IDS 702

Linear Regression - 3 (Checking assumptions)

September 8, 2022
Andrea Lane, PhD

Agenda

1. Standard Error Demonstration
2. MLR in R
3. MLR Activity
4. Checking MLR Assumptions

Learning Objectives

By the end of this class, you should be able to

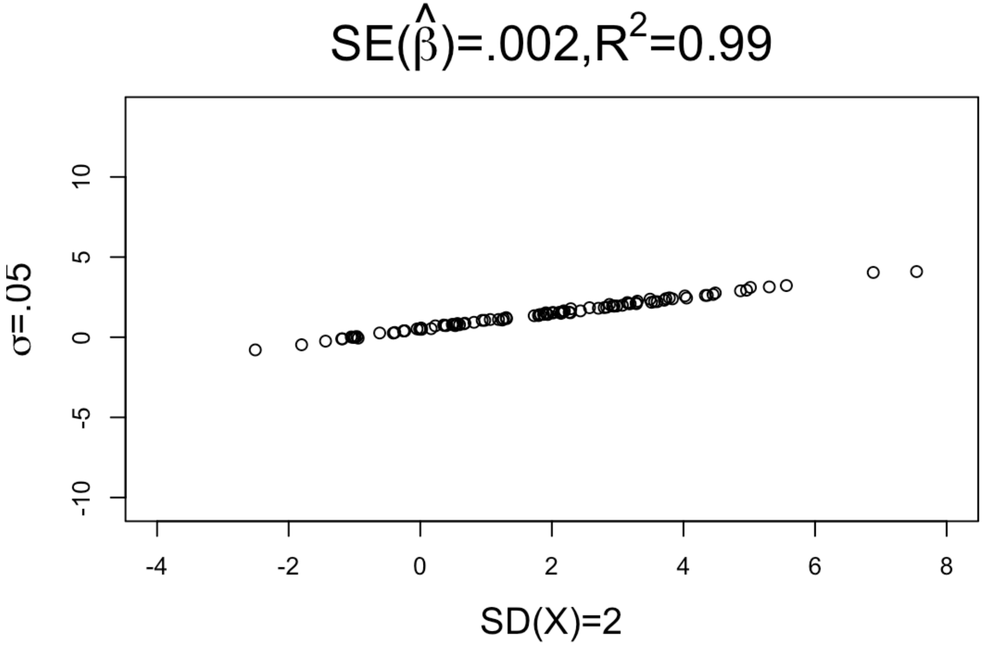
- Generate a MLR model in R
- Check the MLR assumptions

