Make sure Zoom is Recording, Professor K!

Probability Theory and Introductory Statistics

- ALY6010, Section 81229
- Week 6: June 26, 2022
- Professor Dan Koloski (Professor K)
- d.koloski@northeastern.edu
- Roux Institute at Northeastern University

Agenda: 6/26/2022

Time Slot	
6:00-6:15 (15min)	Recap of last week; Quiz Update; Breakouts to Review Bluman Problems
6:15-6:55 (40min)	Section 1: Coefficient of Determination and Standard Error of the Estimate
6:55-7:00 (5min)	Break
7:00-7:40 (40min)	Section 2: Regression Analysis for Categorical Variables and Regression Subset Analysis
7:40-7:45 (5min)	Break
7:45-8:35 (40min)	Section 3: Review Homework; ALY6010 Wrap-Up; Open Discussion
8:35-8:40 (5min)	Wrap-Up

Week 5 Objectives and Task List

Learning Objectives

By the end of this module, you should be able to:

- Assess association between two or more variables
- Calculate covariance, correlation and regression coefficient
- Create scatter plots and line plots in a statistical tool
- Create correlation and regression tables in a statistical tool

Task List

- View Lessons in Canvas
- Read Elementary Statistics, Sections 10.1 & 10.2
- Read R in Action, Chapters 7, 11.
- Complete **primary Discussion post by Thursday (2 secondary by Saturday)**
- Complete Practice Problem Set (not submitted)
- Complete **R Practice assignment by Sunday**
- Take **quiz by Sunday**
- Continue work on Final Project

Breakouts: Week 5 Problem Sets

- In your Zoom Breakout Rooms
  - Share your Blumen assignments and answers with each other

- Identify any outstanding questions the group has and bring back to the larger group

Week 6 Objectives and Task List

Learning Objectives

By the end of this module, you should be able to:

- Use continuous and categorical variables with linear regression
- Apply basics of model selection
- Perform testing of differences using linear regression
- Perform regression diagnostics

Task List

- View Lessons in Canvas
- Read Elementary Statistics, Section 10.3
- Read R in Action, Chapters 7, 11.
- Complete **primary Discussion post by Thursday, 2 secondary by Saturday**
- Complete Practice Problem Set (not submitted)
- **Complete R Practice assignment by Saturday**
- **Submit Final Project by Saturday**

Section 1

Section 1: Coefficient of Determination and Standard Error of the Estimate

(Bluman 10.3)

Review: Correlation → Regression → Fit

- If correlation coefficient *r* is significant, the equation of the regression line can be determined
  - Then, for various values of the independent variable X, the dependent variable Y can be predicted using regression

- Regression on x and actual y, creates a set of predicted values y'
  - y' are the predicted values of y using the regression line equation

- The closer *r* is to -1 or +1, the better the fit will be, and the y' values will be closer to the actual y values

Deviations/Variations Between (Predicted) *y'* and (Actual) *y*

- Total Variation = sum of squares of the difference between y and mean(y)

$$\boldsymbol{Total\ Variation} = \sum (\boldsymbol{y} - \overline{\boldsymbol{y}})^{\boldsymbol{2}}$$
$$= \boldsymbol{Explained\ Var.} + \boldsymbol{Unexplained\ Var.}$$

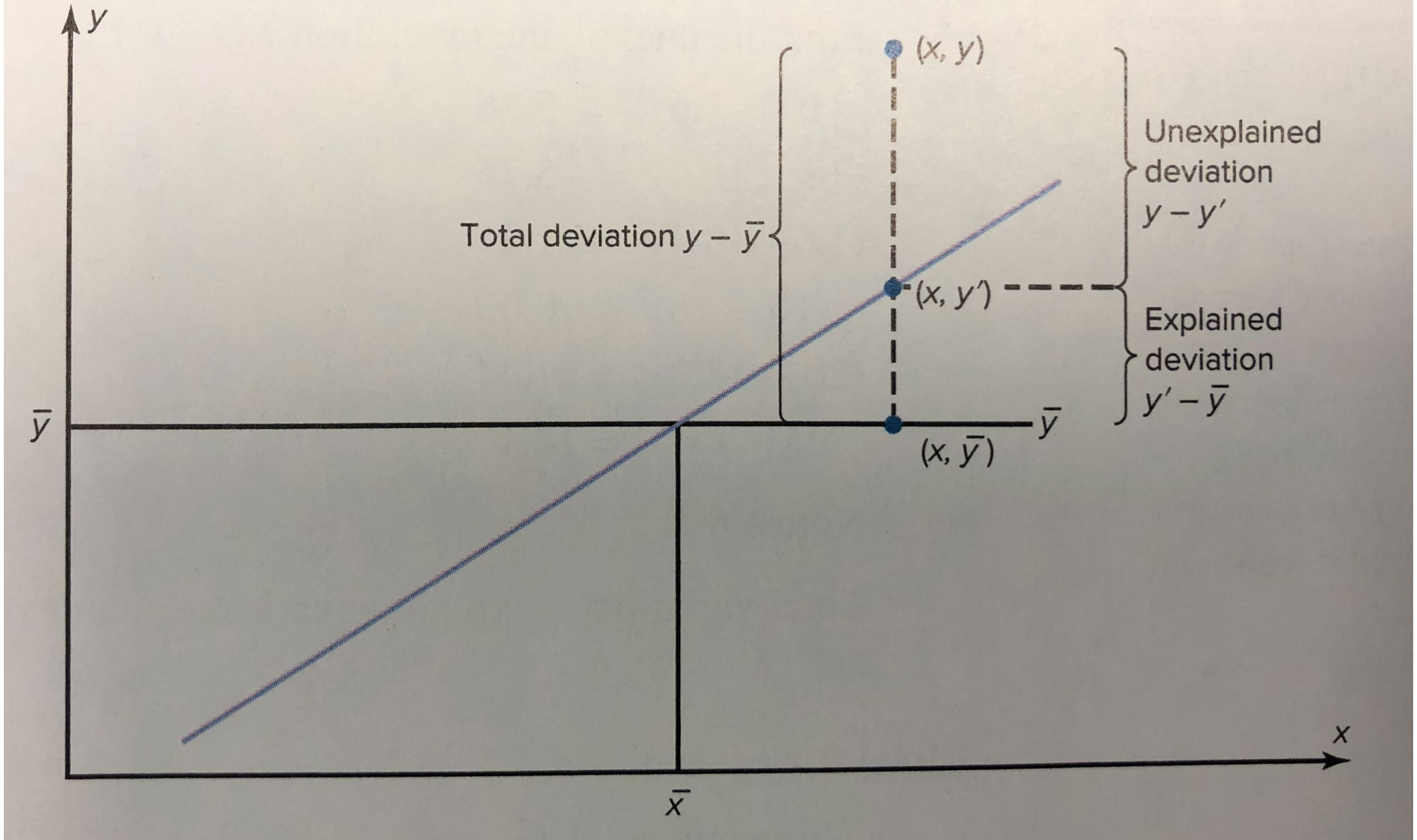
- Explained Variation = portion of total variation due to the relationship between X and Y

$$\boldsymbol{Explained\ Variation} = \sum (\boldsymbol{y'} - \overline{\boldsymbol{y}})^{\boldsymbol{2}}$$

- Unexplained Variation = portion of total variation due to chance

$$\boldsymbol{Unexplained\ Variation} = \sum (\boldsymbol{y} - \boldsymbol{y'})^{\boldsymbol{2}}$$

Section 10–3 Coefficient of Determination and Standard Error of the Estimate

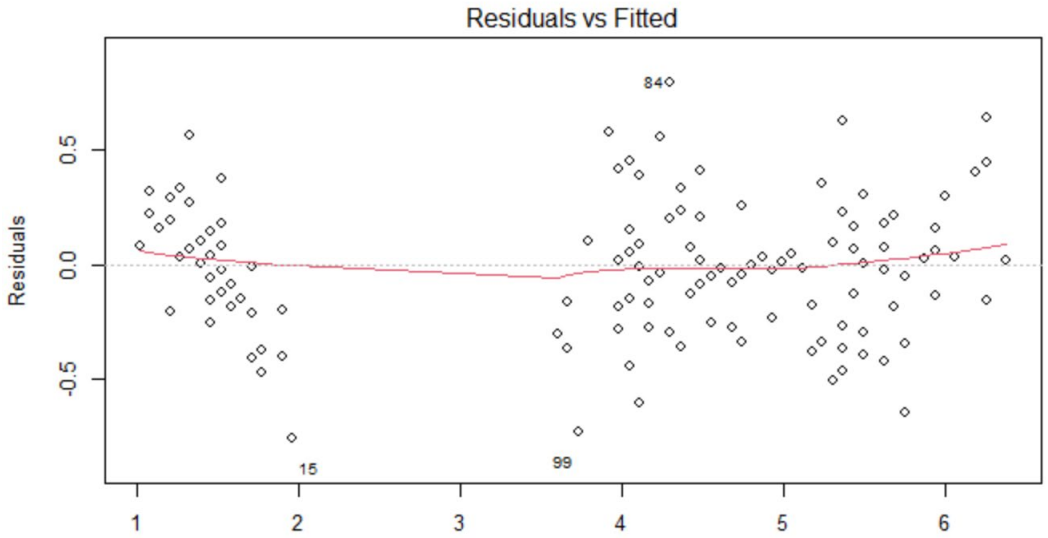


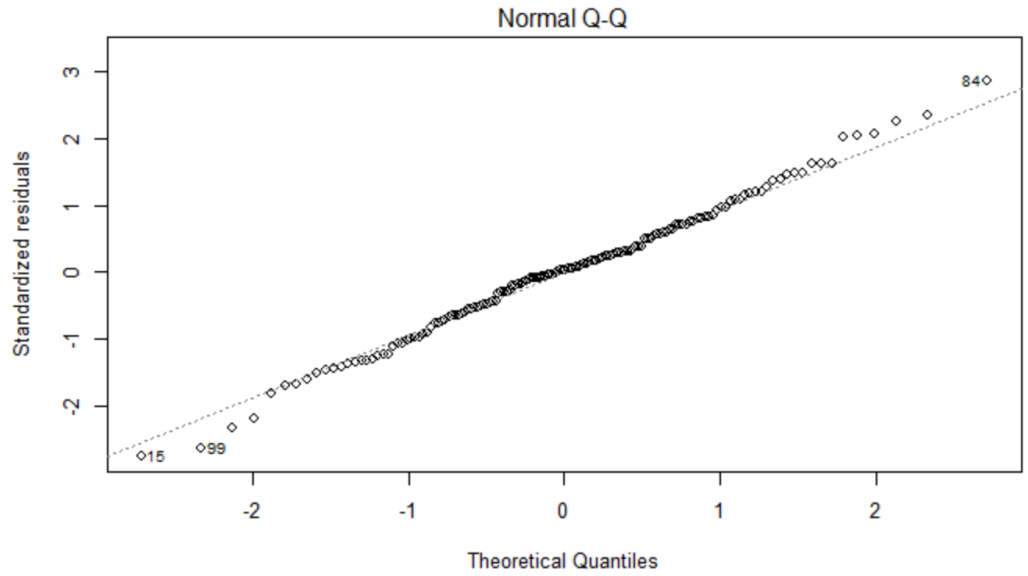
Bluman, 583.

Residuals

- Residuals = differences between actual observed and predicted y values
  - $for\ each\ X,\ the\ residual = y - y'(predicted)$

- Residuals can be plotted to show how accurate the regression equation is for predictions





Coefficient of Determination

- Coefficient of Determination (a.k.a. r^2)
  - Ratio of explained variation to total variation
  - **Measures how much of the dependent variable variation is explained by the regression line and the independent variable versus chance**
  - Will be a value between 0 and 1 inclusive, expressed as a percentage

$$r^2 = \frac{Explained\ Variation}{Total\ Variation}$$

Interpret as the percentage of y's variation that is explained by x

- An R^2 of 0 means that 0% of y's variation that is explained by x
- An R^2 of 0.5 means that 50% of y's variation that is explained by x
- An R^2 of 1 means that 100% of y's variation that is explained by x

  - (and yes, $r^2 = r * r$)

- Coefficient of Nondetermination $= 1.00 - r^2$

Bluman Example in Section 10-3 (p. 580)

- Given a certain set of data
  - find Total, Explained and Unexplained Variation
  - Plot the residuals
  - Determine Coefficient of Determination

x	1	2	3	4	5
y	10	8	12	16	20

Standard Error of the Estimate s_{est}

- The Standard Error of the Estimate or s_{est} is used primarily as a means to calculate a prediction interval about the y' values within which actual y is likely to fall, at a given alpha

- The Standard Error of the Estimate, or s_{est} = the standard deviation of observed y values about the predicted y' values

$$s_{est} = \sqrt{\frac{\sum(y - y')^2}{n - 2}}$$

Prediction Interval About y’

- Using the standard formula format:
  - (obs. value – standard error) < exp. value < (obs. value + standard error)

$$y' - t_{\frac{\alpha}{2}} * s_{est} * \sqrt{1 + \frac{1}{n} + \frac{n(x - \bar{X})^2}{n\sum x^2 - (\sum x)^2}}$$

$$< y <$$

$$y' + t_{\frac{\alpha}{2}} * s_{est} * \sqrt{1 + \frac{1}{n} + \frac{n(x - \bar{X})^2}{n\sum x^2 - (\sum x)^2}}$$

