Linear Regression
(Simple Linear Case)

Terms

- Y : response/ dependent variable (DV)
  - Variable of interest
- X : predictor/ covariate/ explanatory/ independent variable (IV)
  - Variable used to fit the model for our interest Y

Review: ANOVA Models

- A continuous response **Y**
- One or more categorical explanatory variables X
- Errors assumed to be Normal, N(0,σ^2)
- Interested in differences of expected values (mean) of response variables between groups

Linear Regression Model

- Continuous **response**
- One or more **explanatory** variables; categorical or continuous predictors
- Errors assumed to be Normal, iid $N(0, \sigma^2)$

- Example of regression questions:
  - Does advertising expenditures affect corporate sales?
  - Does the number of hours studied predict course performance?

Exploratory analysis

- Scatter plot (Y vs. X) -> examples on the next page

- The **population correlation coefficient ρ** (rho) measures the **<u>strength</u>** (absolute value) and **<u>direction</u>** (sign) of the association between the variables

✓ **Correlation** is always between -1 and +1 (no unit; standardized measure)

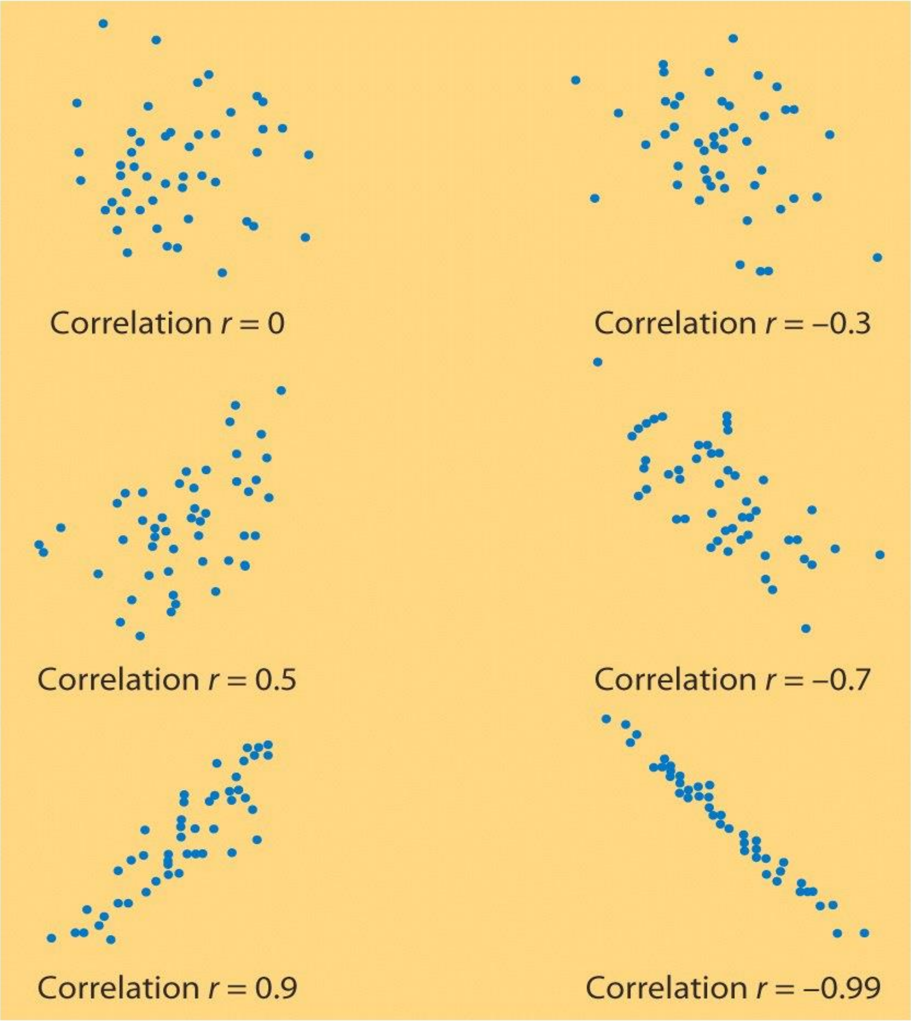
✓ **ρ =0 implies no <u>linear</u> relationship**

But, zero **ρ** does not imply no relationship

✓ Pearson correlation is sensitive to outliers

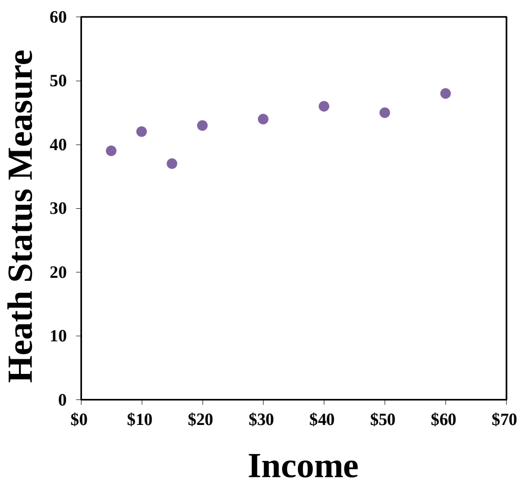
✓ Spearman correlation is robust to outliers

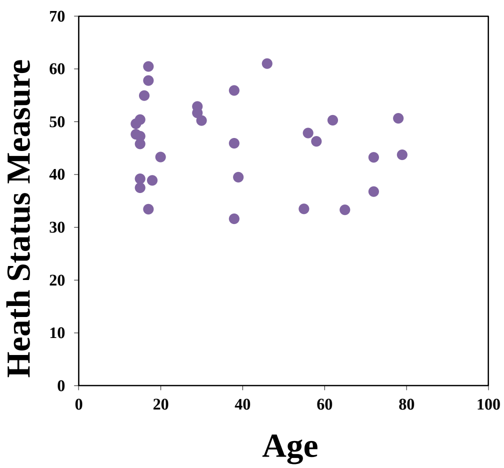
Examples of scatter plots and Correlations

