P8130: Biostatistical Methods I

Lecture 17: Multiple Linear Regression
Model Diagnostics

Cody Chiuzan, PhD
Department of Biostatistics
Mailman School of Public Health (MSPH)

Diagnostics in MLR

- Remember the assumptions we made regarding residuals:
  - Residuals are normally distributed
  - Variance of residuals is constant across the range of X(s)
  - Residuals are independent of one another
- Model diagnostics are important and should always be checked!

Diagnostic Considerations

Regarding residuals:

- Residuals are not normally distributed
- Residuals are not independent
- Residuals do not have constant variance (homoscedasticity vs heteroscedasticity)

Other considerations:

- The regression function is not linear
- The model fits well, but there are some 'unusual' observations
  - Outliers in Y
  - Outliers in X
  - Influential observations
- There is high correlation between predictors (multicollinearity)

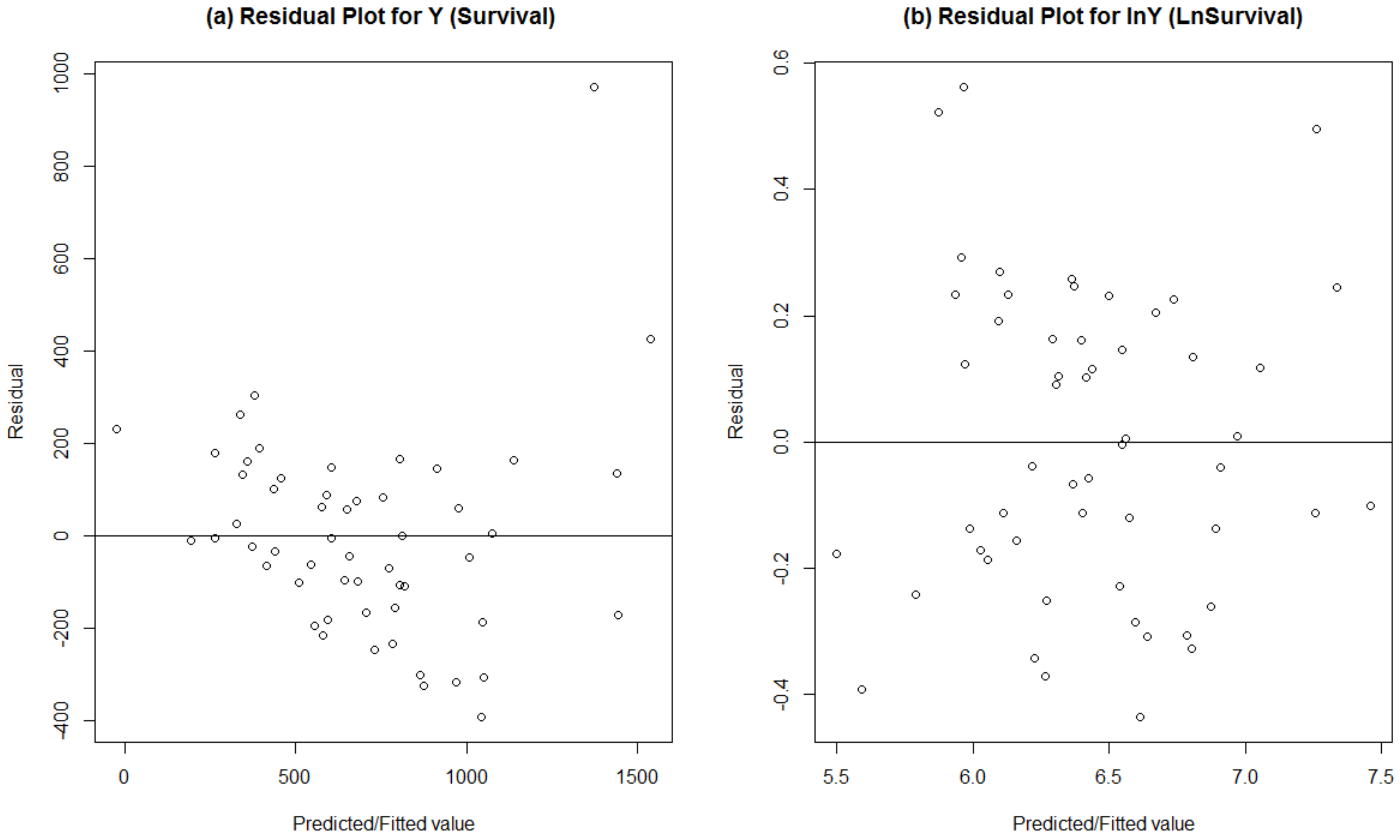
Diagnostic Plots

- Residuals (Y-axis) vs fitted values (X-axis)
- Residuals (Y-axis) vs an observed covariate (X-axis)
- Residuals boxplot
- Normality probability plot and quantile-quantile (QQ plot)

Residuals vs Fitted/Predicted Values Plot

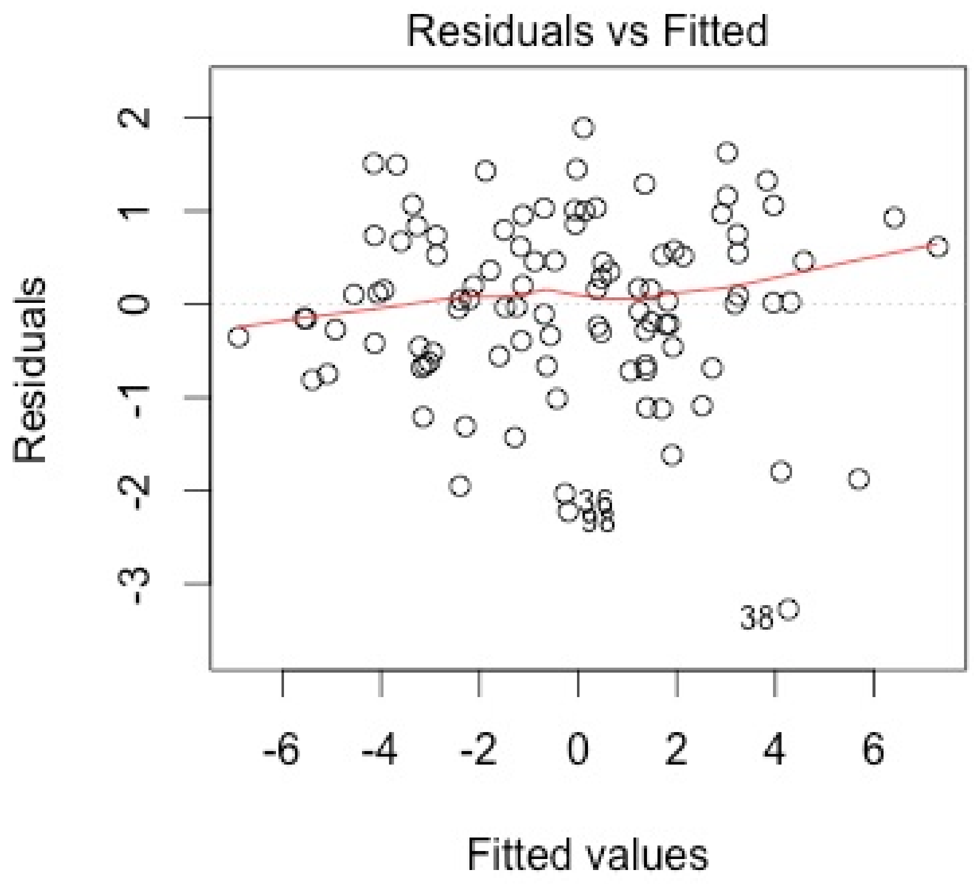
- The plot is used to detect unequal error variance (heteroscedasticity) and outliers
- Ideally, we would like to see that:
  - Residual values bounce around 0 (the expected value is 0, right?)
  - Residuals form a horizontal (linear) ‘band’ around zero: above and below (indication of equal variance)
  - No ‘unusual’ values stand out from the random pattern (indication of no potential outliers)
- Do not over-interpret these plots and be careful about small data sets!