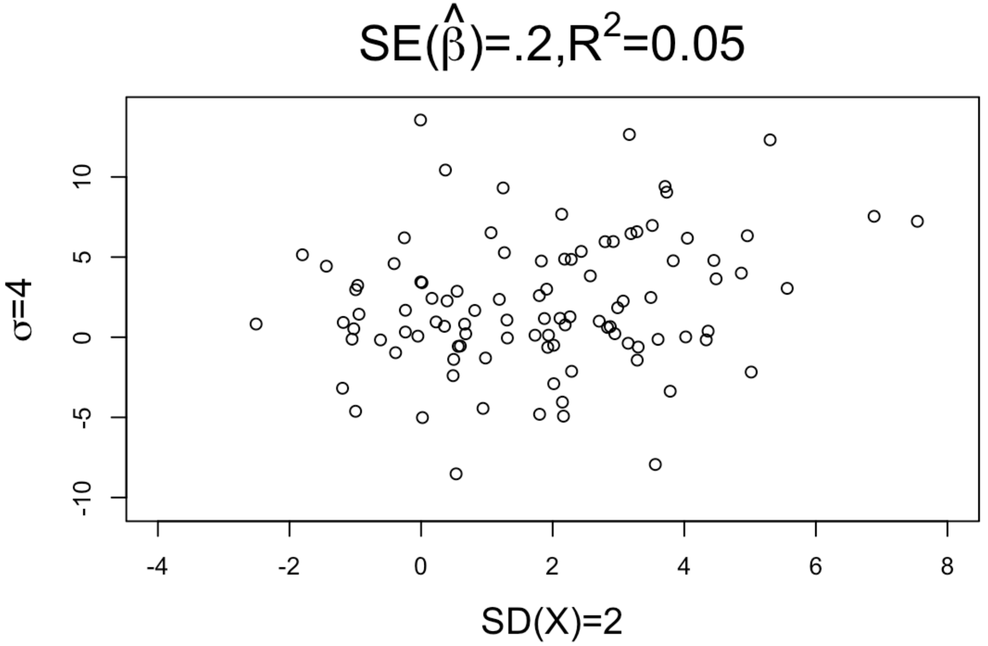
1. Standard Error Demonstration

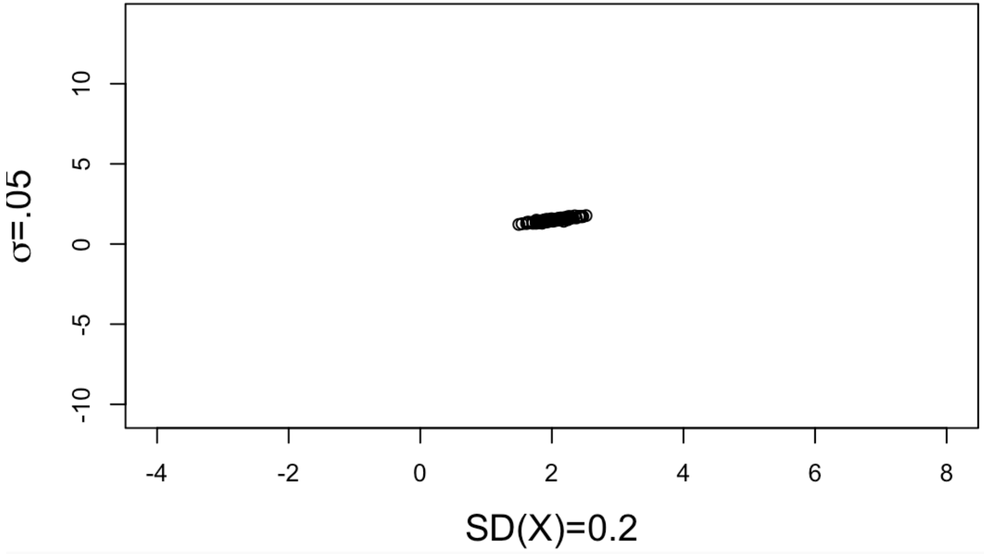
Coefficient standard error

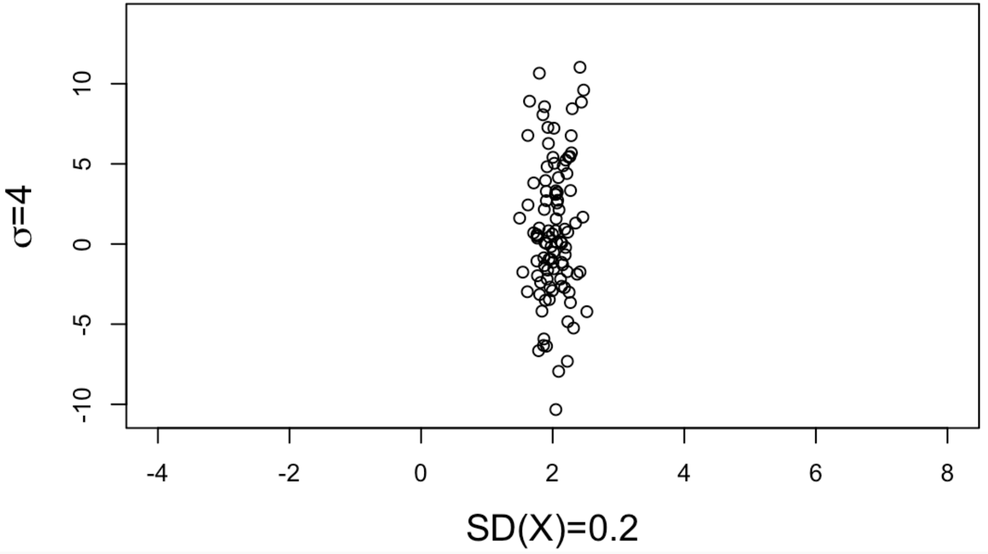
Look back at SLR

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma^2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$


SE($\hat{\beta}$)=.002,R^2=0.99
σ=.05
SD(X)=2
-10
-5
0
5
10
-4
-2
0
2
4
6
8


SE($\hat{\beta}$)=.2,R^2=0.05
σ=4
SD(X)=2
-10
-5
0
5
10
-4
-2
0
2
4
6
8


SE($\hat{\beta}$)=.02,R^2=0.83
σ=.05
SD(X)=0.2
-10
-5
0
5
10
-4
-2
0
2
4
6
8


SE($\hat{\beta}$)=2,R^2=0.005
σ=4
SD(X)=0.2
-10
-5
0
5
10
-4
-2
0
2
4
6
8

2. MLR in R

3. MLR Activity

4. Checking MLR Assumptions

MLR Assumptions

- Linear relationship between EACH X and Y
- Independence of errors
- Equal variance of errors
- Normality of errors
- No multicollinearity

