S: small
L: large

- Then, what are our hypotheses now?
  - H_0: $\beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0$
  - H_1: at least one β_j is non-zero.
- $F = \frac{(\mathrm{RSS}_S - \mathrm{RSS}_L)/(p-q)}{\mathrm{RSS}_L/(n-p-1)} = \frac{\text{between-group variability}}{\text{within-group varability}} \sim F_{p-q,n-p-q}$

  where $\mathrm{RSS}_S = \sum_{i,\hat{y}_i \in \hat{Y}_S} (y_i - \hat{y}_i)^2$ and $\mathrm{RSS}_L = \sum_{i,\hat{y}_i \in \hat{Y}_L} (y_i - \hat{y}_i)^2$

Four important questions

F-test in R

```r
reg.price.small = lm(medv ~ rm, data=df)
reg.price.large = lm(medv ~ rm + age + crim, data=df)

anova(reg.price.small, reg.price.large)
```

```
Analysis of Variance Table

Model 1: medv ~ rm
Model 2: medv ~ rm + age + crim
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    504 22062
2    502 18640  2    3421.8 46.077 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

P-value

F-test is a subset of ANOVA

Four important questions

Relation between *t*-test and *F*-test

- *F*-test is a generalization of the *t*-test.
- Remember, *t*-test is conducted to test for a single variable:

$$H_0: \beta_j = 0 \text{ vs } H_0: \beta_j \neq 0$$

- We can do the same test with the F-test:

$$F \sim F_{df_1, df_2} \text{ where } df_1 = 1,\ df_2 = n - p - 1$$

- It is known that

$$F_{1, n-p-1} = t^2_{n-p-1}$$

Four important questions

Relation between *t*-test and *F*-test

F-test for a single estimate, $\hat{\beta}_{age}$

```r
reg.price.small = lm(medv ~ rm, data=df)
reg.price.large = lm(medv ~ rm + age, data=df)

anova(reg.price.small, reg.price.large)
```

```
Analysis of Variance Table

Model 1: medv ~ rm
Model 2: medv ~ rm + age
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    504 22062
2    503 20065  1      1997 50.062 5.025e-12 ***
```

$$F_{1,\,n-p-1} = t^2_{n-p-1}$$

***t*-test**

```r
reg.price.large = lm(medv ~ rm + age, data=df)
summary(reg.price.large)
```

```
Call:
lm(formula = medv ~ rm + age, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-20.555  -2.882  -0.274   2.293  40.799

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -25.27740    2.85676  -8.848  < 2e-16 ***
rm            8.40158    0.41208  20.388  < 2e-16 ***
age          -0.07278    0.01029  -7.075 5.02e-12 ***
```

Four important questions

3. How well does the model fit the data?

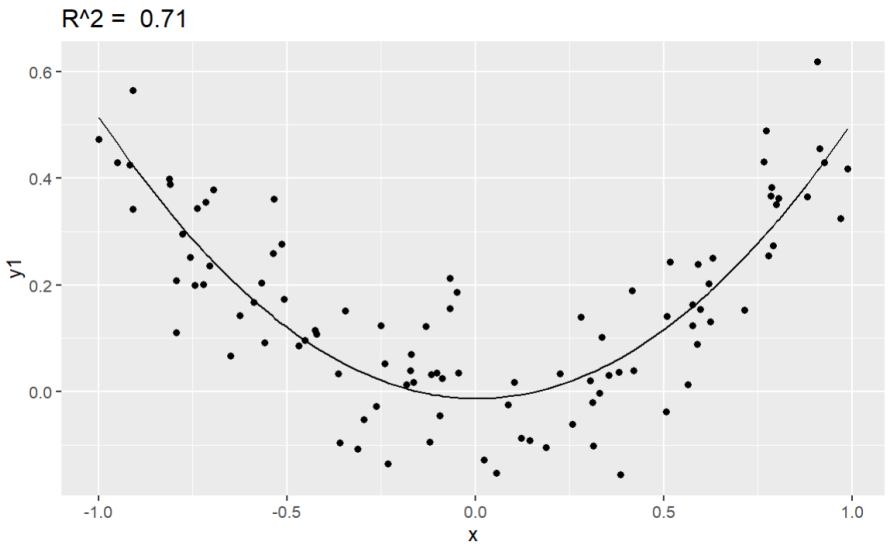
- $\mathrm{RSE} = \sqrt{\frac{\mathrm{RSS}}{n-p-1}}$

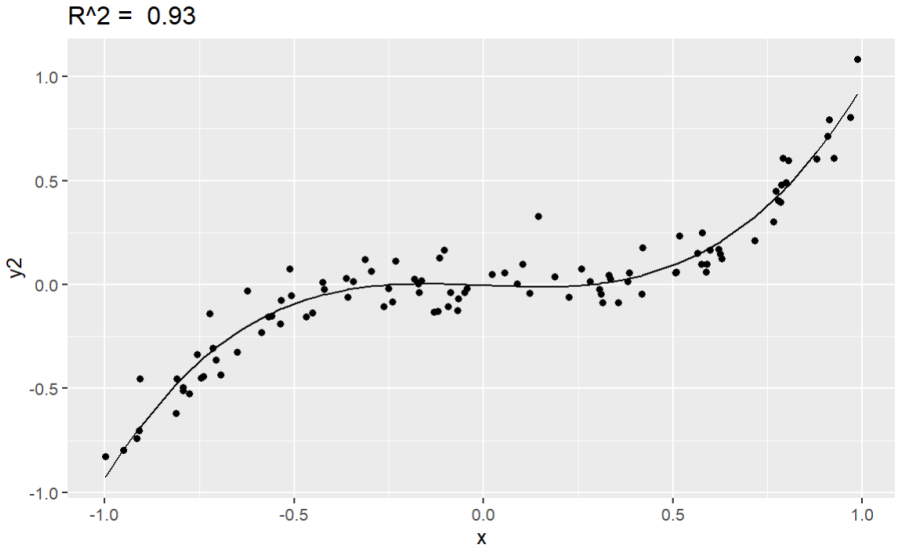
RSS: Residual Sum of Squares

RSE: Residual Standard of Error

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

Typically, it ranges between 0 and 1. The higher, the better.





Four important questions

Adjusted R^2

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

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

- R_a^2 **penalizes for adding more input features**.
- R_a^2 says *"add useful input features only"*

  Score = performance – model size

Four important questions

Adjusted R^2

```r
reg.price = lm(medv ~ rm + age, data=df)
summary(reg.price)
```

```
Call:
lm(formula = medv ~ rm + age, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-20.555  -2.882  -0.274   2.293  40.799