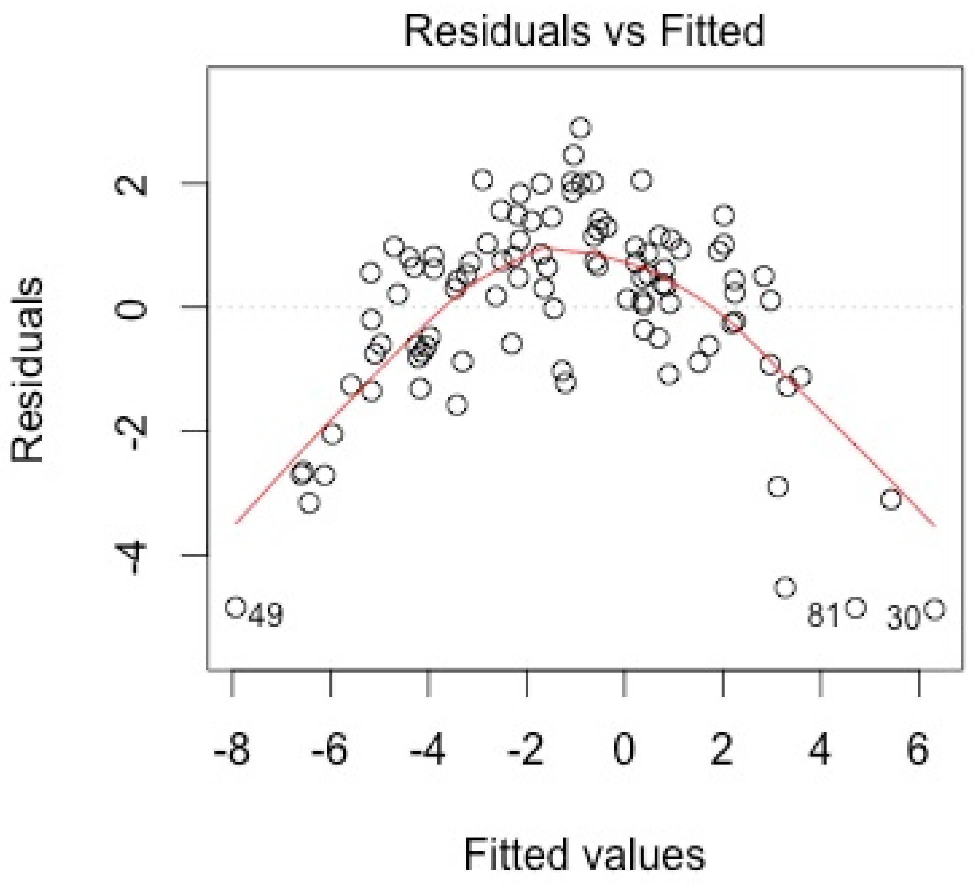
Residuals vs Fitted/Predicted Values



Plot b) is to be preferred – random pattern, evenly distributed around 0.

Add’l Examples



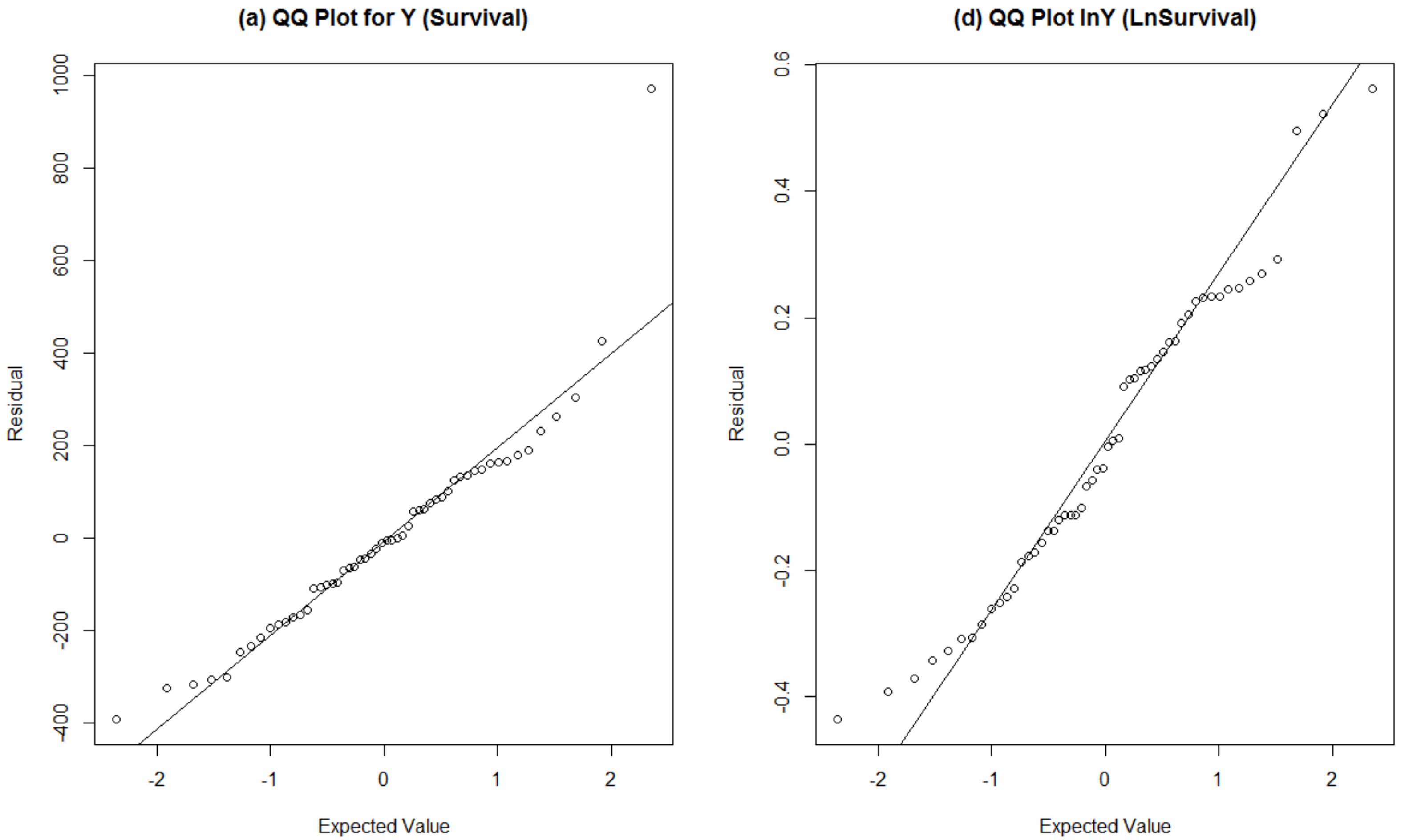


Compare the two plots – obviously, the plot on the right suggests a potential curvilinear (quadratic?) trend.