To use this technique:

- Set degrees of freedom as n-2

Bluman Example in Section 10-12/14 (p. 585)

- Given a certain set of data about copy machine age and years
  - Find the standard error of the estimate
  - Find the 95% prediction interval for the monthly maintenance cost of a machine that is 3 years old

Machine	Age in Years	Monthly Cost $
A	1	$62
B	2	78
C	3	70
D	4	90
E	4	93
F	6	103

[Break]

Section 2

Section 2: Regression Analysis for Categorical Variables and Regression Subset Analysis

Regression With Categorical Variables

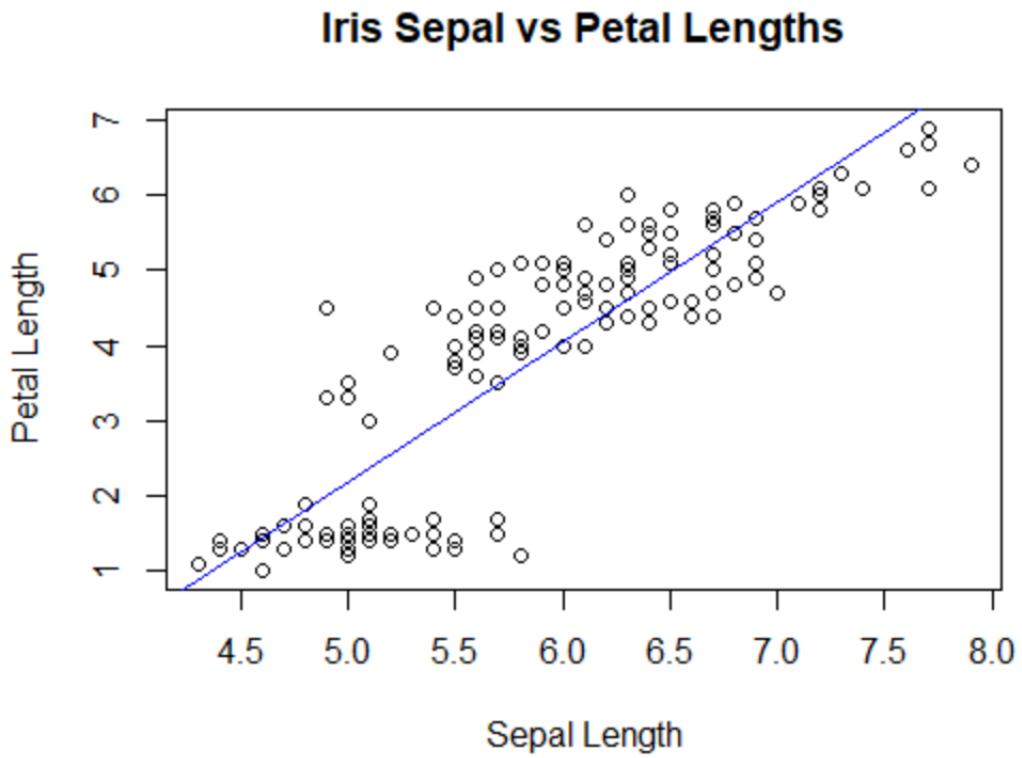
- Categorical variables can affect the value of a dependent variable
  - Example (cats): Regression of Hwt vs Bwt … differentiated by Gender

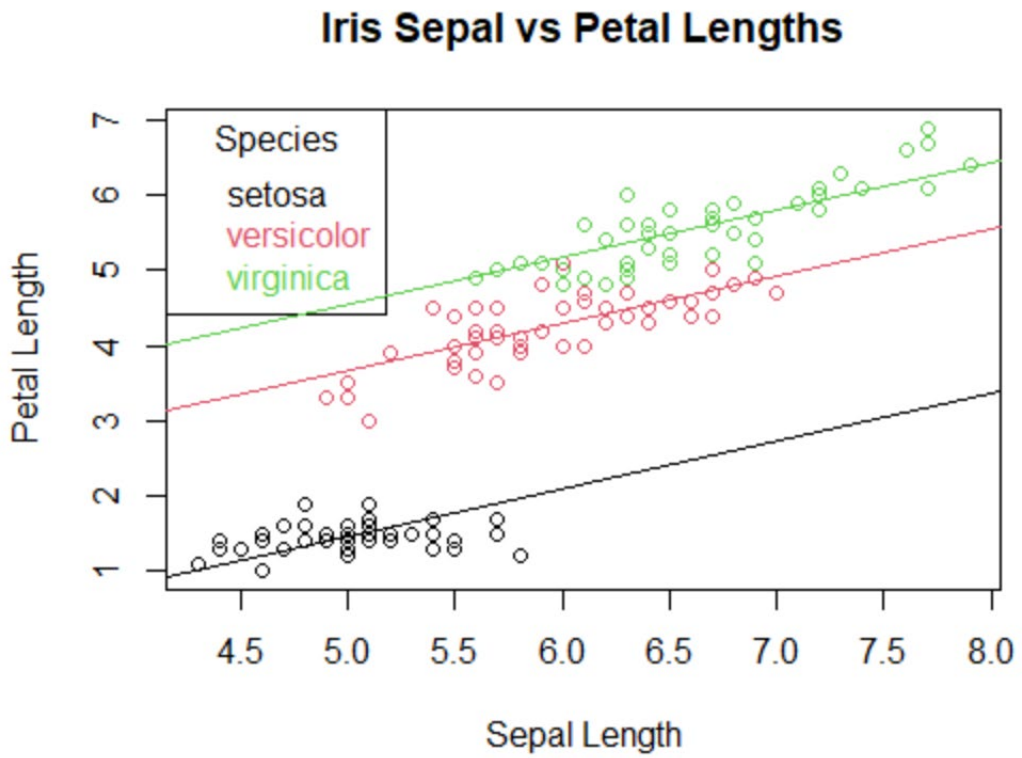
- To calculate and display separate regressions by category, we add "Dummy Variables" to individual regression lines for each category
  - Dummy Variable: conversion of categorical variable into a numeric variable
    - (Similar to how R calculates factors)
    - Provides an additional increment to y-intercept by category for display purposes. The slope of the regression lines stay the same
    - Dummy variable contains 2 numbers, 0 and 1. 0="Excluded category"

Using Dummy Variables

- Linear regression line: $\boldsymbol{y' = a + bX}$
- With dummy variables: $\boldsymbol{y' = a + bX + a_1D_1 + a_2D_2 + a_3D_3, \ldots}$
  - $\boldsymbol{for\ each\ D_i, either\ D_i = 0\ for\ the\ excluded\ category}$
    $\boldsymbol{or}$
    $\boldsymbol{D_i = 1\ for\ the\ other\ catgeory}$
  - For each category, create a new $\boldsymbol{D_i}$ comparing to the base/excluded category
  - For each category, crate a new $\boldsymbol{a_i}$, which is the difference in the y-intercept calculation for each category
- Procedure if you are doing it manually (R does this very easily for you)
  - Convert categories to dummy or indicator variables (0 or 1).
  - Indicator variables are special case of categorical variables with two levels (0 or 1)
  - Use k-1 indicator variables to represent k categories, where one variable will have 0 value for all dummy variables
  - The resulting regression lines all have the same slope, but different y-intercepts