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -25.27740    2.85676  -8.848  < 2e-16 ***
rm            8.40158    0.41208  20.388  < 2e-16 ***
age          -0.07278    0.01029  -7.075 5.02e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.316 on 503 degrees of freedom
Multiple R-squared:  0.5303,    Adjusted R-squared:  0.5284
F-statistic: 283.9 on 2 and 503 DF,  p-value: < 2.2e-16
```

Four important questions

How much uncertainty is in our estimated model and prediction?

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

poses uncertainty to our prediction.

carries uncertainty.
Remember the $\mathrm{SE}(\hat{\beta}_j)$!

Two types of uncertainties here

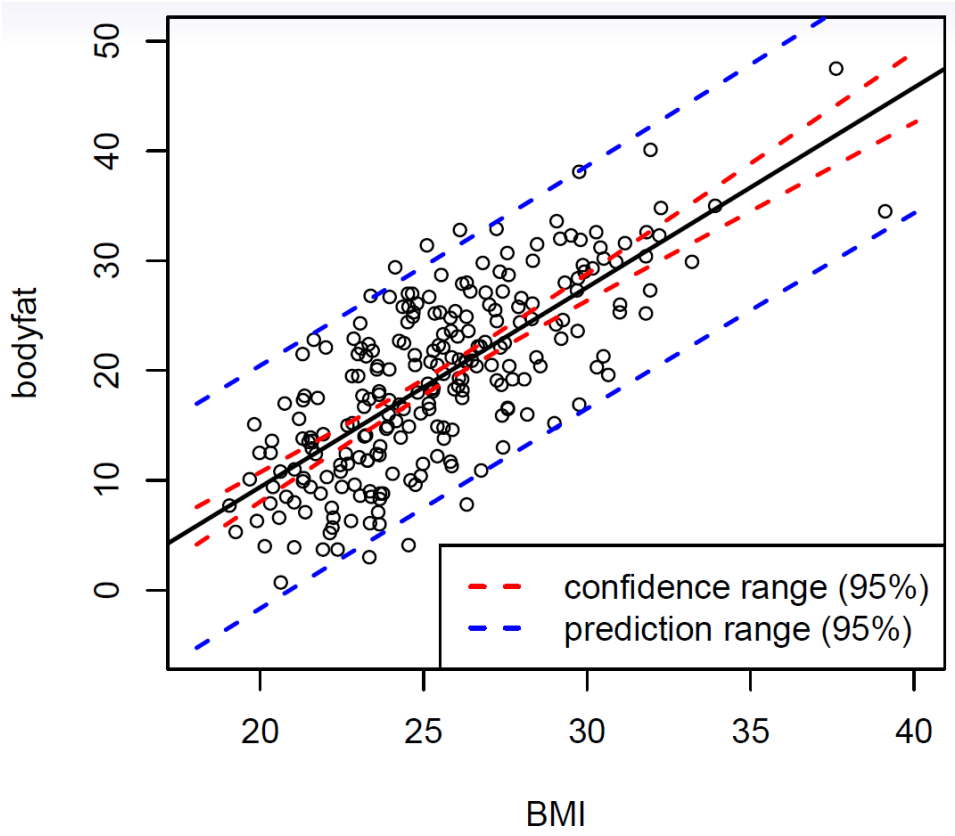
- Uncertainty in $\hat{f}$ due to $\mathrm{SE}(\hat{\beta})$
  → *confidence interval*
- Uncertainty in $\hat{Y}$ due to $\mathrm{SE}(\hat{\beta})$ and ϵ
  → *prediction interval*

```{r}
reg.price = lm(medv ~ rm + age, data=df)
summary(reg.price)
```

```
Call:
lm(formula = medv ~ rm + age, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-20.555  -2.882  -0.274   2.293  40.799

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -25.27740    2.85676  -8.848  < 2e-16 ***
rm            8.40158    0.41208  20.388  < 2e-16 ***
age          -0.07278    0.01029  -7.075 5.02e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.316 on 503 degrees of freedom
Multiple R-squared:  0.5303,    Adjusted R-squared:  0.5284
F-statistic: 283.9 on 2 and 503 DF,  p-value: < 2.2e-16
```

Four important questions

How much uncertainty is in our estimated model and prediction?



Two types of uncertainties here

- Uncertainty in $\hat{f}$ due to $\mathrm{SE}(\hat{\beta})$
  → *confidence interval*
- Uncertainty in $\hat{Y}$ due to $\mathrm{SE}(\hat{\beta})$ and ϵ
  → *prediction interval*

Q. Why is the prediction range wider than the confidence range?

Four important questions

Confidence range and Prediction range in R

```r
reg.price = lm(medv ~ rm + age, data=df)

newobs = df[1,]  # taking the first row from the dataset
```

```r
predict(reg.price, newdata=newobs, interval="confidence")
```

```
       fit      lwr      upr
1 25.21795 24.62086 25.81504
```

```r
predict(reg.price, newdata=newobs, interval="prediction")
```

```
       fit      lwr      upr
1 25.21795 12.79482 37.64108
```

Extensions of the linear model

Extensions of the linear model

Extensions that make the linear model very powerful!

- Binary covariates (*e.g.*, male/female, smoker/non-smoker)
- Categorical covariates (*e.g.*, green/brown/yellow)
- Interaction terms (*e.g.*, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_1 X_2$)
- Non-linear terms (*e.g.*, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_1^2$)

Covariates == Input feature

Extensions of the linear model

Binary predictors (covariate / input feature)

Let's look at some example

- $y_{weight} = \beta_0 + \beta_1 x_{gender} + \epsilon$
- $x_{gender} \in \{\text{male}, \text{female}\}$
- Then, can we do

$$y_{weight} = \beta_0 + \beta_1 x_{male} + \beta_2 x_{female} + \epsilon$$

$$x_{male} = \begin{cases} 1 & \text{If so,} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{female} = \begin{cases} 1 & \text{If so,} \\ 0 & \text{otherwise} \end{cases}$$

- The answer is STRONG NO!!
- *The model becomes non-interpretable.*

  So, both x_{male} and x_{female} have valid values of 0 and 1.

  What does $y_{weight} = \beta_0$ mean, then? (the translation would be: "when it's neither male or female???")
- We need a reference category!

Extensions of the linear model

Binary predictors (covariate / input feature)

- Then, can we do

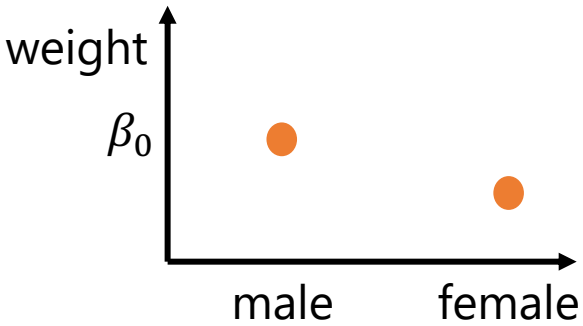
$$y_{weight} = \beta_0 + \beta_1 x_{male} + \beta_2 x_{female} + \epsilon$$

$$x_{male} = \begin{cases} 1 & \text{If so,} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{female} = \begin{cases} 1 & \text{If so,} \\ 0 & \text{otherwise} \end{cases}$$

- The answer is STRONG NO!!
- *The model becomes non-interpretable.*
  So, both x_{male} and x_{female} have valid values of 0 and 1.
  What does $y_{weight} = \beta_0$ mean, then? (the translation would be: "when it's neither male or female?") (ignore ϵ for now)
- We need a reference category! (*i.e.*, we need a baseline)
- We set "male" as the reference category (= baseline), then we're actually saying
  $y_{weight} = \beta_0$ represents the male weight.
- Then, when x is female, we have $y_{weight} = \beta_0 + \beta_1 x_{female.compared.to.male}$

$$x_{female.compared.to.male} = \begin{cases} 1 & \text{if female,} \\ 0 & \text{otherwise} \end{cases}$$



Extensions of the linear model

Qualitative predictors (input features) with more than 2 levels

Let's look at some example

- $x_{education} \in \{\text{below. collge}, \text{Bachelor}, \text{Master}, \text{PhD}\}$
- We set "below.college" as the reference category! Then,
- $y_{income} = \beta_0 + \beta_1 x_{Bachelor \leftrightarrow below.college} + \beta_2 x_{Master \leftrightarrow below.college} + \beta_3 x_{PhD \leftrightarrow below.college} + \epsilon$

$$x_{Bachelor \leftrightarrow below.college} = \begin{cases} 1 & \text{if Bachelor,} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{Master \leftrightarrow below.college} = \begin{cases} 1 & \text{if Master,} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{PhD \leftrightarrow below.college} = \begin{cases} 1 & \text{if PhD,} \\ 0 & \text{otherwise} \end{cases}$$

- That means, when $y_{income} = \beta_0$,
  β_0 refers to y_{income} when the education level is below college.
- Remember a number of β for k-level qualitative variable is $k - 1$.

Terminology: $x_{red}, x_{blue}, x_{green}$ are called *dummy variables*.

Extensions of the linear model

Example

- We're now using a new dataset, named *Credit* from the ISLR library

