Topic 4:
Normal Linear Regression

Recap: t-Test

lm(mass ~ region)

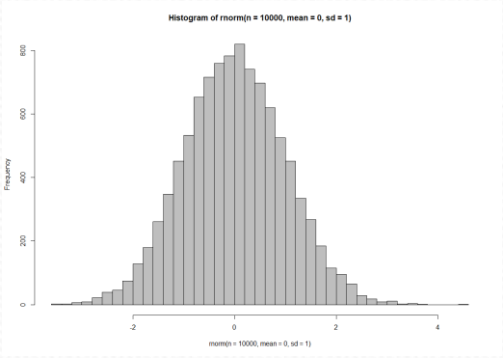
$$\mathbf{mass}_i = \boldsymbol{\alpha} + \boldsymbol{\beta} * \mathbf{region}_i + \boldsymbol{\varepsilon_i}$$

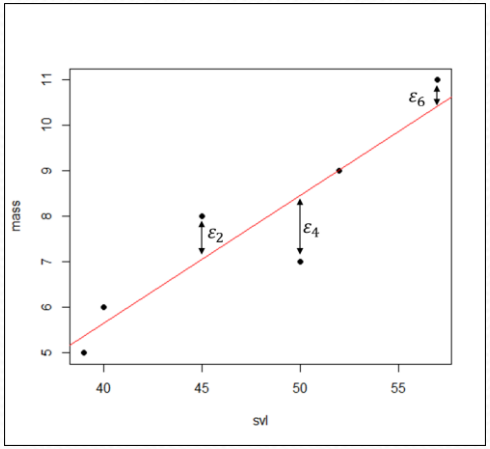
α = constant

β = constant to be multiplied with region indicator

ε_i = residual for snake i

$\varepsilon_i \sim$ Normal$(0, \sigma^2)$





Recap: (simple) normal linear regression

lm(mass ~ svl)

$$\mathbf{mass}_i = \boldsymbol{\alpha} + \boldsymbol{\beta} * \mathbf{svl}_i + \boldsymbol{\varepsilon}_i$$

α = constant

β = constant to be multiplied with snout–vent leng.

ε_i = residual for snake i

$\varepsilon_i \sim$ Normal(o,σ^2)

Only difference: the variable **svl** does not just take on two possible values to indicate group membership; rather, it is a measurement that can take on any possible value.

Simple/multiple linear regression

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{im} + \varepsilon_i$$

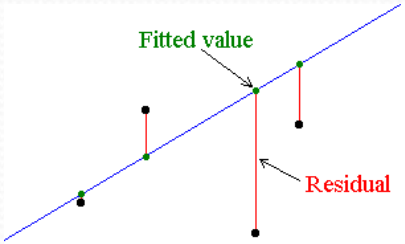
and $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, iid

A relationship between variables involving:

- A *variable of interest* Y_i *(dependent variable, response variable);* hard to measure
- "easy to measure" variables (also called *predictor*, *independent*, or *explanatory variables*) - related to the variable of interest and here labeled x_{i1} , x_{i2},.....x_{im}

- a **single error term**; a number of **assumptions** about the error term have been met

Normal linear regression



How to solve the equation?

- **Ordinary Least Squares, OLS** (linear least squares) a method for estimating the unknown parameters in a linear regression model
- OLS chooses the parameters of a linear function of a set of explanatory variables by minimizing the **sum of the squares** of the residuals

- Linear model does not imply a straight line!
  E.g. a quadratic equation produces a curved line but the relationship between response and explanatory variables is still linear.

Model validation

Other key words: assumptions, residual analysis, diagnostics

Before you trust your model, you have to examine whether the assumptions on which the model is based are valid!