R Examples Using Dummy Variables

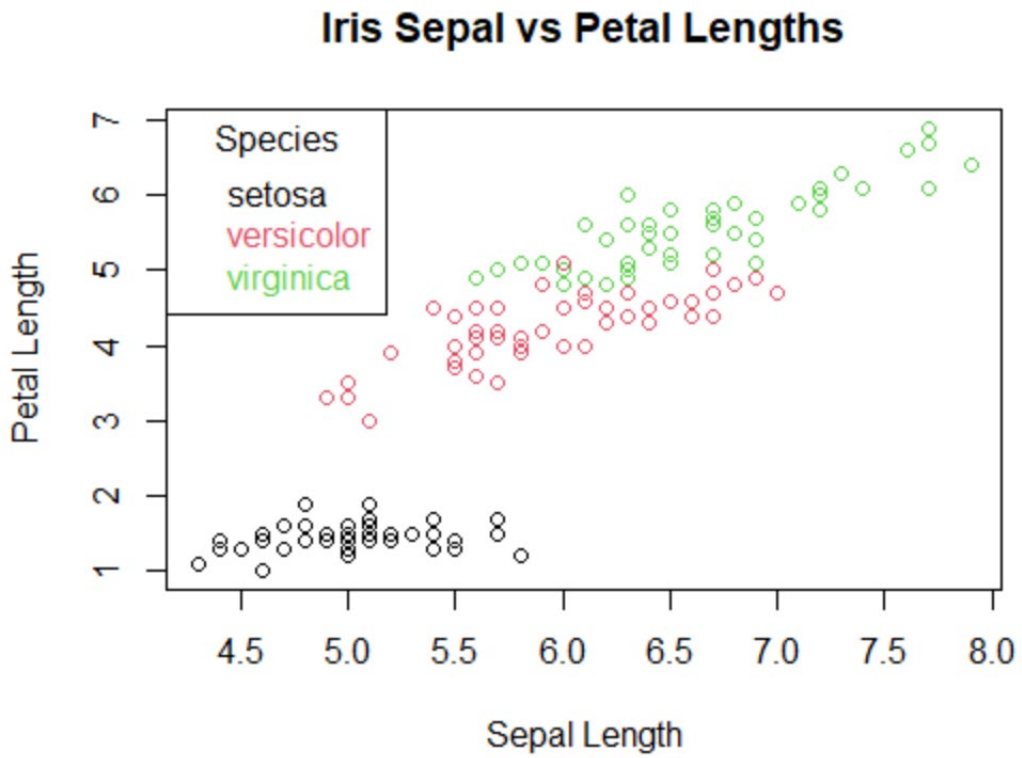


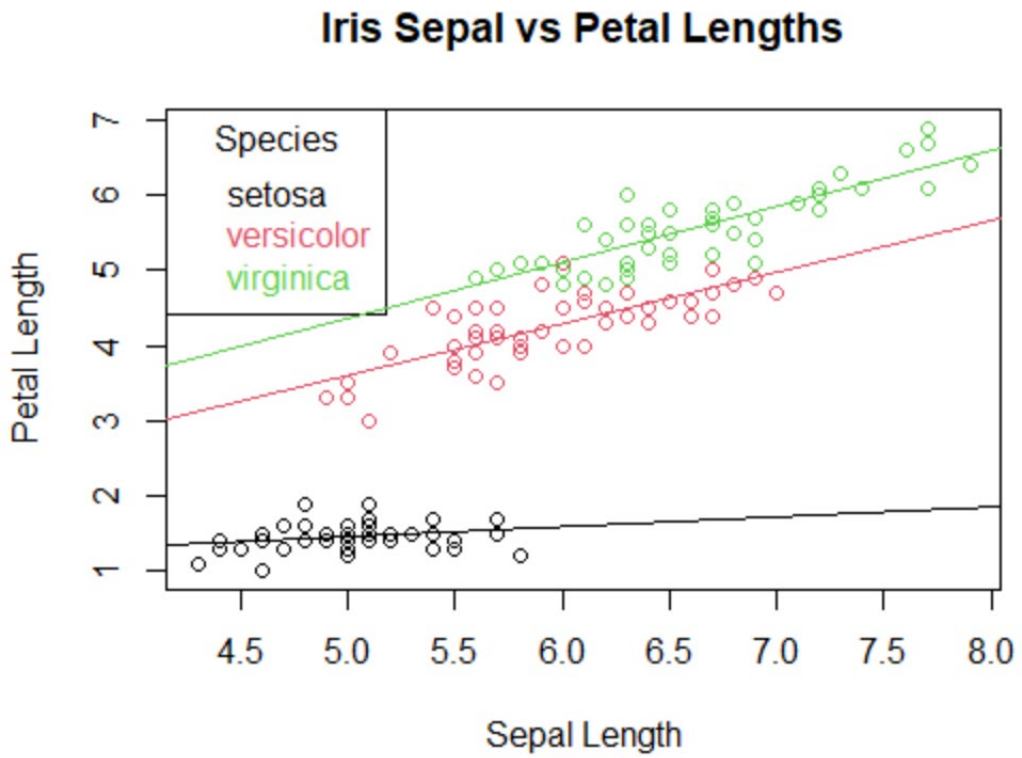


Linear Regression Using Data Subsets

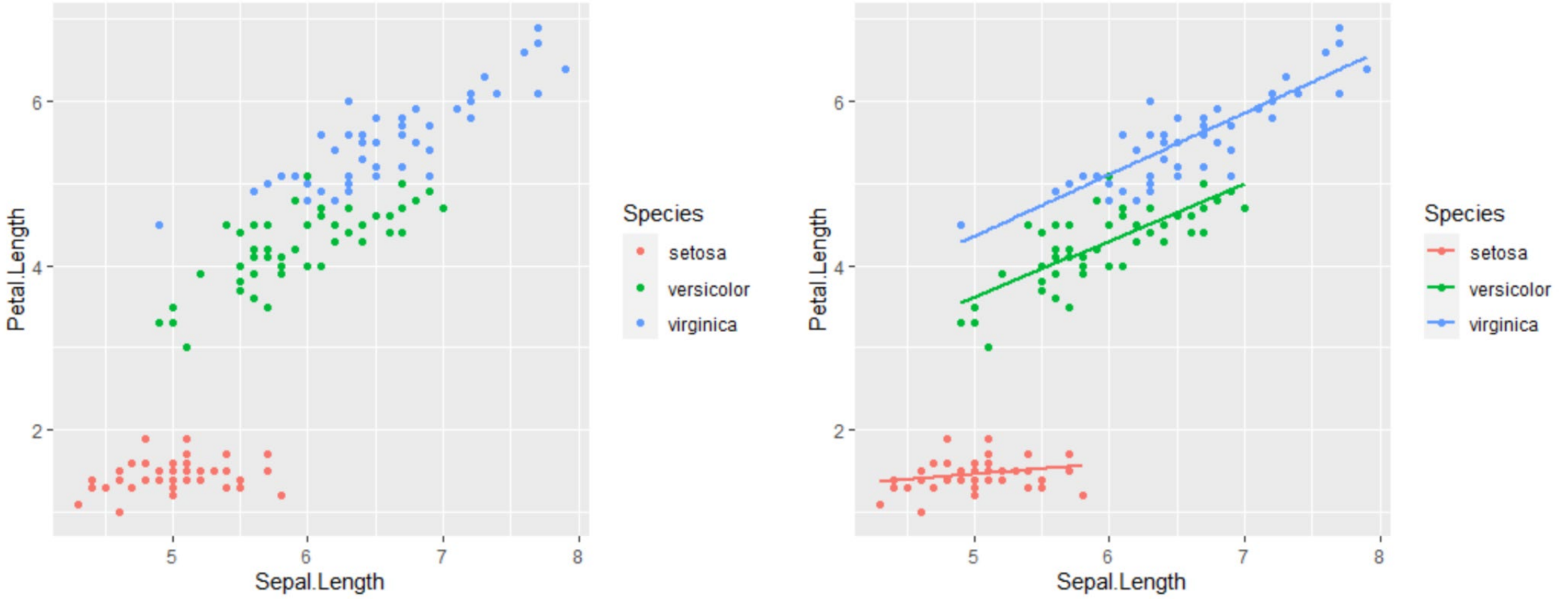
- Subset analysis is an analytical technique to subset/partition a data set prior to linear regression
- Used when doing regression across the whole data is not meaningful
  - Maybe there's too much dispersion/multi-modal data.
  - One can drop as outliers, but that removes potentially large parts of the data set
- Possible to get a better model if we separate our data into individual subsets and do a separate linear regression for each subset in isolation
  - Results in multiple lines with different slopes
    - Ex
      - $y' = \beta_1 + \beta_2 X + \varepsilon$, for first category
      - $y' = \alpha_1 + \alpha_2 X + \varepsilon$, for second category
      - $y' = \gamma_1 + \gamma_2 X + \varepsilon$, for third category (note γ = Greek "gamma")

R Examples Using Data Subsets



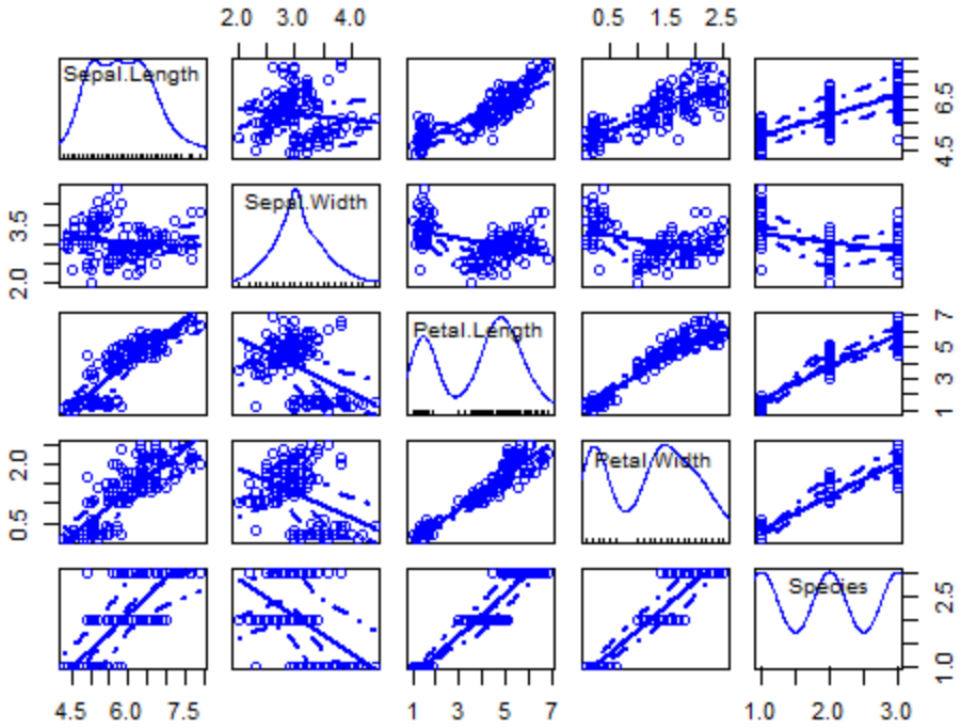


R Examples Using Data Subsets



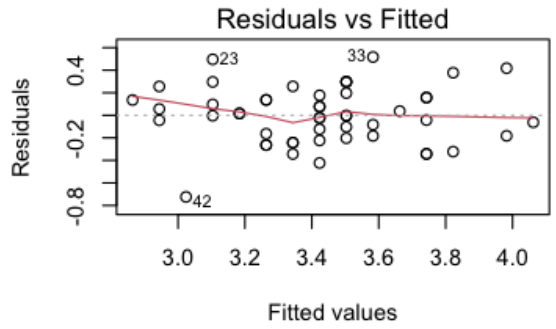
Regression Diagnostics in R

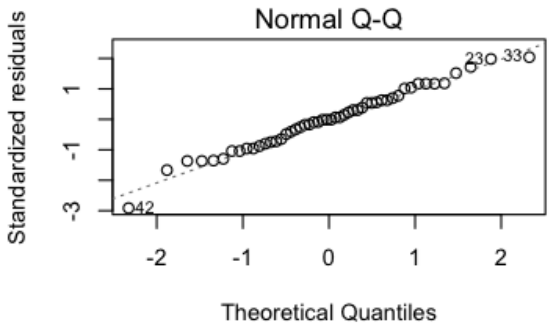
- Regression Diagnostics help you understand how useful/accurate your regression might be
  - …and whether to use subsets/dummy variables or a different (non-linear?) model
- **`plot(yourlmmodel)`** when used with **par(mfrow=c(2,2))**
  - Primarily examines behavior of residuals
- **`scatterplotMatrix()`** from "car" package
  - Quick pairwise scatterplots with lm() and non-linear models



Does our model satisfy the assumptions of linear regression?

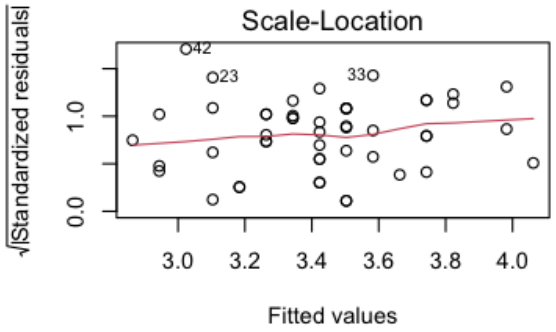
Linearity Assumption
Ideally all points fall equally above/below 0 line with no apparent pattern.

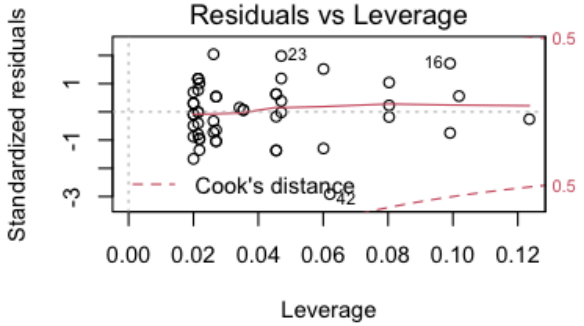




Normality Assumption
All points should fall on dotted line

Equal (Constant) Variance Assumption
Ideally all points fall in a random band around a horizontal line (Homoscedasticity); failure to do so is Heteroscedasticity





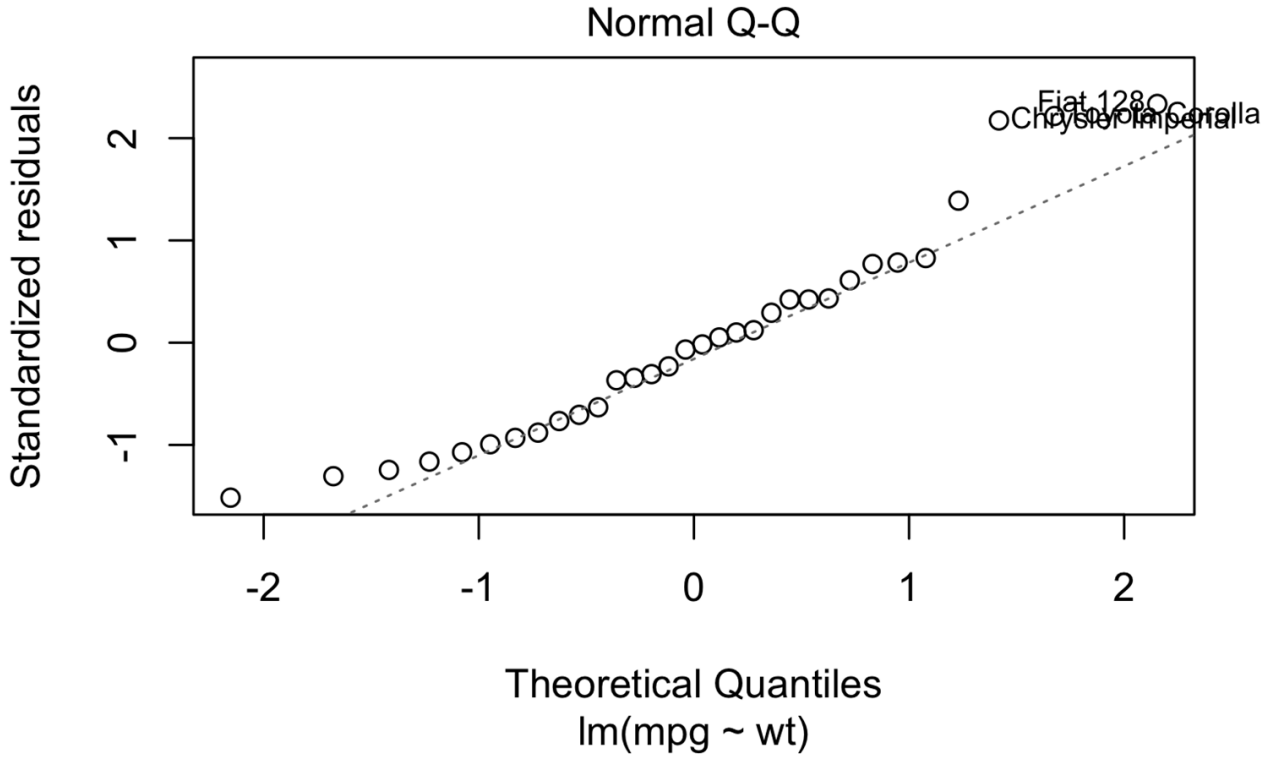
Identifying
Outliers: large residual
High-leverage points: large leverage
Influential observations: Beyond Cook's distance

```
par(mfrow = c(2,2)) # make a 2x2 grid of plots
plot(fit) # will by default give the 4 key graphs
```

Reference reading: Kabacoff Unit 8.3 Regression Diagnostics

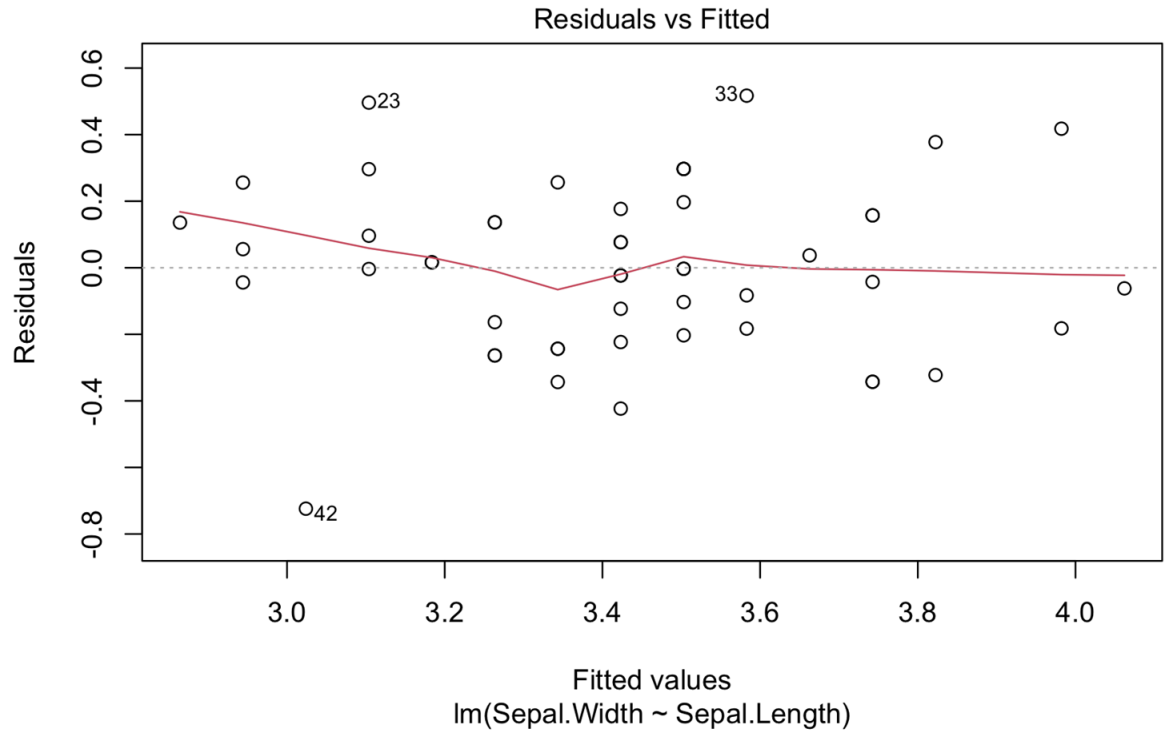
Q-Q Plot to check for normality

- X-axis is where the data would lie if it were perfectly normally distributed
- Y-axis is there the data really lies
- If points fall on a straight line, the data is normally distributed