```r
library(ISLR)

head(Credit)
```

	ID <int>	Income <dbl>	Limit <int>	Rating <int>	Cards <int>	Age <int>	Education <int>	Gender <fctr>	Student <fctr>
1	1	14.891	3606	283	2	34	11	Male	No
2	2	106.025	6645	483	3	82	15	Female	Yes
3	3	104.593	7075	514	4	71	11	Male	No
4	4	148.924	9504	681	3	36	11	Female	No
5	5	55.882	4897	357	2	68	16	Male	No
6	6	80.180	8047	569	4	77	10	Male	No

Extensions of the linear model

Example: Introducing a new dataset

- We're now using a new dataset, named *Credit* from the ISLR library

ID <int>	Income <dbl>	Limit <int>	Rating <int>	Cards <int>	Age <int>	Education <int>	Gender <fctr>	Student <fctr>	Married <fctr>	Ethnicity <fctr>	Balance <int>
1	14.891	3606	283	2	34	11	Male	No	Yes	Caucasian	333
2	106.025	6645	483	3	82	15	Female	Yes	Yes	Asian	903
3	104.593	7075	514	4	71	11	Male	No	No	Asian	580
4	148.924	9504	681	3	36	11	Female	No	No	Asian	964
5	55.882	4897	357	2	68	16	Male	No	Yes	Caucasian	331
6	80.180	8047	569	4	77	10	Male	No	No	Caucasian	1151

Column names:
"ID" "Income" "Liit" "Rating" "Cards" "Age" "Education" "Gender" "Student" "Married" "Ethnicity" "Balance"

Extensions of the linear model

Example: Linear regression model with a qualitative input variable.

```{r}
lm.credit <- lm(Rating ~ Student, data=Credit)
summary(lm.credit)
```

```
Call:
lm(formula = Rating ~ Student, data = Credit)

Residuals:
    Min      1Q  Median      3Q     Max
-261.00 -107.04  -11.04   82.97  626.96

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  355.044      8.165   43.48   <2e-16 ***
StudentYes    -1.044     25.820   -0.04    0.968
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 154.9 on 398 degrees of freedom
Multiple R-squared:  4.111e-06, Adjusted R-squared:  -0.002508
F-statistic: 0.001636 on 1 and 398 DF,  p-value: 0.9678
```

$\hat{y}_{rating} = \hat{\beta}_0 + \hat{\beta}_1 x_{student.yes}$

The refernece category is ”not student”

When x is not a student,
$\hat{y}_{rating} = \hat{\beta}_0 = 355.04$

When x is a student,
$\hat{y}_{rating} = \hat{\beta}_0 + \hat{\beta}_1 x_{student.yes} = 355.04 - 1.04 \times 1$

Extensions of the linear model

Example: Linear regression model with a qualitative input variable.

```
```{r}
lm.credit <- lm(Cards ~ Ethnicity, data=Credit)
summary(lm.credit)
```
```

```
Call:
lm(formula = Cards ~ Ethnicity, data = Credit)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9706 -0.9596  0.0404  1.0404  6.0503

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.959596   0.138162  21.421   <2e-16 ***
EthnicityAsian     0.010992   0.193949   0.057    0.955
EthnicityCaucasian -0.009847   0.169072  -0.058    0.954
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.375 on 397 degrees of freedom
Multiple R-squared:  3.98e-05,  Adjusted R-squared:  -0.004998
F-statistic: 0.007901 on 2 and 397 DF,  p-value: 0.9921
```

- There are three ethnicities:
  African American, Asian, Caucasian
- Cards: a number of credit cards

The refernece category is "African American"

$$\hat{y}_{cards} = \hat{\beta}_0 + \hat{\beta}_1 x_{Asian \leftrightarrow American} + \hat{\beta}_1 x_{Caucasian \leftrightarrow American}$$

When x is Afican American,

$$\hat{y}_{cards} = \hat{\beta}_0 = 2.96$$

When x is Asian,

$$\hat{y}_{cards} = \hat{\beta}_0 + \hat{\beta} x_{Asian \leftrightarrow American} = 2.96 + 0.01 \times 1$$

When x is Caucasian,

$$\hat{y}_{cards} = \hat{\beta}_0 + \hat{\beta}_1 x_{Caucasian \leftrightarrow American} = 2.96 - 0.01 \times 1$$

Extensions of the linear model

Testing for a qualitative input variable (=categorical predictor)

Question: Is *Ethnicity* a useful feature to predict *Cards*?

- Statistical testing can be used to answer the question

$$H_0\colon \beta_1 = \beta_2 = 0$$

vs

$$H_1\colon \text{at least one } \beta_j \text{ is non-zero.}$$

- We can use F-test.

```
```{r}
lm.credit <- lm(Cards ~ Ethnicity, data=Credit)
anova(lm.credit)
```
```

```
Analysis of Variance Table

Response: Cards
            Df Sum Sq Mean Sq F value Pr(>F)
Ethnicity    2   0.03 0.01493  0.0079 0.9921
Residuals  397 750.25 1.88979
```

Extensions of the linear model

Testing for a qualitative input variable (=categorical predictor)

- We can use F-test.