If the assumptions are not satisfied, then your model will be unreliable! (at least to some extent)

Key Assumptions

Assumptions are about the error term -> Residual analysis

Residuals are independently, identically distributed (iid)

1. The mean of the residuals is (close to) zero along the regression line, i.e. there is a linear relationship between dependent and independent variable

2. The residuals are normally distributed

3. The variances of the residuals are homogenous (homoscedasticity)

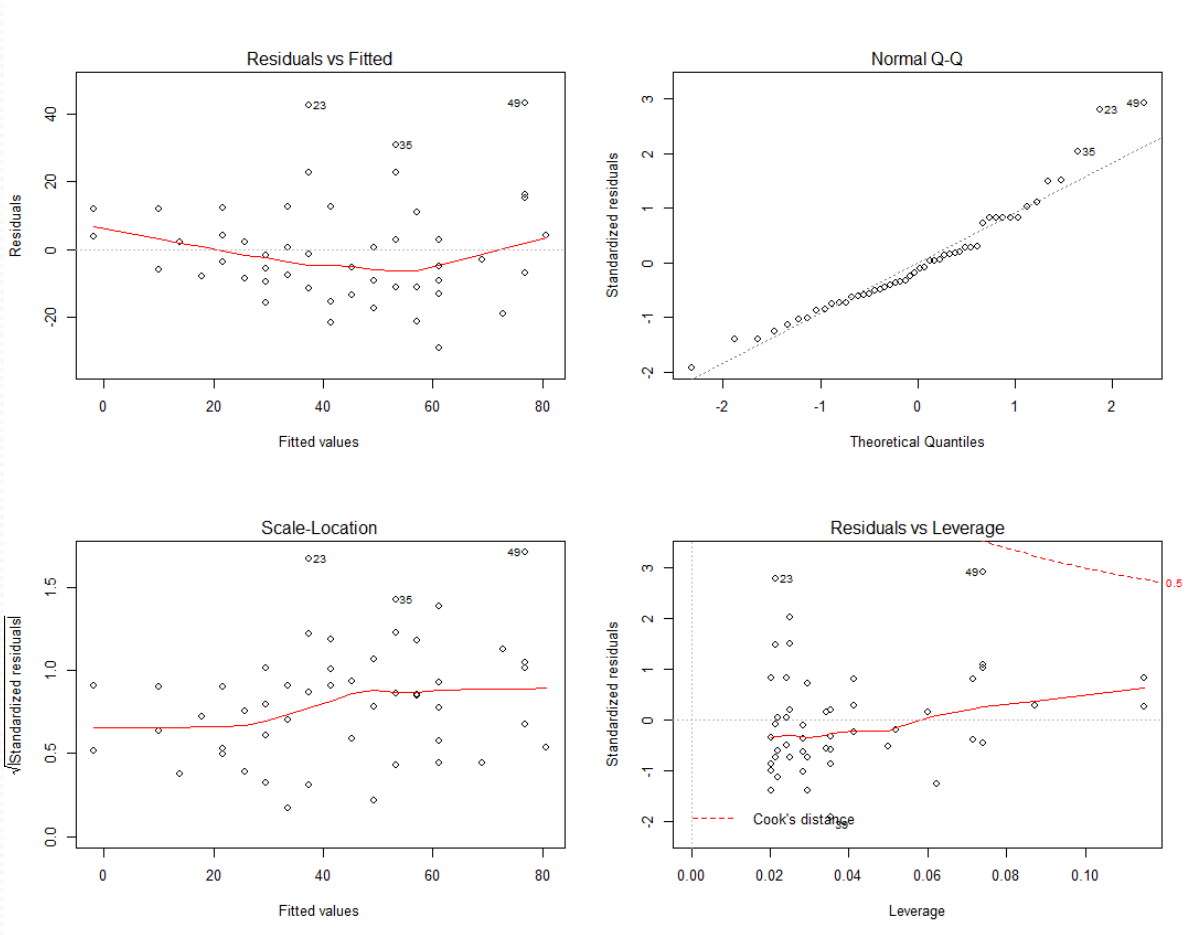
4. The model is not biased by unduly influencial observations

