Linear regression assumptions

- Linear relationship between the independent and the dependent variable
  - Usually independent variable is continuous (scale)
  - Check for outliers, Use scatterplots
- Multivariate normality - normally distributed data
  - Use histogram, a fitted normal curve or a Q-Q-Plot → if normal, points should be close to the diagonal reference line
- No or little multicollinearity - independence of independent variables
  - Correlation matrix, Variation Inflation Factor (VIF)
  - Correlation is high when < 0.6
- Homoscedasticity - error terms along the regression are equal, constant variance of the errors → Scatterplot
- No auto-correlation - residuals should be independent from each other
  - A residual is a measure of how well a line fits an individual data point → it is the vertical distance of data points from the regression line - fits better when close to 0
- Minimum of 20 cases recommended

(Binomial) logistic regression

- Dependent variable - dichotomous, binary, with two outcomes
  - Logistic regression predicts the *probability* of independent variable taking a specific value → “success” over “failure”
  - Binomial distribution (Bernoulli is one form of this)
  - (With large enough data sample normal and binomial may *look* the same)
- One or more independent variables (continuous or categorical)
- Categories must be mutually exclusive
- Minimum of 50 cases recommended - need for maximum likelihood estimation (MLE)