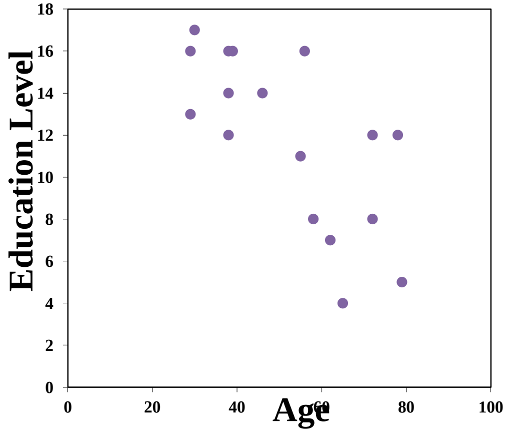


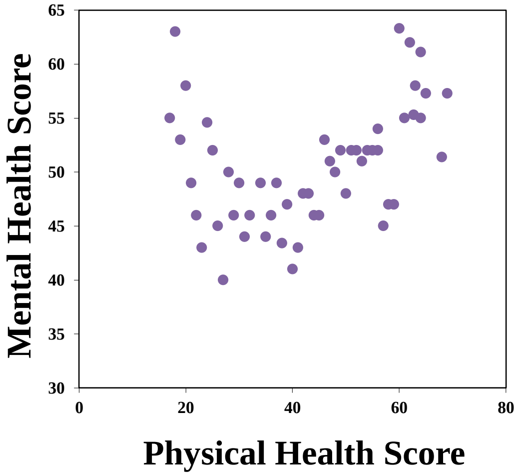
Examples of Relationships

: positive or negative association/ linear or quadratic association etc.









Linear Regression

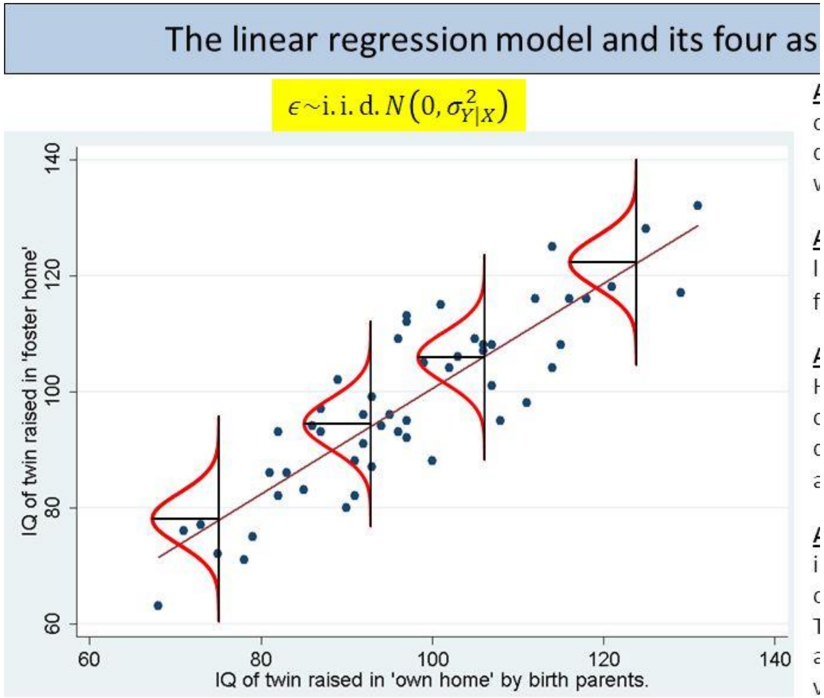
- <u>Objective 1</u>: To quantify the linear relationship between an explanatory variable (X) and response variable (Y) and **test** whether or not it is significant

- <u>Objective 2</u>: to **predict** the **average response** for subjects with a given value of the explanatory variable.

- Not necessarily a causal relationship

- This is parametric approach and what are assumptions?

Assumptions on Linear Regression



1. Linear relationship between X and Y
2. Conditionally **Normal** distribution of Y at each given X=x
3. Homoscedasticity (equal variance assumption on Y at each give X=x)
4. Conditional independence: for given X=x, Y’s are independent each other

Common Linear Models

- Simple linear regression
- Multiple linear regression
- Regression through the origin
- Polynomial regression
- Weighted linear regression
- Linear regression with transformation
- ANOVA model
- ANCOVA model

Least Squares

- Determine the "best" line
- Line as close as possible to the data points in the vertical (*Y*) direction
- **Least Squares Estimator**: Use the line that minimizes the sum of the squares of the vertical distances of the data points from the line
- True (unknown): $Y = \beta_0 + \beta_1 X$+ ERROR
- Regression (estimated) equation: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
- **Interpretation: On average, Y is predicted to have an increase (or decrease) of $\widehat{\beta}_1$ when X increases by 1 unit.**

R output

- Use **lm()** in R to run linear regression
- F-statistics and p-value
- T- statistics and confidence intervals
- If there is significiant linear relationship between Y and X, we can check the direction and magnitude through estimated coefficient
- Able to get the information:
  - Significance of the model (F-test)
  - Significance of individual predictor (t-test)
  - Write down regression equation ($\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X$)
  - How much of variation explained by model: R^2

R output

$$\hat{y} = -5.99 + 1.97X$$

On average, cirrhosis is predicted to have an increase 1.97, with one unit increase on alcohol

```
##
## Call:
## lm(formula = cirrhosis ~ alcohol, data = drinking)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5635 -2.3508  0.1415  2.6149  5.3674
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.9958     2.0977  -2.858   0.0134 *
## alcohol       1.9779     0.2012   9.829  2.2e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1
##
## Residual standard error: 4.17 on 13 degrees of freedom
## Multiple R-squared:  0.8814, Adjusted R-squared:  0.8723
## F-statistic: 96.61 on 1 and 13 DF,  p-value: 2.197e-07
```

t-test on individual term

H0: all β's are zero vs. Ha: at least one β nonzero

(Model is not useful) (Model is useful)

NOTE- we do not test intercept. β's in null hypothesis do not include intercept term

Model Hypothesis Test (F-test in simple linear regression)