```
```{r}
lm.credit <- lm(Cards ~ Ethnicity, data=Credit)
anova(lm.credit)
```
```

```
Analysis of Variance Table

Response: Cards
           Df Sum Sq Mean Sq F value Pr(>F)
Ethnicity   2   0.03 0.01493  0.0079 0.9921
Residuals 397 750.25 1.88979
```

```
```{r}
lm.credit <- lm(Cards ~ Ethnicity, data=Credit)
summary(lm.credit)
```
```

```
Call:
lm(formula = Cards ~ Ethnicity, data = Credit)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9706 -0.9596  0.0404  1.0404  6.0503

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.959596   0.138162  21.421   <2e-16 ***
EthnicityAsian     0.010992   0.193949   0.057    0.955
EthnicityCaucasian -0.009847   0.169072  -0.058    0.954
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.375 on 397 degrees of freedom
Multiple R-squared:  3.98e-05,  Adjusted R-squared:  -0.004998
F-statistic: 0.007901 on 2 and 397 DF,  p-value: 0.9921
```

Extensions of the linear model

Interaction terms

Let's say, we were initially modeling with the following two input variables: {Student, Age}

$$y_{rating} = \beta_0 + \beta_1 x_{student.yes} + \beta_2 x_{age} + \epsilon$$

where the reference category for *Student* is "not student" and x_{age} is a quantitative variable.

Then, we want to add some "interaction term" as:

$$y_{rating} = \beta_0 + \beta_1 x_{student.yes} + \beta_2 x_{age} + \beta_3 x_{student.yes} x_{age} + \epsilon$$

Then,

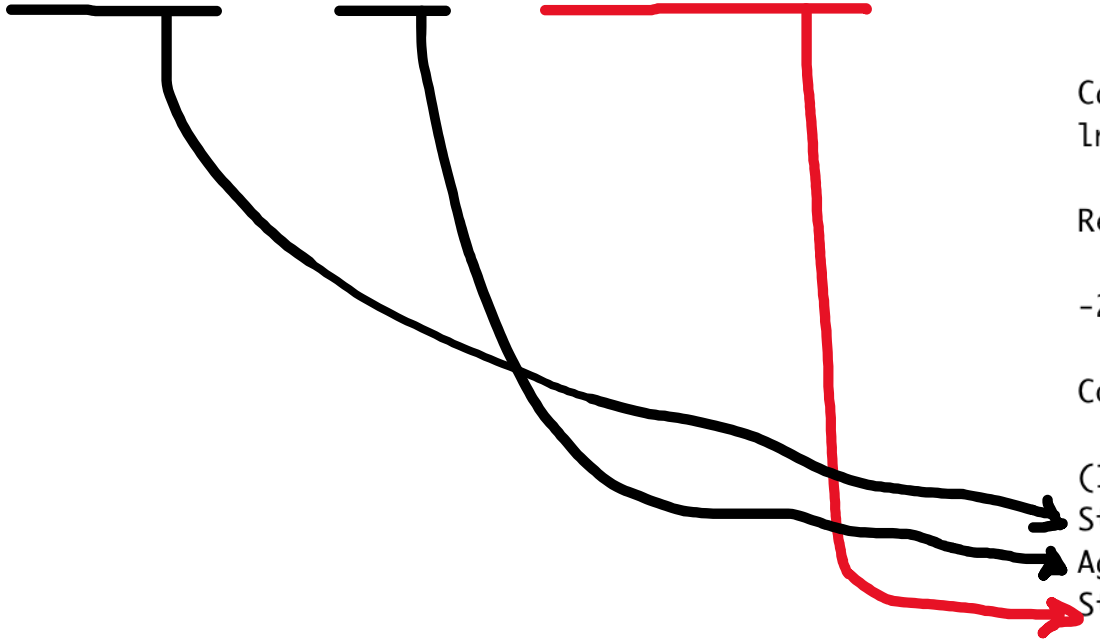
$$y_{rating} = \begin{cases} \beta_0 + \beta_1 + (\beta_2 + \beta_3) x_{age} + \epsilon & \text{if student,} \\ \beta_0 + \beta_2 x_{age} + \epsilon & \text{otherwise,} \end{cases}$$

NB! We now allow different slops and intercepts for the different two groups (*i.e.,* students and non-students).

Extensions of the linear model

Interaction terms

$$y_{rating} = \beta_0 + \beta_1 x_{student.yes} + \beta_2 x_{age} + \beta_3 x_{student.yes} x_{age} + \epsilon$$



```r
lm.credit <- lm(Rating ~ Student * Age, data=Credit)
summary(lm.credit)
```

```
Call:
lm(formula = Rating ~ Student * Age, data = Credit)

Residuals:
    Min      1Q  Median      3Q     Max
-258.45 -106.69   -9.86   85.01  607.24

