Contents

[Unit 2: Simple Linear Regression 3](#_Toc89892857)

[STOR 455 - Class 2 – Review of Inference 4](#_Toc89892858)

[STOR 455 Class 3 Linear Models and assessing conditions 4](#_Toc89892859)

[STOR 455 Class 4 assessing conditions and transformations 4](#_Toc89892860)

[STOR 455 Class 5 Transformations 4](#_Toc89892861)

[STOR 455 Class 6 Outliers and Points of Influence 4](#_Toc89892862)

[STOR 455 Class 7 Outliers and Points of Influence Again 4](#_Toc89892863)

[STOR 455 Class 8 Inference for Regression Slope 4](#_Toc89892864)

[STOR 455 Class 8 Inference for Regression Slope 4](#_Toc89892865)

[STOR 455 Class 9 Partitioning Variability - ANOVA 4](#_Toc89892866)

[STOR 455 Quiz 1 Solutions 4](#_Toc89892867)

[**Quiz A** 4](#_Toc89892868)

[Unit 3: Multiple Regression 3](#_Toc89892869)

[STOR 455 Class 11 R Assessing Multiple Linear Regression Models 4](#_Toc89892870)

[STOR 455 Class 12 R Correlated Predictors & Model Selection Methods 4](#_Toc89892871)

[STOR 455 - Class 13 - R Model Section Methods 4](#_Toc89892872)

[STOR 455 Homework #2 4](#_Toc89892873)

[Homework #3 Individual Part 4](#_Toc89892874)

[STOR 455 Homework #3 Group Part 4](#_Toc89892875)

[Unit 4: Multiple Regression II 3](#_Toc89892876)

[STOR 455 - Class 17 – Comparing two regression lines 4](#_Toc89892877)

[STOR 455 - Class Coding categorial variables 4](#_Toc89892878)

[STOR 455 - Class 19 – Testing a subset of predictors 4](#_Toc89892879)

[STOR 455 - Class 20 – new predictors from old 4](#_Toc89892880)

[STOR 455 - Class 21 - R Cross Validation 4](#_Toc89892881)

[STOR 455 Class 22 Influential Points in Multiple Regression 4](#_Toc89892882)

[STOR 455 Homework #4 4](#_Toc89892883)

[Homework 5 4](#_Toc89892884)

[Quiz 2A 4](#_Toc89892885)

[Unit 5: Logistic Regression 3](#_Toc89892886)

[Class 26 R Logistic Regression 4](#_Toc89892887)

[Class 27 R Logistic Regression: Odds Ratio And Inferences 4](#_Toc89892888)

[Class 28 R Logistic Regression Assessing the model 4](#_Toc89892889)

[STOR 455 Class 29 R Multiple Logistic Regression 4](#_Toc89892890)

[STOR 455 Class 30 Multiple Logistic Regression (Again) 4](#_Toc89892891)

[STOR 455 Class 31 R Multiple Logistic Regression Again Agian 4](#_Toc89892892)

[**STOR 455 Homework 6** 4](#_Toc89892893)

[STOR 455 Practice Exam Solutions 4](#_Toc89892894)

[**STOR 455 Exam #2** 4](#_Toc89892895)

[**Should STOR have Undergraduate Learning Assitants? (YES! Starting Fall 2022…)** 4](#_Toc89892896)

[Unit 6: ANOVA 3](#_Toc89892897)

[STOR 455 Class 34 R One Way ANOVA 3](#_Toc89892898)

[STOR 455 Class 35 One Way ANOVA (Again) 3](#_Toc89892899)

[STOR 455 Class 36 Two Way ANOVA 3](#_Toc89892900)

[STOR 455 - Class 37 – ANCOVA and Homogeneity of Variances 3](#_Toc89892901)

[STOR 455 Class 38 Homogeneity of Variances and Contrasts 3](#_Toc89892902)

[Unit 7: Conclusion 14](#_Toc89892903)

[Simple Linear Regression 14](#_Toc89892904)

[Building a Model & Assessing Conditions 14](#_Toc89892905)

[Transformations 14](#_Toc89892906)

[Outliers & Influential Points 14](#_Toc89892907)

[Inference for Regression Slope/Correlation 14](#_Toc89892908)

[ANOVA for simple linear regression 14](#_Toc89892909)

[Confidence and Prediction Intervals 14](#_Toc89892910)

[Multiple Regression 14](#_Toc89892911)

[Building a Model & Assessing Conditions 14](#_Toc89892912)

[Methods for Choosing Predictors 14](#_Toc89892913)

[Interaction Terms 14](#_Toc89892914)

[Polynomial Models 14](#_Toc89892915)

[Second Order Models 14](#_Toc89892916)

[Multicollinearity & VIF 14](#_Toc89892917)

[Nested F-Test 14](#_Toc89892918)

[Cross Validation Correlation 14](#_Toc89892919)

[Logistic Regression 14](#_Toc89892920)

[Building a Model with a single predictors 14](#_Toc89892921)

[Building a Model with a single predictors 14](#_Toc89892922)

[Building a Model with multiple predictors 14](#_Toc89892923)

[Tests of significance for a model (single or multiple predictors) 14](#_Toc89892924)

[Test of significance for nested models 14](#_Toc89892925)

[Methods for Choosing Predictors 14](#_Toc89892926)

[Assessing linearity of the logit model 14](#_Toc89892927)

[Confidence interval for Odds Ratio 14](#_Toc89892928)

[Interpreting Odds Ratio 14](#_Toc89892929)

[Predicting Probabilities 14](#_Toc89892930)

[ANOVA for Means 14](#_Toc89892931)

[One Way ANOVA 14](#_Toc89892932)

[Two Way ANOVA 14](#_Toc89892933)

[Two Way ANOVA with Interaction 14](#_Toc89892934)

[Tukey HSD 14](#_Toc89892935)

[Levene’s Test 14](#_Toc89892936)

# Unit 2: Simple Linear Regression

## STOR 455 - Class 2 – Review of Inference

# loads packages needed  
# install a package before first using it for the first time  
  
library(readr)  
library(mosaic)  
  
# loads the DistanceHome dataframe into the environment from github  
  
DistanceHome <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/DistanceHome.csv")  
  
# Alternative way to load dataframe (remove # to use)  
# DistanceHome.csv must be saved in the same folder as this notebook!  
  
#DistanceHome <- read\_csv("DistanceHome.csv")  
  
# Shows the variables and first 6 cases (by default)  
head(DistanceHome)

## # A tibble: 6 x 3  
## Distance Hours Introvert   
## <dbl> <dbl> <chr>   
## 1 7606 21 Introversion  
## 2 7606 21 Introversion  
## 3 3800 20 Introversion  
## 4 7102 20 Introversion  
## 5 6000 20 Introversion  
## 6 7756 18 Introversion

**Example: Distance to Home** - *Question:* How can we predict the distance from campus to home for Carolina students? - *Data:* Estimated distance to home (in miles) for students taking STOR 455 in a previous semester. - *Predictor variables:* Start with none.