5. The independent variables are independent of each other (no collinearity)

6. The dataset does not contain serial auto-correlation

Residual analysis with diagnostic plots

R function for lm and glm object: plot(model)

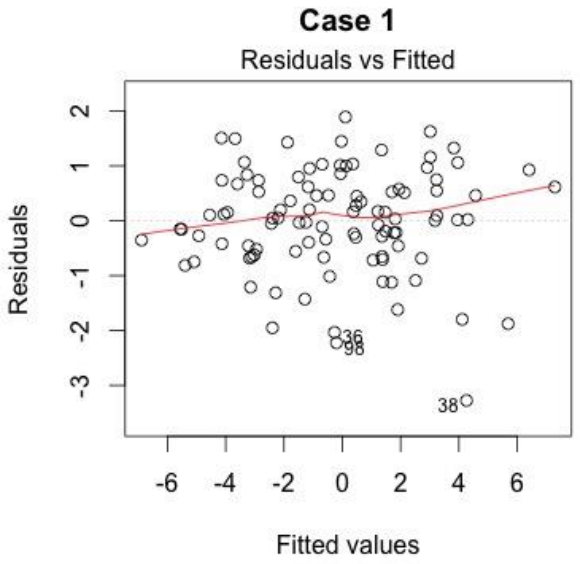


1. Residuals versus fitted values

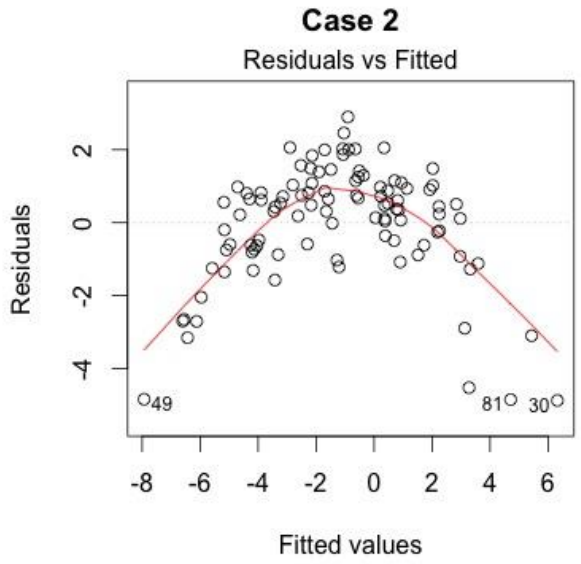
We want to see…

- no pattern and the red line is around zero (Patterns indicate non-linear relationship)
- equally spreaded residuals around a horizontal line without distinct patterns

Good model



Bad model



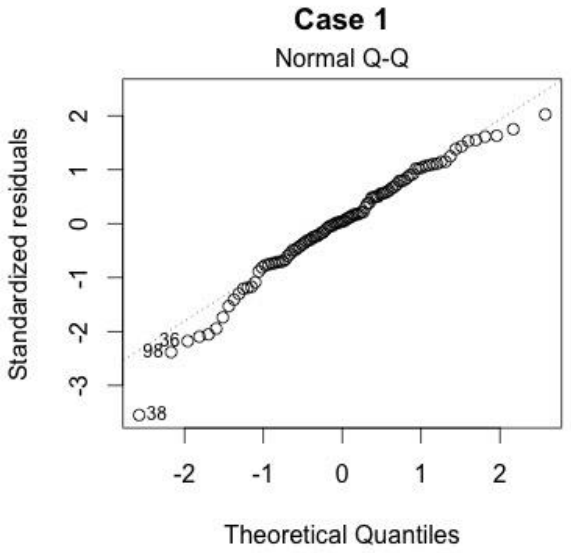
http://data.library.virginia.edu/diagnostic-plots/