Quantile – Quantile Plot (QQplot)

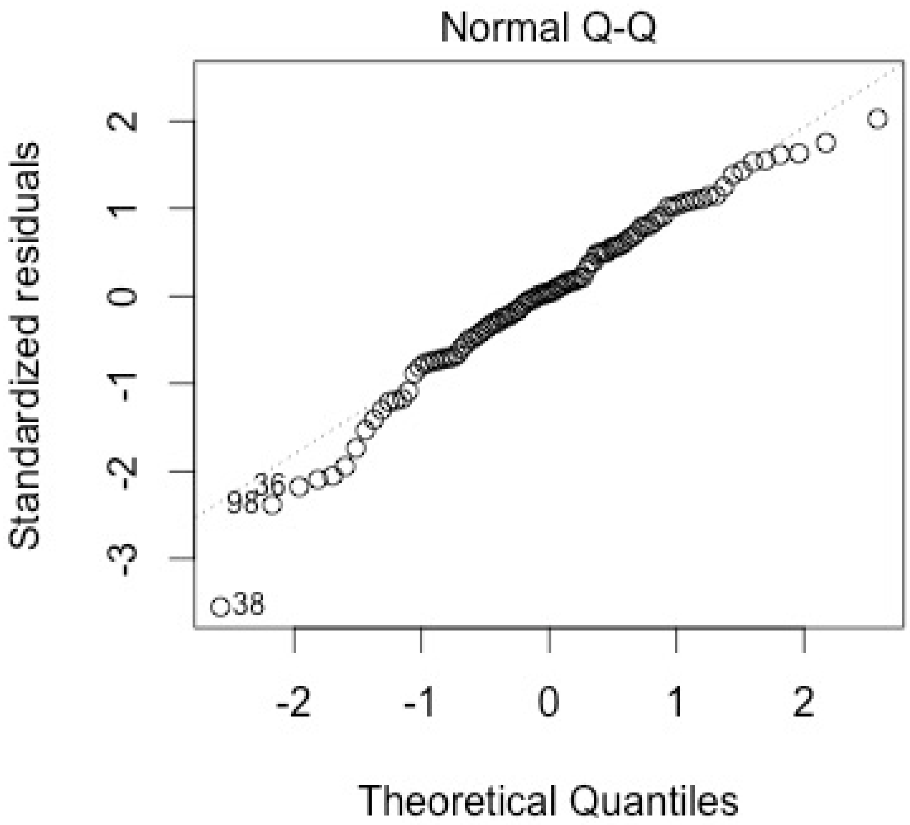
- The plot is used to detect <u>(non) normality of residuals and outliers</u>
- It plots the quantiles of a standard normal vs quantiles of the observed data
- Ideally, we would like to see a straight line (residuals are normal)
- Small departures from normality are not concerning
- Heavy tails indicate the presence of outliers

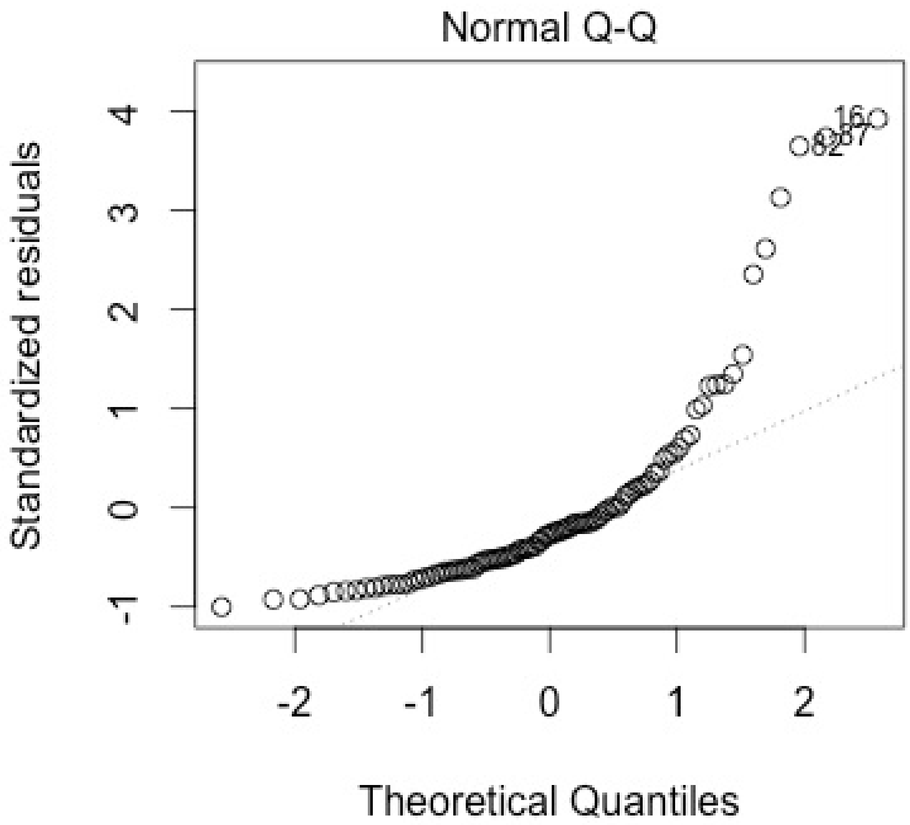
Quantile-Quantile Plot (QQ-Plot)



Which of these plots is more in line with the ‘normality’ assumption?

