Regression Error

Grinnell College

Spring 2025

Review

- Regression models a linear relationship between response variable y and explanatory variable X of the form

$$y = \beta_0 + \beta_1 X + \epsilon$$

- Can expand this to include combinations of explanatory variables (quant. and cat.)

Error Terms

$$y = \beta_0 + X\beta_1 + \epsilon$$

Assumptions:

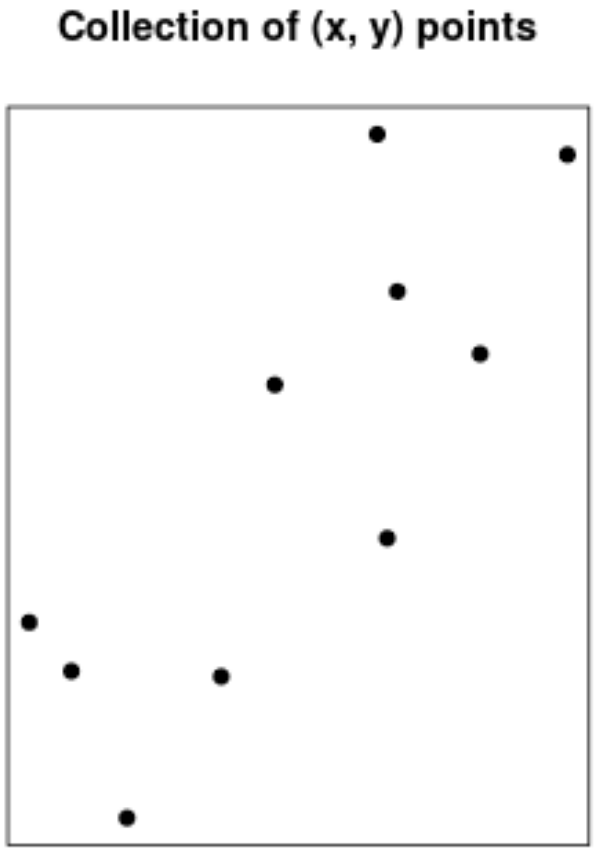
- Linear relationship between X and y
- Error term is normally distributed, $\epsilon \sim N(0, \sigma)$
- Error should be the same for all values of X, i.e., error same for all observations

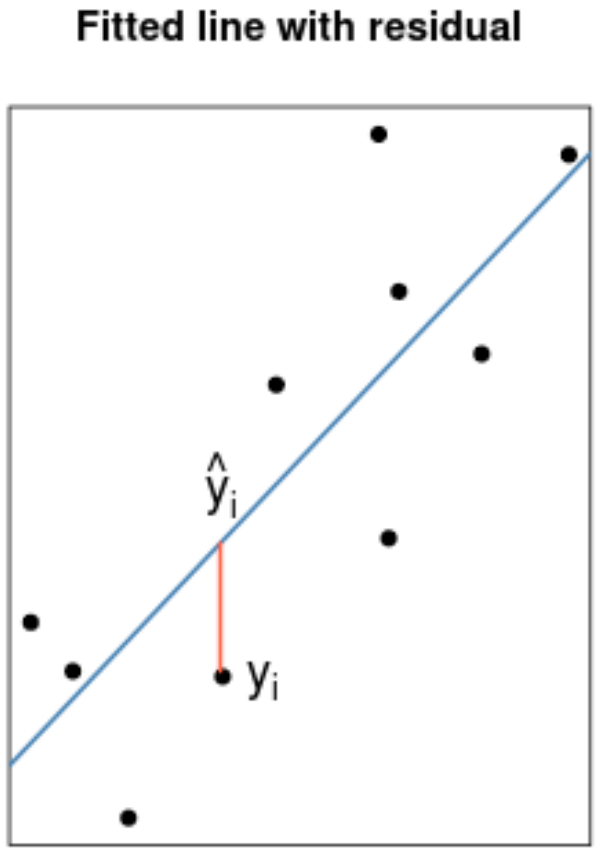
Analyzing the error terms gives us a way to test the assumptions of our model