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    305.0486    27.3448  11.156   <2e-16 ***
StudentYes     -20.0890    93.5180  -0.215   0.8300
Age              0.8954     0.4675   1.915   0.0562 .
StudentYes:Age   0.3802     1.6568   0.229   0.8186
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 154.5 on 396 degrees of freedom
Multiple R-squared:  0.01078,   Adjusted R-squared:  0.003282
F-statistic: 1.438 on 3 and 396 DF,  p-value: 0.2313
```

Extensions of the linear model

The hierarchical principle

- If we include an interaction term in a model, we should always include the main effects.
- For instance, if you're modeling

$$y_{rating} = \beta_0 + \beta_1 x_{student.yes} + \beta_2 x_{age} + \beta_3 x_{student.yes} x_{age} + \epsilon$$

- You can't remove the main term(s) of the interaction term:

$$y_{rating} = \beta_0 + \beta_1 x_{student.yes} + \beta_3 x_{student.yes} x_{age} + \epsilon$$

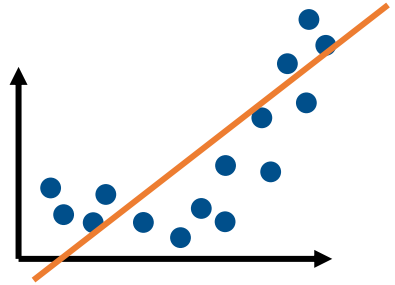
- The further explanation is given on p. 89 in the ISLR book.

Extensions of the linear model

Non-linear terms

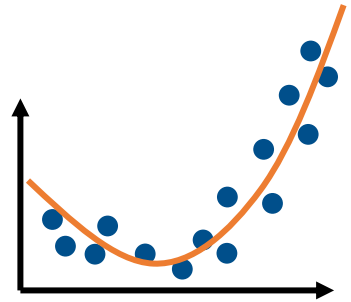
- Let's say we have a basic linear model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon$$



- Then, we want to add *non-linear terms by transforming the existing variable*,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon$$



- More of it?

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon$$

- Other common transformations are:
  - log
  - $\sqrt{\dots}$