Add'l Examples



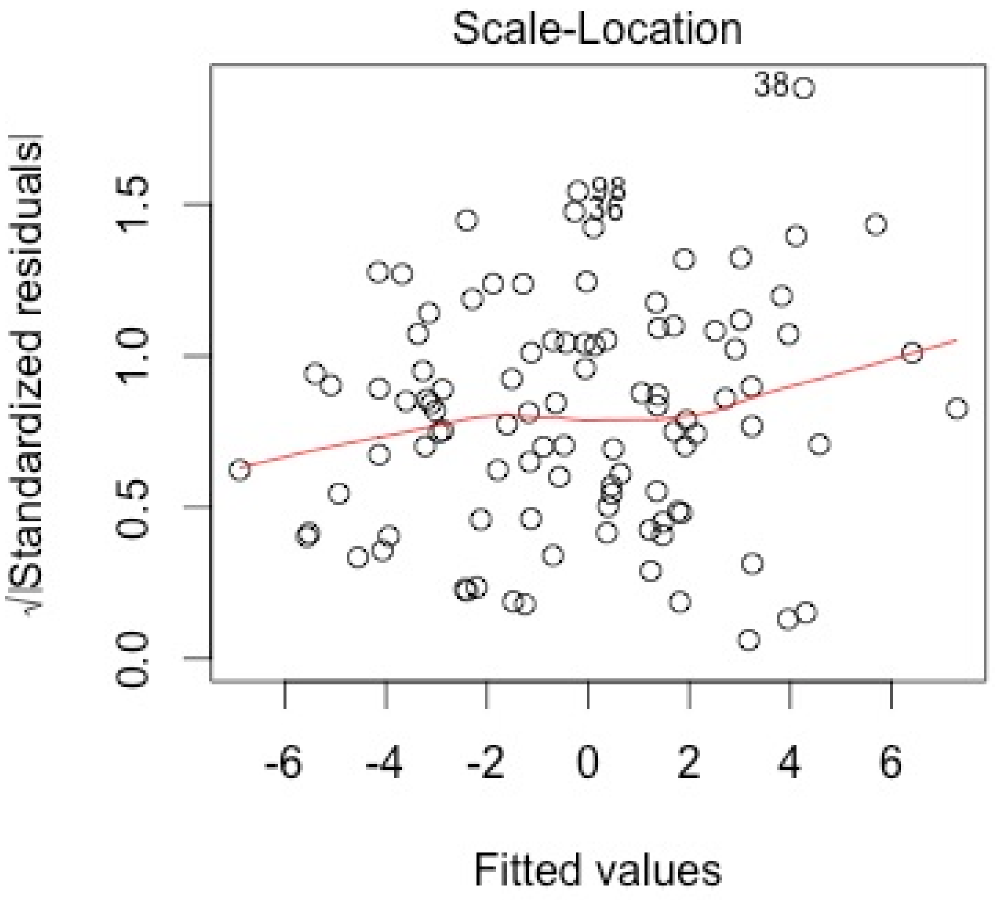


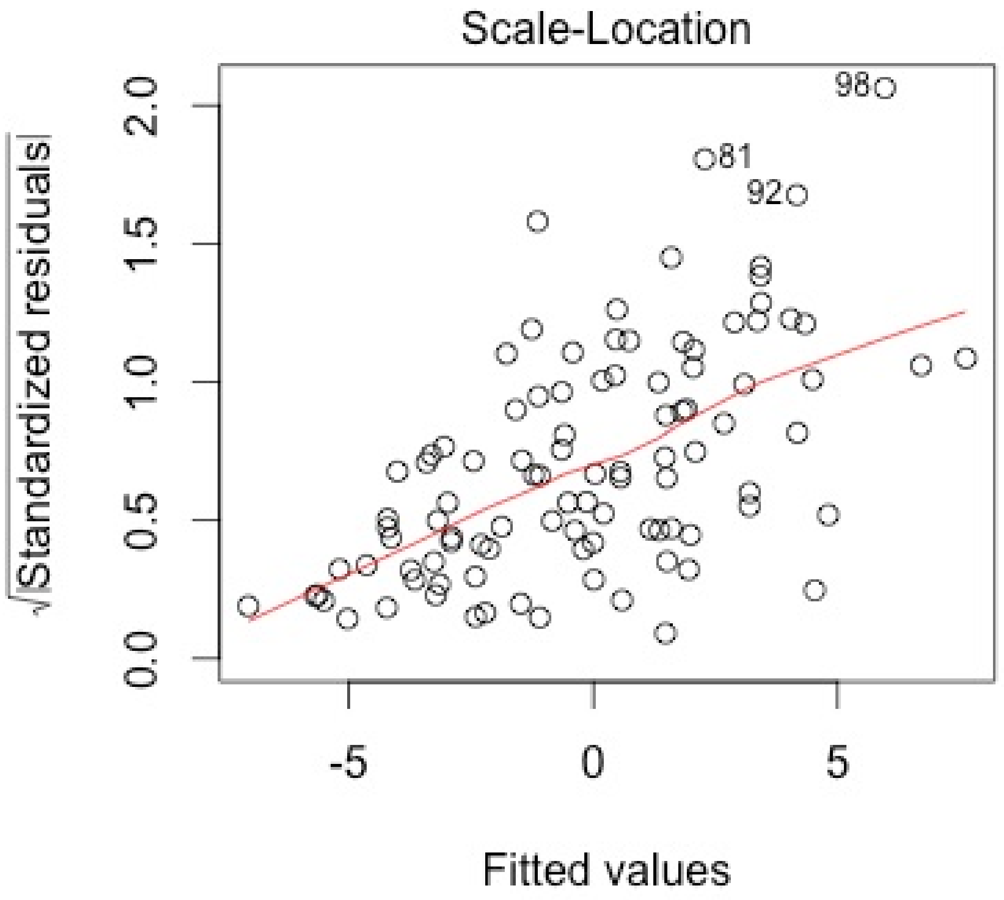
Compare the two plots – any deviations from normality?

Other Plots

- **Scale-Location** plot shows if the residuals are spread equally along the range of predictors
  - The plot helps checking the assumption of equal variance
  - Ideally, we would like to see a horizontal line with equally spread points
- **Residuals vs Leverage**
  - The plot helps identify influential cases
  - Not all outliers are influential, i.e., the results do not differ much if we include/exclude from analysis
  - We watch for outlying values at the upper right or lower right corner (cases outside of a dashed line, Cook's distance)