- Often want to look at **residual plots** to check assumptions

Linearity

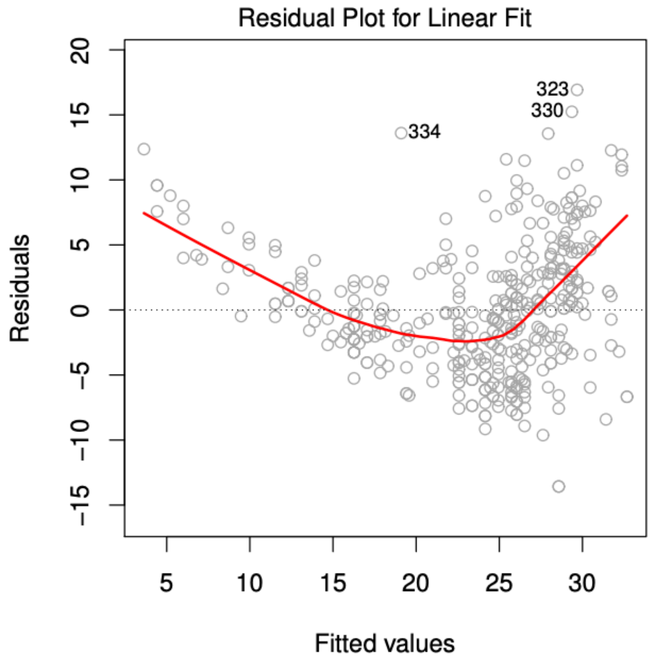
- Plot the residuals vs each predictor (or vs predicted values)
- Residuals contain information about the response variable that has NOT been explained by the predictors
- Expect to see no pattern: some pattern is usually an indication of a relationship (often nonlinear) between the response and a predictor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

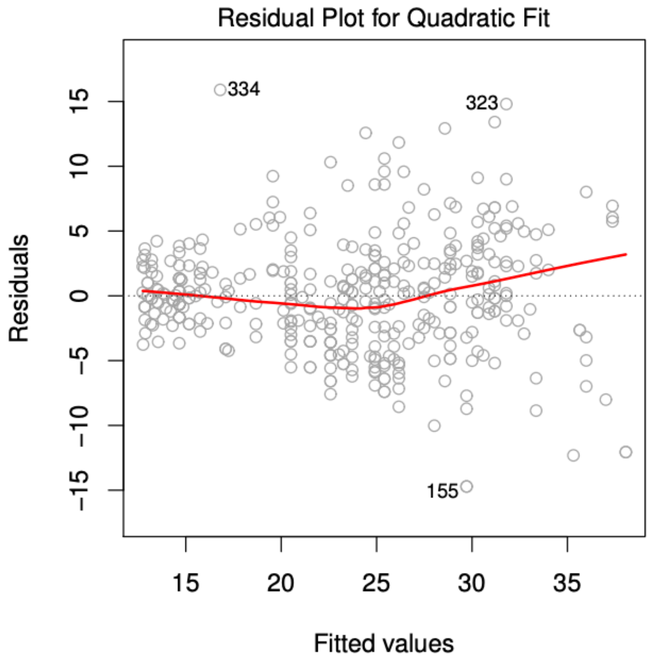
Variable transformations

- Natural log transformation is most common
- Quadratic terms
- Consider interpretation
- May take some trial and error

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

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

Log transformation of the outcome variable

$$ln(y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$$

Then

$$y_i = e^{(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i)} = e^{\beta_0} e^{\beta_1 x_{i1}} \ldots e^{\beta_p x_{ip}} e^{\epsilon_i}$$

The predictors have a multiplicative effect on y

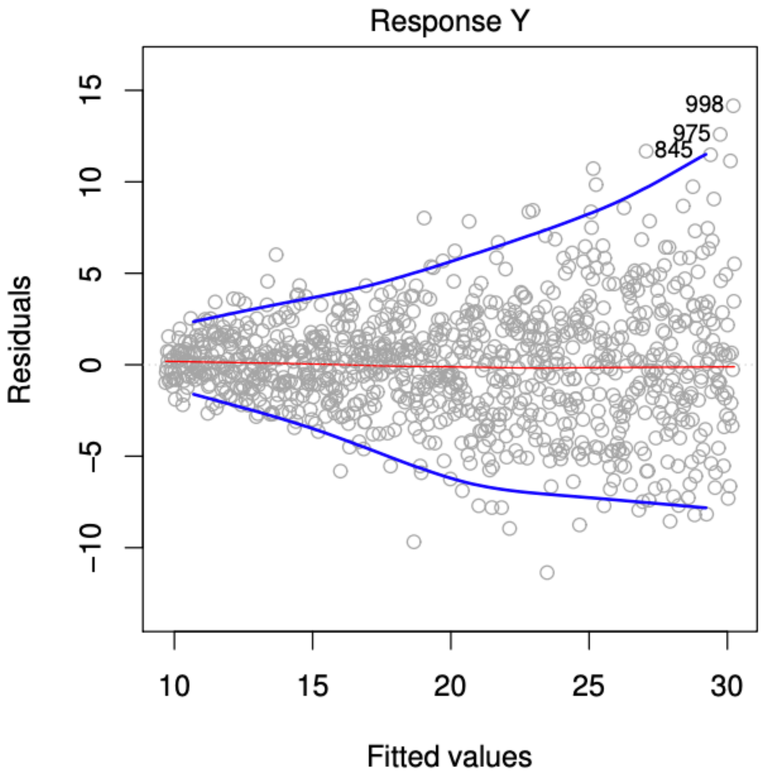
Log transformation of the outcome variable