**Example: Constant Model** Y = c + Error

Where c = an unknown constant

**Terminology** - “The constant c” = **parameter** of the model - “Sample estimate” - use data to provide a sample estimate of c

*How should we estimate 𝑐 from data?*

*Below:* Summarize the Distance variable - Numerical: mean and median

# dataframe$variable\_name  
  
mean(DistanceHome$Distance)

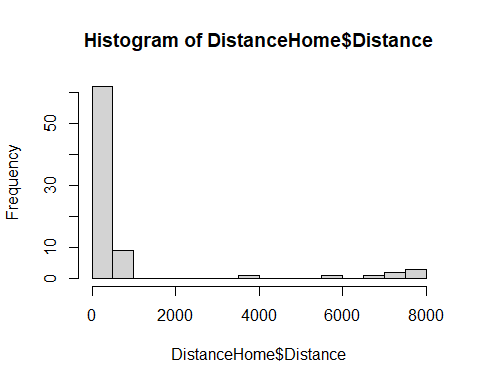
## [1] 844.6234

median(DistanceHome$Distance)

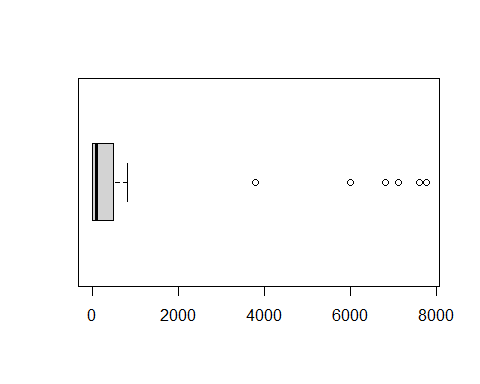
## [1] 113

*Below:* Summarize the Distance variable - Graphical: histogram, boxplot

hist(DistanceHome$Distance, breaks=20)



boxplot(DistanceHome$Distance, horizontal = TRUE)



**Predicted Value for Response** - Get an estimate for Y using the predictors and the model with estimated parameters.

*Notation:* The predicted y is denoted yhat

For the constant Model: yhat = chat

Examples: yhat = chat = ybar (\*sample mean) yhat = chat = m (sample median)

**Can we use a predictor to improve the model?**  X = Hours to travel home? X = Introvert? - Two sample t test for a difference in means

**Model with a Binary Predictor** Y = f(x) + Error where X = introversion, mu1 = mean distance for Extroverts mu2 = mean distance for Introverts

mean(Distance~Introvert, data=DistanceHome)

## Extroversion Introversion   
## 365.6026 1288.5939

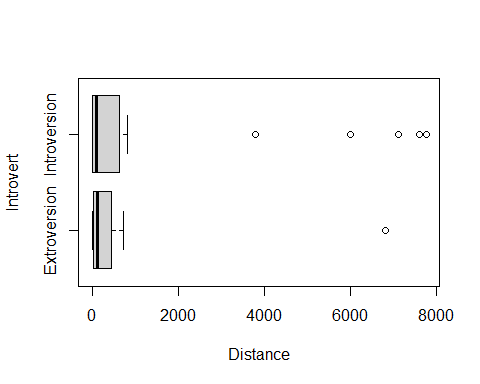
tally(~Introvert, data=DistanceHome)

## Introvert  
## Extroversion Introversion   
## 38 41

sd(Distance~Introvert, data=DistanceHome)

## Extroversion Introversion   
## 1094.443 2559.412

boxplot(Distance~Introvert, data=DistanceHome, horizontal=TRUE)



**Two-sample T-Test Difference in Means** *Hypothesis* Ho: Mu1 = Mu2 Ha: Mu1 != Mu2

*Compare to a t-dist*

**P-value** - The p-value is the proportion of samples, when the H0 is true, that would be as (or more) extreme as the observed sample.

*Below, Conclusion:* Decision: Reject H0 only when the p-value is small.

t.test(Distance~Introvert, data=DistanceHome)

##   
## Welch Two Sample t-test  
##   
## data: Distance by Introvert  
## t = -2.1103, df = 55.025, p-value = 0.03939  
## alternative hypothesis: true difference in means between group Extroversion and group Introversion is not equal to 0  
## 95 percent confidence interval:  
## -1799.48957 -46.49298  
## sample estimates:  
## mean in group Extroversion mean in group Introversion   
## 365.6026 1288.5939

**Normality?** - The two-sample t-test assumes both samples are from normal populations

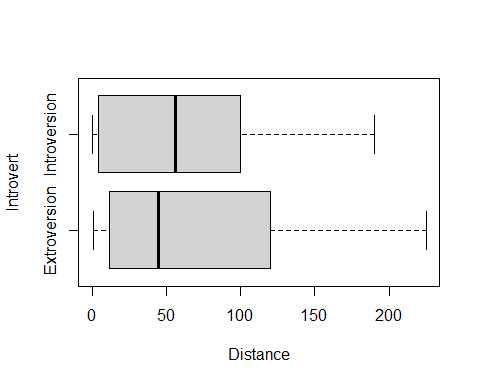
**Domestic Distance** - Suppose that we look only at distances that are really “drivable”? *See below*

Domestic = subset(DistanceHome, Distance<250)  
head(Domestic)

## # A tibble: 6 x 3  
## Distance Hours Introvert   
## <dbl> <dbl> <chr>   
## 1 225 3.5 Extroversion  
## 2 120 3 Extroversion  
## 3 190 3 Introversion  
## 4 172 3 Introversion  
## 5 167. 3 Extroversion  
## 6 190 3 Introversion

* For distance home (only including students less than 250 miles from home) the Introvert variable does not improve the model significantly.

boxplot(Distance~Introvert, data=Domestic, horizontal=TRUE)



t.test(Distance~Introvert, data=Domestic)

##   
## Welch Two Sample t-test  
##   
## data: Distance by Introvert  
## t = 0.09629, df = 51.968, p-value = 0.9237  
## alternative hypothesis: true difference in means between group Extroversion and group Introversion is not equal to 0  
## 95 percent confidence interval:  
## -33.17687 36.52132  
## sample estimates:  
## mean in group Extroversion mean in group Introversion   
## 67.16667 65.49444

**Inference Review: Hypothesis Testing** - Suppose that we look only at distances that are really “drivable”? *Test* Ho: mu1 = Mu2 Ha: mu1 != Mu2

There is a 93.3% chance that we would receive a samples with a difference as extreme as we did if the null hypothesis is true. p-value = 0.933

Since the p-value is greater than 0.05, we fail to reject the null hypothesis. There is not evidence to suggest that there is a difference in the number of miles from home Carolina students are (of those students 250 miles or less) based on if they are introverts or extroverts.

There is a 93.3% chance that we would receive a samples with a difference as extreme as we did if the null hypothesis is true. p-value = 0.933

Domestic=subset(DistanceHome,Distance<250)

**Domestic Distance** For distance home (only including students less than 250 miles from home) the Introvert variable does not improve the model significantly.

t.test(Distance~Introvert, data=Domestic)

##   
## Welch Two Sample t-test  
##   
## data: Distance by Introvert  
## t = 0.09629, df = 51.968, p-value = 0.9237  
## alternative hypothesis: true difference in means between group Extroversion and group Introversion is not equal to 0  
## 95 percent confidence interval:  
## -33.17687 36.52132  
## sample estimates:  
## mean in group Extroversion mean in group Introversion   
## 67.16667 65.49444

## STOR 455 Class 3 Linear Models and assessing conditions

# message=FALSE, warning=FALSE supress warnings and messages from appearing in knitted html  
  
library(readr)  
  
DistanceHome <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/DistanceHome.csv")  
  
# If notebook and csv file are saved in the same folder  
# DistanceHome <- read\_csv("DistanceHome.csv")  
  
Domestic=subset(DistanceHome,Distance<250)

**Single Quantitative Predictor Model** - Notation:  
– Y = Response variable – X = Predictor variable

*Assume (for now) that both Y and X are quantitative variables.* Y = f(x) + Error

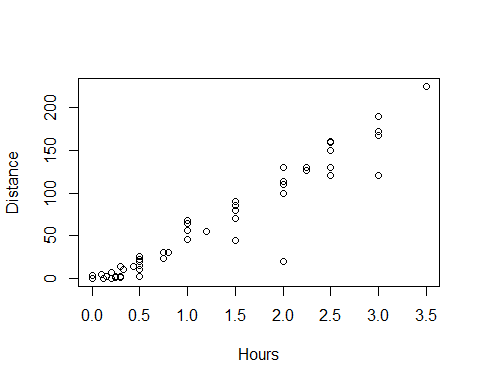
*Simple Linear Model* - X = Single quantitative predictor - Y = Quantitative response

*Goal*: Find a line that best summarizes the trend in the data.