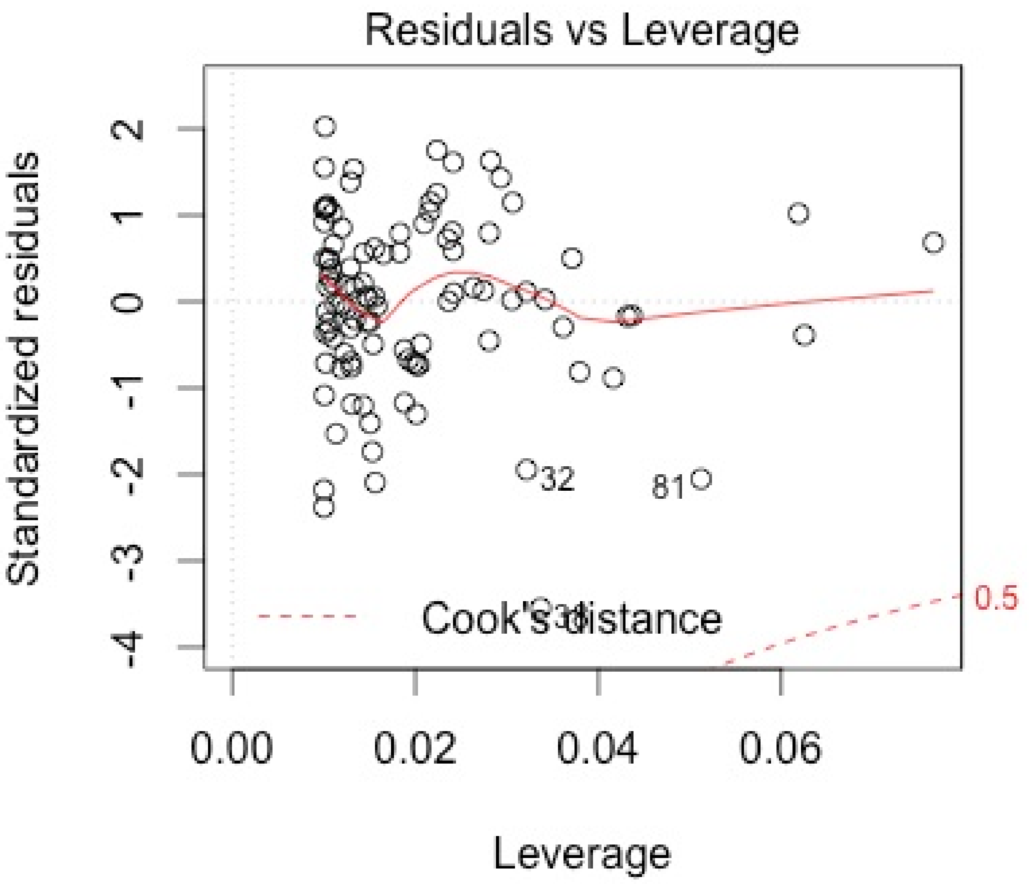
Add'l Examples

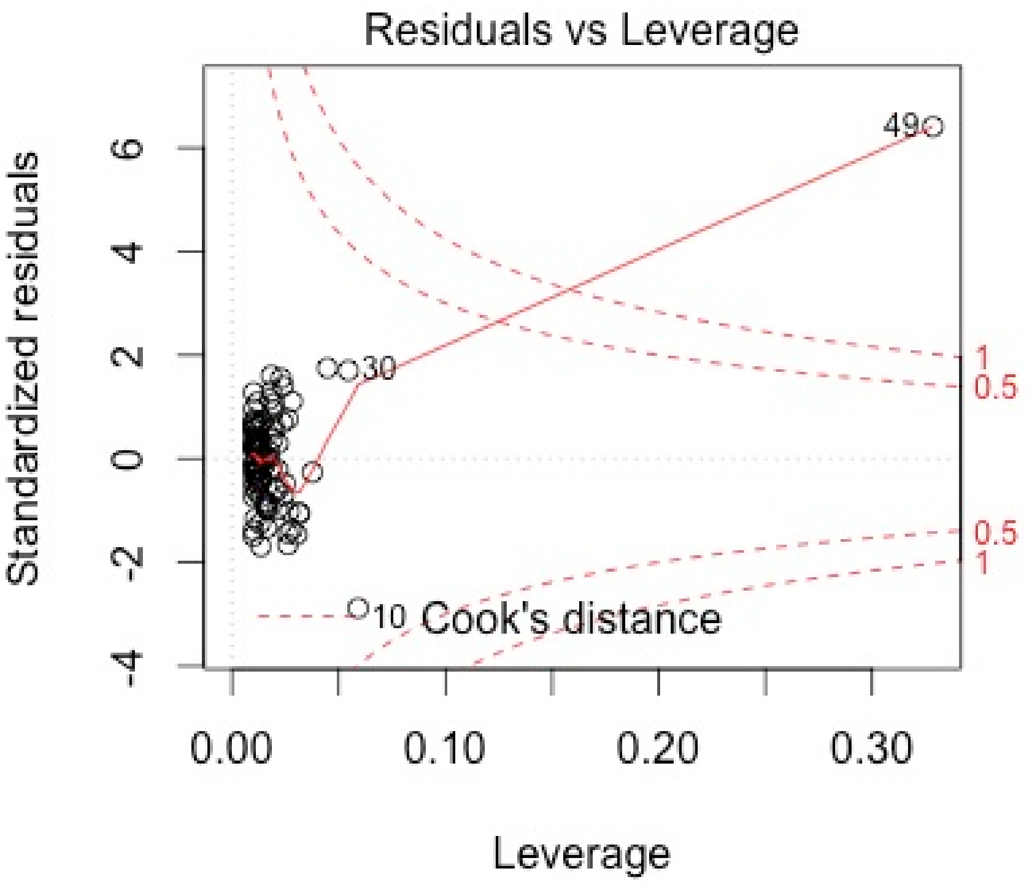




What can you see in the two plots?

Add'l Examples

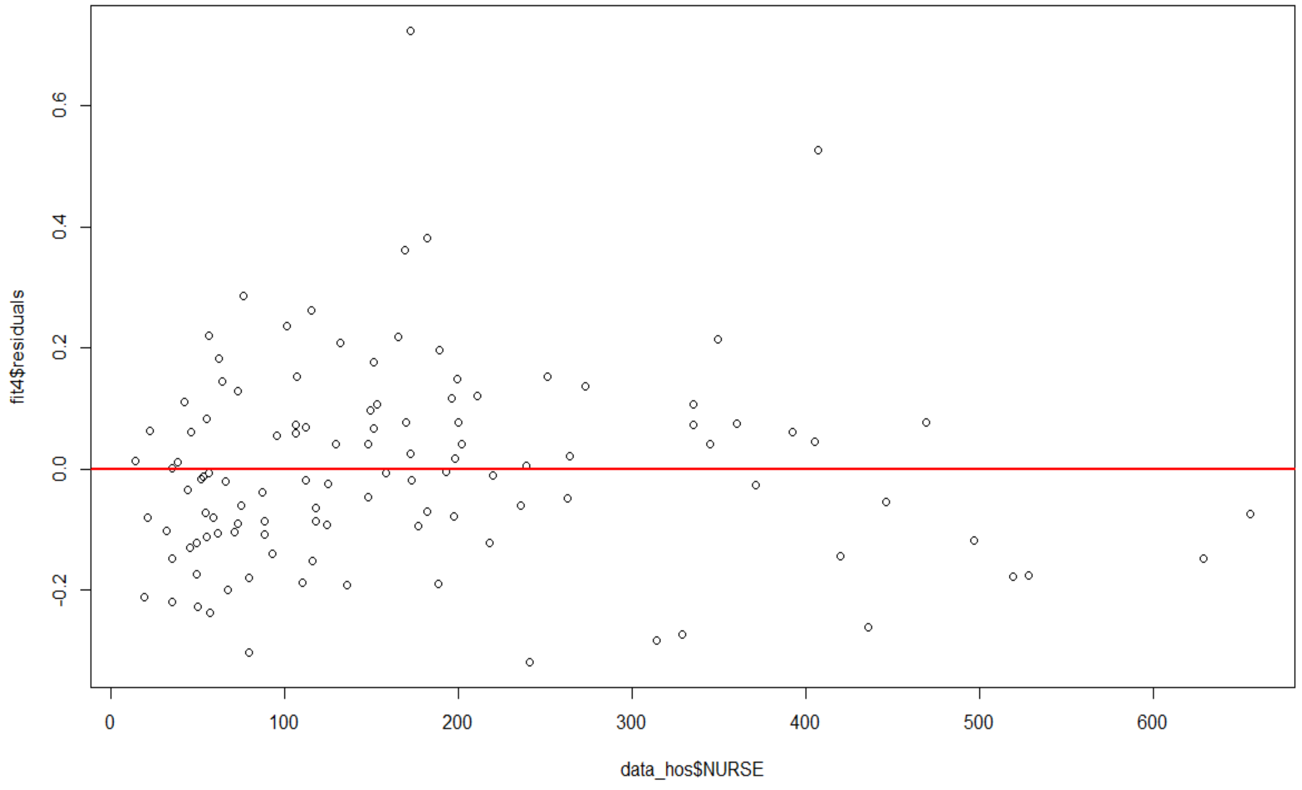




What can you see in the two plots?

Residuals vs Covariate Plot

- The plot shows if the spread of residuals around the *Y=0* line is approximately constant over the range of X
- Can also observe the 'fanning of the data', i.e., large variance as X increases
- Is there a linear relationship b/w residuals and X? A curved pattern indicates that the predictor should enter the model in a curvilinear manner.