- **Null Hypothesis:**
  - ✓ **Conceptually:** The simple linear regression model does ***not*** fit the data better than the baseline model.
  - ✓ **Conceptually:** There is no linear relationship between x and y
  - ✓ **Statistically:** $\beta_1 = 0$

- **Alternative Hypothesis:**
  - ✓ **Conceptually:** The simple linear regression model does fit the data better than the baseline model.
  - ✓ **Conceptually:** There is some linear relationship between x and y
  - ✓ **Statistically:** $\beta_1 \neq 0$

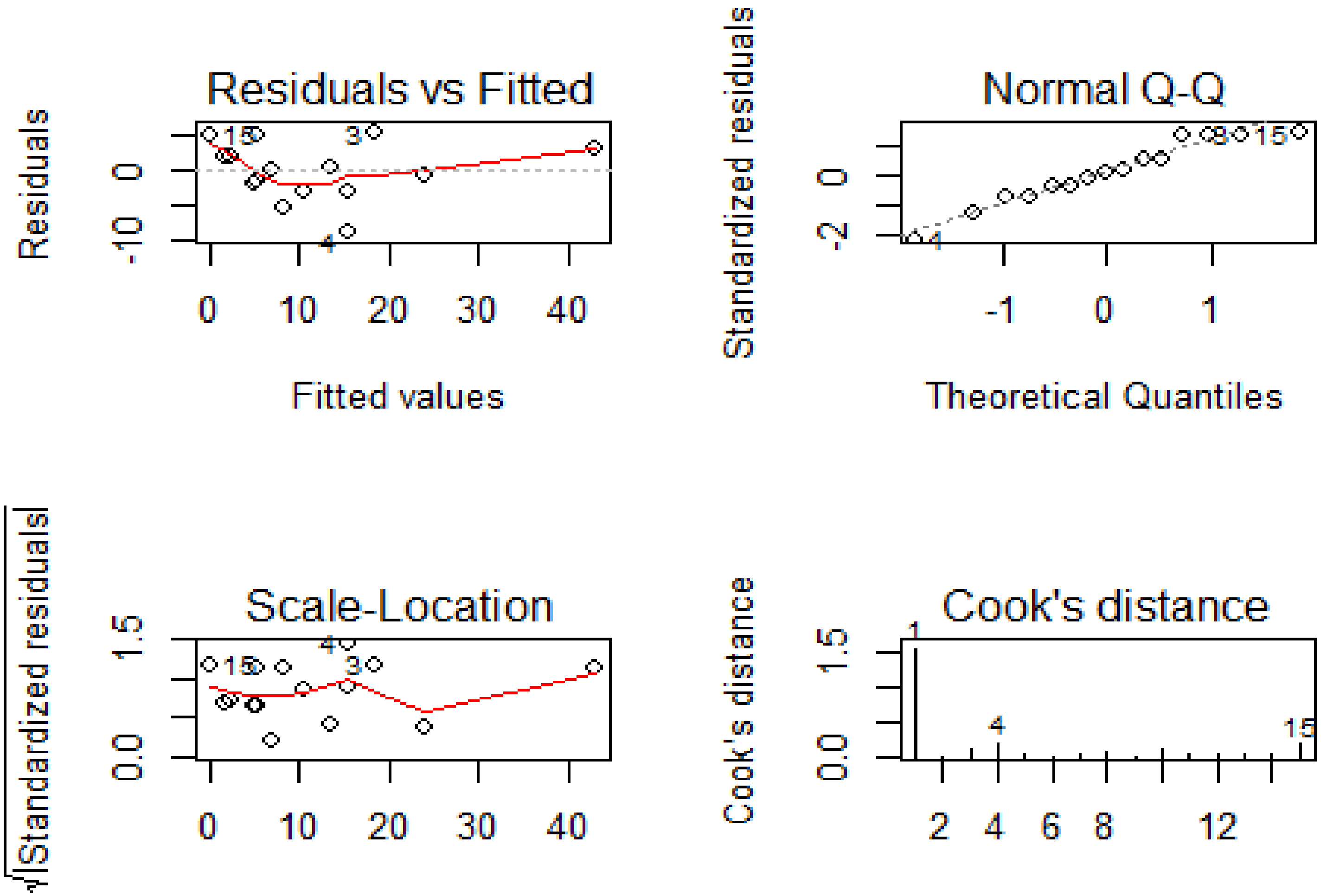
Goodness-of-fit

- Goodness-of-fit of a statistical model **describes how well it fits a set of observations**
- R^2 can be a measure
- Model significance and goodness-of-fit
  - Insignificant model can have good fit (large R^2)
  - Significant model can have lack of fit (small R^2)
- In general, we need a "good fit model" for prediction purpose
- Small R^2 but significant model can be still useful
  - E.g., tobacco use vs. lung cancer

Residuals

- Residuals: $Y - \hat{Y}$
- Able to check **Normality assumption** through residuals
- Plot against predictors or fitted value to see trends to check **homoscedasticity assumption**
- Check suspicious outliers

Diagnostics plots



Diagnostic plots; When model assumptions are valid

- **Residual vs. Fitted values**
  - Distributed within (0, 1.5) band on sqrt(standardized residuals)
    - *Default plot in R*
  - Distributed within (-2,+2) band on standardized residuals
    - *Manual generation in R (see class code)*
  - Check normality assumption
  - No pattern found
    - check homogeneity and linear assumptions
- **Normal QQ plot**
  - almost Straight line
    - checks normality of errors

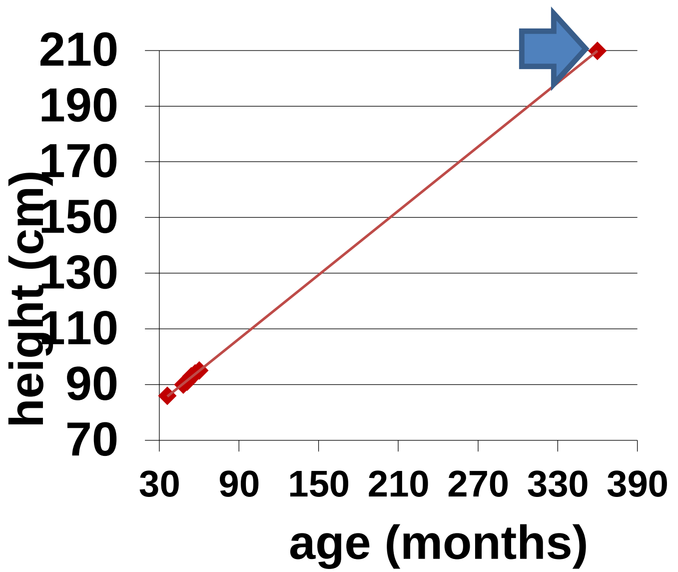
- Rigorous Normality test (shapiro test) can be applied to residuals but, in practice, generally check diagnostics plots
- Others like **cook'sD** ; not for assumption but for data

