Linear Regression (Review)

assumptions

1. linearity: $E[Y] = X\beta$ — linear combination of the predictors
2. independence: $cov(\varepsilon_i, \varepsilon_j) = 0$ $\quad i \neq j$
3. Homoscedasticity: $var(Y_j) = var(\varepsilon_j) = 0$
4. Normality: $\varepsilon_j \overset{iid}{\sim} N(0, \sigma^2)$

what if assumptions are not met? We can't use linear regression

<u>case study</u>:
- response is not Normal
- constant variance (homoscedasticity) is not met

If the response is Binomial

$$Y_i \sim Bin(n, p_i) \qquad i = 1, \dots, n$$

1. PMF $\quad P(Y_i = y_i) = \binom{n}{y} p_i^y (1-p_i)^{n-y}$
2. mean $\quad E[Y_i] = \mu_i = np_i$
3. variance $\quad var(Y_i) = \sigma_i^2 = np_i(1-p_i)$

from definition of expectation and variance for discrete r.v.

1

This variance can change for each y_i, so assumption 3 is not met.

On residuals vs fitted values we can see if assumptions are met. Here, not the case.

GLM (Generalized Linear Models)

if there is a correlation structure, mixed models can be used.