Y = Bo + B1x + Error Response = Intercept + Slope*Predictor + Random Error* Assumptions:\* - Assume: Error ~ Follows a normal distribution and independent - There are 3 parameters to estimate: Bo, B1, and std error

**Scatterplot in R** *See below*

plot(Distance~Hours, data=Domestic)



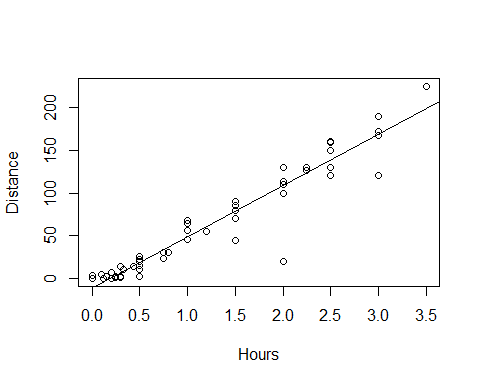
**Least Square Regression in R** *Syntax:* Syntax: lm(Response~Predictor,data= )

lm(Distance~Hours, data=Domestic)

##   
## Call:  
## lm(formula = Distance ~ Hours, data = Domestic)  
##   
## Coefficients:  
## (Intercept) Hours   
## -11.06 59.98

*R – Plot with Fitted Line*

mod1=lm(Distance~Hours, data=Domestic)  
plot(Distance~Hours, data=Domestic)  
abline(mod1)



*Simple Linear Model- Conditions* **Model:** 1. Linearity: The means for Y vary as a linear function of X. **Error:** 2.Zero Mean: The distribution of the errors is centered at zero. 3.Constant variance: The variance for Y is the same at each X. (Homoscedasticity) 4.Independence: No relationships among errors. 5.Normality: - Residuals are normally distributed - (sometimes) At each X, the Y’s follow a normal distribution.

*Linear* Look for consistent curvature or non-linear patterns

*Constant Variance* Look for “fan-shaped” pattern - Fan-shaped is **bad**

summary(mod1)

##   
## Call:  
## lm(formula = Distance ~ Hours, data = Domestic)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -88.892 -4.680 2.172 7.082 26.141   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -11.063 4.056 -2.727 0.00868 \*\*   
## Hours 59.977 2.484 24.144 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.26 on 52 degrees of freedom  
## Multiple R-squared: 0.9181, Adjusted R-squared: 0.9165   
## F-statistic: 582.9 on 1 and 52 DF, p-value: < 2.2e-16

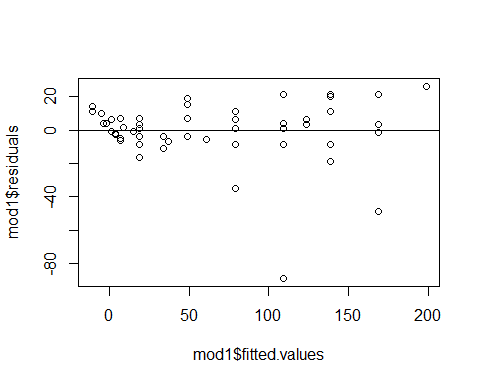
mod1$residuals

## 1 2 3 4 5 6   
## 26.1414099 -48.8698461 21.1301539 3.1301539 -1.7698461 21.1301539   
## 7 8 9 10 11 12   
## -8.8811021 20.1188979 -18.8811021 11.1188979 21.1188979 6.1132699   
## 13 14 15 16 17 18   
## 3.1132699 -8.8923581 21.1076419 -88.8923581 4.1076419 1.1076419   
## 19 20 21 22 23 24   
## 6.0963859 6.0963859 -34.9036141 -8.9036141 6.0963859 11.0963859   
## 25 26 27 28 29 30   
## 1.0963859 -5.9103678 -3.9148702 19.0851298 15.0851298 7.0851298   
## 31 32 33 34 35 36   
## -6.9193726 -3.9204982 -10.9204982 -16.4261262 3.0738738 7.0738738   
## 37 38 39 40 41 42   
## -3.9261262 -8.9261262 1.0738738 -0.7277020 1.2700468 -4.9306286   
## 43 44 45 46 47 48   
## 7.0693714 -6.4306286 -2.9317542 -1.9317542 -1.9317542 -0.7328798   
## 49 50 51 52 53 54   
## 6.4671202 4.0659946 4.0153193 10.0648690 14.0626178 11.0626178

mod1$fitted.values

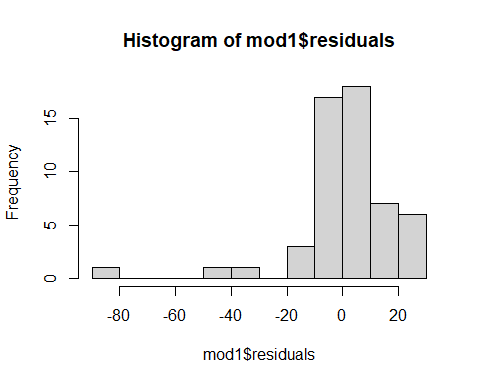
## 1 2 3 4 5 6   
## 198.8585901 168.8698461 168.8698461 168.8698461 168.8698461 168.8698461   
## 7 8 9 10 11 12   
## 138.8811021 138.8811021 138.8811021 138.8811021 138.8811021 123.8867301   
## 13 14 15 16 17 18   
## 123.8867301 108.8923581 108.8923581 108.8923581 108.8923581 108.8923581   
## 19 20 21 22 23 24   
## 78.9036141 78.9036141 78.9036141 78.9036141 78.9036141 78.9036141   
## 25 26 27 28 29 30   
## 78.9036141 60.9103678 48.9148702 48.9148702 48.9148702 48.9148702   
## 31 32 33 34 35 36   
## 36.9193726 33.9204982 33.9204982 18.9261262 18.9261262 18.9261262   
## 37 38 39 40 41 42   
## 18.9261262 18.9261262 18.9261262 14.7277020 8.7299532 6.9306286   
## 43 44 45 46 47 48   
## 6.9306286 6.9306286 3.9317542 3.9317542 3.9317542 0.9328798   
## 49 50 51 52 53 54   
## 0.9328798 -2.0659946 -3.8653193 -5.0648690 -11.0626178 -11.0626178

plot(mod1$residuals~mod1$fitted.values)  
abline(0,0)



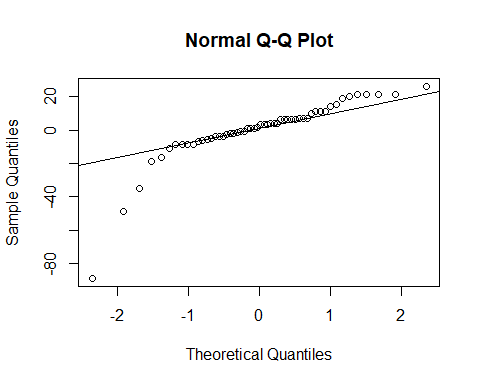
*Residuals* Look at a histogram of the residuals Look for clear skewness and outliers - skew and outliers are **bad**

hist(mod1$residuals, breaks=10)