  - sin, cos, etc.

Extensions of the linear model

Non-linear terms

- How can you call that a *linear* regression model?

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon$$

- The word *linear* refers to the *linearity in the coefficients*,
  If we consider $X_2 = X_1^2$, then

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- Often, our world is not linear, and we need some non-linearity to express that.
- We'll learn about fancier non-linear methods in a later module!

Challenges – for model fit

Challenges – for model fit

Recap of modeling assumptions in linear regression

- The assumptions in linear regression is that the residuals ϵ follow a normal distribution,

$$\epsilon \sim N(0, \sigma^2)$$

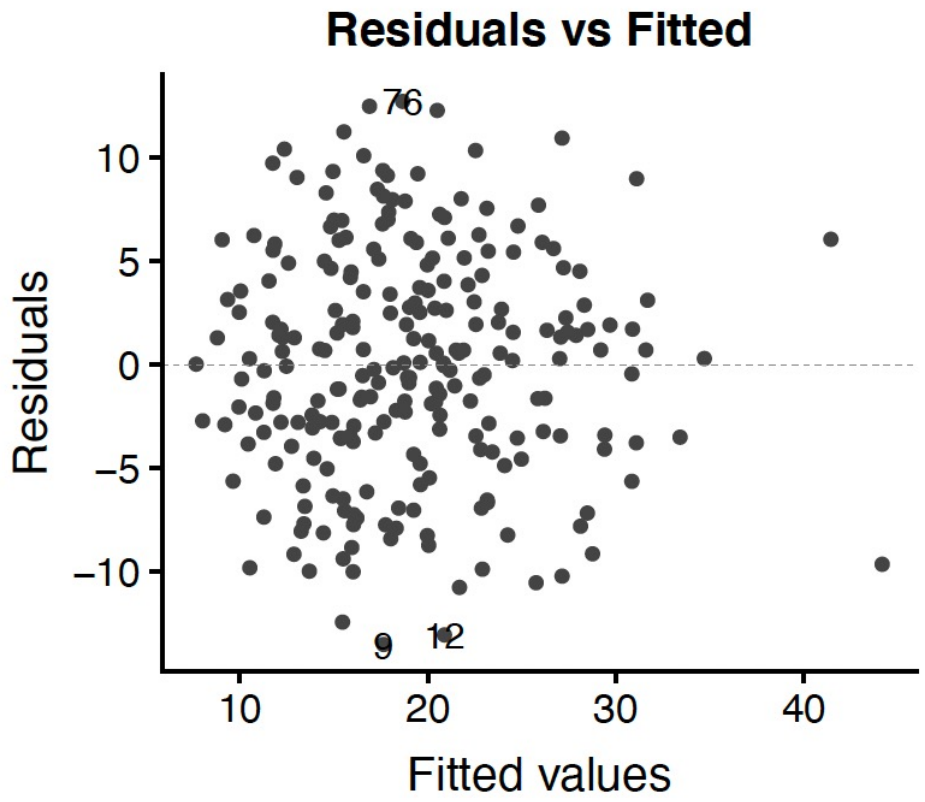
  implying that:

  1. $\mathrm{E}(\epsilon_i) = 0$
  2. $\mathrm{Var}(\epsilon_i) = \sigma^2$
  3. ϵ_i are normally distributed.
  4. ϵ_i are independent of each other.

- To make valid inference from our model, our model assumptions should be fulfilled!

Challenges – for model fit

Modeling checking tool 1: Tukey-Anscombe diagram



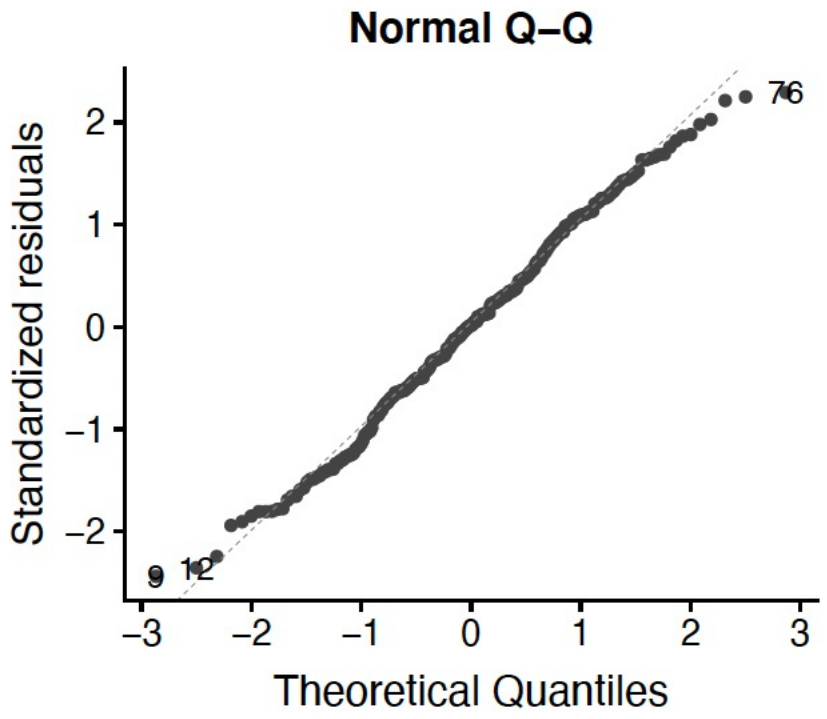
Modeling assumptions:

1. $\mathrm{E}(\epsilon_i) = 0$
2. $\mathrm{Var}(\epsilon_i) = \sigma^2$
3. ϵ_i are normally distributed.
4. ϵ_i are independent of each other.

- We can check 1, 2 (and sort of 4)

Challenges – for model fit

Model checking tool 2: The QQ-diagram



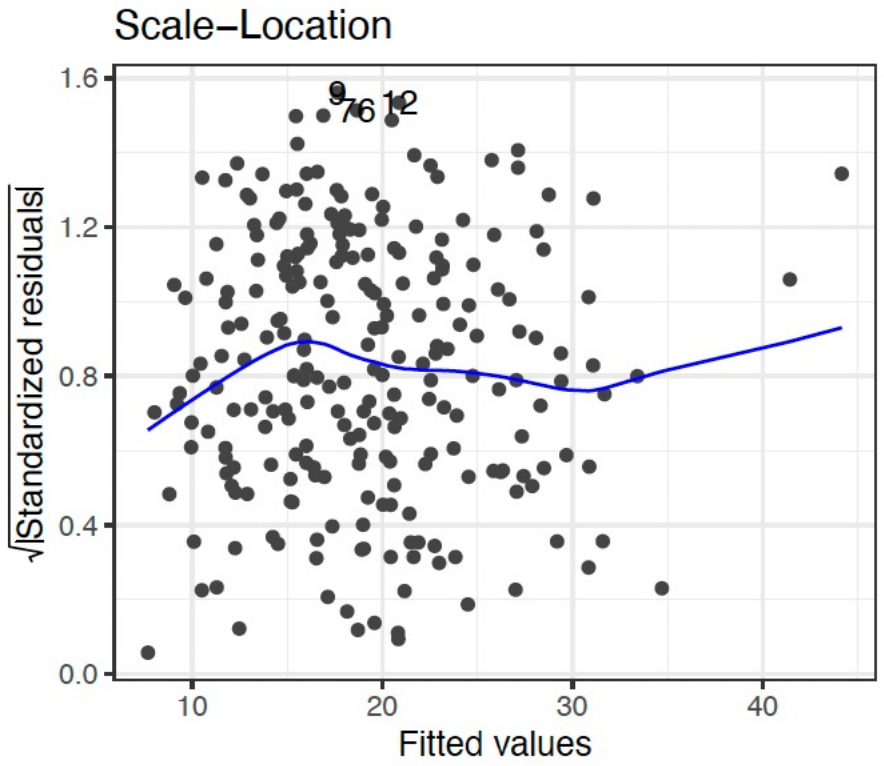
Modeling assumptions:

1. $\mathrm{E}(\epsilon_i) = 0$
2. $\mathrm{Var}(\epsilon_i) = \sigma^2$
3. ϵ_i are normally distributed.
4. ϵ_i are independent of each other.

- We can check 3.
- If the points lie roughly on a straight line, the data is fairly normally distributed.

Challenges – for model fit

Modeling checking tool 3: The scale-location plot



Modeling assumptions:

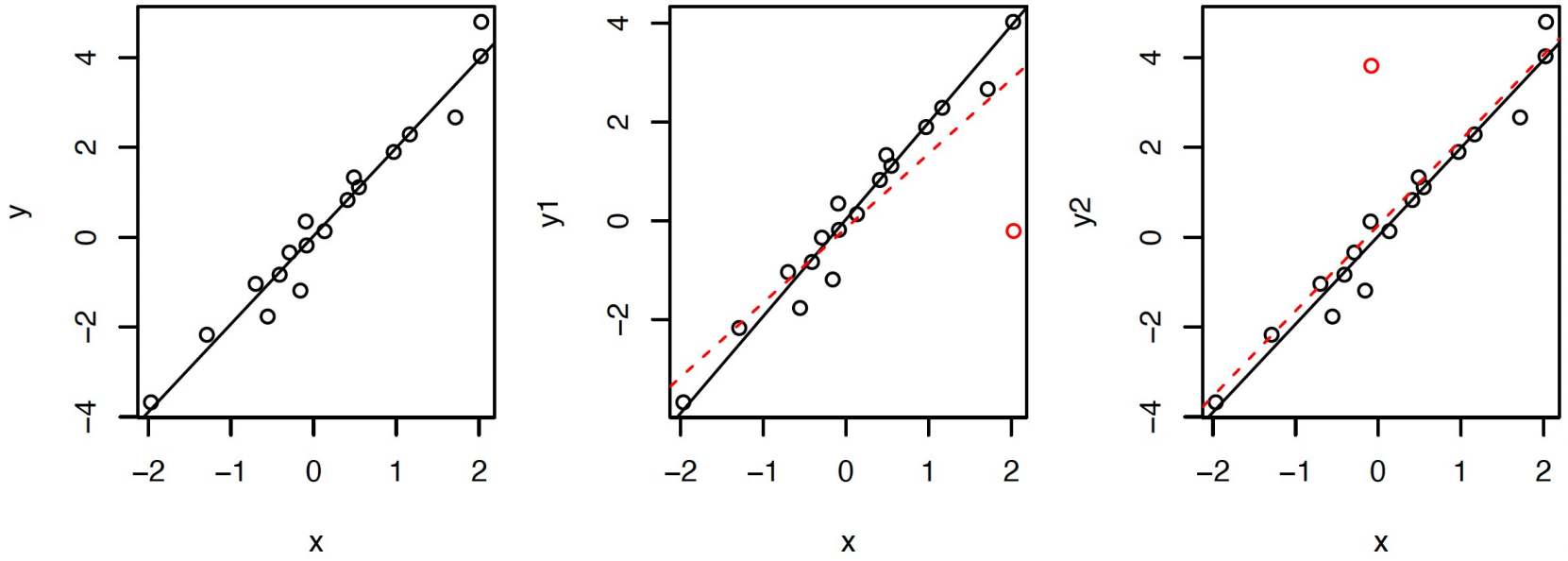
1. $\mathrm{E}(\epsilon_i) = 0$
2. $\mathrm{Var}(\epsilon_i) = \sigma^2$
3. ϵ_i are normally distributed.
4. ϵ_i are independent of each other.

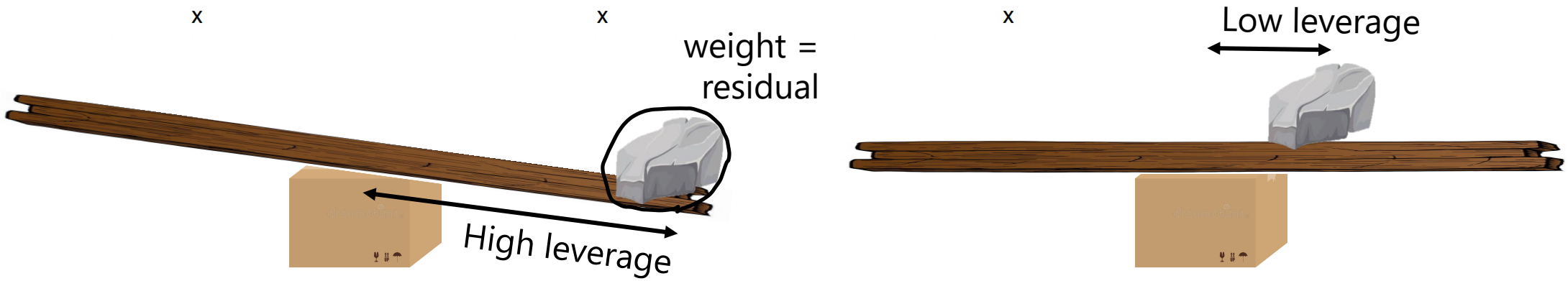
- We can check 2.
- We also want the blue line to be straight (*i.e.*, no pattern in the residual)

Challenges – for model fit

Modeling checking tool 4: The leverage plot

- What is the leverage?



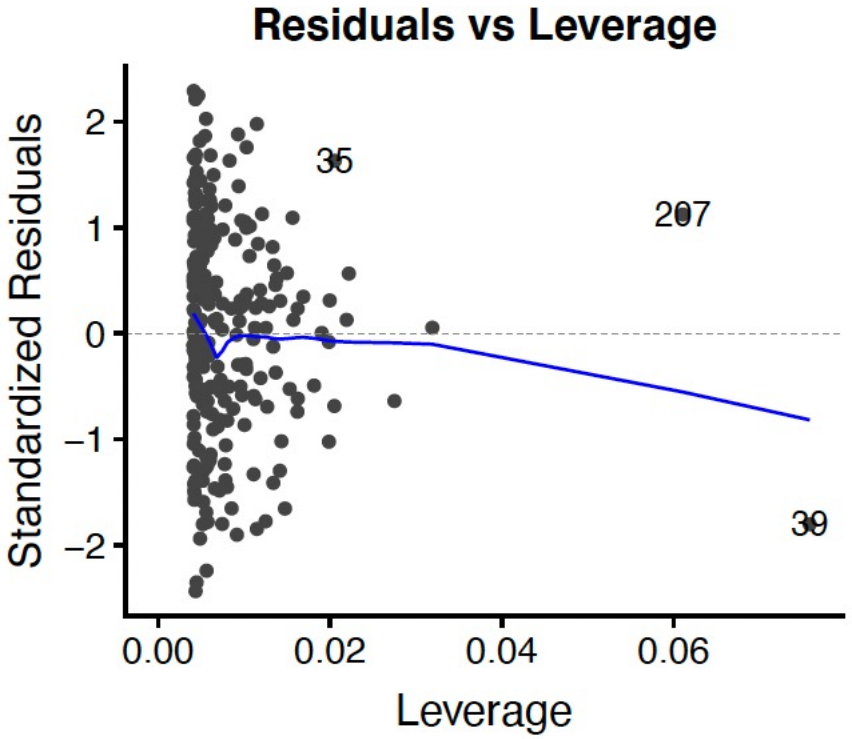


Challenges – for model fit

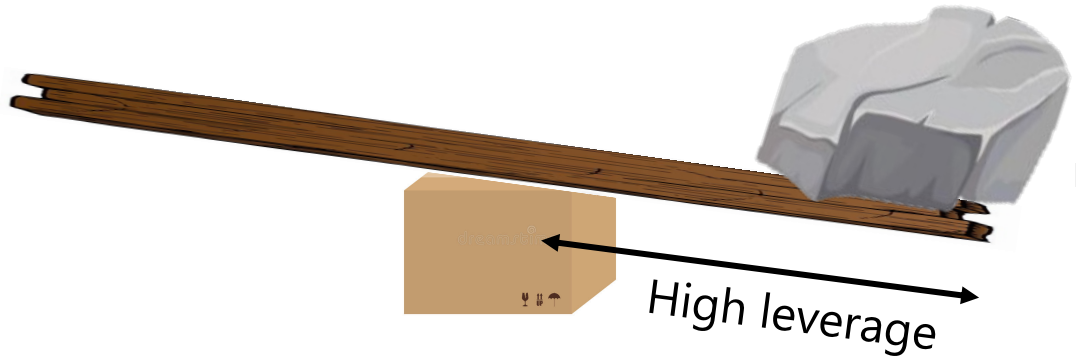
Modeling checking tool 4: The leverage plot

- In simple regression, the leverage is defined as

$$H_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'}(x_{i'} - \bar{x})^2}$$



- It's useful to determine outliers.
- Watch out the samples with *high leverage* and *high residual*.



Challenges – for model fit

Interactive material to see the leverage effects

- https://seeing-theory.brown.edu/regression-analysis/index.html

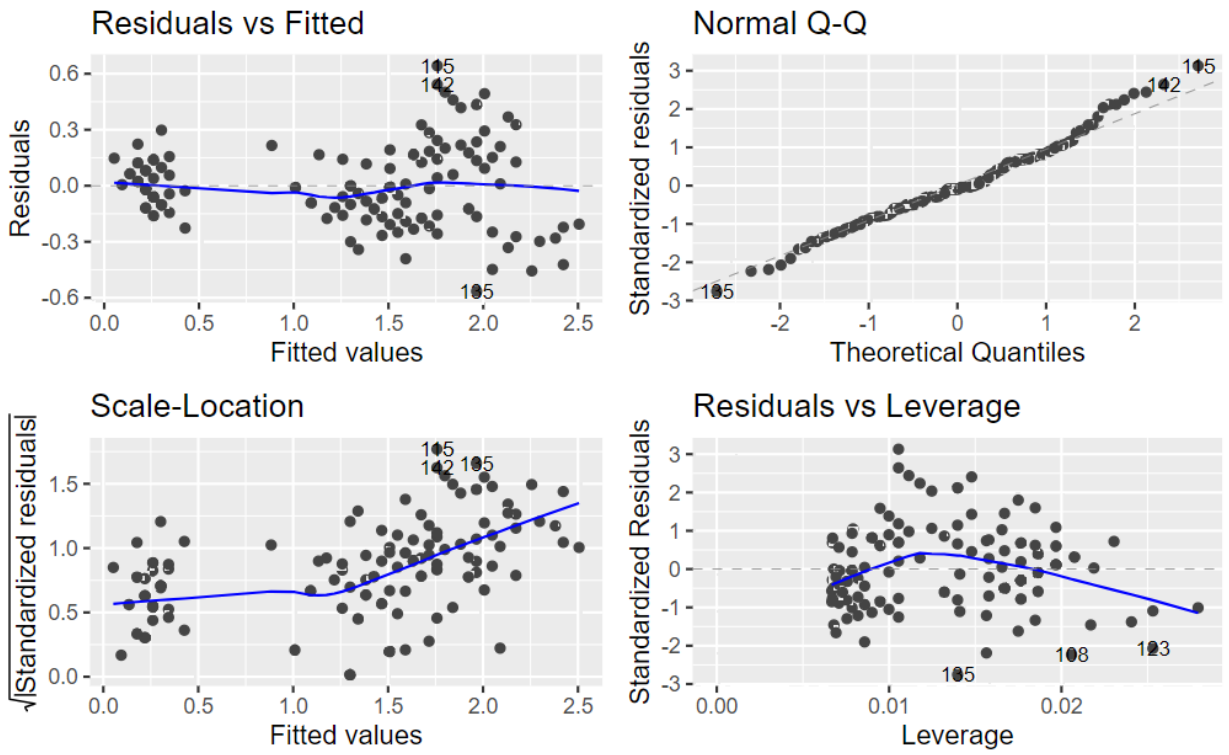
Challenges – for model fit

Diagnostic plots in R

- We can use autoplot() from the ggfortify package

9.4 ggfortify