Residuals

Visually, let's review what residuals look like

- residuals represent how far off our prediction is

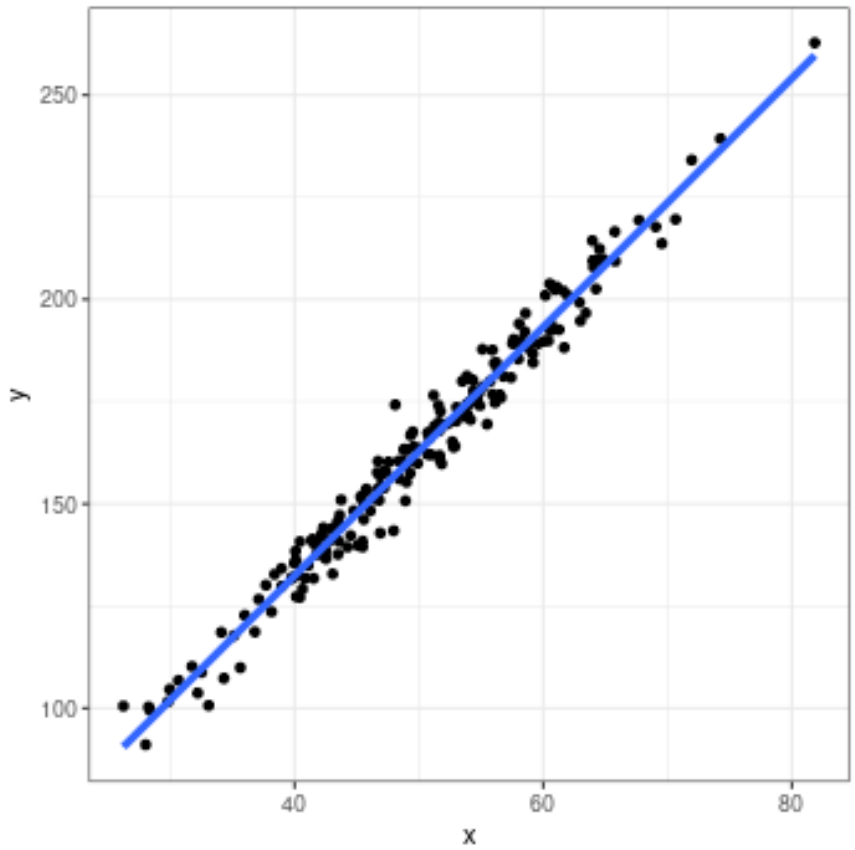




Residuals and assumptions

Three common ways to investigate residuals visually:

1. Plot histogram of residuals (normality)
2. Plot residuals against covariate (linear trend, changing variance)
3. Normal Quantile Plot (compares quantiles of residuals to quantiles of Normal distribution to see if they match)