Diagnostics - Remedies

- Residuals heteroscedasticity (non-constant variance) and non-normality
  - Transformations of the outcome (Y) or of the covariates (X)
  - Common transformations
    - Natural logarithm
    - Square root
    - Inverse: 1/Y

- How to determine the ‘best’ transformation?
  - Box-Cox transformation: finds the ‘best’ power transformation to achieve normal residuals

Box-Cox Transformation

- Box-Cox method applies a transformation by raising Y to different powers: Y^a
- The solution or the recommended transformation is the 'a' value that maximizes the likelihood and stabilizes the variance or makes the data more 'normal'

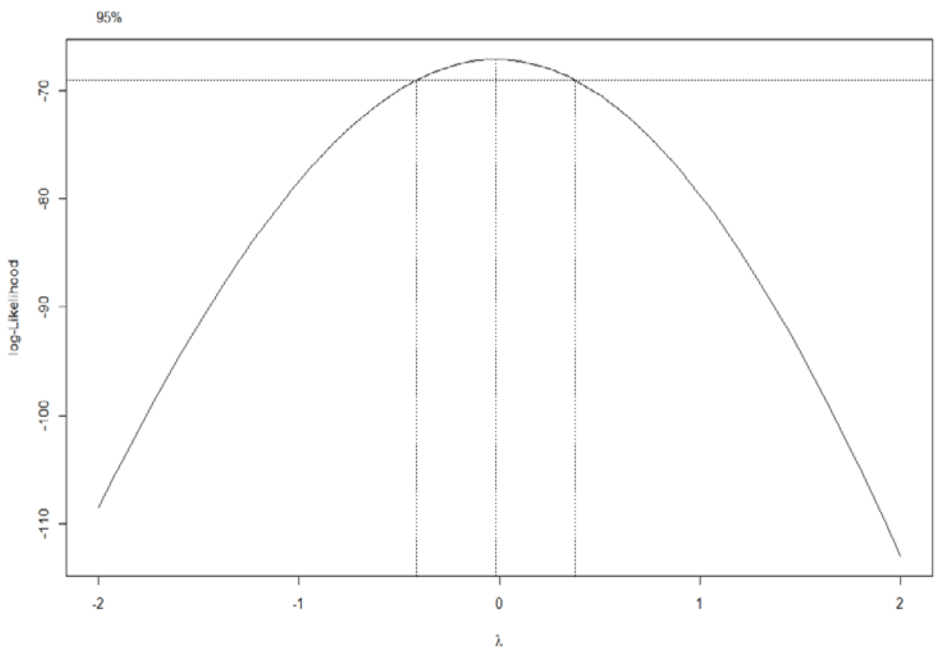
- Common powers:
    - Natural logarithm: a=0
    - Square root: a=1/2
    - Inverse: a=-1
    - Identity: a=1 (no transformation)

Box-Cox Transformation

Simple Linear Regression

fit1 <- lm(Survival ~ Bloodclot, data=data_surg)

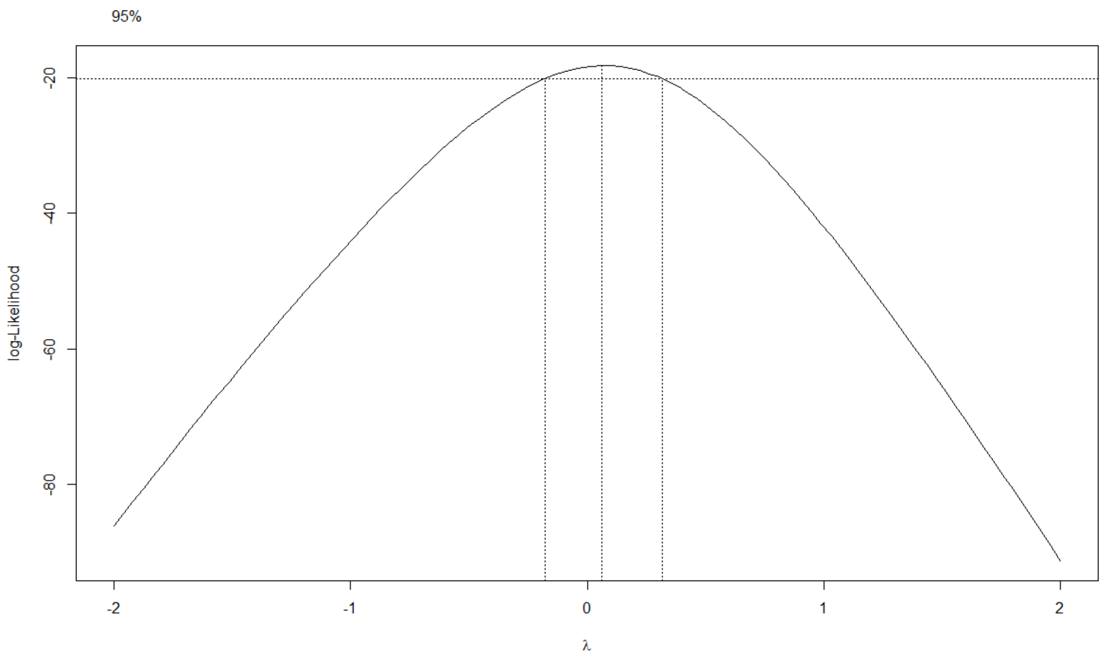
boxcox(fit1)



Multiple Linear Regression (MLR)

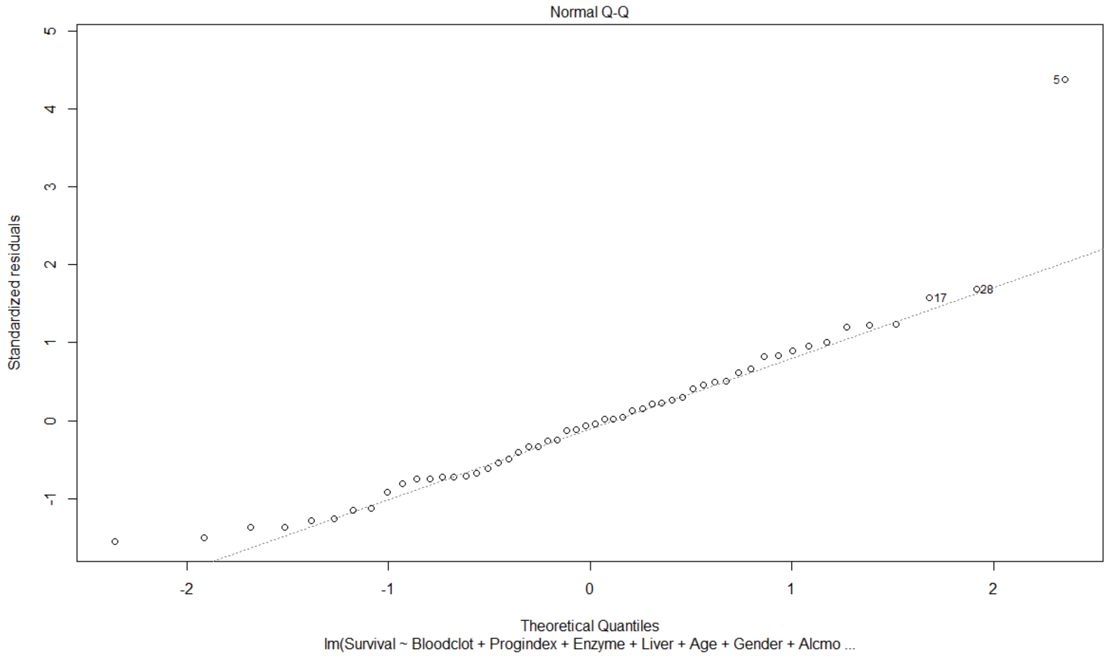
mult.fit1 <- lm(Survival ~ Bloodclot + Progindex + Enzyme + Liver + Age + Gender + Alcmod + Alcheav, data=data_surg)

boxcox(mult.fit1)