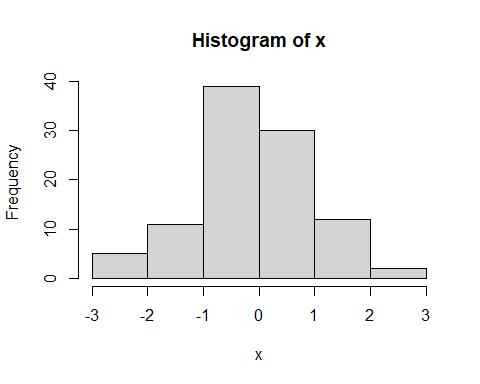


*How much Variability is Expected?*

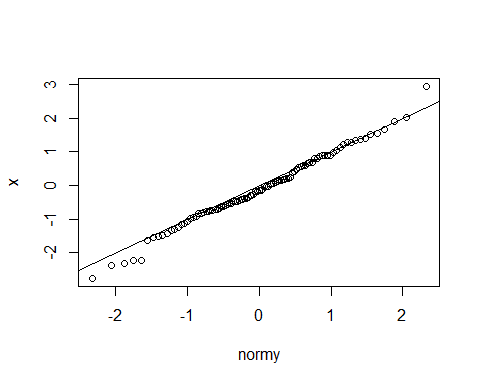
qqnorm(mod1$residuals)  
qqline(mod1$residuals)



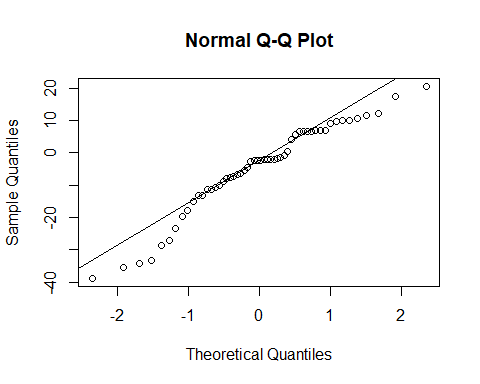
#Sets randomization  
set.seed(455001)  
  
#Sample of 99 values from a Normal distribution; mean=0; sd=1; sorted ascending  
x = sort(rnorm(99,0,1))  
hist(x)



#list of integers 1 through 99.  
y = c(1:99)  
  
#z-scores of dataset of 99 values if perfectly normally distributed  
normy = qnorm(y/100)  
  
plot(x~normy)  
abline(0,1)



x <- rnorm(54, 0, 18.26)  
qqnorm(x)  
qqline(x)



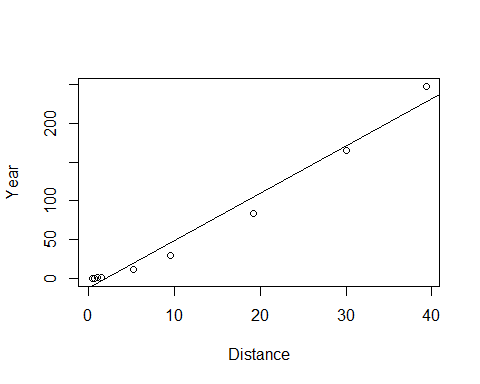
## STOR 455 Class 4 assessing conditions and transformations

# message=FALSE, warning=FALSE suppress warnings and messages from appearing in knitted html  
  
library(readr)  
library(Stat2Data)  
  
Planets <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data//Planets.csv")

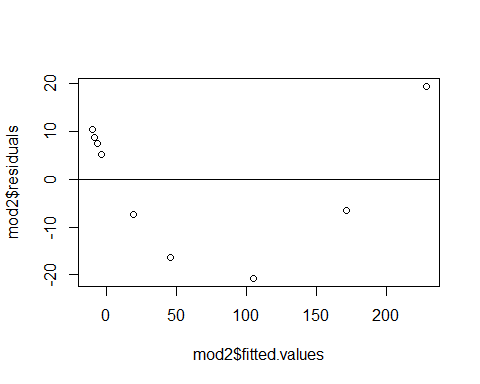
head(Planets, 9)

## # A tibble: 9 x 7  
## Planet Distance Year Mass Day Diameter Gravity  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Mercury 0.39 0.24 0.055 1408. 3.04 0.37  
## 2 Venus 0.72 0.61 0.815 5832. 7.52 0.88  
## 3 Earth 1 1 1 24 7.92 1   
## 4 Mars 1.52 1.88 0.107 24.6 4.22 0.17  
## 5 Jupiter 5.2 11.9 318. 9.9 88.8 2.64  
## 6 Saturn 9.52 29.5 95.2 10.2 74.6 1.15  
## 7 Uranus 19.2 84.1 14.5 17.2 31.6 1.15  
## 8 Neptune 30.0 165. 17.2 16.1 30.2 1.12  
## 9 Pluto 39.3 248. 0.003 153. 1.86 0.04

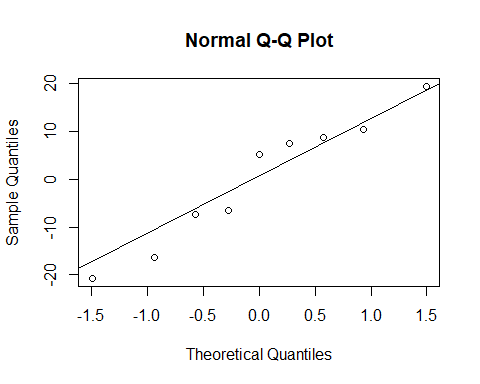
plot(Year~Distance, data=Planets)  
mod2=lm(Year~Distance, data=Planets)  
abline(mod2)

 *Simple Linear Model- Conditions* **Model:** 1. Linearity: The means for Y vary as a linear function of X. **Error:** 2.Zero Mean: The distribution of the errors is centered at zero. 3.Constant variance: The variance for Y is the same at each X. (Homoscedasticity) 4.Independence: No relationships among errors. 5.Normality: - Residuals are normally distributed - (sometimes) At each X, the Y’s follow a normal distribution.

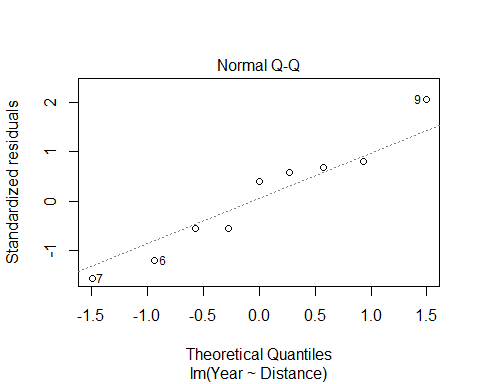
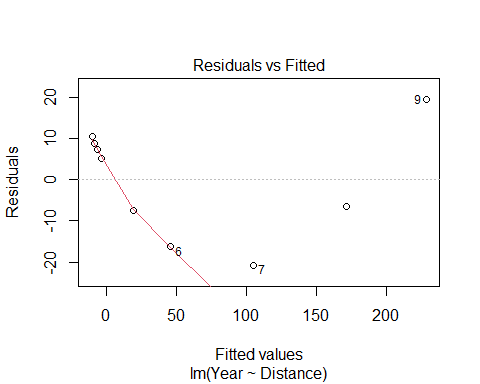
plot(mod2$residuals~mod2$fitted.values)  
abline(0,0)



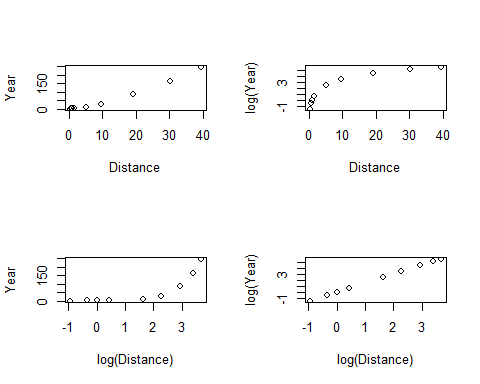
qqnorm(mod2$residuals)  
qqline(mod2$residuals)



plot(mod2, 1:2)

 **What to do when regression assumptions are violated?** *Examples:* 1.Nonlinear patterns in residuals 2.Heteroscedasticity (nonconstant variance) 3.Lack of normality in residuals 4.Outliers: influential points, large residuals

par(mfrow=c(2,2))  
  
plot(Year~Distance, data=Planets)  
plot(log(Year)~Distance, data=Planets)  
plot(Year~log(Distance), data=Planets)  
plot(log(Year)~log(Distance), data=Planets)

