IDS 702

Linear regression assumptions

MLR Assumptions

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i; \epsilon_i \overset{\text{iid}}{\sim} N(0,\sigma^2), i = 1,...n$$

Linearity

Independence of errors

Normality of errors

Equal variance of errors

To check the assumptions, we look at the residuals

Linearity

Independence of errors

Normality of errors

Equal variance of errors

Linearity

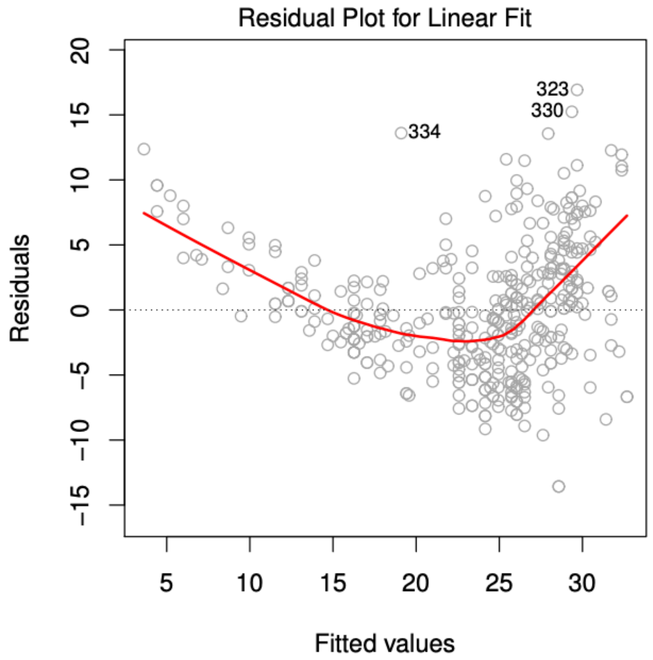
- Plot the residuals vs each predictor (or vs fitted values)
- Expect to see no pattern: some pattern is usually an indication of a relationship (often nonlinear) between the response and a predictor which has not been captured in the model
- What to do? Can consider a transformation in the predictor variable

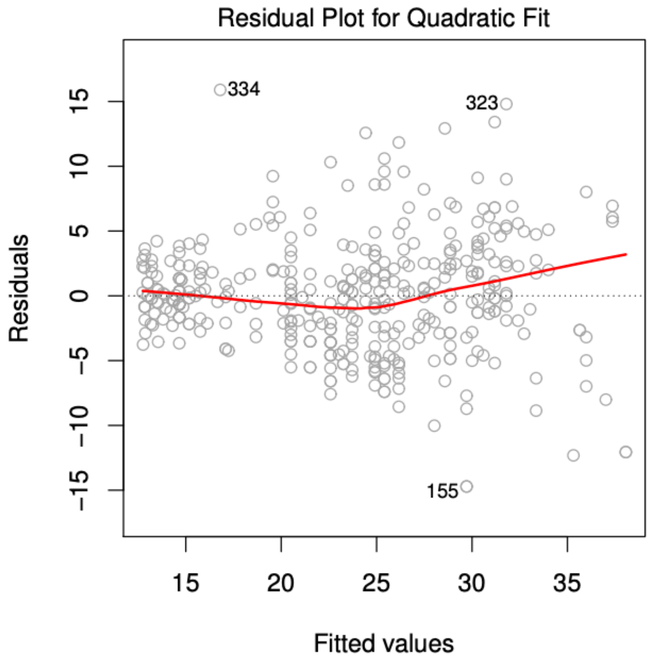
Variable transformations

- Natural log transformation is most common
- Quadratic terms
- Consider interpretation