2. Normal Q-Q plot

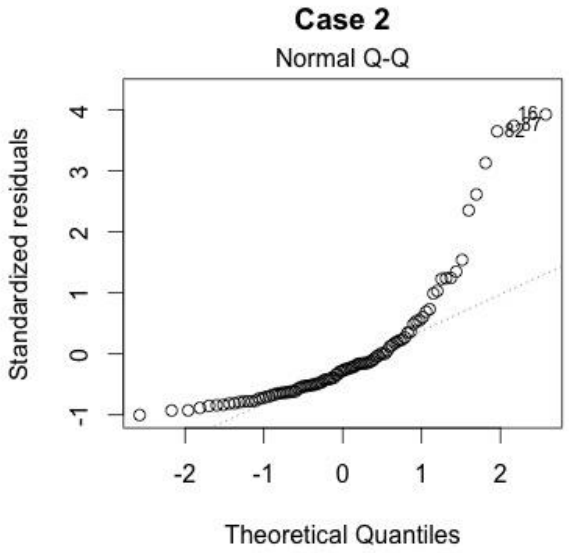
Are residuals normally distributed?

- Do residuals follow a straight line or do they deviate severely?
- It's good if residuals are lined well on the straight dashed line

Good model
But #38 might be a Problem!



Bad model



http://data.library.virginia.edu/diagnostic-plots/

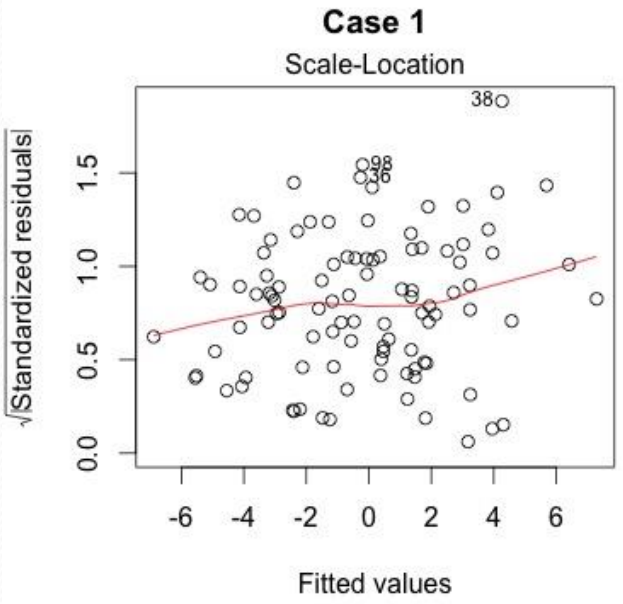
3. Scale-Location plot

Are residuals spread equally along the ranges of predictors?

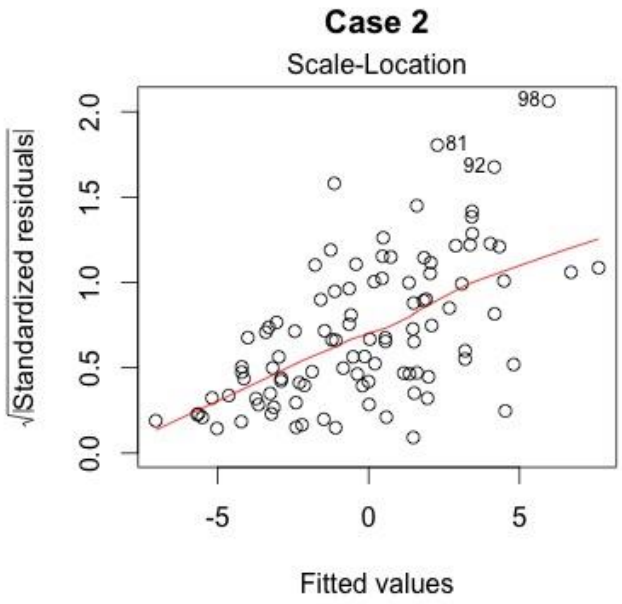
Checks the assumption of equal variance (homoscedasticity).