 **Data Transformations** *Can be used to:* 1.Address non-linear patterns 2.Stabilize variance 3.Remove skewness from residuals 4.Minimize effects of outliers

**Common Transformations** For either the response (Y) or predictor (X)… - Logarithm: 𝑌→l𝑜𝑔⁡(𝑌) - Square root: 𝑌→√𝑌 - Exponentiation: 𝑌→𝑒^Y - Power function: 𝑌→𝑌^3 - Reciprocal: 𝑌→1/𝑌

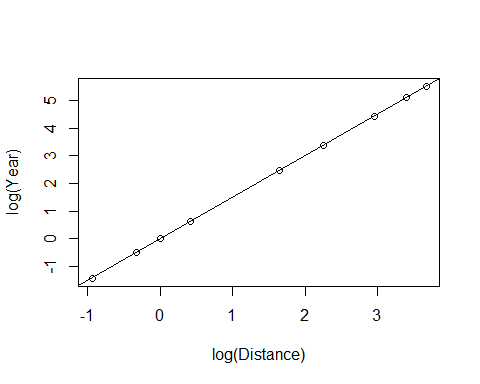
**Example: Planets**

Y = Length of the “year” for planets X = Distance from the sun

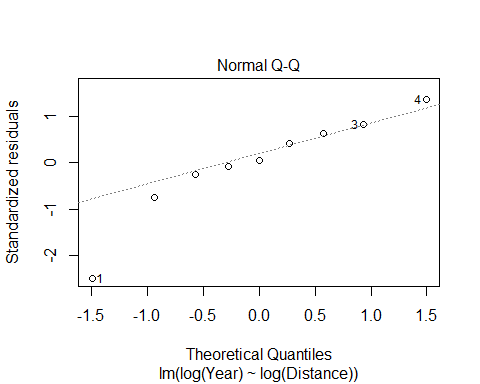
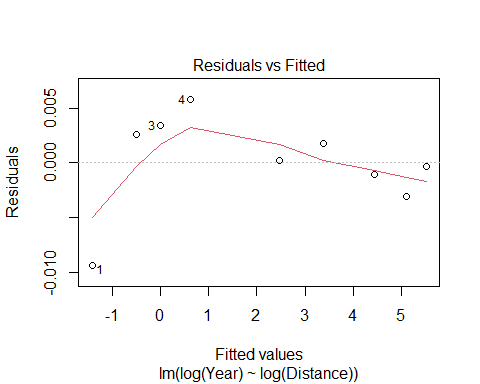
Try scatterplots and LM with Year vs. Distance log(Year) vs. Distance Year vs. log(Distance) log(Year) vs. log(Distance)

*Which transformation gives the best linearity?*

mod3 = lm(log(Year)~log(Distance), data=Planets)  
  
plot(log(Year)~log(Distance), data=Planets)  
abline(mod3)



plot(mod3, 1:2)



summary(mod3)

##   
## Call:  
## lm(formula = log(Year) ~ log(Distance), data = Planets)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.0093289 -0.0010233 0.0002193 0.0025708 0.0057772   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.0034339 0.0020852 -1.647 0.144   
## log(Distance) 1.5020611 0.0009567 1570.016 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.004662 on 7 degrees of freedom  
## Multiple R-squared: 1, Adjusted R-squared: 1   
## F-statistic: 2.465e+06 on 1 and 7 DF, p-value: < 2.2e-16

log(Year) = -0.0034399 + 1.5020611\*log(Distance)

Year = e ^(-0.0034399 + 1.5020611(log(Distance))

Year = e ^(-0.0034399) e ^((1.5020611)(log(Distance))

Year = e ^(-0.0034399) e ^(log(Distance ^1.5020611))

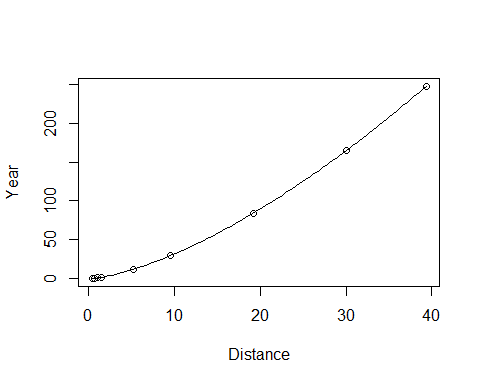
Year = e ^(-0.0034399) (Distance^1.5020611)

exp(-0.0034339)

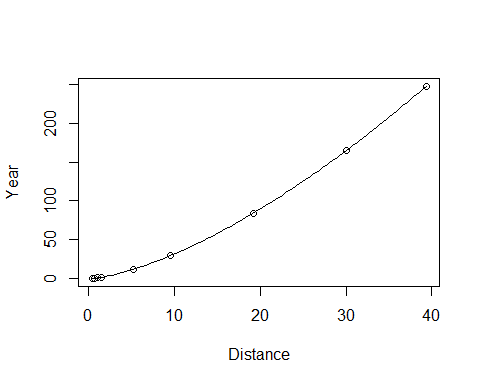
## [1] 0.996572

Year = 0.996572(Distance^1.5020611)

plot(Year~Distance, data=Planets)  
curve(0.996572\*(x^1.5020611), add=TRUE)



B0 = summary(mod3)$coefficients[1,1]  
B1 = summary(mod3)$coefficients[2,1]  
  
plot(Year~Distance, data=Planets)  
curve(exp(B0)\*x^B1, add=TRUE)



## STOR 455 Class 5 Transformations

# message=FALSE, warning=FALSE suppress warnings and messages from appearing in knitted files  
  
library(readr)  
library(Stat2Data)

**What to do when regression assumptions are violated?** *Examples of violations:* 1. Nonlinear patterns in residuals 2. Heteroscedasticity (nonconstant variance) 3. Lack of normality in residuals 4. Outliers: influential points, large residuals

**Data Transfomrations** Can be used to: 1. Address non-linear patterns 2. Stabilize variance 3. Remove skewness from residuals 4. Minimize effects of outliers

**Common Transformations** - Log - Square root - Exponentiation - Power function - Reciprocal

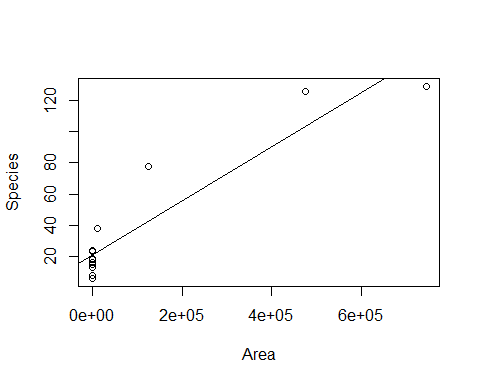
*Example: Mammal Species* Y = Number of mammal species on an island X = Area of the island

Data on fourteen islands in Southeast Asia are stored in SpeciesArea (in Stats2Data)

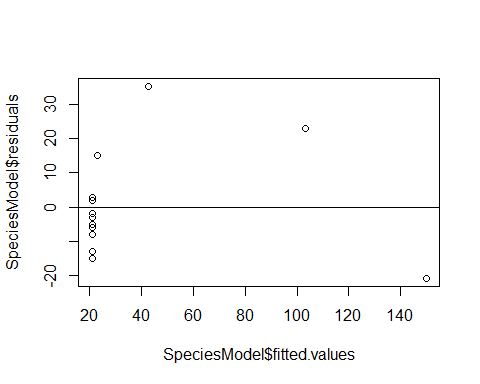
*Notes* - Make sure you check your working environment