Influence Diagnostics

- Leverage / Outlier
- Cook's distances – influence of individual data points on the fitting cooks.distance()
- DFFITS – influence of individual data points on the predicted values
- DFBETAS – influence of individual data points on the parameter estimates

Caution

- Be aware of Extrapolation



- Correlation does not imply Causation
  - Strong correlations do not correspond to a causal relationship (change in X causes change in Y)

Example 1: Cirrhosis and Alcohol

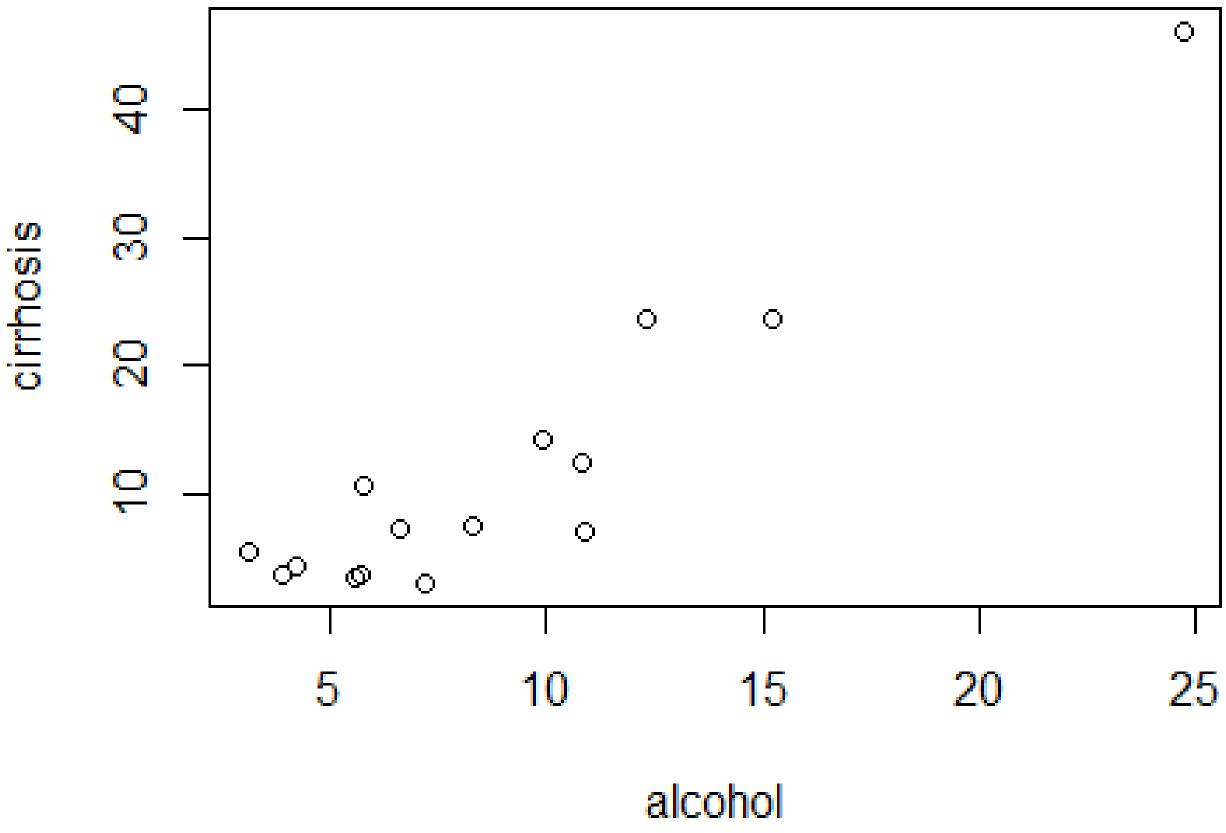
Data set:

- Data from 15 countries
- Cirrhosis deaths per 100,000 people
- Annual alcohol consumption (in litres per person per year)

Example 1: Linear trend?

- Considering cirrhosis deaths as a function of alcohol consumption
- Create a scatter plot of the data
- Linear trend reasonable?
- Any indications of problems with the data?

Example 1: Linear trend?



Exercise 1: Linear regression

- Fit linear regression model with cirrhosis deaths as response
- Comment on quality of model
- Comment on any problems noticed in diagnostics
- Relationship between alcohol consumption and cirrhosis related death rate?

R output

$$\hat{y} = -5.99 + 1.97X$$

```
##
## Call:
## lm(formula = cirrhosis ~ alcohol, data = drinking)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5635 -2.3508  0.1415  2.6149  5.3674
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.9958     2.0977  -2.858   0.0134 *
## alcohol       1.9779     0.2012   9.829  2.2e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1
##
## Residual standard error: 4.17 on 13 degrees of freedom
## Multiple R-squared:  0.8814, Adjusted R-squared:  0.8723
## F-statistic: 96.61 on 1 and 13 DF,  p-value: 2.197e-07
```