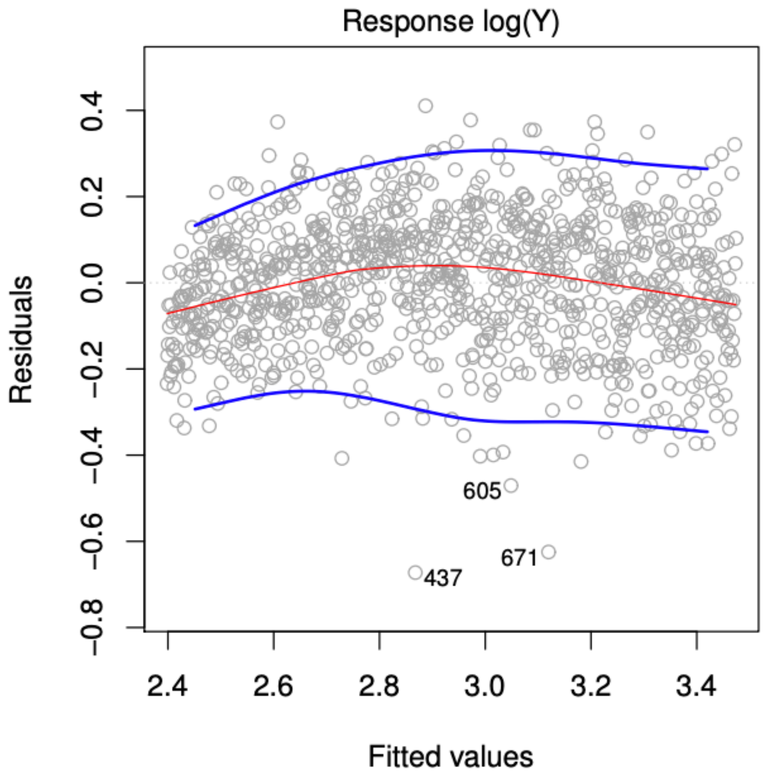
- The estimated coefficients can be interpreted in terms of approximate proportional differences

$$\beta_1 = 0.1 \rightarrow e^{\beta_1} = 1.1052$$

- A difference of 1 unit in x_1 corresponds to an expected positive difference of approximately 11% in y

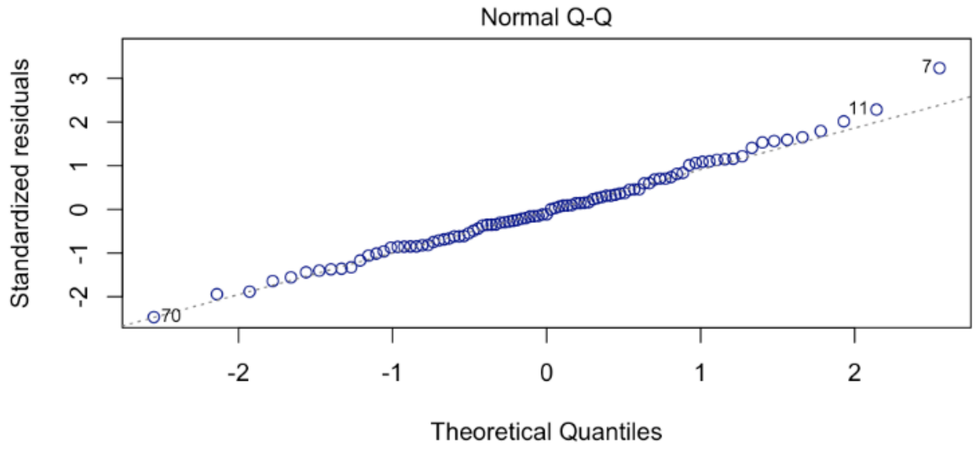
Equal variance of errors





Normality of errors

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



Wrap-up

- Data analysis assignment 1 due Sept 16 11:55 PM