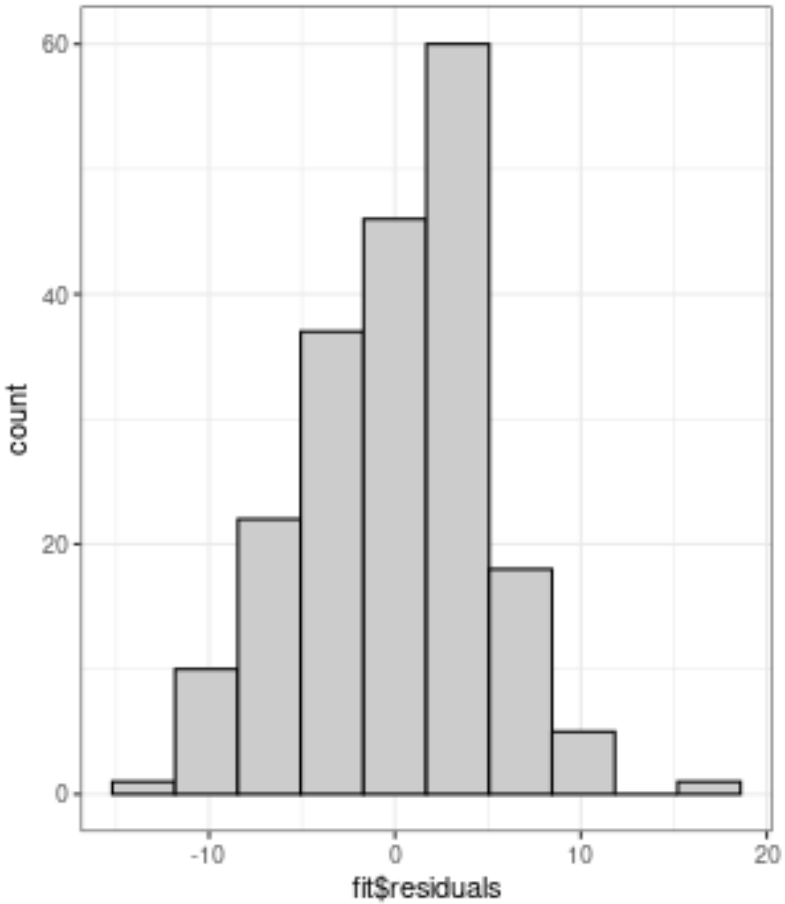


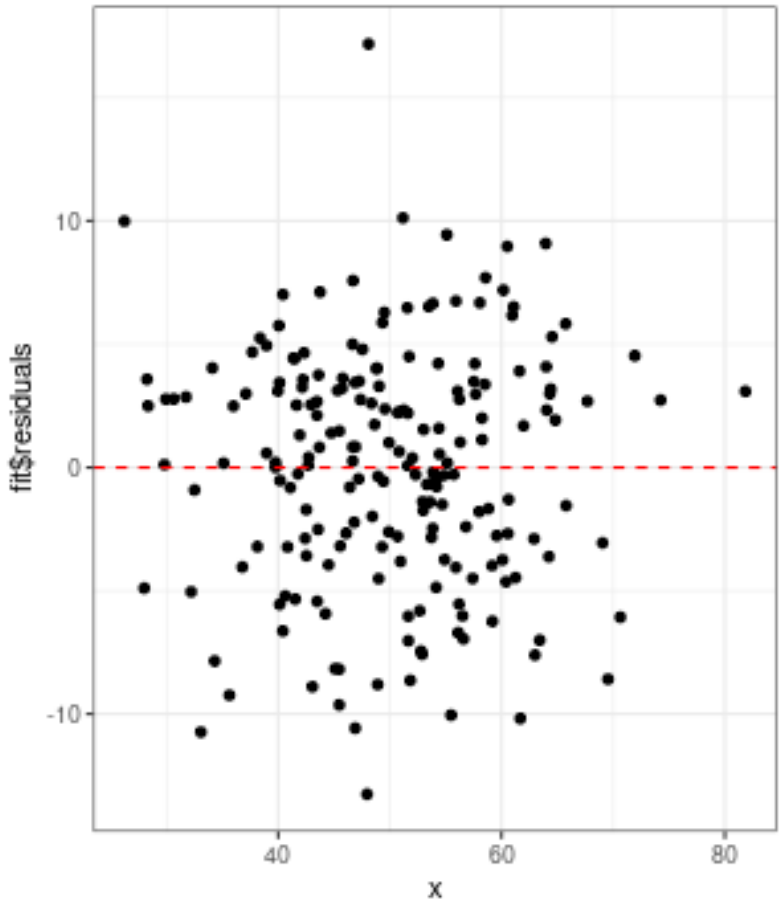
Checking Normality

Histogram of Residuals should be ≈Normal if our model is doing well

Residuals should not have a pattern other than 'blob of points' in a Resid. vs. Expl. Var. scatterplotplot