t-tset on individual term

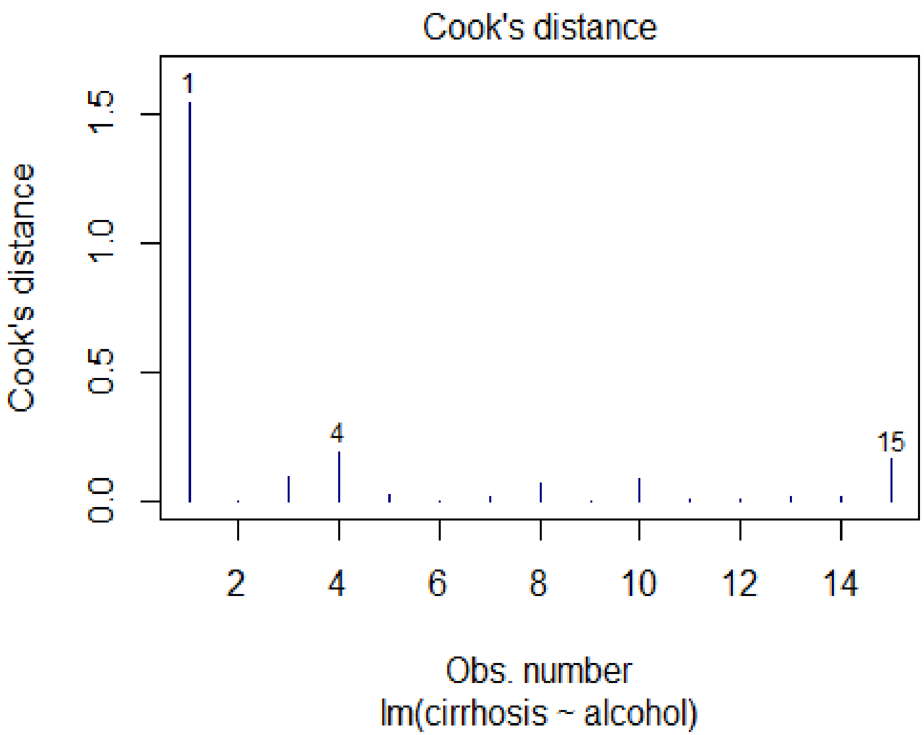
H0: all β's are zero vs. Ha: at least one β nonzero
(Model is not useful) (Model is useful)

NOTE- we do not test intercept. β's in null do not include intercept term

Exercise 1: Undue influence

- Points too influential based on Cook’s distance?
- Remove points with Cook’s distance greater than 1 and refit the model
- How do the results change?
- Any remaining problems with the model?

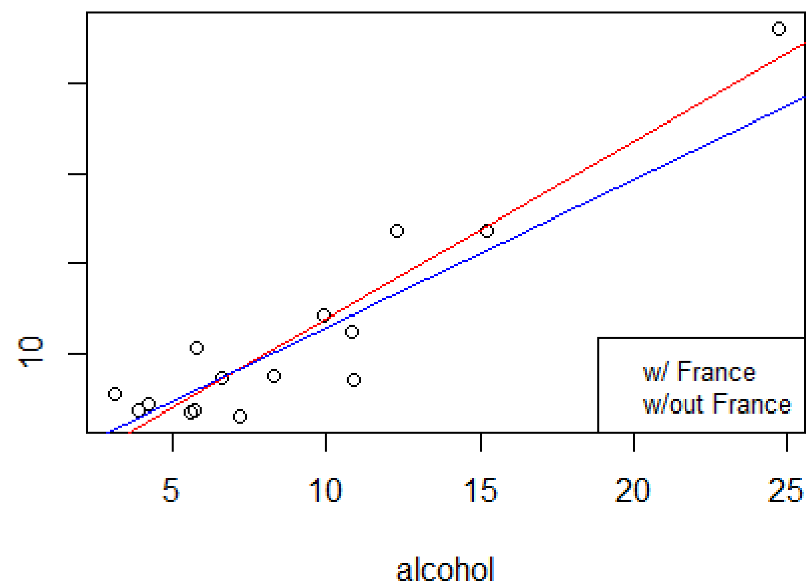
Exercise 1: Undue influence



Obs1= France

Red line: estimated regression line with France

Blue line: estimated regression line without France



Example 2: crime data

- Using **crime.csv**
- We want to check if there is a relationship between **crime numbers** and **rate of poverty**.

```
head(crime)

##    sid state crime murder pctmetro pctwhite pcths poverty single
## 1    1    ak   761    9.0     41.8     75.2  86.6     9.1   14.3
## 2    2    al   780   11.6     67.4     73.5  66.9    17.4   11.5
## 3    3    ar   593   10.2     44.7     82.9  66.3    20.0   10.7
## 4    4    az   715    8.6     84.7     88.6  78.7    15.4   12.1
## 5    5    ca  1078   13.1     96.7     79.3  76.2    18.2   12.5
## 6    6    co   567    5.8     81.8     92.5  84.4     9.9   12.1
```

