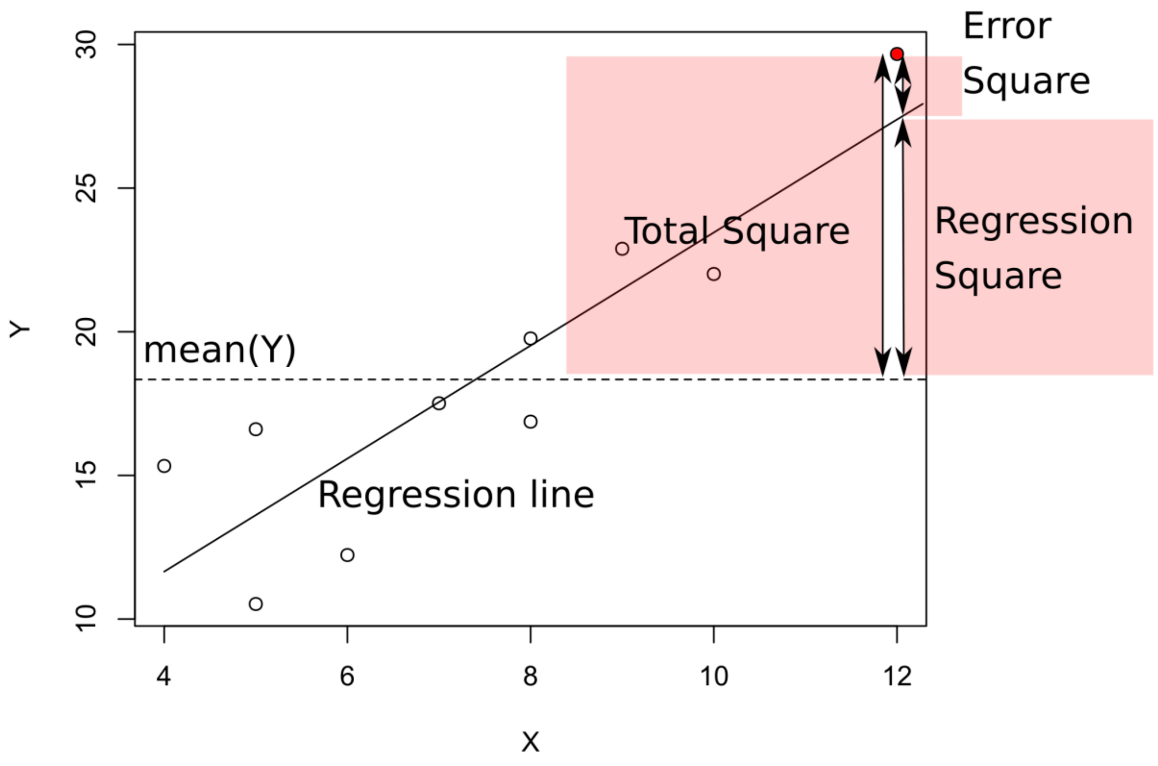
Chapter 9

Linear regression

Correlation

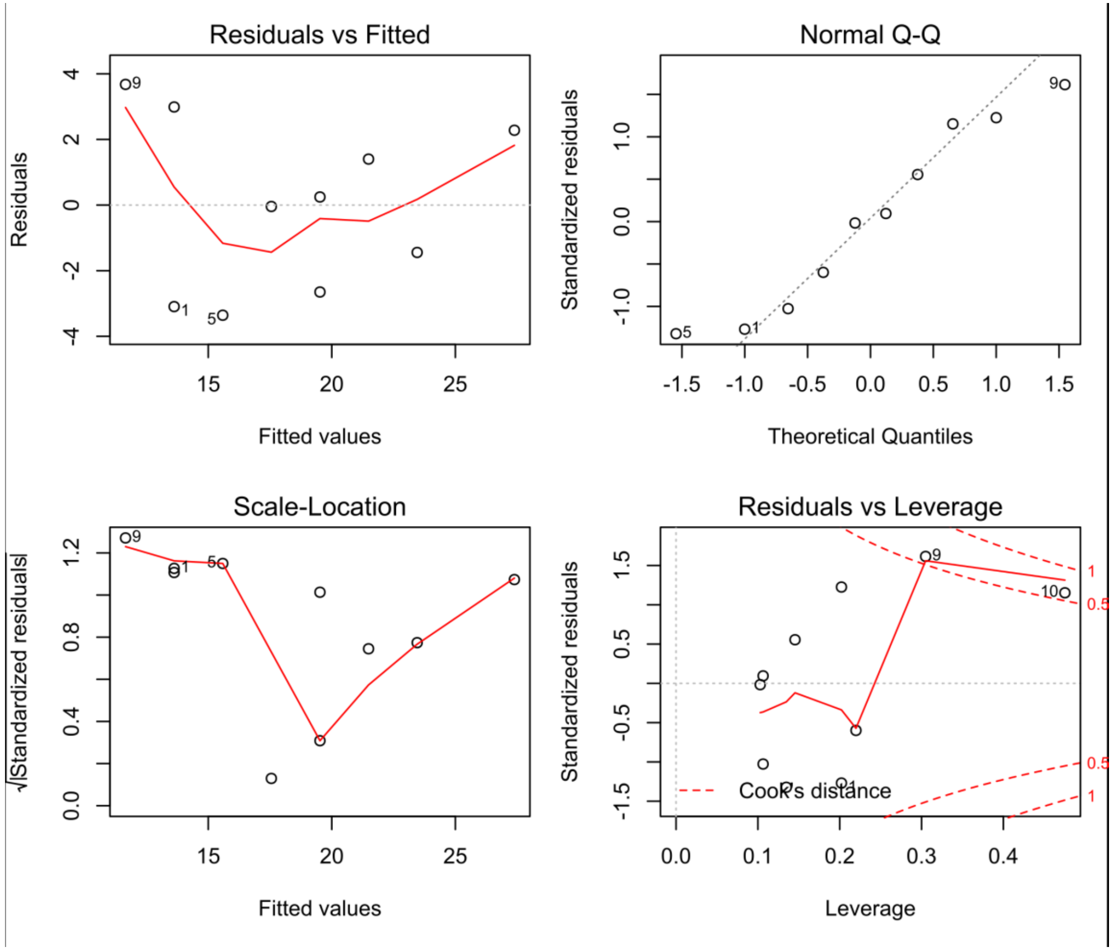
Linear regression

- Describes asymetric dependence of two quantitative variables
  - Response variable depends on predictor(s)
- General formula: Y = a + bX + ε
  - Y = response
  - X = predictor
  - **a = intercept**
  - **b = slope**
  - ε = residuals (error)
- Decomposition of total Sum Sq. into Regression Sum Sq. and Error Sum. Sq. as in ANOVA
- Significance testing by F-test
  - $F_{DFregr,DFerror} = MS_{regr} / MS_{error}$
- DFregr = number of predictors (1 in simple regression)
- Dferror = number of observations – DFregr – 1
- Coefficient of determination R^2 = SSregr/SStotal
- Adjusted $R^2 = 1 - MS_{error}/MS_{total}$
  - Accounts for the estimate nature of the R^2



Regression assumptions

- Normality of residuals
- Indepdendence between residuals and fitted values
- Linear relationship between X and Y
- Check by Regression diagnostics



Correlation

- Symmetric association between two quantitative variables
- Pearson correlation coefficient: $r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$
  - $r > 0$ (max. 1): positive correlation
  - $r < 0$ (min. -1): negative correlation
  - $r = 0$: independence
- Can be tested for significance (i.e. difference from 0) by a single sample *t*-test with DF = n – 2
- r^2 = proportion of shared variability = regression R^2

Correlation and causality

- Causality = if X changes, Y also changes
- Correlation = association between two variables
  - A change caused by a manipulation in one does not imply a necessary change in the other
- Associations are mostly analyzed by regression
  - Numerical equivalence between correlation and regression
  - Significant regression does not mean causality
- Causality can only be demonstrated by **manipulative experiments**!