**Why is log of something useful?** - Log goes up and then trails off forever - If have a big right skew, or extreme values or outliers in the right side, it helps squeeze the outliers back in - If we have constant variance issues, with fanning patterns, use log - DOesn’t effect low values very much, but helps smoosh bigger data

# Pull in the data   
data("SpeciesArea")  
  
# The log of the species are going to be the most useful   
# Plot the data, make a linear model   
# Want to predict the number of species on each island, based on teh island area  
plot(Species~Area, data=SpeciesArea) # Just the scatterplot, see that is appears to follow a log, but we'll keep make the other models to jsut see how bad it violates the other conditions of linear  
  
SpeciesModel=lm(Species~Area, data=SpeciesArea)  
abline(SpeciesModel) # See that the line deosn't work for us



# Doing Residual analysis   
plot(SpeciesModel$residuals~SpeciesModel$fitted.values)  
abline(0,0)



# When looking at plots, you can see that it's pretty bad   
# Residual analysis can also be done with plot(mod1, 1:2)

# Tells you which point is biggests value of residual   
max(SpeciesModel$residuals)

## [1] 35.24152

# Tells you where the value is in the table   
which.max(SpeciesModel$residuals)

## 3   
## 3

# Gives you the row of the max, this is the number that you got from which.max above  
SpeciesArea[3,]

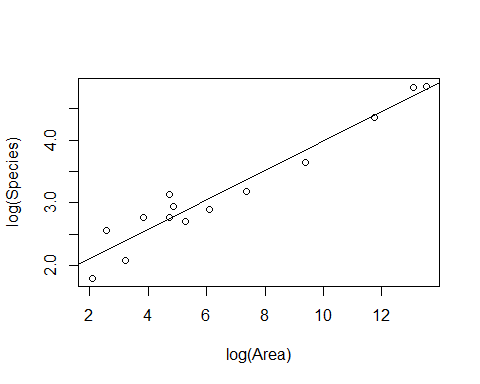
## Name Area Species logArea logSpecies  
## 3 Java 125628 78 11.7411 4.35671

# Just another way to call SpeciesArea[3,]  
SpeciesArea[SpeciesArea$Name=="Java",]

## Name Area Species logArea logSpecies  
## 3 Java 125628 78 11.7411 4.35671

**New Transformation Model**

# New transformation model   
plot(log(Species)~log(Area), data=SpeciesArea)  
SpeciesModel2=lm(log(Species)~log(Area), data=SpeciesArea)  
abline(SpeciesModel2) # Plotting the linear regression line on the scatterplot of the data

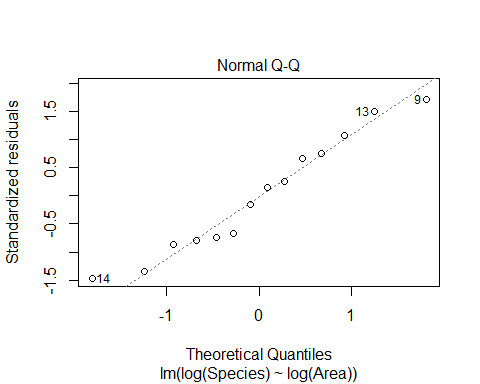
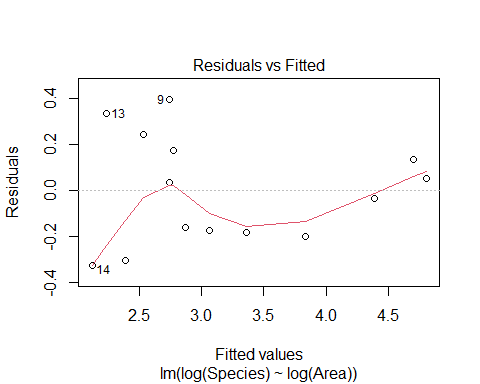


**HOw to interpret the summary table of a transformed linear model** - *Interpret:* For every 1 unit increase in the log(area), there is a 0.2355 increase in log(species)

summary(SpeciesModel2) # Gives you the output of the linear model

##   
## Call:  
## lm(formula = log(Species) ~ log(Area), data = SpeciesArea)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.32280 -0.18071 0.00079 0.16356 0.39534   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.6249 0.1326 12.26 3.81e-08 \*\*\*  
## log(Area) 0.2355 0.0175 13.46 1.34e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2427 on 12 degrees of freedom  
## Multiple R-squared: 0.9379, Adjusted R-squared: 0.9327   
## F-statistic: 181.1 on 1 and 12 DF, p-value: 1.335e-08

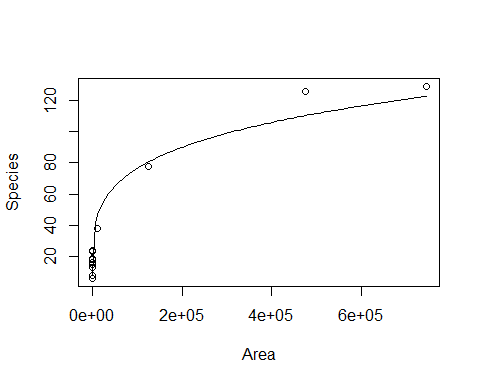
# Look at the coefficients table: Intercept and slope for the Log(area)   
# Intercept (B0) for Log(area) = 1.6249, intercept of the reg line   
# Slope (B1): Log(Area): 0.2335  
# For every 1 unit increase in the log(area), there is a 0.2355 increase in log(species)  
  
# Checking the conditions of the transformed linear model   
plot(SpeciesModel2, 1:2)



**Pulling out the coeffecients of the linear model** - Also shows how to plot the linear model - You need to solve for the same variables before you plot the linear model curve on the base plot because otherwise, it’s in different variables

*BELOW: HOW TO SOLVE OUT FOR LOG OF BOTH SIDES* **IMPORTARNT**

B0 = summary(SpeciesModel2)$coefficients[1,1] # Intercept  
B1 = summary(SpeciesModel2)$coefficients[2,1] # Slope  
  
plot(Species~Area, data=SpeciesArea)  
curve(exp(B0)\*x^B1, add=TRUE) # This is the linear model curve, on the normal data, but solved so that they are in the same units.



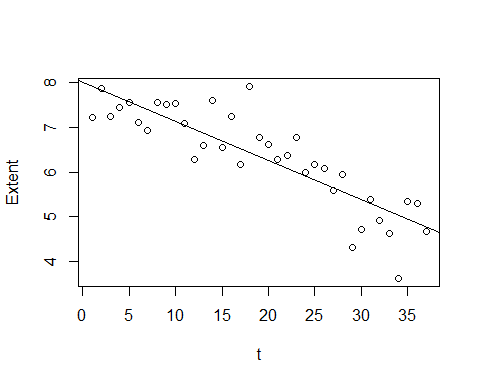
**Artic Sea Ice** The SeaIce data gives information about the amount of sea in the arctic region as measured in September (the time when the amount of ice is at its least) since 1979. The basic research question is to see if we can use time to model the amount of sea ice.