Example 2: crime data

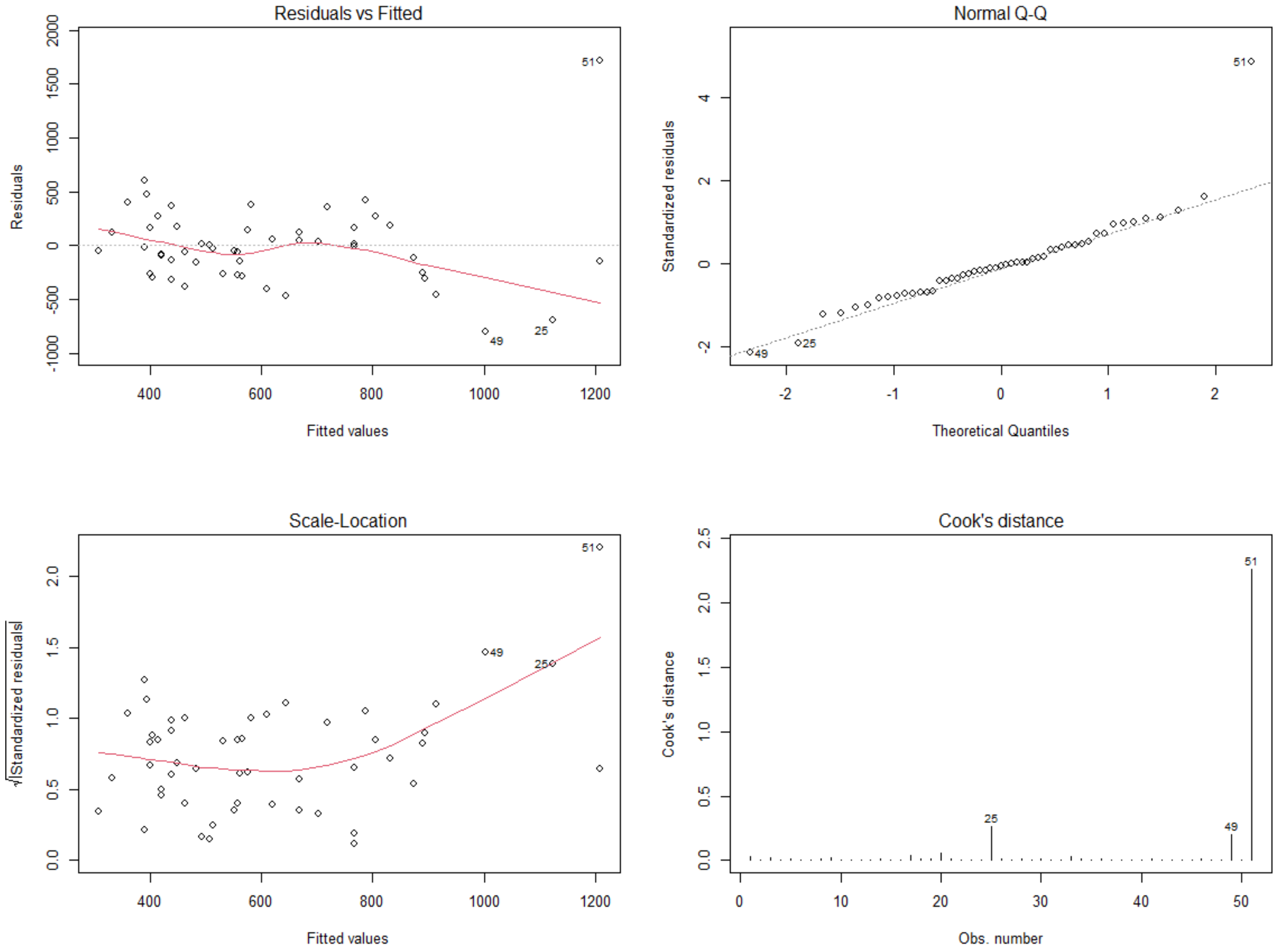
```
lm.crime <- lm(crime ~ poverty, data=crime)
summary(lm.crime)
```

```
Call:
lm(formula = crime ~ poverty, data = crime)

Residuals:
   Min       1Q  Median      3Q     Max
-794.16 -257.78  -22.91  165.50 1713.93
•
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -86.20     176.99  -0.487 0.628403
poverty        49.03      11.83   4.145 0.000134 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 383.4 on 49 degrees of freedom
Multiple R-squared:  0.2596,	Adjusted R-squared:  0.2445
F-statistic: 17.18 on 1 and 49 DF,  p-value: 0.0001342
```

Example 2: crime data

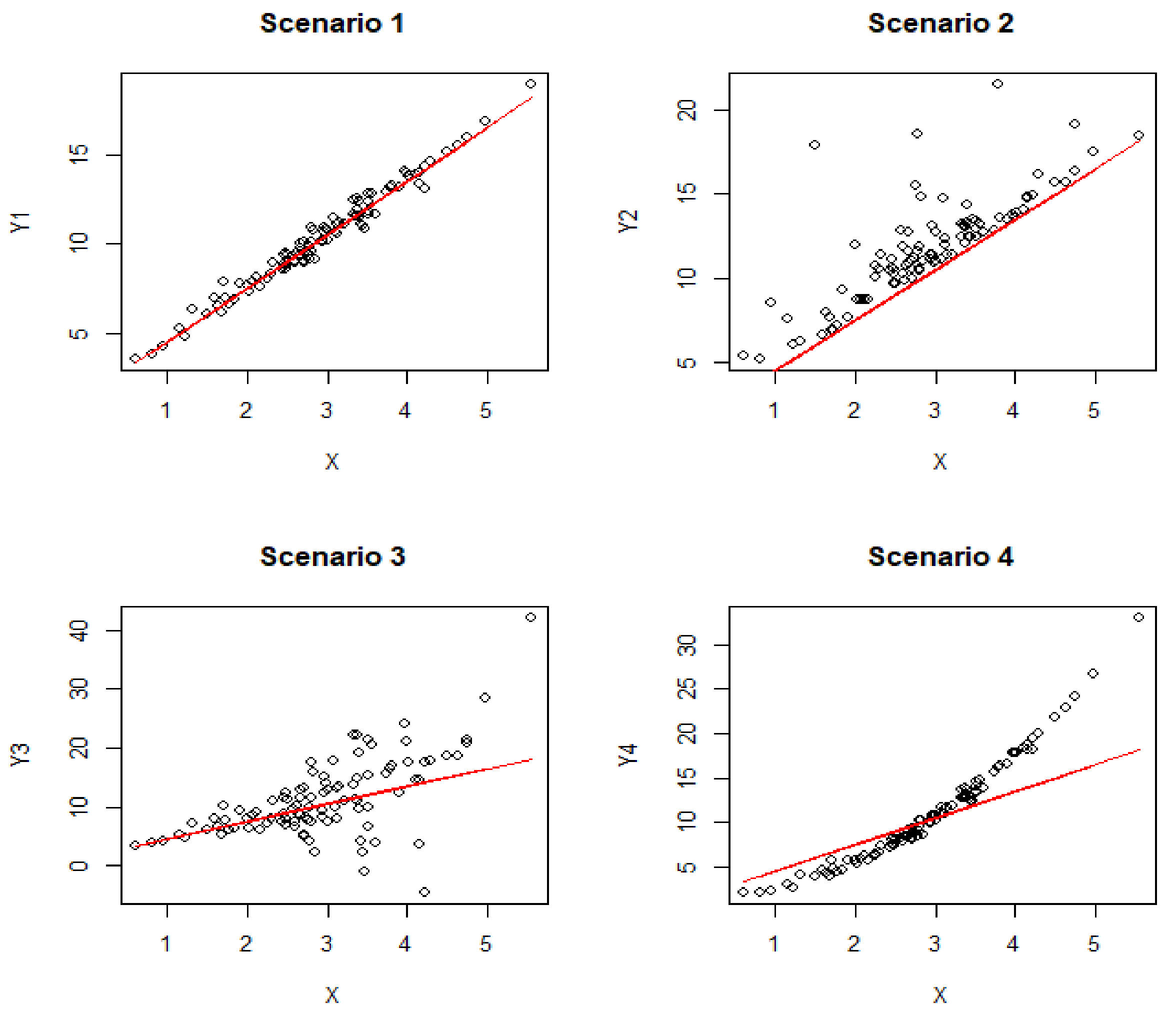


Simulation study:
What happens if assumption is violated?