Linearity

1. Non-linearity of the Data



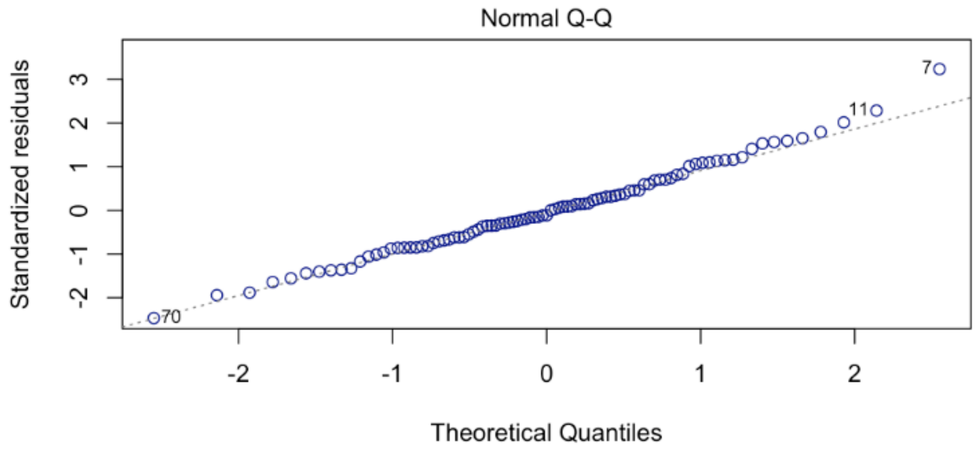


Independence of errors

- Can plot residuals vs fitted values or residuals vs index of observations (should look random)
- Generally enough to think about study design
- What to do? Consider a different model

Normality of errors

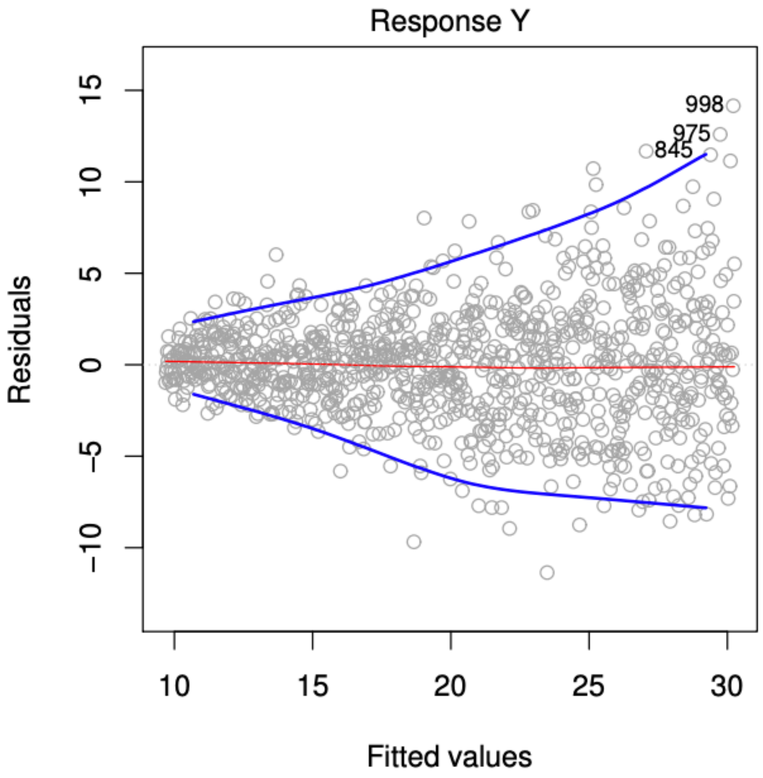
- qq-plot (quantile-quantile plot) compares the distribution of standardized residuals to a standard normal distribution
- Clustering of the points around the 45 degree line usually implies normality assumption is not violated
- Generally the least important assumption