In fact, there are two ways to measure the amount of sea ice: Area and Extent. Area measures the actual amount of space taken up by ice. Extent measures the area inside the outer boundaries created by the ice. If there are areas inside the outer boundaries that are not ice (think about a slice of swiss cheese), then the Extent will be a larger number than the Area. In fact, this is almost always true.

data("SeaIce")  
head(SeaIce)

## Year Extent Area t  
## 1 1979 7.22 4.54 1  
## 2 1980 7.86 4.83 2  
## 3 1981 7.25 4.38 3  
## 4 1982 7.45 4.38 4  
## 5 1983 7.54 4.64 5  
## 6 1984 7.11 4.04 6

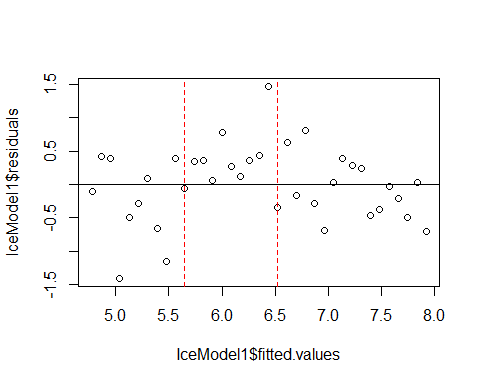
plot(Extent~t, data = SeaIce) # Basic plot of the oringial data   
IceModel1=lm(Extent~t, data = SeaIce) # Linear model of extent by time   
abline(IceModel1) # Plotting the line on the plot



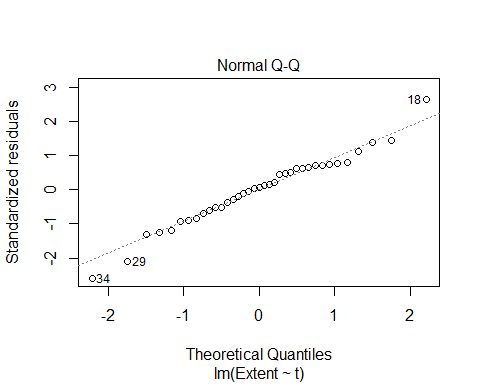
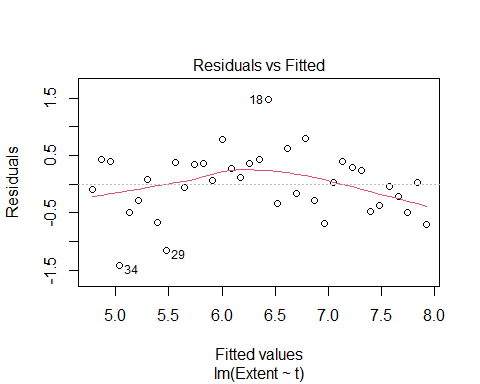
# We see that th eline look spretty good on the data

**Looking at Residuals** - The residual plot has a slight curve - The red line = the benefit of using the plotted of the model itself, rather than the residuals by fitted separate; shows that there is some pattern and curve there - We can see this really well at the middle region - There is some region bt 5 and 6.5 where all the residuals are above the line - So your prediction in that range will always be below what it should be

plot(IceModel1$residuals~IceModel1$fitted.values)  
abline(0,0) #Could also write plot(mod1, 1:2)  
  
# Below used the line to draw two vertical lines where the residual plot was looking weird  
# Will draw 2 vertical lines, one at X of 5.65 and the other 6.52  
abline(v=c(5.65,6.52),   
 col=c("red", "red"),   
 lty=c(2,2), # Look like a dash line   
 lwd=c(1, 1)) #Draws red dashed vertical lines; width

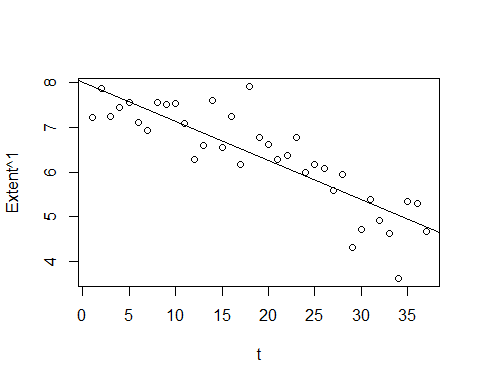


# The abline above shows you where the plot is under predicting   
  
plot(IceModel1, 1:2)

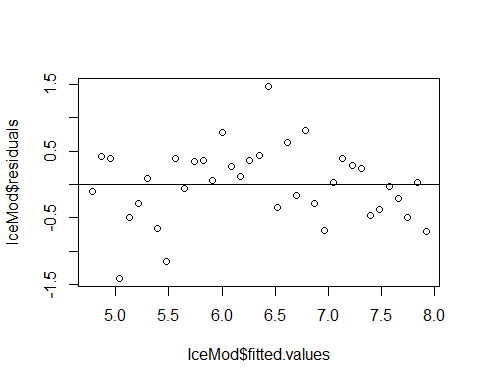


**How does doing transformations change the output?** - Trying exponential to 1

plot(Extent^1~t, data = SeaIce) # Basic plot of the oringial data   
IceMod=lm(Extent^1~t, data = SeaIce) # Linear model of extent by time   
abline(IceMod)



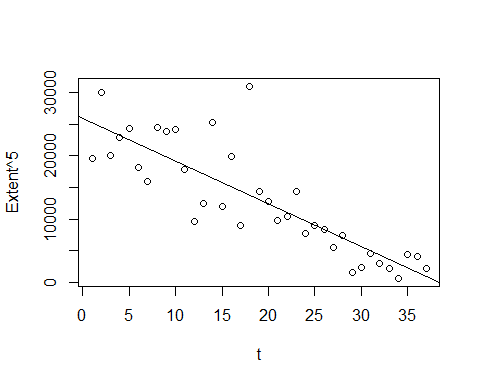
plot(IceMod$residuals~IceMod$fitted.values)  
abline(0,0) #Could also write plot(mod1, 1:2)



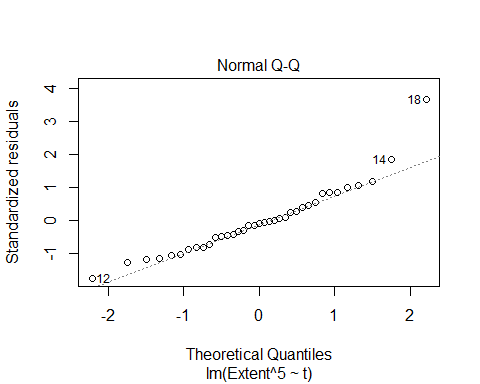
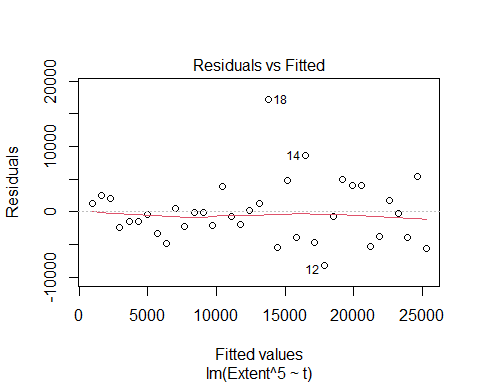
**Trying to raise to the 5th power** - Notice: This looks a bit better on the residual plot - Subjective - No very defined curvature - Look at teh oringal data, we see that there are not some region anymore where all the dots are above or below the line - This might have helped

* We may have also made other problems. See that one middle point that is well above the line? The transformation might have made things worse in different ways

plot(Extent^5~t, data = SeaIce) # Basic plot of the oringial data   
IceMod=lm(Extent^5~t, data = SeaIce) # Linear model of extent by time   
abline(IceMod)



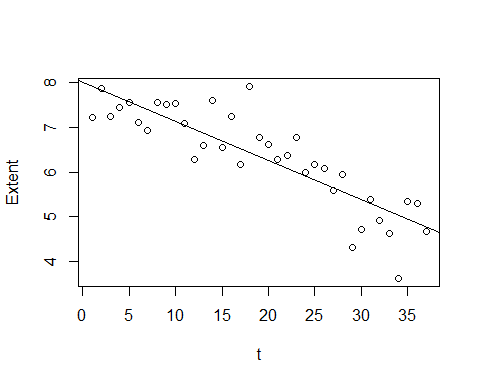
plot(IceMod, 1:2)