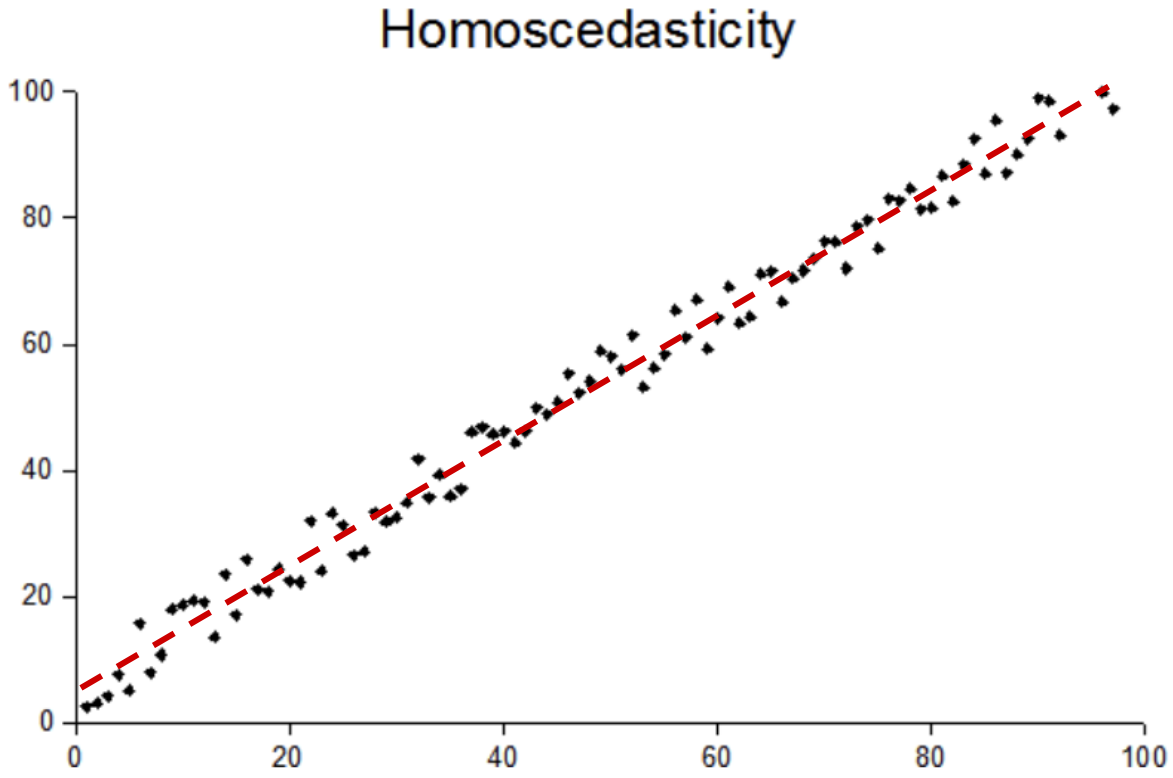
Linearity Assumption

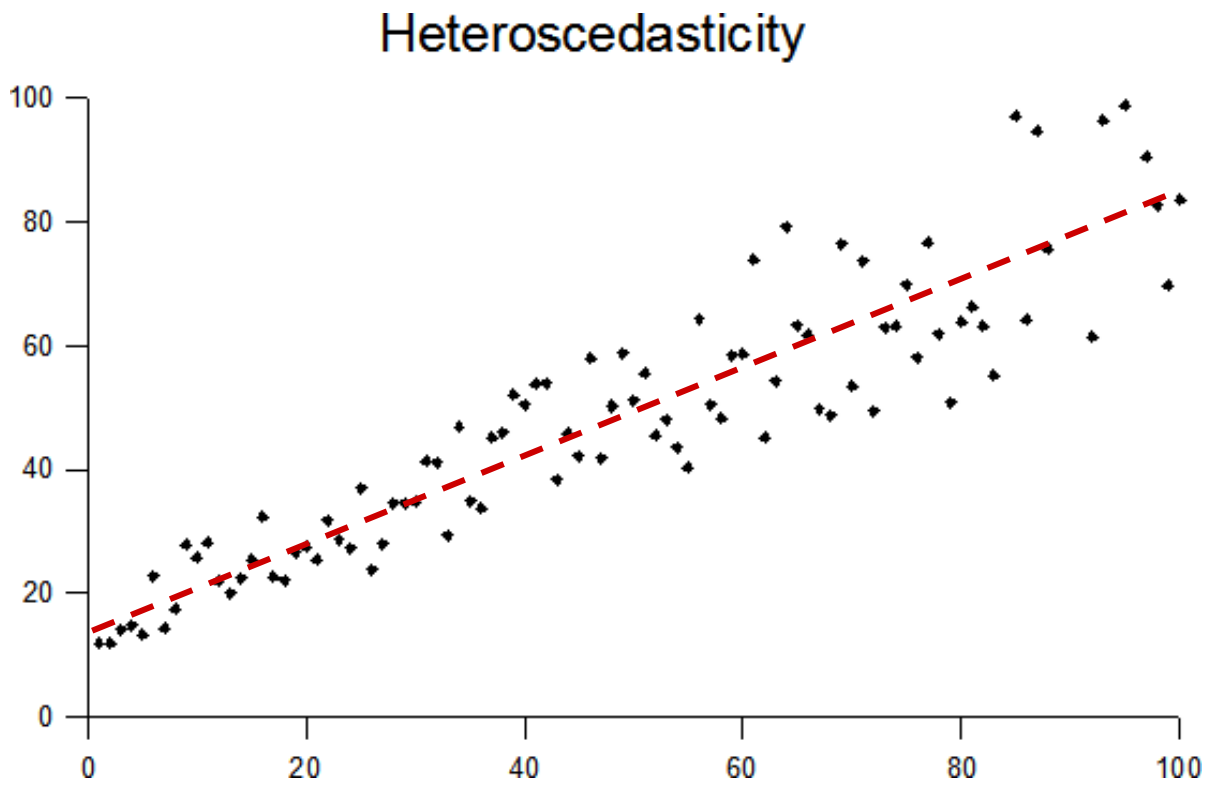
- There shouldn’t be a trend in residuals vs fitted values
  - Red line should be horizontal
  - There should be no visible pattern in the data



Homoscedasticity (equal variance) assumption

- homoscedasticity means “having the same scatter” [1]

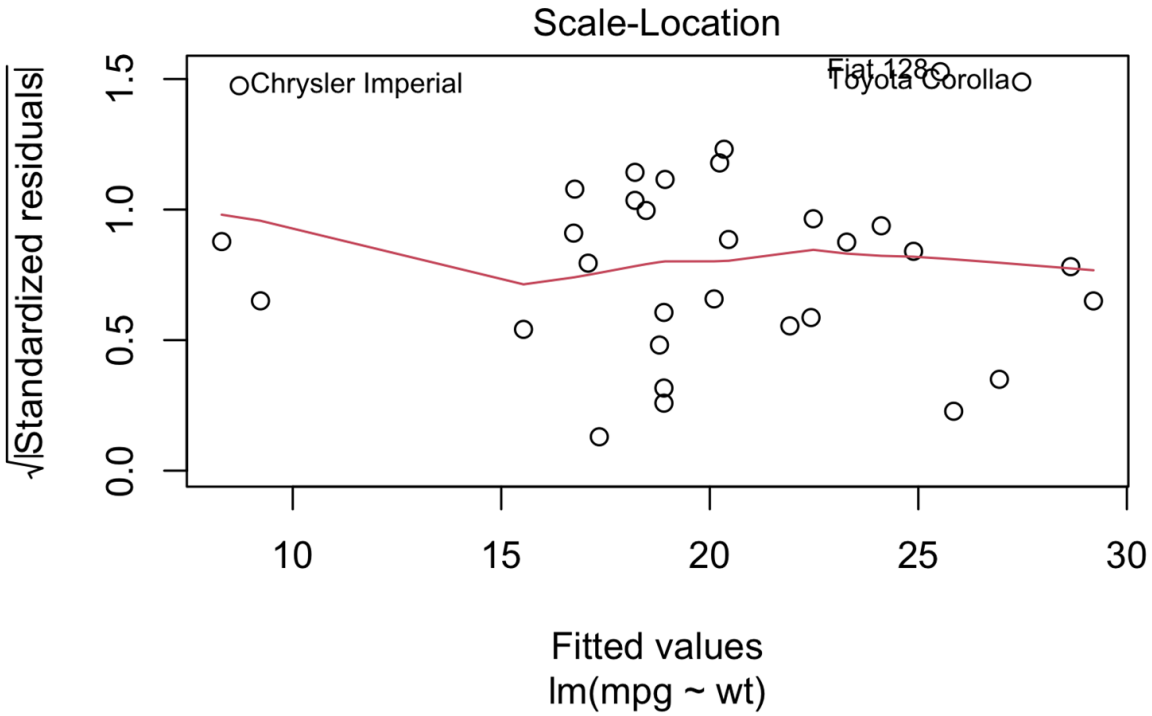




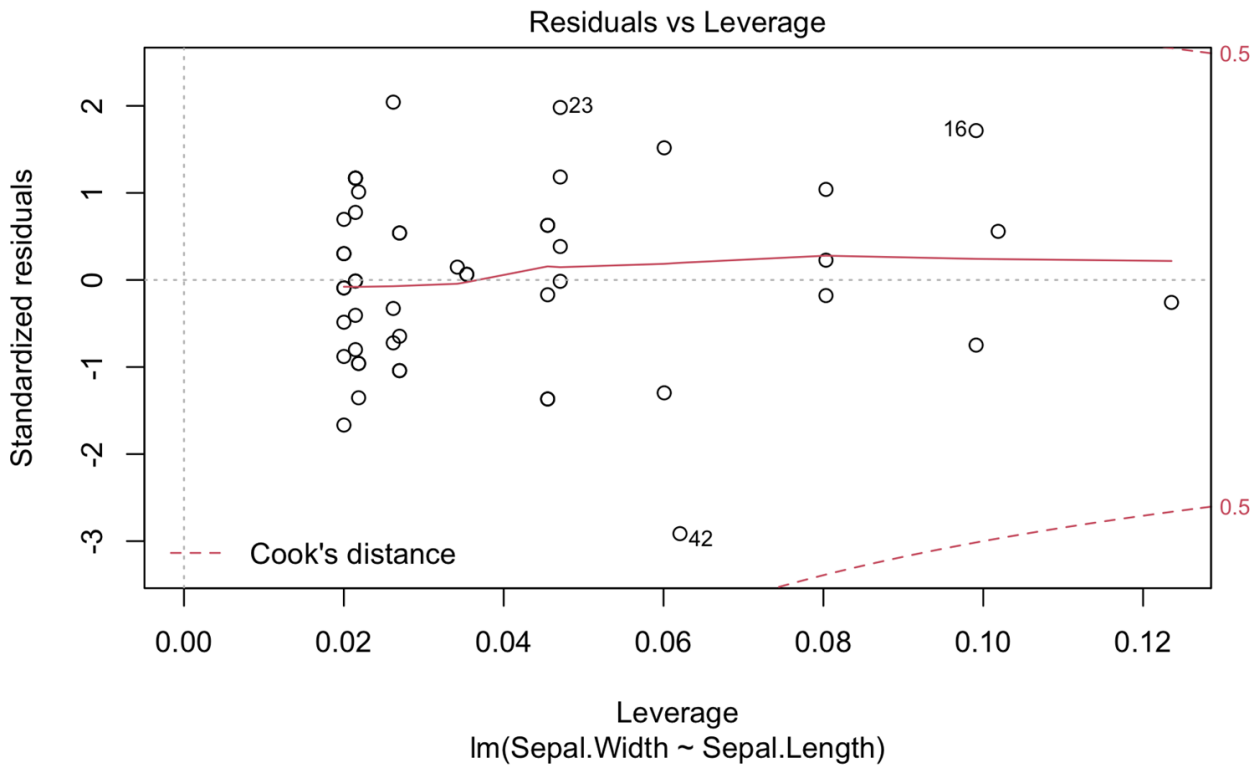
[1] https://www.statisticshowto.com/homoscedasticity/