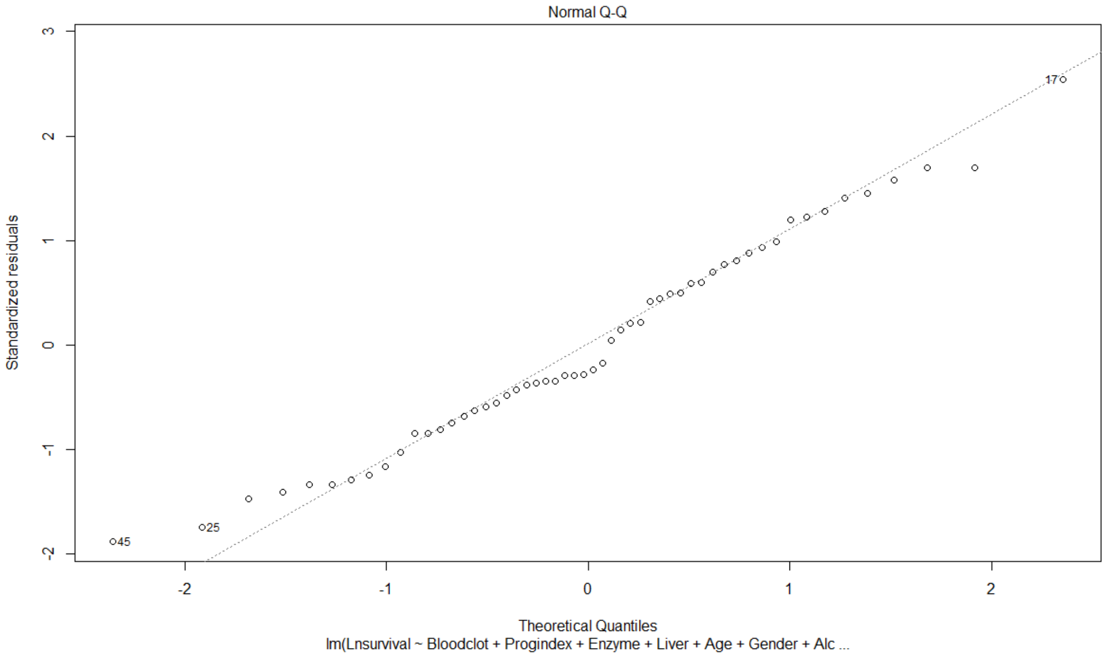


Box-Cox Transformation

MLR untransformed



MLR log transformed



Compare the two plots? Should we even transform at all?

Outliers and Influential Observations

- Outliers are individual observations that are in ‘some way’ different from the bulk of the data set

- Outliers can be found among:
  - The Y values
  - The X values
  - For both X and Y

- In SLR, outliers can be identified with scatter plots
- In MLR, we rely on the complex diagnostics

Outliers in Y

- Outliers in Y can be detected using 'studentized residuals'
- The internally studentized residual for the i^{th} observation is:

$$r_i = \frac{e_i}{s\{e_i\}} = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}}$$

- Where h_{ii} is the diagonal element of the hat matrix **H**:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}', \text{ where } \tilde{\mathrm{e}} = (\mathbf{I} - \mathbf{H})\widetilde{\mathrm{Y}}$$

- Rule of thumb: an observation with $|r_i| > 2.5$ may be considered an outlier (in Y)

Leverage values: Outliers in X

- The diagonal elements h_{ii} of the hat matrix are also called <u>leverage values</u>

- They measure how far each observation is from the center of the X space

$$\sigma^2(\tilde{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$$

$$0 \leq h_{ii} \leq 1, \text{ with } \sum_{i=1}^{n} h_{ii} = p,$$

p represents the number of model parameters (includes intercept).

Leverage values: Outliers in X

- If a h_{ii} leverage value is high, this means that the i^{th} observation might have an influence on the fitted equation (model parameter estimates).
- Rules of thumb:

$$0.2 < h_{ii} > 0.5\text{, moderate leverage}$$
$$h_{ii} > 2\mathrm{p}/\mathrm{n}\text{, high leverage}$$
$$h_{ii} > 0.5\text{, very high leverage}$$

- This rule only applies to large data sets, relative to the number of parameters
  - Imagine *n = 4* cases and *p = 3*?
  - For *p = 2*, take 0.2 as cutoff

Influential Observations

- After identifying cases that are outlying with respect to their Y and/or X values, we need to determine if they are *influential*

- An observation is influential if its exclusion or inclusion causes major changes in the regression function (estimates)

- We focus on two measures of influence that both measure the difference b/w the fitted line with observation ‘*i*’ included and the fitted line with observation ‘*i*’ deleted

Influential Observations

- DFFITS: DF stands for the difference between the fitted value for the i^{th}case when all observations are used to fit the regression line and the predicted value of the i^{th}case, when the i^{th}observation is omitted:

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)}\, h_{ii}}}$$

- Rules of thumb of concern:

$|DFFITS_i| > 1$, for small data sets

$|DFFITS_i| > 2\sqrt{p/n}$, for large data sets

Influential Observations

- Cook's distance D: considers the influence of the ith case on all fitted values:

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y_j} - \widehat{Y_{j(i)}})^2}{pMSE}$$

- Using Cook's distance, an observation can be influential if:
  - Has a high residual e_i and moderate h_{ii}
  - Has high h_{ii} and moderate e_i
  - Or both

- Rule of thumb: $D_i > 1 \text{ (0.5 in R) } or\ D_i > 4/n$ is of concern.

Handling ‘Unusual’ Observations

- Always be true to the data
- Examine the observations before excluding them from the analysis
- If they are truly representative of the population, it is better to leave them in the dataset
- Compare the model estimates with and without outliers
- If the slope estimates change, then you might deal with influential points

Multicollinearity or Intercorrelation

- Refers to the situation when the predictor variables are (highly) correlated among themselves

- (Multi)Collinearity occurs when essentially the same variable is entered into a regression equation twice, or when two variables contain exactly the same information as two other variables

- Other causes of collinearity
  - One variable is a linear combination of several variables in the data
  - Perfect collinearity is usually ‘by mistake’

Effects of Collinearity

- Assume a linear regression with only one predictor:

$$E(Y|X) = \widehat{\beta_o} + \widehat{\beta_1}X_1$$

- Let us fit another regression with X_1 and $X_2 = 5 + 0.5X_1$

- The fitted response functions can be:

$$E(Y|X) = -87 + X_1 + 18X_2$$

OR

$$E(Y|X) = -7 + X_1 + 2X_2$$

Effects of Collinearity

- The fitted response functions have entirely different response surfaces, intersecting only on one line, which is:

$$X_2 = 5 + 0.5X_1$$

- Because the two variables are perfectly related implies that a unique solution might not exist
  - Design matrix $\boldsymbol{X}$ is not full rank, i.e., the columns are linearly dependent
  - Matrix $\boldsymbol{X}'\boldsymbol{X}$ is singular: $(\boldsymbol{X}'\boldsymbol{X})^{-1}$does not exist and a generalized inverse will be used instead: $(\boldsymbol{X}'\boldsymbol{X})^{-}$

Effects of Collinearity

- R example: generate a dataset containing the following variables:
  - Height (response variable)
  - Age
  - Age_copy = Age (purely a copy of Age)
  - Age_new = Age + small error