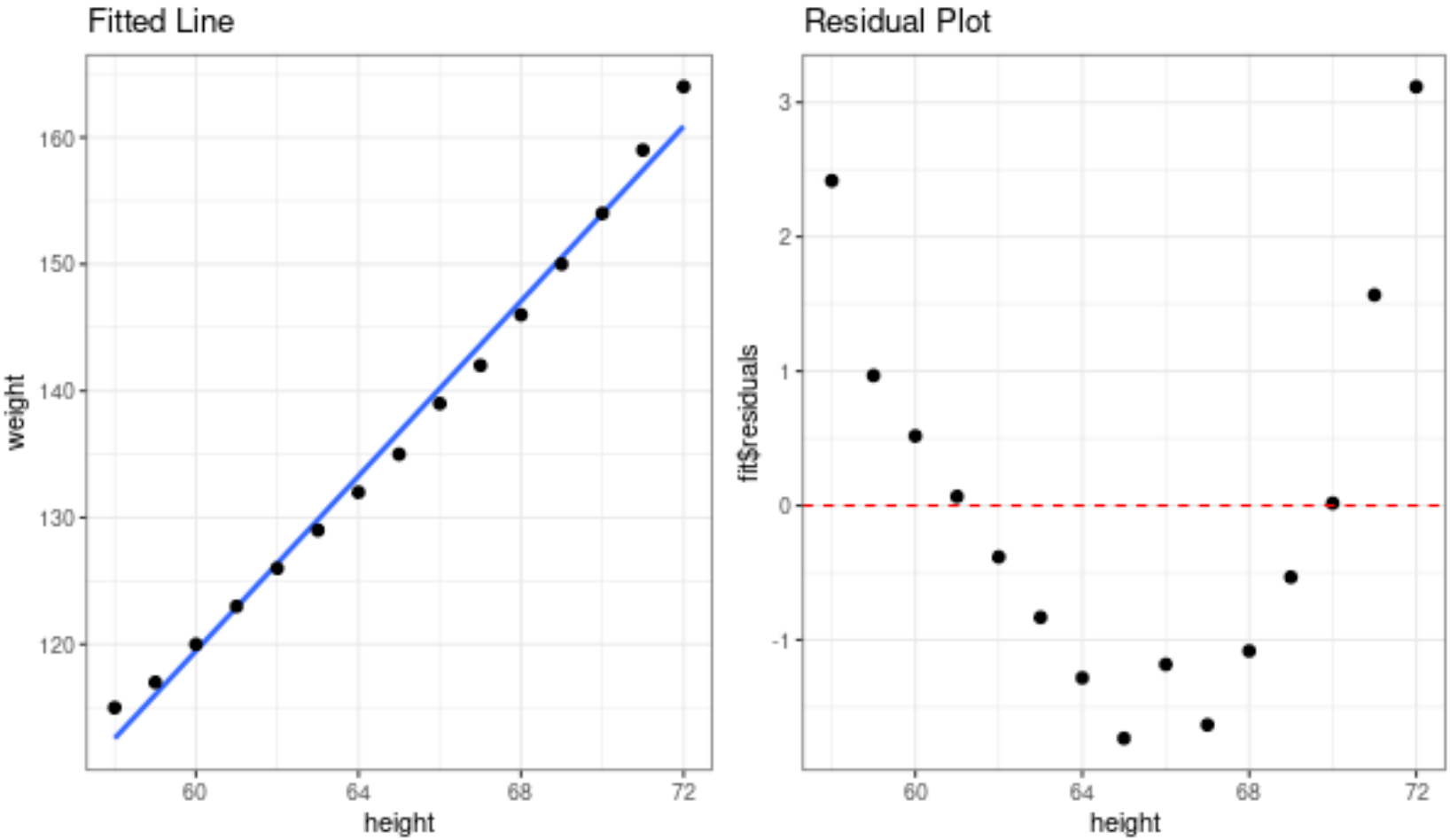
- don't want correlation between residuals and explanatory variables