It's good if you see a horizontal line with equally (randomly) spread points.

Good model



Bad model

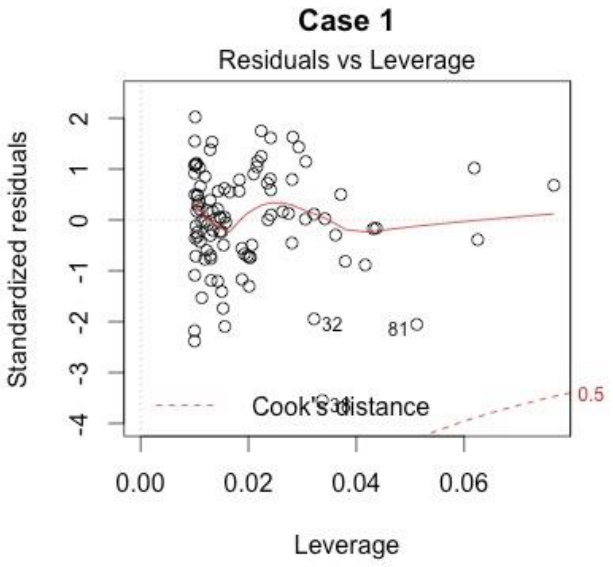


http://data.library.virginia.edu/diagnostic-plots/

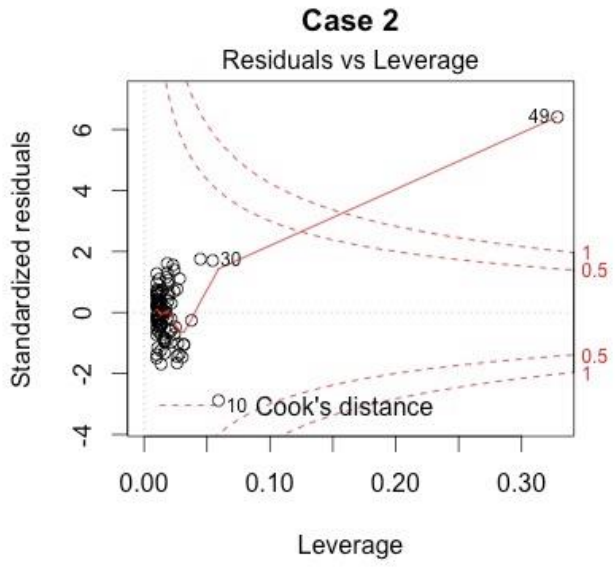
4. Residuals versus Leverage

- helps to find influential cases (not all outliers are influential in linear regression analysis) - Here, patterns are not relevant
- Watch out for outlying values at the upper right corner or at the lower right corner. -> high Cook's distance scores: cases are outside of a dashed line. Such cases are influential to the regression results, which will be altered if we exclude those cases.

Good model

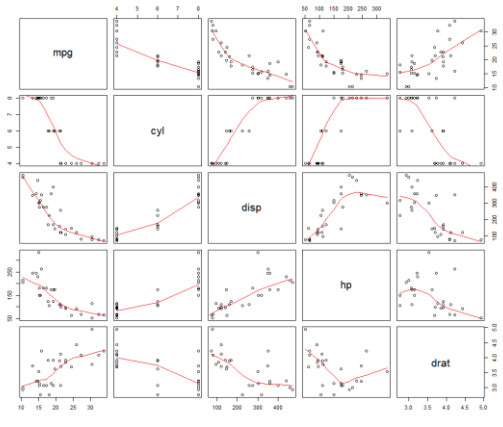


Bad model



http://data.library.virginia.edu/diagnostic-plots/

5. Check for Collinearity



Collinearity is where two independent variables are strongly correlated with each other

Should be checked aready before running a model!