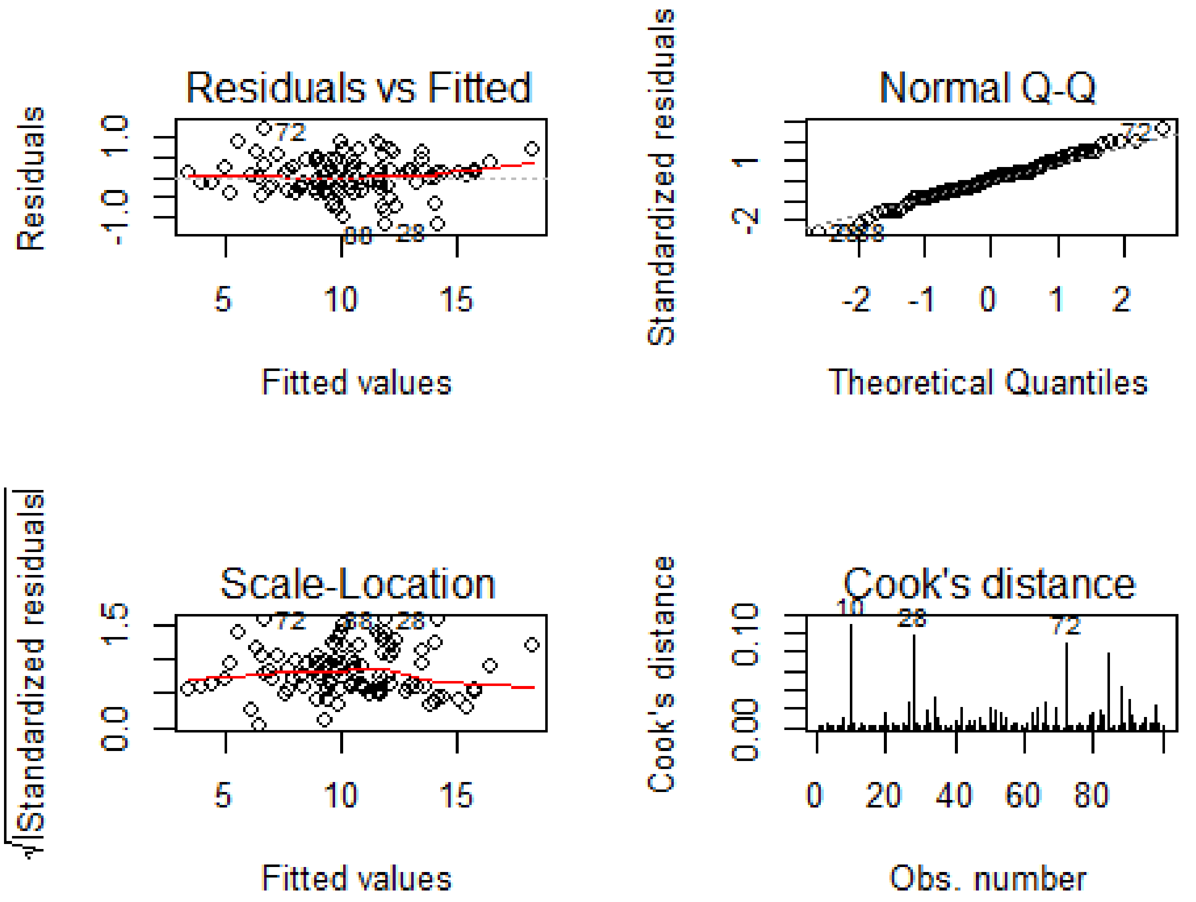
- SimData.csv (simulated data in R)
- Y1=1.5+3*x1+**e**
- Y2=1.5+3*x1+**e***
- Y3=1.5+3*x1+**e’**
- Y4=1.5+3***X1^2**+**e**
  - **e**: normal error
  - **e***: log-normal error (highly right skewed)
  - **e’**: heterogeneous errors proportional to X1

- Which assumption is violated at Y2-Y4?
- <u>How diagnostic plot looks like</u>?