Scale-location Plot: check for homoscedasticity

- shows if residuals are spread equally along the ranges of predictors
- Ideally, the line should be horizontal with evenly distributed points



High Leverage Points



- Outliers - data point whose response (y) does not follow the general trend of the rest of the data
- High Leverage - the point's x-value is extreme
- Influential Points - unduly influences the regression analysis
  - Outliers and high leverage data points are not always influential
- The 3 most extreme leverage points are highlighted (23, 12, 42)
- You want all of your points to be inside the Cook's distance lines
  - It's good if they don't show up on the plot - that means the points are well below Cook's distance

https://online.stat.psu.edu/stat462/node/170/

Multivariate Regression and Adjusted $\boldsymbol{r^2}$ (r^2_{adj})

- *(See Bluman 10-4: not covered in required course material)*
- Multivariate Regression measures the *effect of several independent variables* on a dependent variable
  - $\boldsymbol{y' = a + bX + b_1X_1 + b_2X_2 + b_3X_3, \ldots}$
  - Equation for R among 2 ind./1 dep. Vars: $R = \sqrt{\frac{r^2_{yx(1)} + r^2_{yx(2)} - 2r_{yx(1)} * r_{yx(2)} * r_{x(1)x(2)}}{1 - r^2_{x(1)x(2)}}}$
- Statistical tools will also often output "Adjusted R^2 ", which calculates the value of on the basis of degrees of freedom (n-k-1) instead of just the number of ordered pairs (n-1)
  - R^2_{adj} is especially useful in multivariate regression
- Syntax in R for multivariate regression: **`lm(y ~ x1 + x2 + x3 + …)`**

[Break]

Section 3

Section 3: Review Homework; ALY6010 Wrap-Up; Open Discussion

Module 6 R Practice: *mtcars* (DUE FRIDAY)

- Using an appropriate variable, create dummy variables to subset your dataset. Then rerun your regression line for your dependent variable. How many subsets did you create? How many lines are there? Create a scatterplot with multiple regression lines. How does this impact your understanding of the impact of the categorical variable on the regression?
- Using the appropriate subsetted data from step 1, create separate regression lines for each subset. How do these regression lines differ from the regression lines in step 1? How does this method of looking at the data impact your understanding of the data?

Final Project (DUE FRIDAY)

Assignment Overview

- Now that you have completed your initial EDA and basic hypothesis testing for specific variables, it's time to look deeper into the data to examine relationships between variables. You will also combine all the information you gathered about the data and write a final report.

Instructions

- Using the dataset and what you learned about the data in Assignment 1, identify at least 2 -3 questions you have about the relationships between variables in the data. These questions should employ inferential statistics and regression testing to find answers. Then find the answer to the questions using hypothesis testing. Be sure that you employ scatterplots and linear regression.
- For each test, document the dependent and independent variable, hypothesis testing steps and the results at each step. Finally, provide an analysis of the final results with an explanation of your interpretation.

What to Submit

You must submit a 5-7 page report that includes:

- A summary of your initial EDA
- All the questions you explored (link these to your initial EDA. Why did you ask these questions?)
- Your null and alternative hypotheses (including an explanation of how the hypothesis test was constructed)
- Explanation of the hypothesis testing you completed
- The results
- Your interpretation of the results

Also submit your R code.

Week 6 Objectives and Task List

Learning Objectives

By the end of this module, you should be able to:

- Use continuous and categorical variables with linear regression
- Apply basics of model selection
- Perform testing of differences using linear regression
- Perform regression diagnostics

Task List

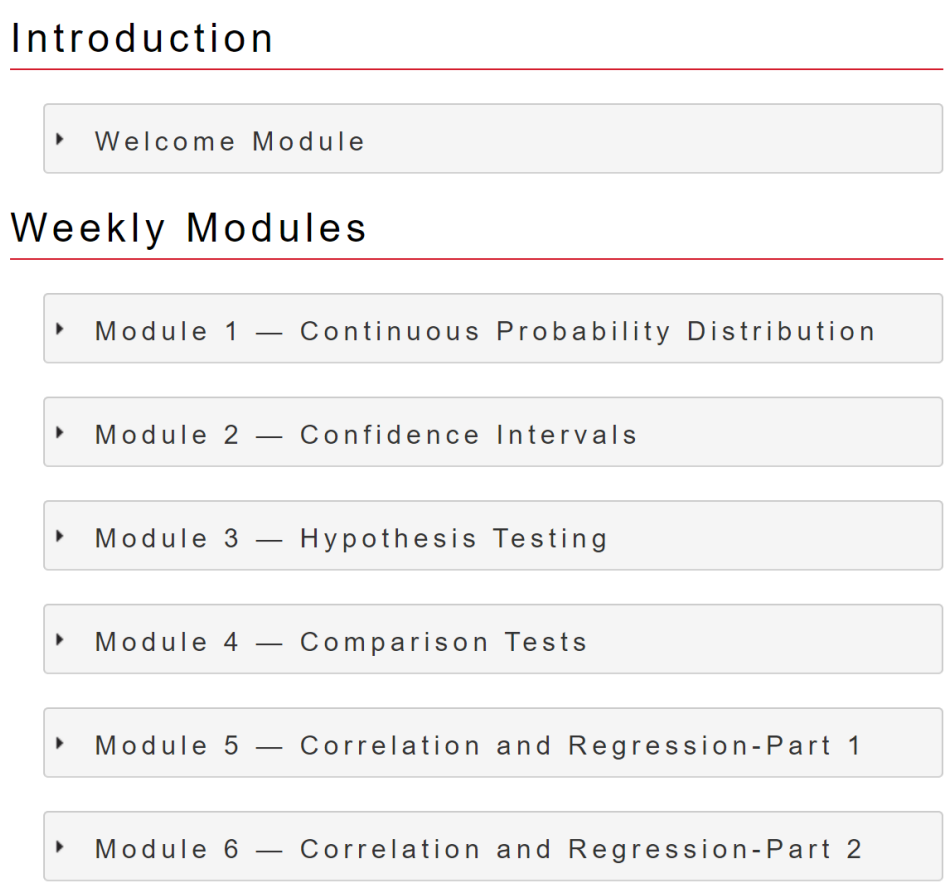
- View Lessons in Canvas
- Read Elementary Statistics, Section 10.3
- Read R in Action, Chapters 7, 11.
- Complete **primary Discussion post by Thursday, 2 secondary by Saturday**
- Complete Practice Problem Set (not submitted)
- **Complete R Practice assignment by Friday**
- **Submit Final Project by Friday**

Wrap Up

Course Wrap-Up

Where We Started

- CLO1: Develop **strategic and operational questions based on the data** and the need of the organization
- CLO2: Use **data analysis techniques (hypothesis testing, correlation, t-testing, variance, and single variable linear regression)** to answer research questions
- CLO3: **Use "R"** to perform computations and a wide range of statistical techniques
- CLO4: **Interpret the results** from data analysis techniques
- CLO5: **Explain the implications** of the results from data analysis for the purpose of answering essential business questions



Feedback Survey – Very Important

- At the end of this course, please take the time to complete the evaluation survey at https://neu.evaluationkit.com.
- Your survey responses are **completely anonymous and confidential**.
- For courses 6 weeks in length or shorter, surveys will be open one week prior to the end of the courses; for courses greater than 6 weeks in length, surveys will be open for two weeks.
- **An email will be sent to your HuskyMail account notifying you when surveys are available.**

Thank you for a great course!

Appendix: Regression Steps

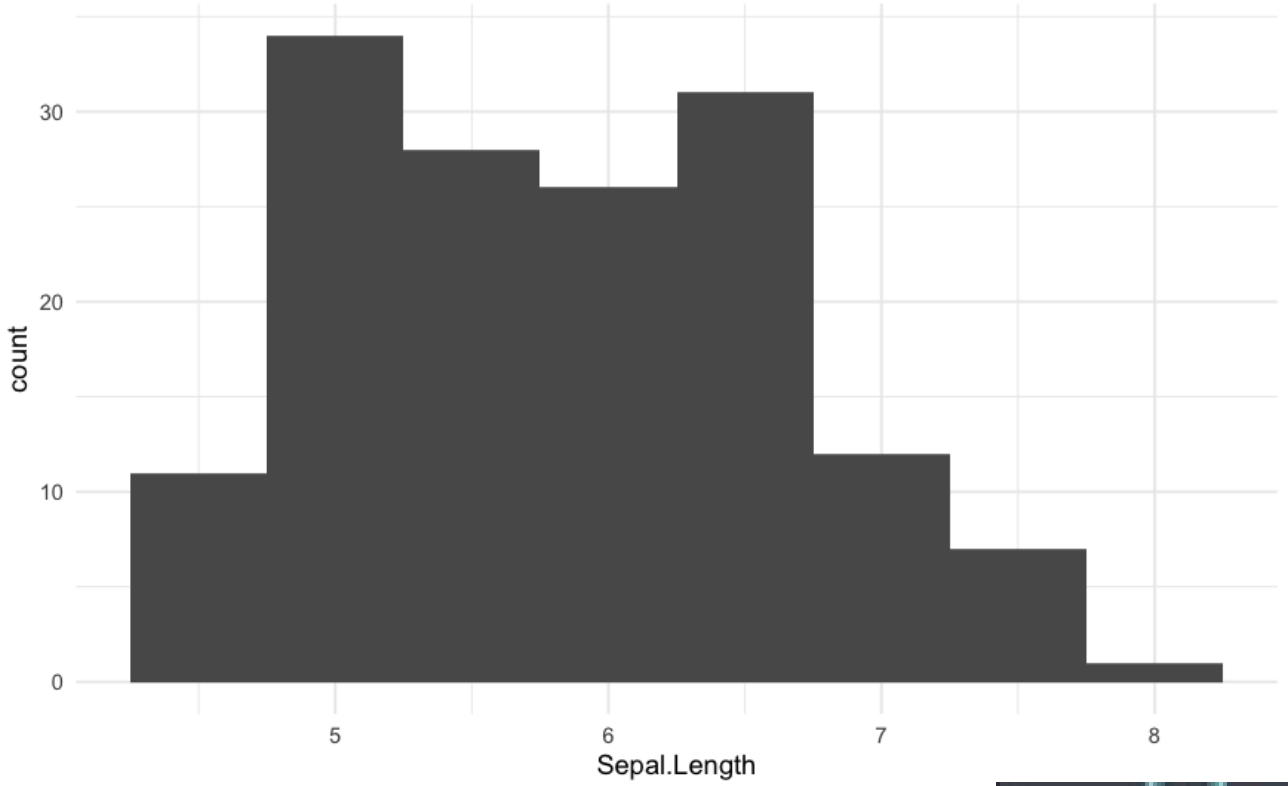
Excellent Step-By-Step Walkthrough of Regression Diagnostics

h/t to Professors Joy-El Talbot and Berkeley Almand-Hunter

Steps in analyzing potential linear relationships

1. (if needed, eg. with time data) Summarize data to get to numeric variables
2. Create histograms of numeric variables
3. Create scatterplots of combinations of numeric variables
4. Identify potential linear relationships
5. Calculate correlation coefficient (r)
6. Test significance of linear correlation coefficient
7. Create linear regression model ($y' = a + bx$)
8. Calculate coefficient of determination (r^2)
9. (optional) Make point predictions with prediction intervals ($y = y'$ +/- prediction interval).
10. Interpret results in context

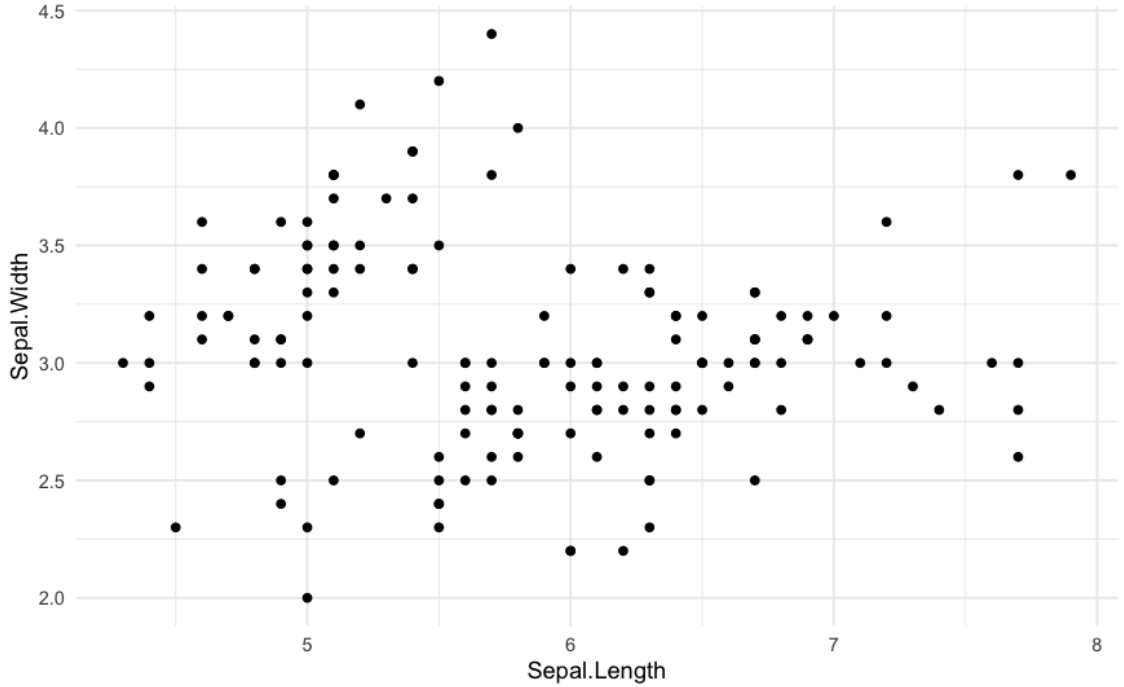
1. Create histograms of numeric variables