- Graphically: plot(dataframe, panel=panel.smooth)
  pairs(dataframe)
- Pearson correlation coefficient: cor(dataframe)
  however unclear when is correlation too large, e.g. r > 0.6 or 0.8?

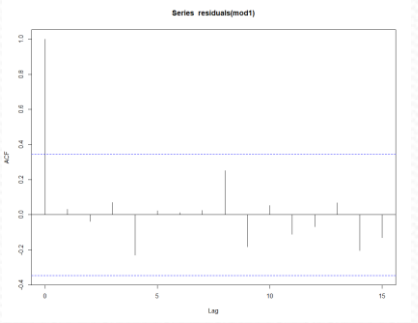
After running a model: Variance Inflation Factor (VIF)
How much variation is already explained by other variables?

- library(car)
  vif(model)

VIF should be < 4, the smaller the better

6. Autocorrelation

The correlation between successive datapoints in the dataset (or residuals in the model) - especially applicable for time series



Graphical check: Auto-Correlation Function (ACF)

acf(residuals(model))

acf(variable)

X axis = lags of residuals/variable, increasing in steps of 1; first line left is the correlation of residual with itself (Lag 0)

When ACF score for a lag value exceeds horizontal dotted lines -> significant auto-correlation at that time-lag

Statistical test: durbinWatsonTest(model) from library(car)
p-value < 0.05 indicates significant auto-correlation

Interpreting model coefficients

```
> summary(mod_cars)

Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Example with dataset cars: Can „dist“ (Stopping distance) be predicted by „speed“?

Interpreting model coefficients

```
> summary(mod_cars)

Call:
lm(formula = dist ~ speed, data = cars)
```

- Output in R from function summary() after using e.g. lm()
- Formular call - the formula R used to fit the data.

Interpreting model coefficients

```
Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201
```

- Rough summary of residual distribution
- First impression whether residuals are normally distributed

Interpreting model coefficients

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- $\mathbf{y}_i = \boldsymbol{\alpha} + \boldsymbol{\beta} * \mathbf{x}_i + \boldsymbol{\varepsilon}_i$
- „**Estimate**“ = estimate for the intercept $\boldsymbol{\alpha}$ and the slope $\boldsymbol{\beta} * \mathbf{x}_i$
- „**Std.Error**“ = estimated standard errors of the estimates
- The last two columns contain **t** and **p-value** for the classical t-test for the null hypothesis that the estimate equals zero.

Here: The Intercept is physically nonsense; it refers to a stopping distance of -17.6 feet at a speed of zero. This illustrates that a linear model has often only a useful characterization of a relationship over a restricted range of the explanatory variables. Hereby important is the slope: With a 1 mph increase in the speed of a car, the required distance to stop goes up by **3.9324** feet, which can vary by **0.4155128** feet.

Interpreting model coefficients

```
Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Residual Standard Error is a measure of the *quality* of a linear regression fit.

Average amount that the response (dist) will deviate from the true regression line.

Here: the actual distance required to stop can deviate from the true regression line by approximately **15.38** feet, on average.

Degrees of freedom are the number of data points that went into the estimation of the parameters (after taking into account these parameters). Here: 50 data points and two parameters (intercept and slope).

Interpreting model coefficients

```
Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

- The **R-squared statistic** provides a measure of how well the model is fitting the actual data.
- R^2 is a measure (0-1) of the linear relationship between our predictor variable (speed) and our response variable (dist).

Here: Roughly 65% of the variance found in the response variable (dist) can be explained by the predictor variable (speed).

- In multiple regressions: R^2 will always increase as more variables are included in the model. **Use the adjusted R^2!**

Interpreting model coefficients

```
Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

- **F-statistic** is a good indicator of whether there is a relationship between our predictor and the response variables. The further the F-statistic is from 1 the better it is.
- However, how much larger the F-statistic needs to be depends on both the number of data points and the number of predictors. E.g. in a large dataset the F-statistic only needs to be a little larger than 1 to reject the null hypothesis.