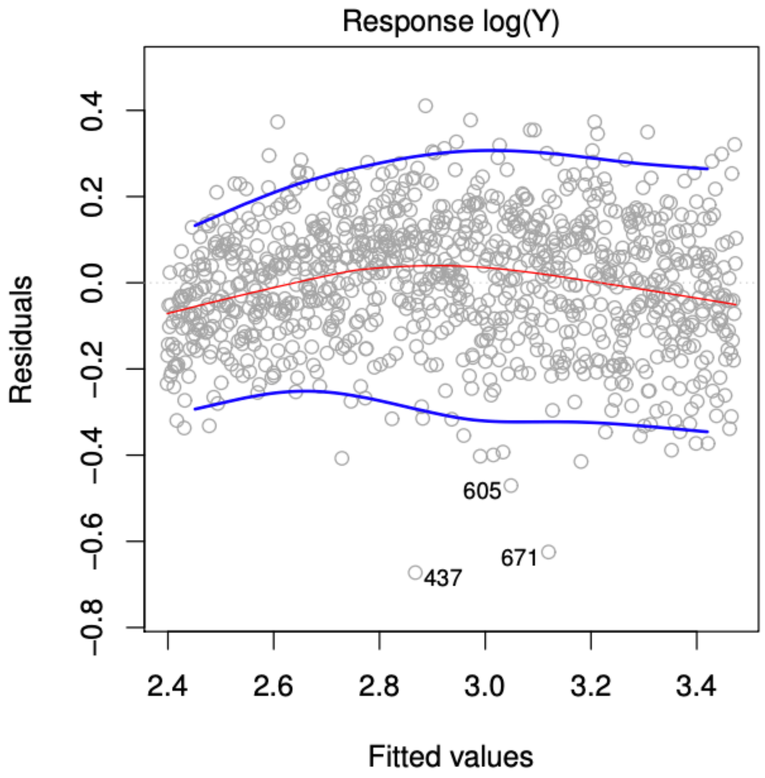


Equal variance of errors (heteroscedasticity)

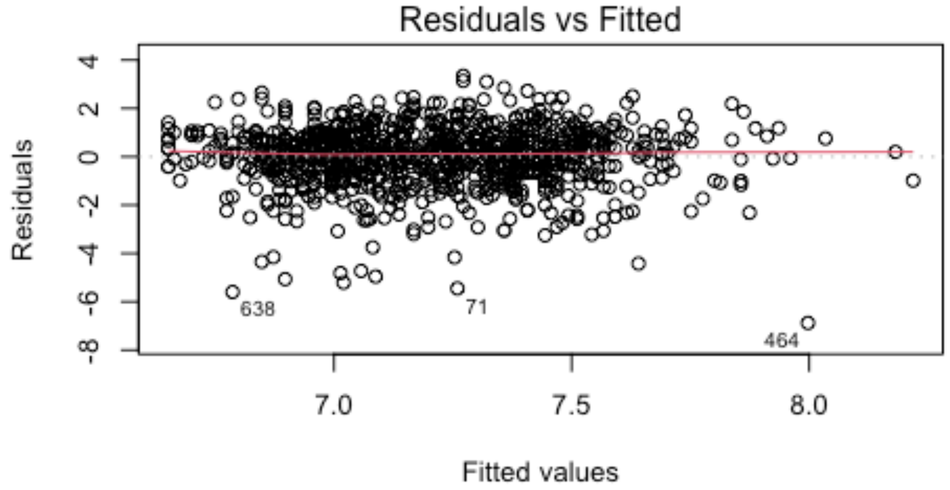
- Can plot residuals vs fitted values or residuals vs index of observations (should be equally spread around 0)
- What to do? Can consider transforming the response variable (natural log most common), or using weighted least squares estimation
- However, the issue is usually minor