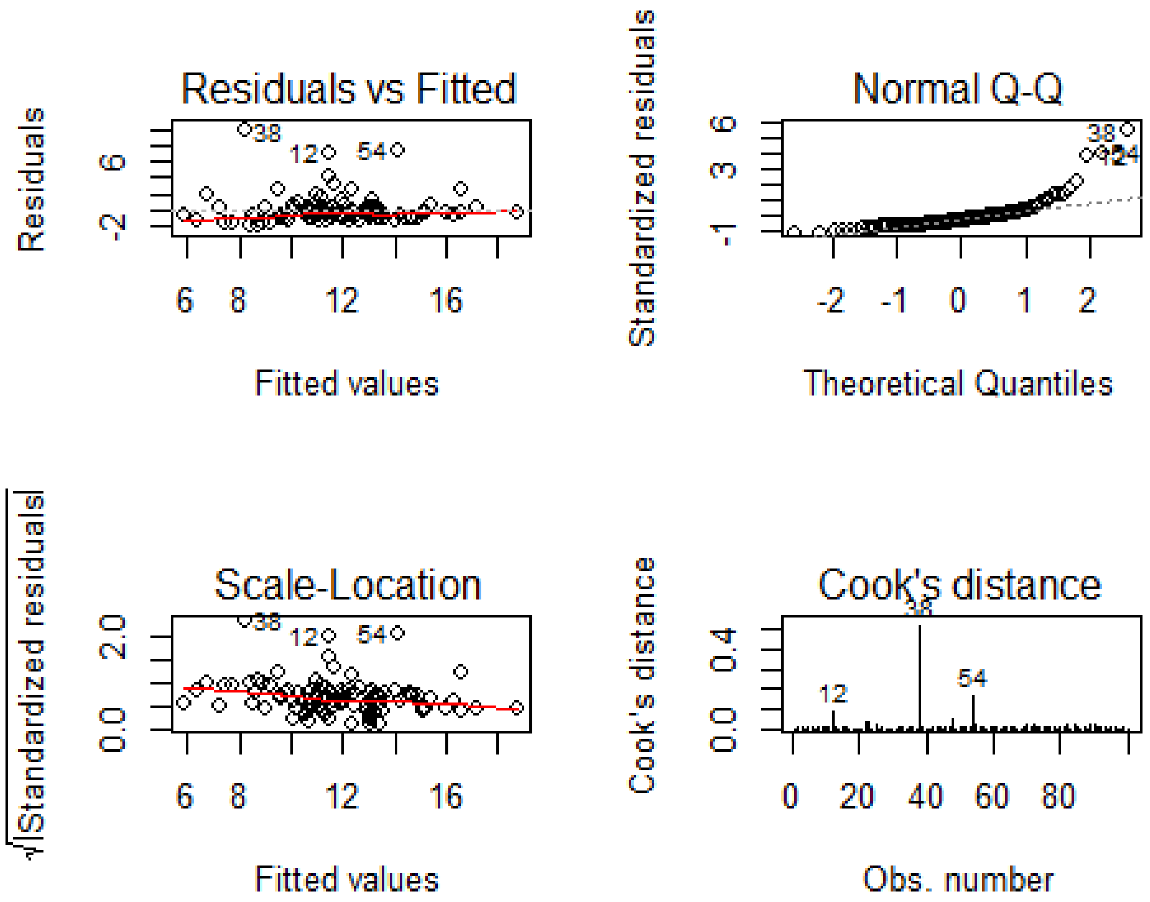
Simulation study:



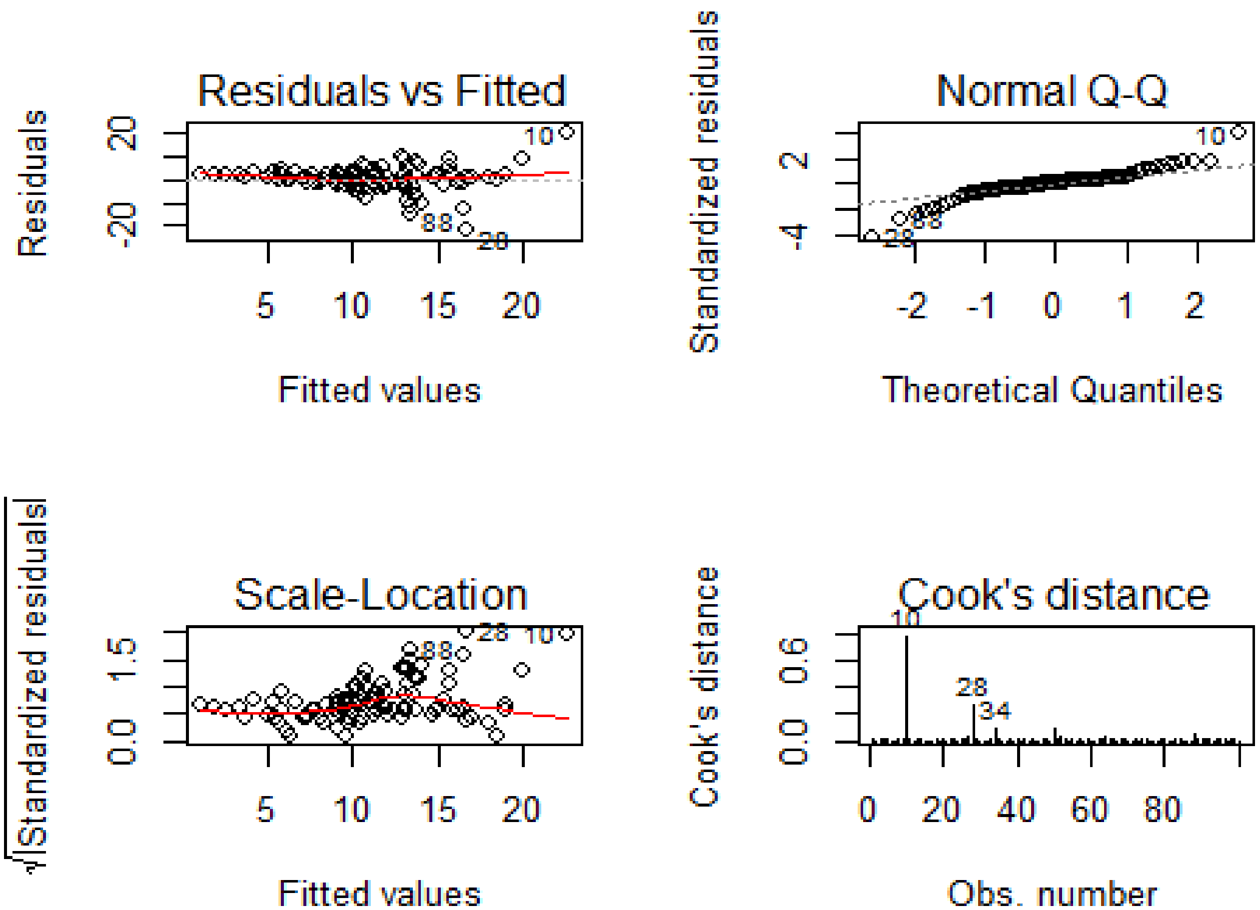
Simulation study: Diagnostics plots of scenario 1



Simulation study: Diagnostics plots of scenario 2



Simulation study: Diagnostics plots of scenario 3



Simulation study: Diagnostics plots of scenario 4