Lesson 7. Conditions for a Simple Linear Regression Model – Part 1

1 The model

- The **simple linear regression model** is

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

2 Conditions for a simple linear regression model

- When is a simple linear regression model reasonable?
  - Are we justified in using our model? How much can we trust predictions that come from the model?
- We check for the following conditions:

1.

- The overall relationship between the variables has a linear pattern
- The average values of the response Y for each value of X fall on a common straight line

The rest of the conditions deal with the distribution of the errors:

2.

- The error distribution is centered at zero
- In other words, the points are scattered at random above and below the line
- Note: least squares regression forces the residual mean to be 0, but other techniques might not

3.

- The variability in the errors is the same for all values of the predictor variable
- In other words, the spread of the points around the line remains fairly constant

4.

- The errors are assumed to be independent from one another
- In other words, one point falling above or below the line has no influence on the location of another point

We need the next two conditions to provide CIs or perform hypothesis tests:

5.

- The data are obtained using a random process
- Most commonly, this arises either from random sampling from a population of interest, or from the use of randomization in a statistical experiment

6.

- In order to use standard distributions for CIs and hypothesis tests, we often need to assume that the random errors follow a Normal distribution

3 Standard deviation of the error term

- What are the unknown parameters in the simple linear regression model?

- For a simple linear regression model, the **estimated standard deviation of the error term** $\hat{\sigma}_\varepsilon$ based on the least squares fit to a sample of n observations is

- The value of $\hat{\sigma}_\varepsilon$ is sometimes called the **regression standard error** or **residual standard error**
  - It is interpreted as
  - It gives us a feel for how far individual cases might lie above or below the regression line

Example 1. Going back to the used Porsche example from Lesson 6...

a. Use the R output to calculate the regression standard error.
b. Using mileage to predict the price of a used Porsche, the typical error will be around what size? Your answer should have dollars as the units.

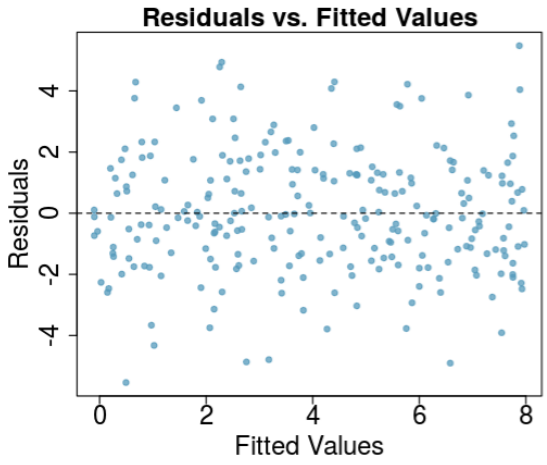
4 Assessing conditions for a simple linear regression model

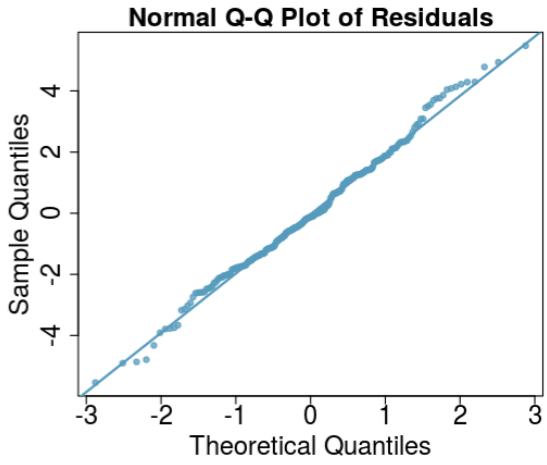
1. The condition that's been automatically met...

2. Think about how data was collected to check for ...

3. We check for linearity, constant variance, and normality using two diagnostic plots:

   - **Residuals versus fitted values plot**
     - This plot reorients axes so that the regression line is represented as a horizontal line through zero
     - Positive residuals are represented by points above the regression line
     - An ideal residual versus flitted values plot looks like
   - **Normal Q-Q plot of the residuals**
     - An ideal Normal Q-Q plot of the residuals looks like
       - The larger the sample size, the more lenient we can be about normality





No model is perfect, and linear regression is fairly robust to slight violations.
We will only be concerned with **blatant** violations.

Example 2. On the blank plots below, sketch the following:

a. A residuals vs. fitted values plot where linearity is met, but constant variance is violated.
b. A residuals vs. fitted values plot where constant variance is met, but linearity is violated.
c. A residuals vs. fitted values plot where both constant variance and linearity are violated.
d. A Normal Q-Q plot that shows dramatic violation of the Normality condition.