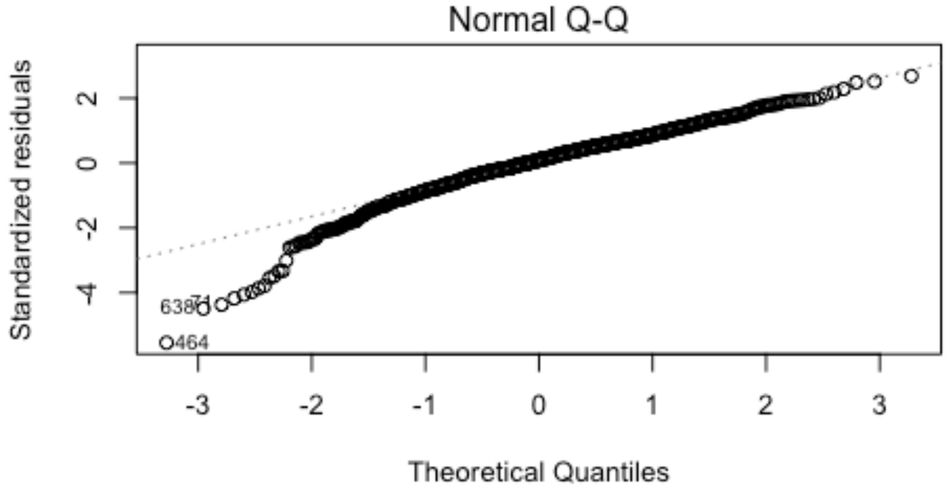
Equal variance of errors

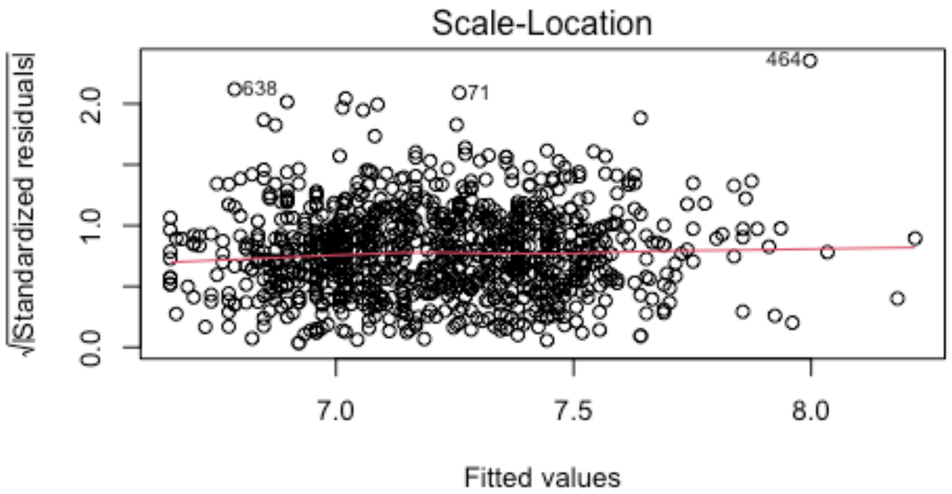


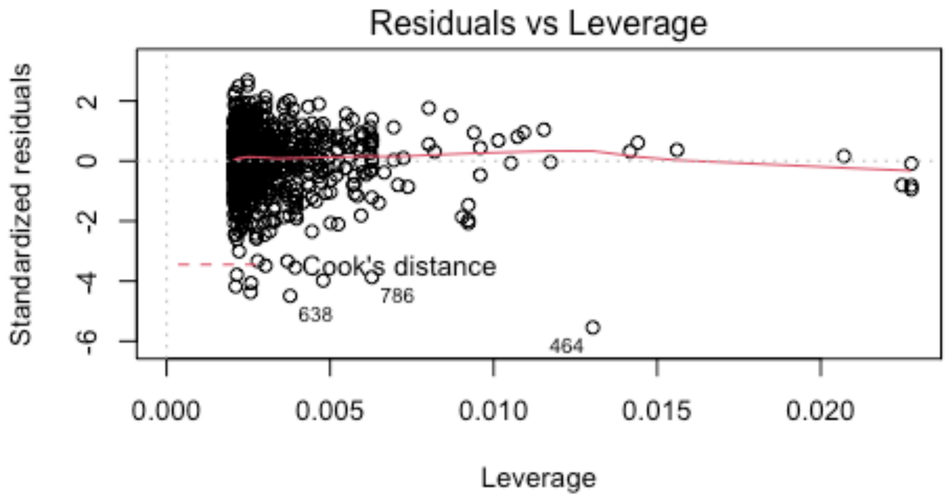


```
births_mod <- lm(weight ~ gained + sex, data=births14)
plot(births_mod)
```









Summary

- Check assumptions by plotting residuals
- Violations of linearity and independence can be “dealbreakers”
- Linear regression robust to violations of normality and equal variance
- Explore data and understand the domain