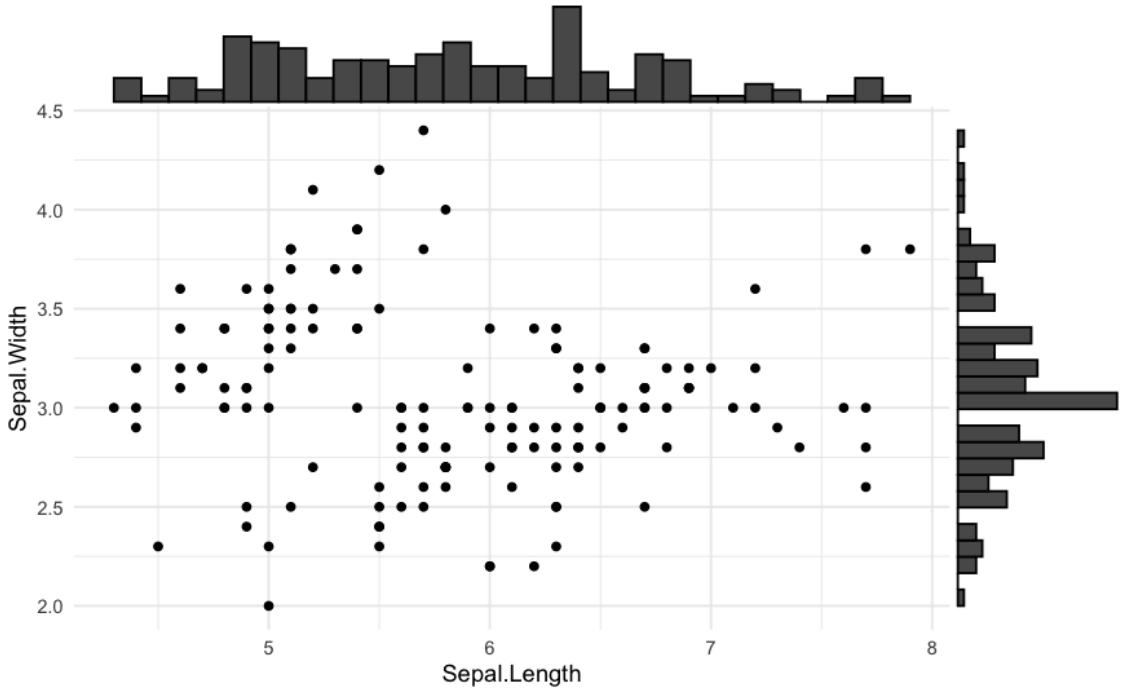

```
ggplot(iris, aes(x = Sepal.Length)) +
  geom_histogram(binwidth = 0.5)
```

3. Create scatterplots of combinations of numeric variables

```
p1 <- ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width)) +
  geom_point()
p1
```



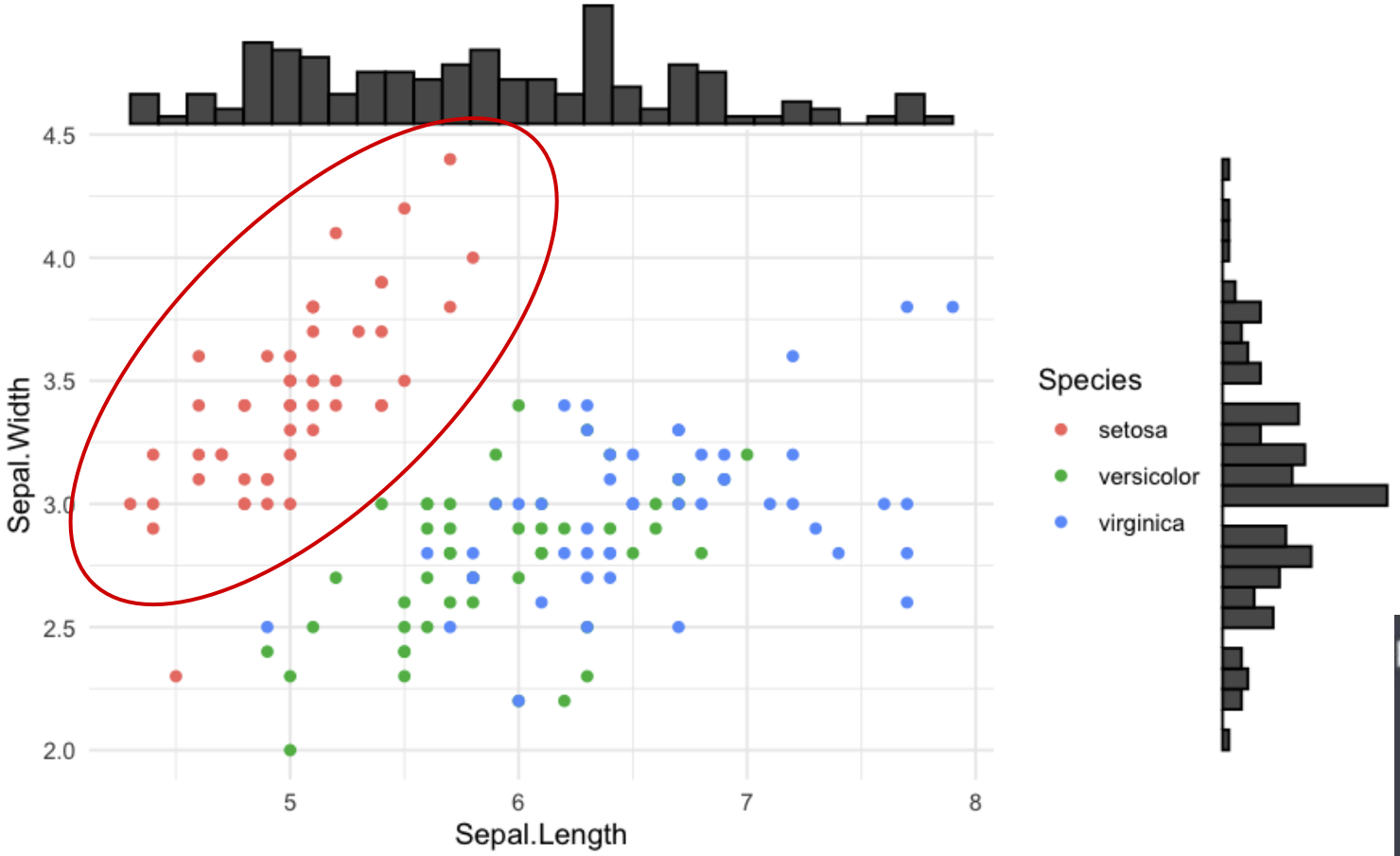

```
p2 <- ggMarginal(p1, type="histogram")
p2
```



There appears to be **no relationship** between Sepal Length and Width based on this scatterplot.

Bonus [Why you should visualize your data - DataSaurus Dozen & Anscombe's Quartet]

4. Identify potential linear relationships
(adding categorical data through color may help)



There may be a **positive linear relationship** between Sepal Length and Width for setosa species.

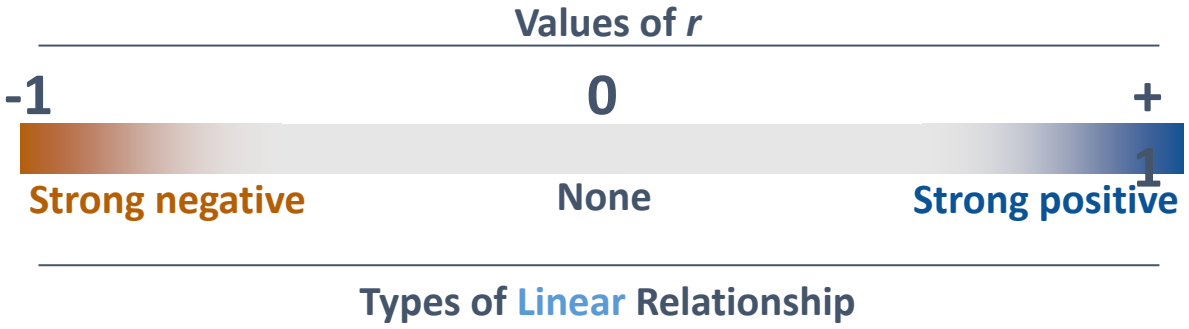
```
p1 <- ggplot(iris, aes(x = Sepal.Length,
                       y = Sepal.Width,
                       color = Species)) +
  geom_point()

ggMarginal(p1, type="histogram")
```

Bonus Why you should visualize your data - DataSaurus Dozen & Anscombe's Quartet

R code demo!

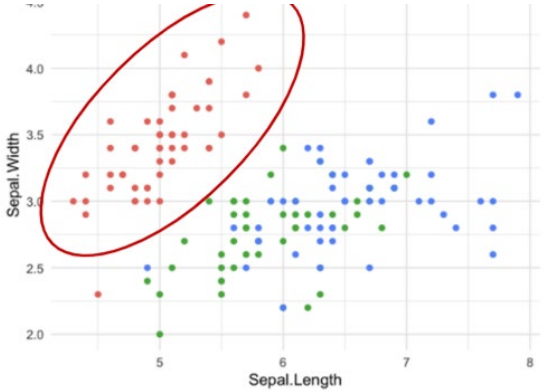
5. Calculate correlation coefficient (*r*)

Population correlation coefficient = ***ρ*** (greek letter "rho")
Linear correlation coefficient (from sample) = ***r***



Assumptions:

- Sample is random
- Data pairs fall approximately on straight line and **measured** at interval or ratio level *(need to be numbers, ideally continuous)*
- Bivariate normal distribution (for a given value in one dimension, the data in the other dimension is normally distributed)




```
df_setosa <- as_tibble(iris) %>%
  filter(Species == "setosa")

cor(df_setosa$Sepal.Width,
    df_setosa$Sepal.Length,
    method = "pearson")
```

setosa (red), r = 0.743
versicolor (green), r = 0.526
virginica (blue), r = 0.457

6. Test significance of linear correlation coefficient (*r*)

H_0: ρ = 0

H_a: $\rho \neq 0$

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

$$df = n - 2, \quad where$$

$$n = number\ of\ orderd\ pairs\ (x,\ y)$$

```
        Pearson's product-moment correlation

data:  setosa$Sepal.Width and setosa$Sepal.Length
t = 7.6807, df = 48, p-value = 6.71e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5851391 0.8460314
sample estimates:
      cor
0.7425467
```

← *r*

Assumptions:

- Sample is random
- Data is quantitative (numerical)
- Scatterplot shows approximately linear relationship
- No outliers in the data
- Variables *x* and *y* come from normally distributed populations

```
# with the cor.test() function
cor.test(setosa$Sepal.Width, setosa$Sepal.Length,
         method = "pearson")
```

PAUSE - Interpreting linear relationships

Possible reasons why your cor.test() returned p-value < α and you reject H_0

1. Direct cause & effect relationship (x causes y)
2. Reverse cause & effect relationship (y causes x)
3. **Relationship caused by third variable (lurking variable)**
4. **Complexity of interrelationships between many variables**
5. Relationship may be coincidental

Address these through detective work

What other factors might be involved? Are they correlated?
Asking subject matter experts about what they have observed.
etc...