Tests of linearity

Residual vs. Explanatory plot makes seeing non-linearity easier

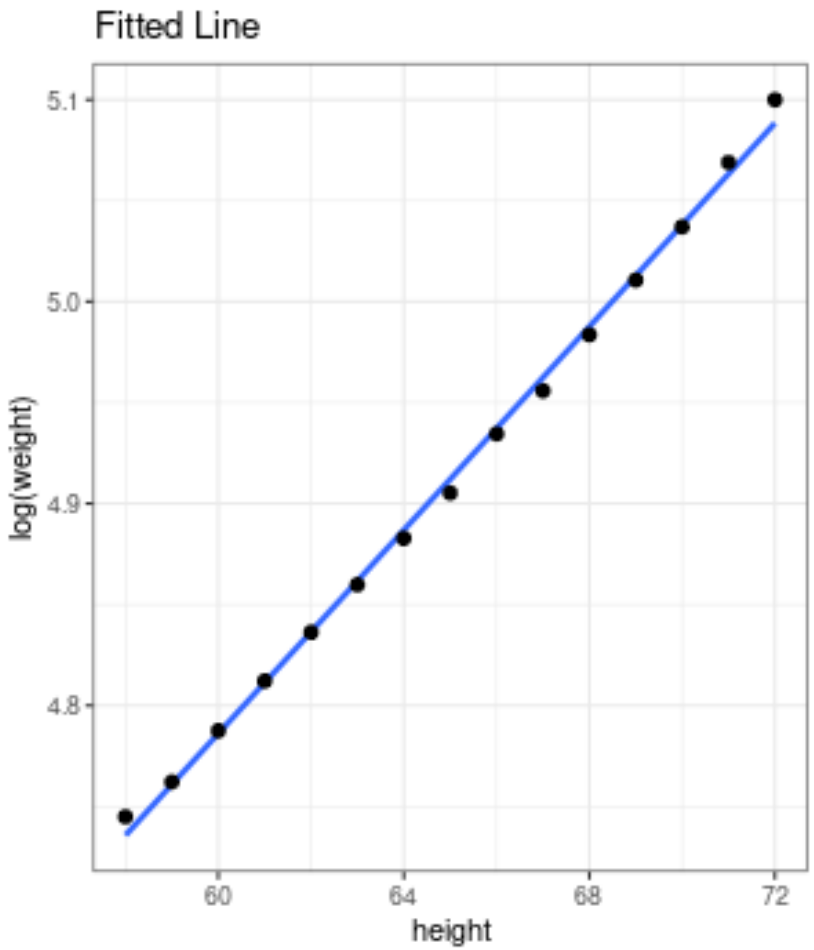


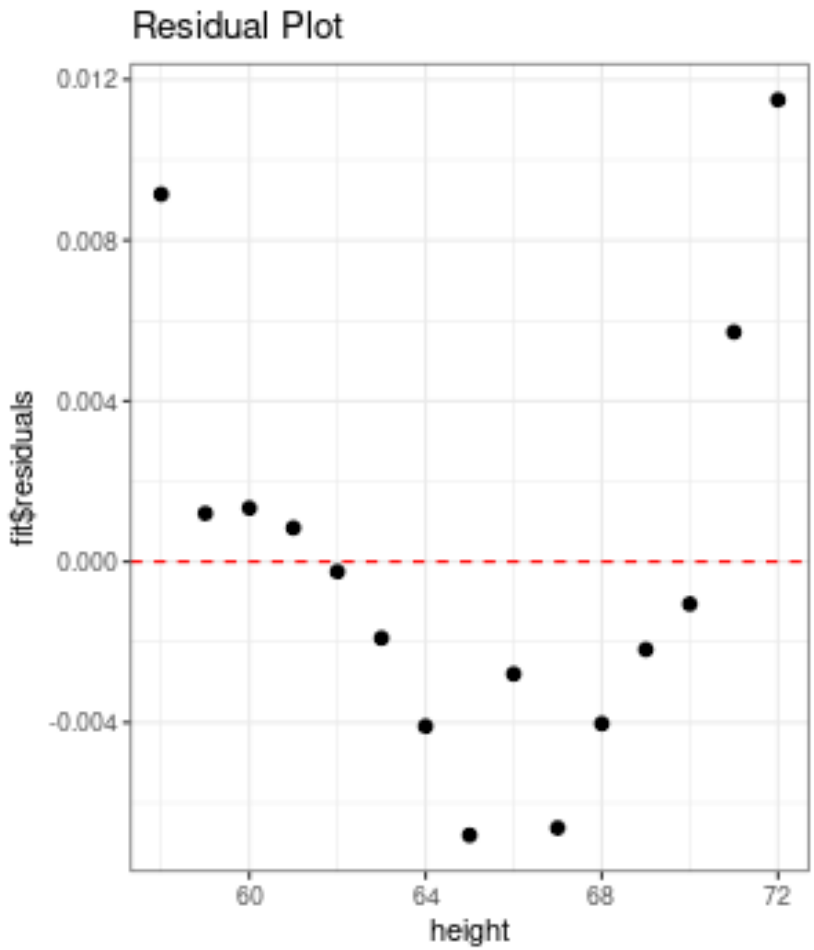
- linear regression could still be useful!
- but we could also look at doing something more complicated if we really cared

Tests of linearity

Sometimes a transformation of a variable can help correct trends $\rightarrow$ log(weight)