```
lm(Petal.Width ~ Petal.Length, data = iris) %>%
  autoplot(label.size = 3)
```



https://bookdown.org/Maxine/r4ds/ggfortify.html

Self-study

Self-study

- Further explanation regarding the *leverages* and *collinearity*.
- The recorded video by Stefanie Muff is available here:
- https://ntnu.cloud.panopto.eu/Panopto/Pages/Viewer.aspx?id=110f5d7b-6d50-4c61-8984-acbc00fe3d02

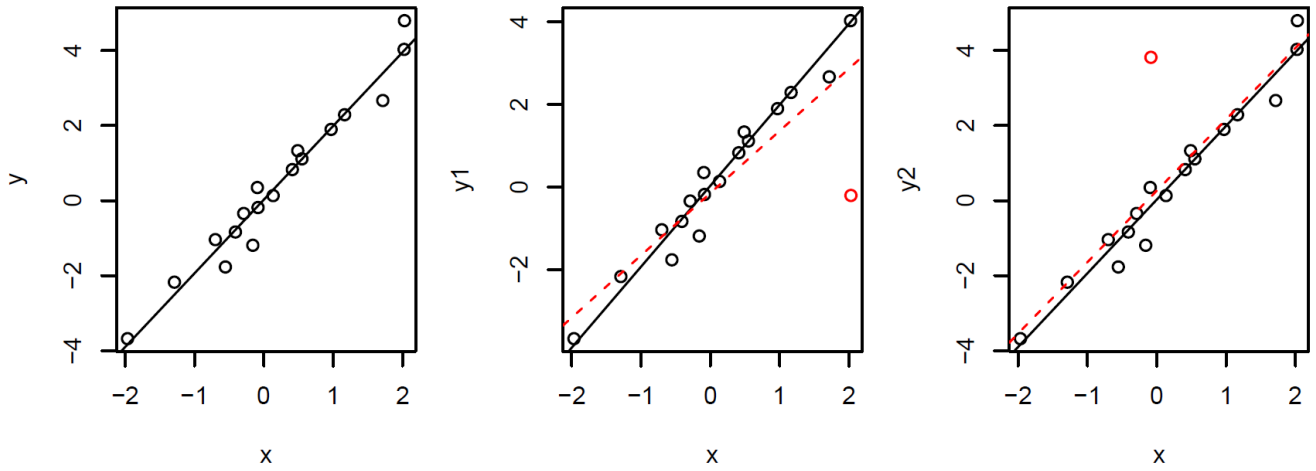
Model checking tool IV: The leverage plot

- Mainly useful to determine outliers.
- To understand the leverage plot, we need to introduce the idea of the **leverage**.
- In simple regression, the leverage of individual i is defined as

$$H_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{i'}(x_{i'} - \overline{x})^2} \ . \tag{4}$$

Q: When are leverages expected to be large/small?

Illustration: Data points with x_i values far from the mean have a stronger leverage effect than when $x_i \approx \overline{x}$:



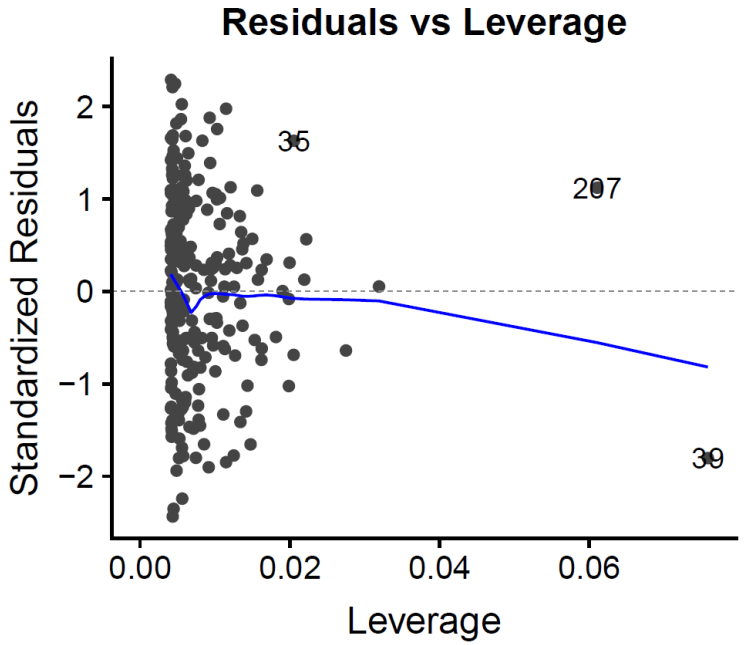
The outlier in the middle plot "pulls" the regression line in its direction and biases the slope.

http://students.brown.edu/seeing-theory/regression-analysis/index.html to do it manually!
[8]

[8] You may choose "Ordinary Least Squares" and then click on "I" for the Ascombe quartett example. Drag points to see what happens with the regression line.

Self-st

In the leverage plot, (standardized) residuals $\tilde{r}_i$ are plotted against the leverage H_{ii} (still for the bodyfat):



Critical ranges are the top and bottom right corners! Why?

Leverages in multiple regression

- Leverage is defined as the diagonal elements of the so-called *hat matrix* $\mathbf{H}$[9], i.e., the leverage of the i-th data point is H_{ii} on the diagonal of $\mathbf{H} = \mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}$.

- Exercise: Verify that formula (4) comes out in the special case of simple linear regression.
- A large leverage indicates that the observation (i) has a large influence on the estimation results, and that the covariate values $(\boldsymbol{x}_i)$ are unusual.

[9] Do you remember why $\mathbf{H}$ is called *hat matrix*?

Different types of residuals?

It can be shown that the vector of residuals, $\mathbf{e} = (e_1, e_2, \ldots, e_n)$ have a normal (singular) distribution with

- $\mathsf{E}(\mathbf{e}) = \mathbf{0}$
- $\mathsf{Cov}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$,

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

This means that the residuals (possibly) have different variance, and may also be correlated.

Q: Why is that a problem?

Self-st

A:

- To check the model assumptions we want to look at the distribution of the error terms ε_i, to check that our errors are independent, homoscedastic (same variance for each observation), and not dependent on our covariates.
- However, we only have the residuals e_i, the "predictions" for ϵ_i.
- It would have been great if the e_i have the same properties as ϵ_i.

$\rightarrow$ To make the e_i more "like ϵ_i", we use *standardized* or *studentized residuals*.

Standardized residuals:

$$\tilde{r}_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - H_{ii}}}$$

where H_{ii} is the ith diagonal element of the hat matrix $\mathbf{H}$.

In R you can get the standardized residuals from an `lm`-object (named `fit`) by `rstandard(fit)`.

Studentized residuals:

$$r_i^* = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1 - H_{ii}}}$$

where $\hat{\sigma}_{(i)}$ is the estimated error variance in a model with observation number i omitted. It can be shown that it is possible to calculated the studentized residuals directly from the standardized residuals.

In R you can get the studentized residuals from an `lm`-object (named `fit`) by `rstudent(fit)`.

Collinearity

In brief, collinearity refers to the situation when two or more predictors are correlated, thus encode (partially) for the same information.

Problems:

- Reduces the accuracy of the estimated coefficients $\hat{\beta}_j$ (large SE!).
- Consequently, reduces power in finding effects (p-values become larger).

Solutions:

- Detect it by calculating the *variance inflation factor* (VIF).
- Remove the problematic variable.
- Or combine the collinear variables into a single new one.

Todo: Read in the course book p.99-102 (self-study).


The End!

Four important questions

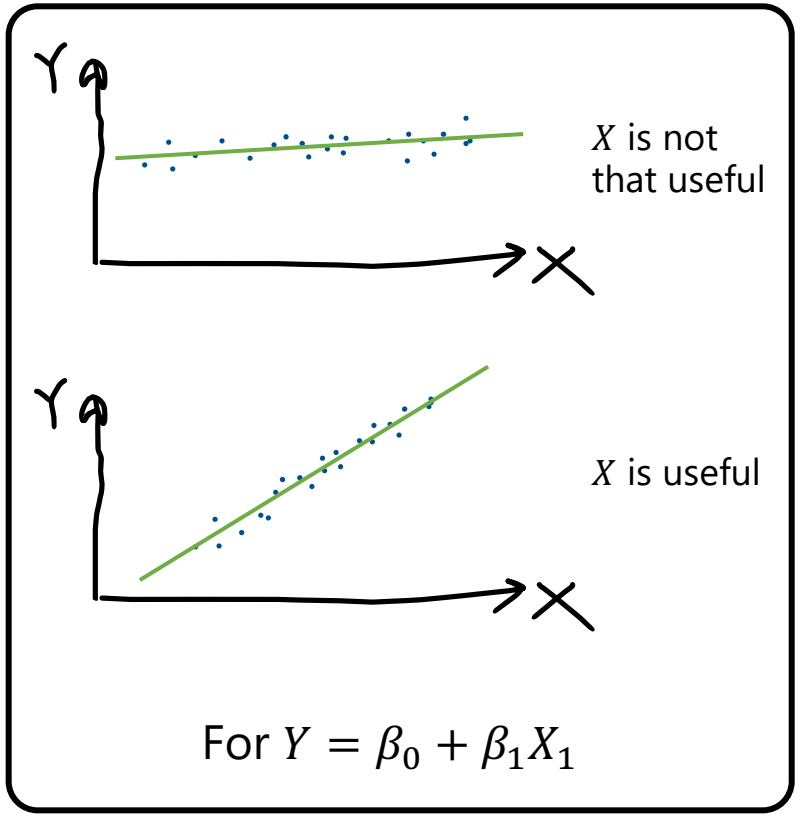
1. Is at least one of the predictors $X_1, \ldots, X_p$ useful in predicting the response?

- To answer the question, we can compare

$$Y = \mathrm{E}[Y] = \beta_0$$

vs

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$



- How do we do that?
  → some sort of statistical test?
- What are our hypotheses?
  - H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$
  - H_1: at least one β_j is non-zero.
- What test can we use? → F-test!
- In case we're a bit fuzzy about it, let's refresh our memory.
  (Even if it's new to you, I'll make sure you get on board!)