6:15 - 6:55

7. Create linear regression model ($y' = a + bx$)

Line of best fit through the data. (Minimizes distance between data points and line.)

b = slope (for every unit of change in x, y' changes by this much)

a = intercept (baseline when x is zero)

Assumptions:

- Sample is random
- For any value x, value of y is normally distributed (*Normality*)
- Standard deviation of y must be the same for each value of x (*Equal variance*)
- There is a linear relationship between x & y (*Linearity*)
- Independent variables not correlated (*Nonmulticollinearity*)
- Values for y variables are independent (*Independence*)

```
# linear regression model
fit <- lm(formula = Sepal.Length ~ Sepal.Width, # y ~ x
          data = setosa)

summary(fit)
```

Assumptions for simple linear regression (Bluman 10-22) and multiple linear regression (Bluman 10-46)

7a. Numerically, do we trust our model?

```
Call:
lm(formula = Sepal.Width ~ Sepal.Length, data = df_setosa)

Residuals:
     Min       1Q   Median       3Q      Max
-0.72394 -0.18273 -0.00306  0.15738  0.51709

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.5694     0.5217  -1.091    0.281
Sepal.Length   0.7985     0.1040   7.681 6.71e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2565 on 48 degrees of freedom
Multiple R-squared:  0.5514,    Adjusted R-squared:  0.542
F-statistic: 58.99 on 1 and 48 DF,  p-value: 6.71e-10
```

$y' = a + bx$

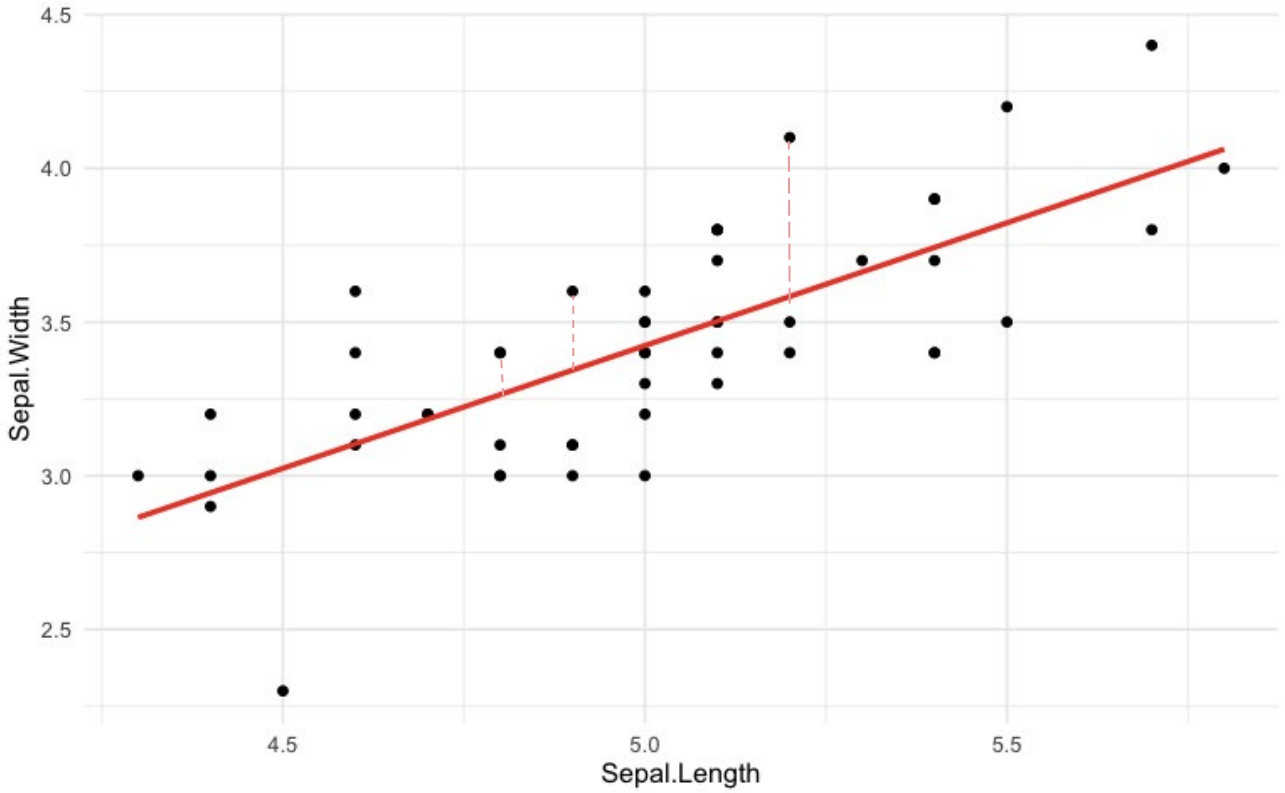
We want the **residuals to be evenly distributed around 0 (zero).** Indicates that remaining unexplained variation is randomly distributed.

We want the **coefficients to be significantly different from zero**. If Pr(>|t|) is less than 0.05 then we can reject that null hypothesis. *If values > 0.05 then we may want to reconsider our formula.*

We want to **maximize** the multiple R-squared to predict as much of the variation of the data as possible. **Prefer adjusted R-squared** as it accounts for over-estimation when number data groups (n) ≈ number independent variables (k).

We want **F-statistic p-values < 0.05** to indicate that the independent variables together predict the dependent variable above random chance.

Regression line plot




```r
df_setosa %>%
  ggplot(aes(x = Sepal.Length, y = Sepal.Width,)) +
    geom_point() +
    geom_smooth(method = "lm", col = "red", se = FALSE)
```

- Residuals are the difference between the predicted and actual y values

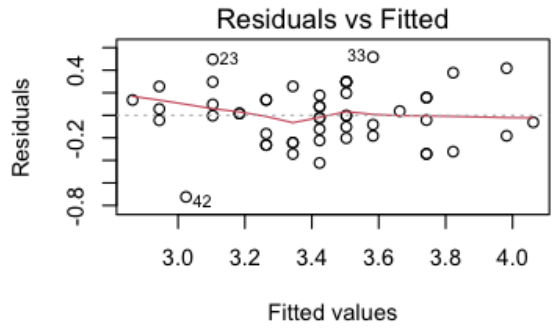
$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

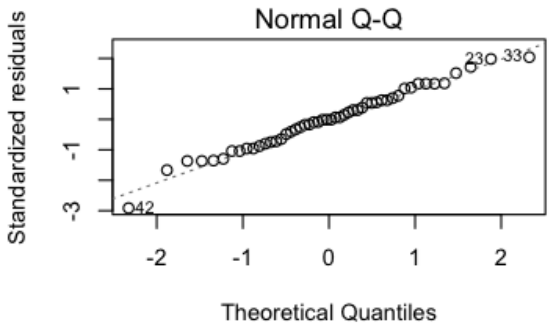
← residual

Checking Regression Assumptions

7b. Does our model satisfy the assumptions of linear regression?

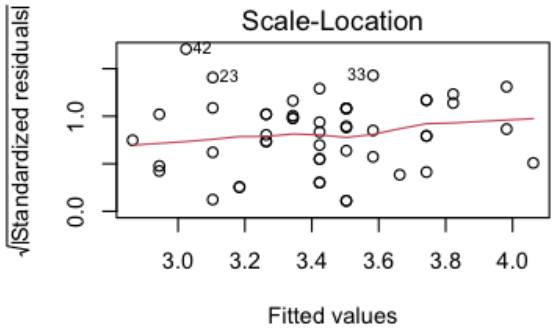
Linearity Assumption
Ideally all points fall equally above/below 0 line with no apparent pattern.

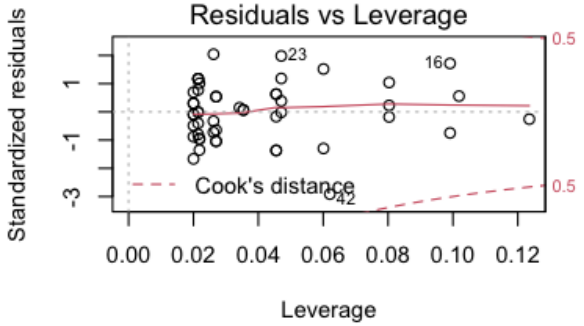




Normality Assumption
All points should fall on dotted line

Equal (Constant) Variance Assumption
Ideally all points fall in a random band around a horizontal line (Homoscedasticity); failure to do so is Heteroscedasticity





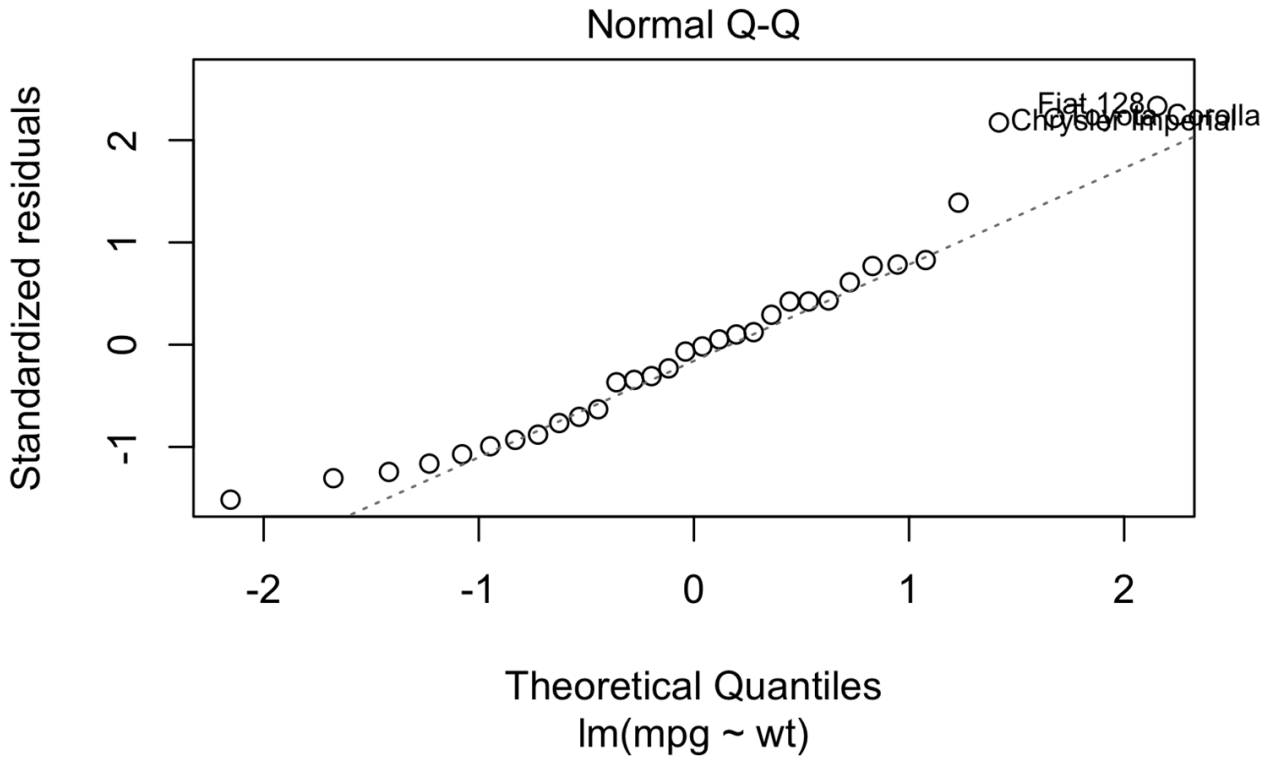
Identifying
Outliers: large residual
High-leverage points: large leverage
Influential observations: Beyond Cook's distance

```
par(mfrow = c(2,2)) # make a 2x2 grid of plots
plot(fit) # will by default give the 4 key graphs
```

Reference reading: Kabacoff Unit 8.3 Regression Diagnostics

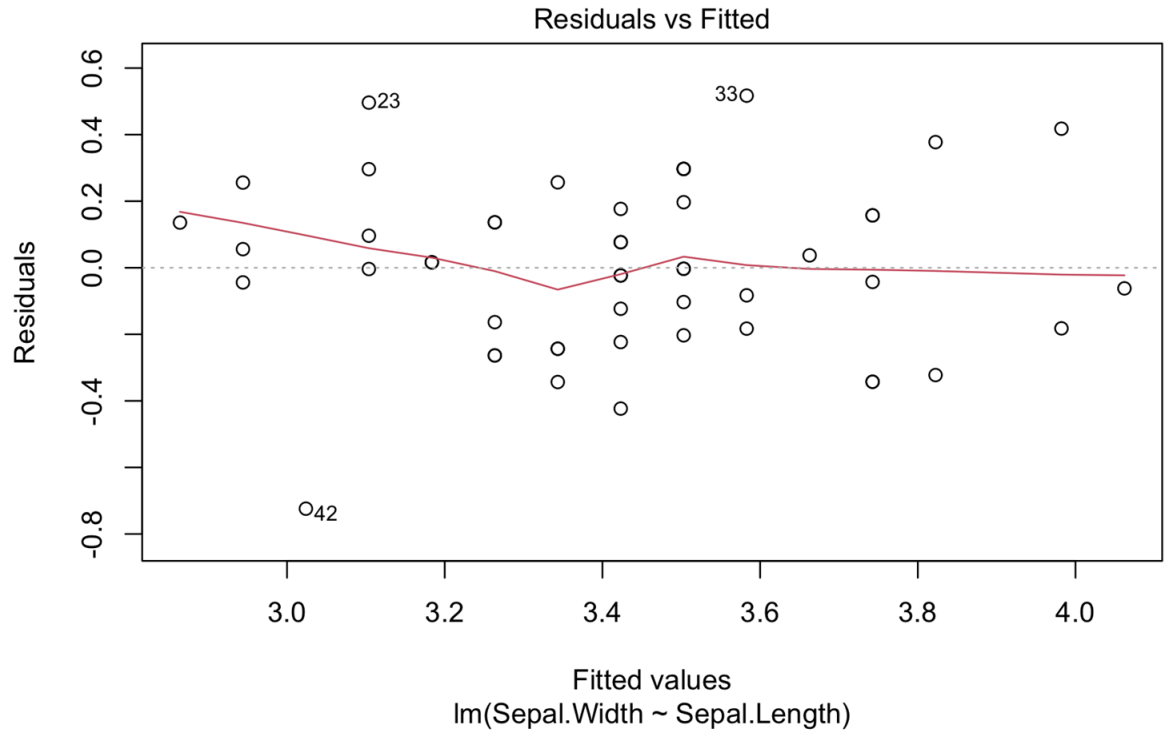
Q-Q Plot to check for normality

- X-axis is where the data would lie if it were perfectly normally distributed
- Y-axis is there the data really lies
- If points fall on a straight line, the data is normally distributed