- Check the correlation/scatter plot matrix for all these variables

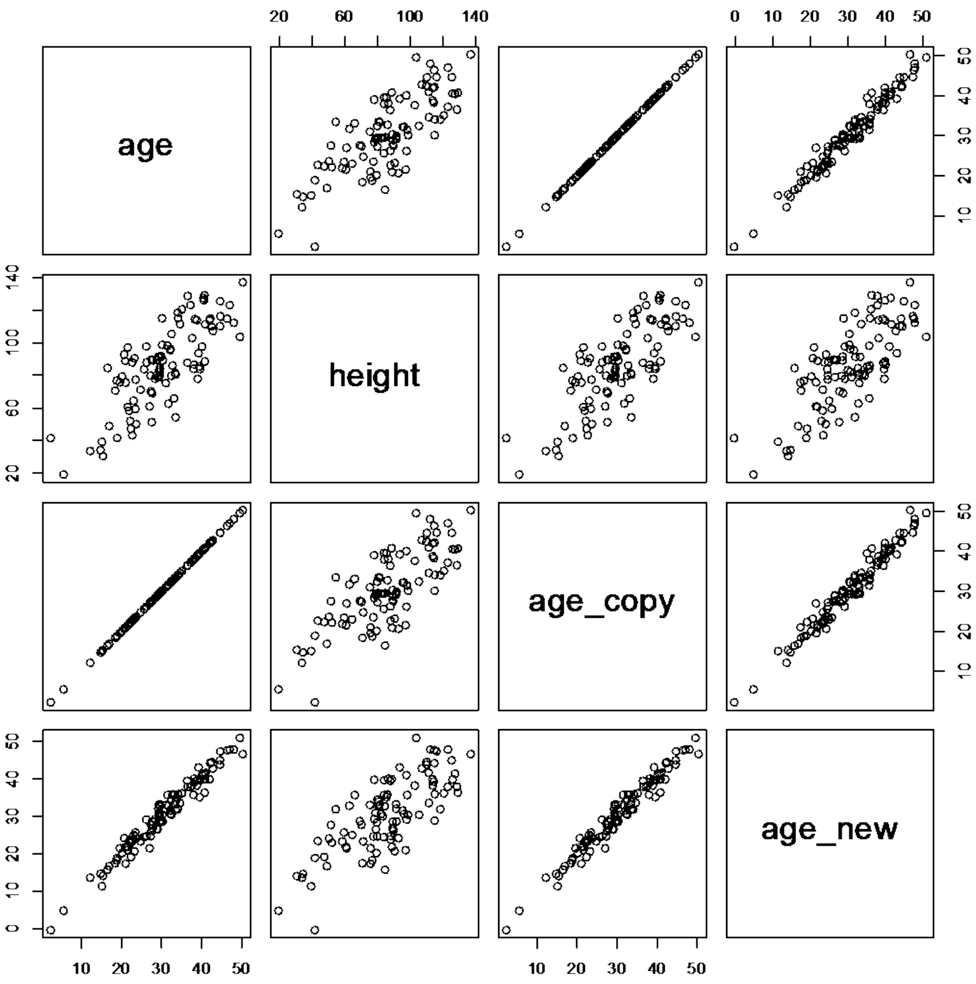
- Fit different regressions and interpret the results

Collinearity: R example

Correlation/scatter plot matrix for all 4 variables in the data

```
>cor(data.multi)
```

	age	height	age_copy	age_new
age	1.00	0.780	1.00	0.980
height	0.78	1.000	0.78	0.762
age_copy	1.00	0.780	1.00	0.980
age_new	0.98	0.762	0.98	1.000



Collinearity: R example

Regression 1:

```
>lm(formula = height ~ age, data = data.multi)
Coefficients:
              Estimate Std. Error  t value  Pr(>|t|)
(Intercept)  22.6639    5.3627    4.226     5.33e-05 ***
age            2.0772    0.1684  12.333  < 2e-16 ***
Residual standard error: 15.88 on 98 degrees of freedom
Multiple R-squared:  0.6082,  Adjusted R-squared:  0.6042
F-statistic: 152.1 on 1 and 98 DF,  p-value: < 2.2e-16
```

Collinearity: R example

Regression 2:

```
>lm(formula = height ~ age + age_copy, data = data.multi)
Coefficients: (1 not defined because of singularities)
            Estimate Std. Error  t value  Pr(>|t|)
(Intercept)  22.6639     5.3627    4.226   5.33e-05 ***
age           2.0772     0.1684   12.333  < 2e-16 ***
age_copy          NA         NA       NA       NA
Residual standard error: 15.88 on 98 degrees of freedom
Multiple R-squared:  0.6082,  Adjusted R-squared:  0.6042
F-statistic: 152.1 on 1 and 98 DF,  p-value: < 2.2e-16
```

Collinearity: R example

Regression 3:

```
>lm(formula = height ~ age + age_new, data = data.multi)
Coefficients:
             Estimate Std. Error  t value   Pr(>|t|)
(Intercept)  22.5716     5.4153    4.168    6.68e-05 ***
age           2.2215     0.8447    2.630    0.00994 **
age_new      -0.1409     0.8085   -0.174    0.86198
Residual standard error: 15.96 on 97 degrees of freedom
Multiple R-squared:  0.6083,  Adjusted R-squared:  0.6002
F-statistic: 75.32 on 2 and 97 DF,  p-value: < 2.2e-16
```

Collinearity: R example

Regression 4:

>lm(formula = height ~ age + age_copy + age_new, data = data.multi)

Coefficients: (1 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	22.5716	5.4153	4.168	6.68e-05 ***
age	2.2215	0.8447	2.630	0.00994 **
age_copy	NA	NA	NA	NA
age_new	-0.1409	0.8085	-0.174	0.86198

Residual standard error: 15.96 on 97 degrees of freedom

Multiple R-squared: 0.6083, Adjusted R-squared: 0.6002

F-statistic: 75.32 on 2 and 97 DF, p-value: < 2.2e-16

R example: Conclusions

- Error message and NAs when Age and Age_copy were modeled together
  - Note: not all software programs give error messages

- Adding Age_new (Regression 4) compared to Age only (Regression 1)
  - The estimated coefficient of Age had a small change, but the standard error was inflated
  - Age_new became non-significant
  - Minimal increase in R-squared, i.e., adding Age_new is redundant

Collinearity: General Conclusions

- If entered together in MLR, one or more correlated variables will become non-significant
- The magnitude/direction of the coefficients will change
- The standard errors will be inflated
    - ‘Flat’ SSR because one variable contains pretty much the same info as the other
- The coefficient of determination will have little increase

Identify Collinearity

- A simple approach is to use the variance inflation factor (VIF)
- A VIF for a single predictor is obtained using the R-squared of the regression of that variable against all other predictors:

$$VIF_j = \frac{1}{1 - R_j}$$

- A VIF is calculated for each of the predictors and variables with 'high' VIFs are removed
  - VIF > 5 suggest that the coefficients might be misleading due to collinearity
  - VIF > 10 implies serious collinearity

Remedies for Collinearity

- Drop one or several correlated variables from the model (which one?)
  - The non-significant variable
  - The variable with less missing data, if the case

- In polynomial regression, use the centered data for predictors

- Use principal components analysis (PCA), based on eigenvalues

- Use Ridge regression or other shrinkage/penalized methods

Readings

Kutner *et al.*, Applied Linear Statistical Models

- Chapter 10, Sections: 10.2 – 10.5