- better, still have a funky Residual vs. Height plot

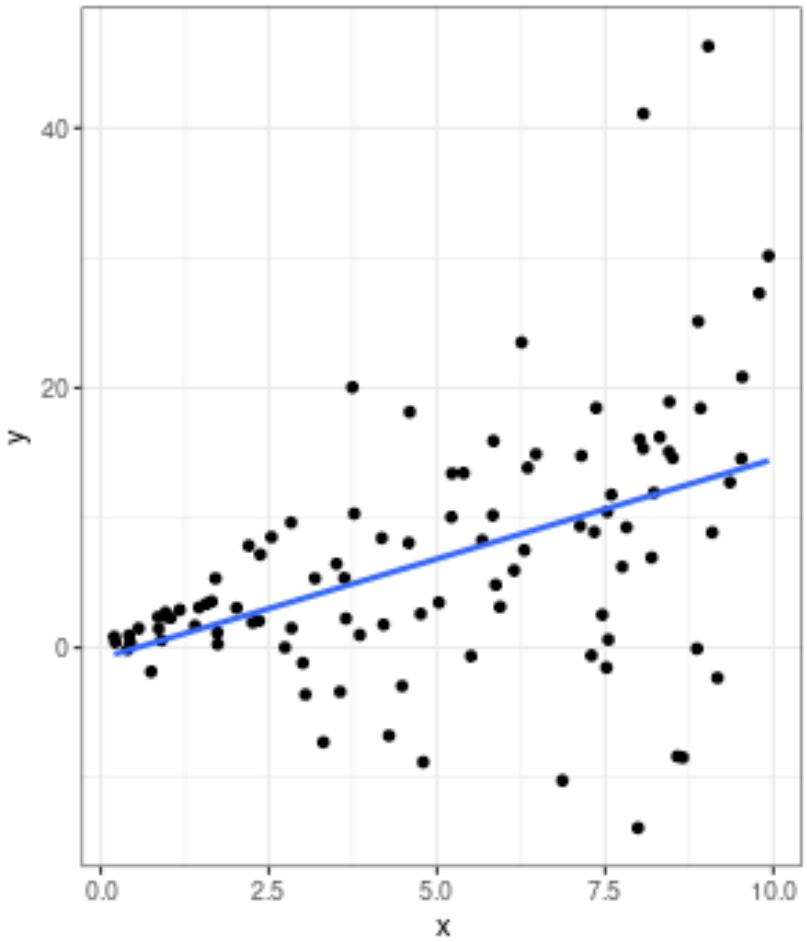


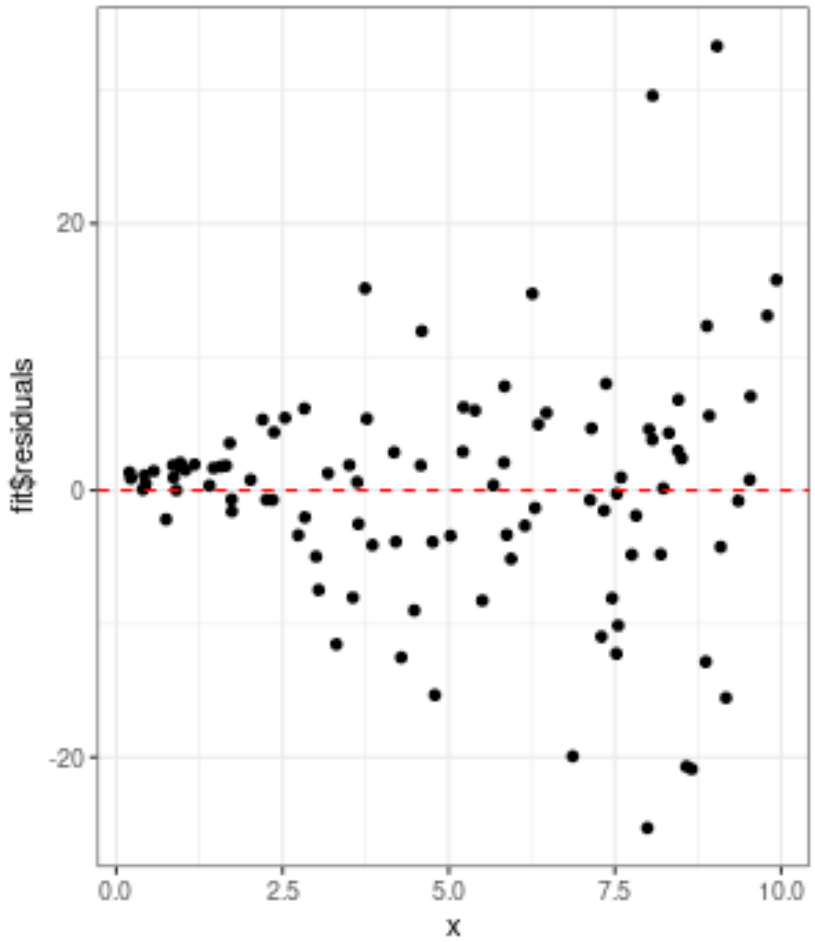


Heteroscedasticity

Hetero = different, scedastic = random
We do not want variance of residuals to increase for really small or really large values of a predictor

- This means are errors start out small but then keep getting bigger → bad!





Normal QQ Plot

A Normal Q-Q plot (Quantile - Quantile) is useful for seeing if our residuals follow a Normal distribution.

- Skewed residuals $\rightarrow$ most of the time residuals are positive/negative (bad), sometimes **really** far off in the other direction (very bad)
- Normal QQ Plot compares the quantiles of our residuals to what we would expect of a Normal distribution that has the same variance as our residuals ($\sigma^2 = \text{MSE}$)
- straight line $\rightarrow$ Normal distribution seems OK