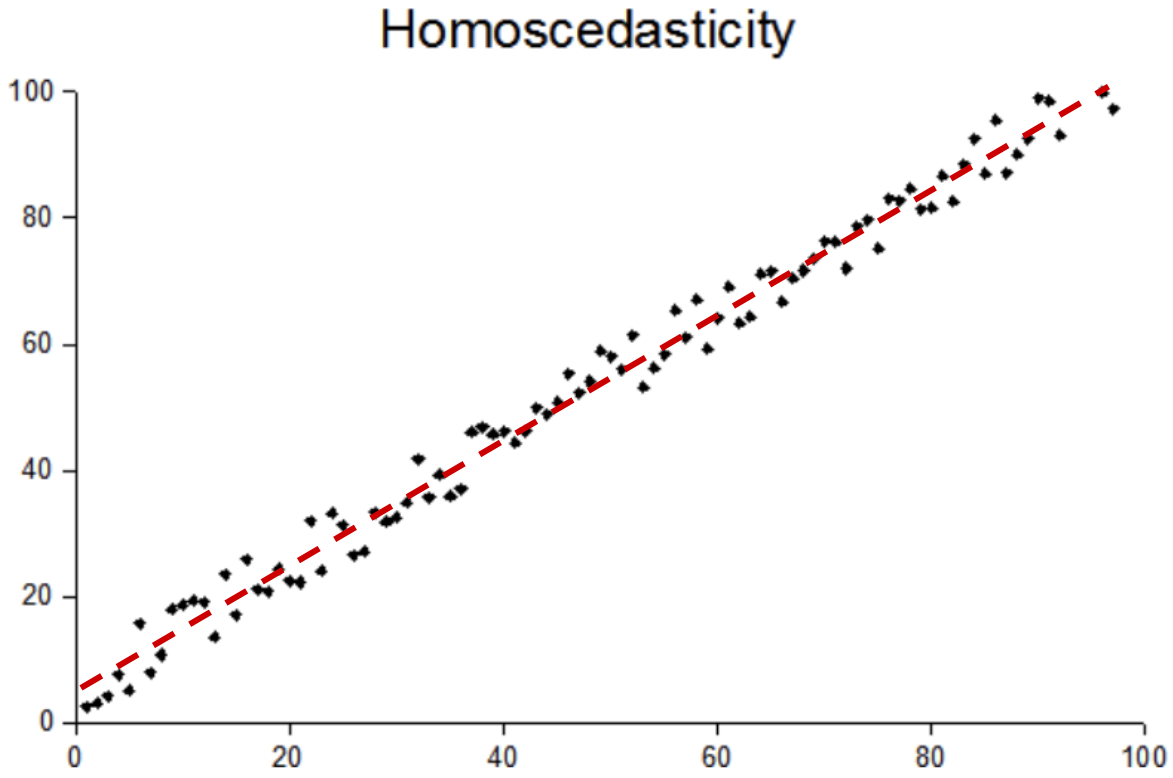
Linearity Assumption

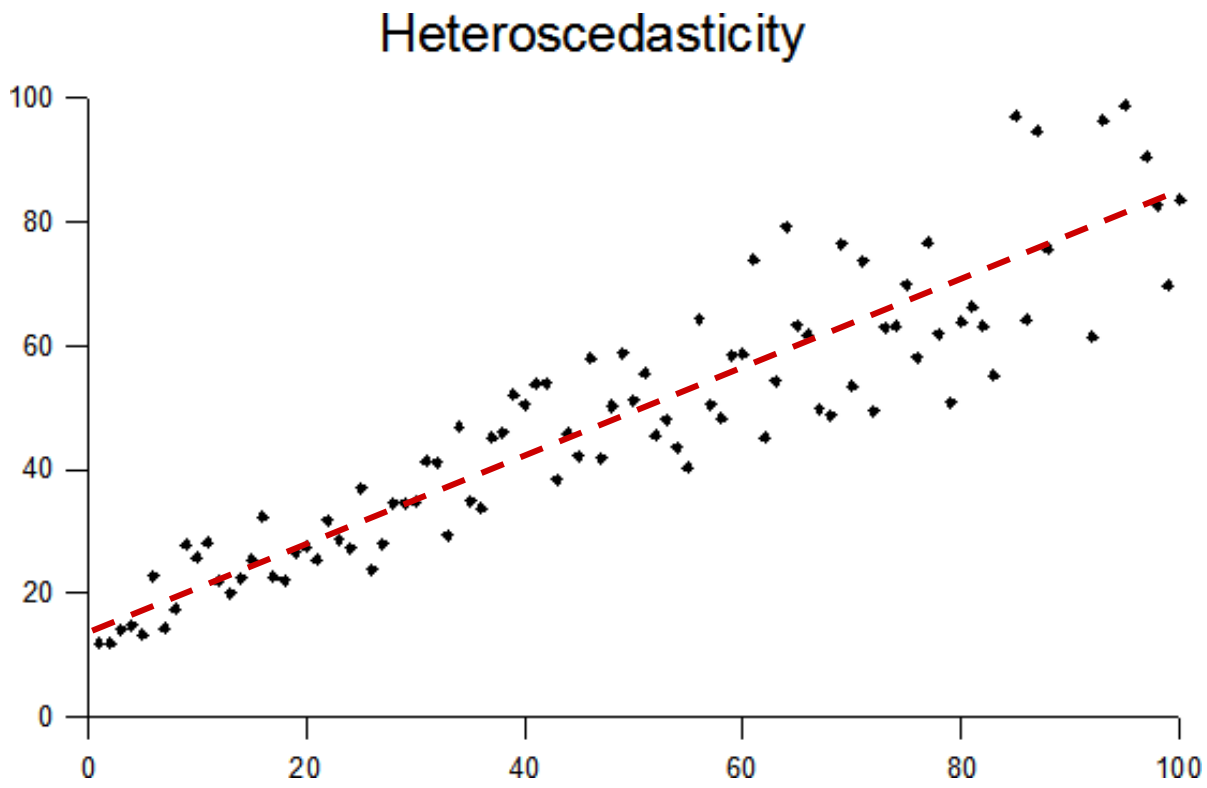
- There shouldn’t be a trend in residuals vs fitted values
  - Red line should be horizontal
  - There should be no visible pattern in the data



Homoscedasticity (equal variance) assumption

- homoscedasticity means “having the same scatter” [1]

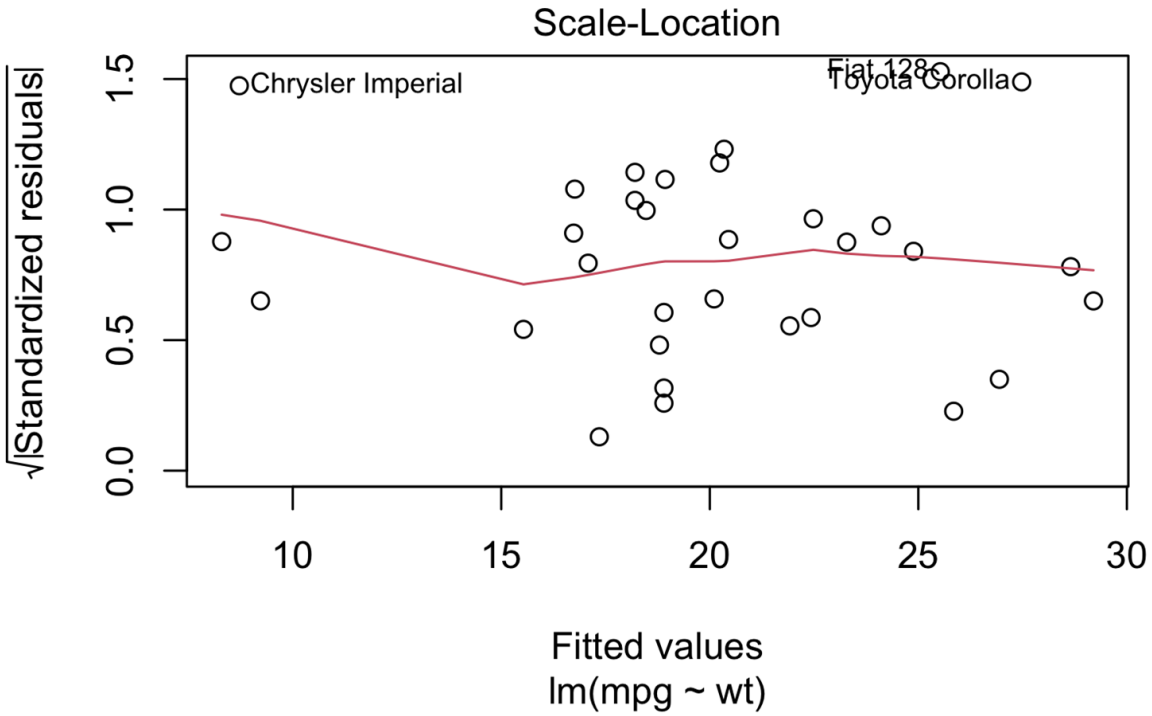




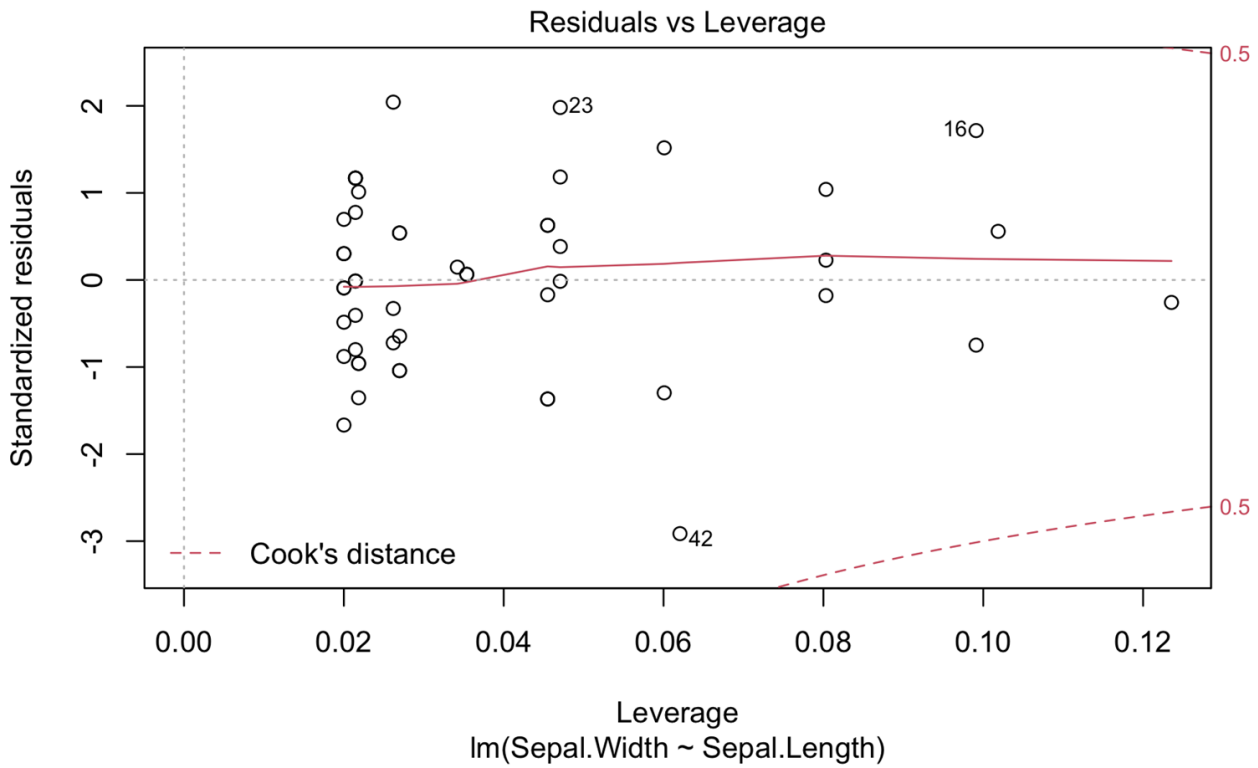
[1] https://www.statisticshowto.com/homoscedasticity/

Scale-location Plot: check for homoscedasticity

- shows if residuals are spread equally along the ranges of predictors
- Ideally, the line should be horizontal with evenly distributed points



High Leverage Points



- Outliers - data point whose response (y) does not follow the general trend of the rest of the data
- High Leverage - the point's x-value is extreme
- Influential Points - unduly influences the regression analysis
  - Outliers and high leverage data points are not always influential
- The 3 most extreme leverage points are highlighted (23, 12, 42)
- You want all of your points to be inside the Cook's distance lines
  - It's good if they don't show up on the plot - that means the points are well below Cook's distance

https://online.stat.psu.edu/stat462/node/170/

8. Calculate coefficient of determination (R^2)

1. Obtain using linear model results in R
2. Ranges from 0 to 1
    a. Interpret as the percentage of y's variation that is explained by x
        i. An R^2 of 0 means that 0% of y's variation that is explained by x
        ii. An R^2 of 0.5 means that 50% of y's variation that is explained by x
        iii. An R^2 of 1 means that 100% of y's variation that is explained by x

6:15 - 6:55

Steps in analyzing potential linear relationships

1. (if needed, eg. with time data) Summarize data to get to numeric variables
2. Create histograms of numeric variables
3. Create scatterplots of combinations of numeric variables
4. Identify potential linear relationships
5. Calculate correlation coefficient (r)
6. Test significance of linear correlation coefficient
7. Create linear regression model ($y' = a + bx$)
8. Calculate coefficient of determination (r^2)
9. (optional) Make point predictions with prediction intervals ($y = y'$ +/- prediction interval).
10. Interpret results in context

6:15 - 6:55