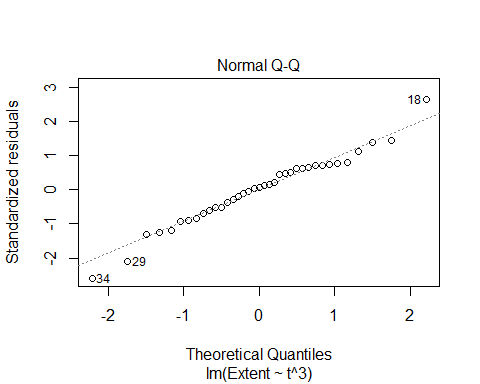
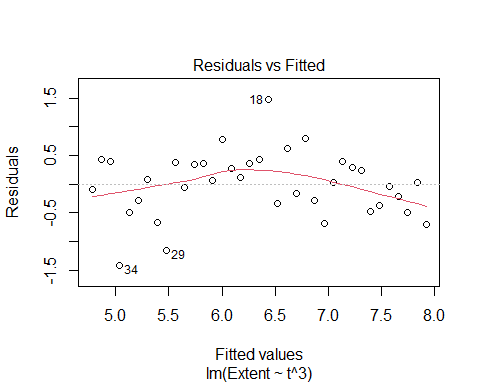
# The reisudal by fitted in teh orinigal plot had a red curve;but in this one the line has appeared to be subdued, but it may be decieving ebcause it plots the range of the data   
# it might just be the extreme case fo 18 appear to stretch it out and tone down what it actualyl looks like when I show it   
# Not as defined as it was before   
  
# The normal QQplot looks pretty good, there the 2 on teh tail that are flying out, but for the small dataset, things appear to fit well   
  
# No constant variance issues here   
  
# Might not be a good model, but it's a better model

**What if we take off the 5th power? and raise time tot he power instead?** - THis gets the curve back in the residual plot, and we don’t want that. so this makes it worse/same - R is ignoring you when you trying to raise the predictor to a power. It thinks that you’re not trying to do that

# THIS IS THE WRONG WAY  
plot(Extent~t^3, data = SeaIce) # Basic plot of the oringial data   
IceMod=lm(Extent~t^3, data = SeaIce) # Linear model of extent by time   
abline(IceMod)



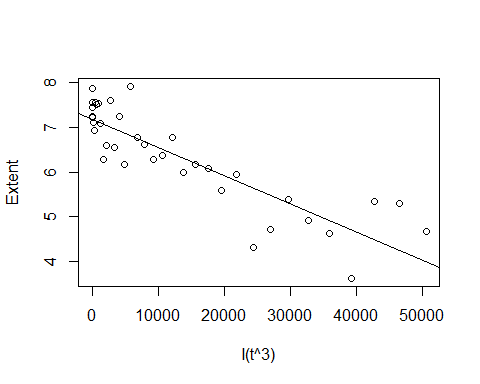
plot(IceMod, 1:2)



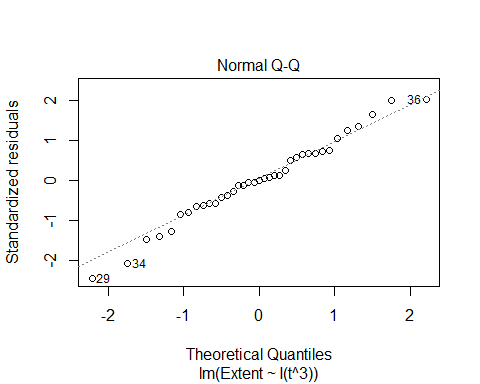
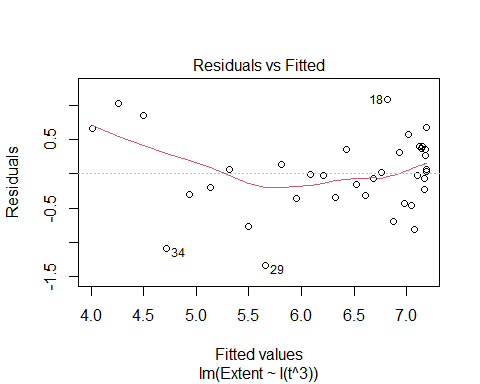
**How to power the predictor**  
-Somethings that work outside of functions may work differently inside of functions. - The carrot that should be raising it, worked on the response, but it won’t work on the predictor, inside of lm - It thinks you’re trying to do an interaction between variables, and you’re not trying to do that.

**Below: HOW TO Power PREDICTOR PROPERLY** - Still doens’t make a good model in this instance, but it could help in the future

# THIS IS THE RIGHT WAY  
plot(Extent~I(t^3), data = SeaIce) # Basic plot of the oringial data   
IceMod=lm(Extent~I(t^3), data = SeaIce) # Linear model of extent by time   
abline(IceMod)

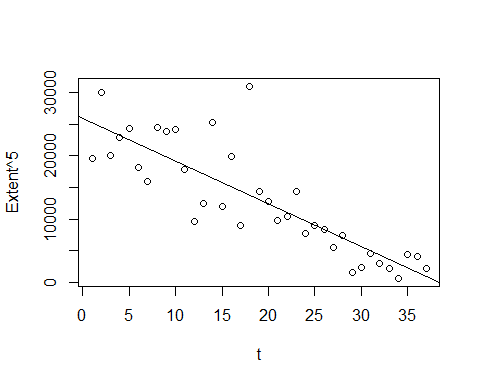


plot(IceMod, 1:2)

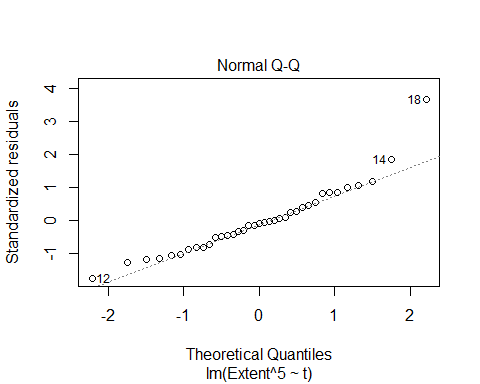
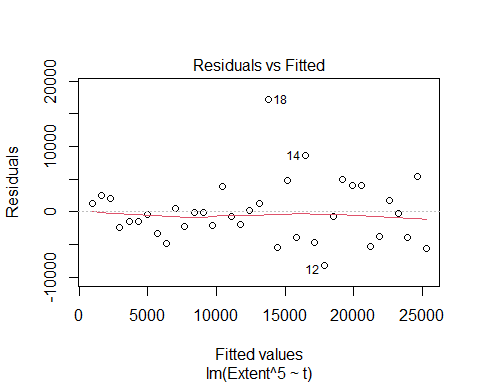


**USING THE BEST MODEL** - WE are using the 5th power model - What if we want to plot this on the orignial raw data?

plot(Extent^5~t, data = SeaIce) # Basic plot of the oringial data   
IceMod=lm(Extent^5~t, data = SeaIce) # Linear model of extent by time   
abline(IceMod)



plot(IceMod, 1:2)

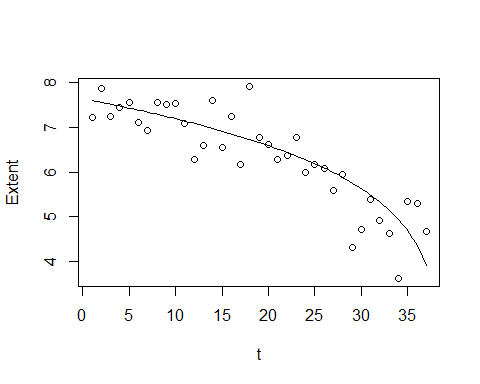


**PLOTTING TRANSFORMATION ON RAW DATA**

plot(Extent~t, data = SeaIce)  
  
summary(IceMod)

##   
## Call:  
## lm(formula = Extent^5 ~ t, data = SeaIce)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8157.2 -3296.7 -438.1 2102.9 17179.4   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 25969.69 1597.27 16.259 < 2e-16 \*\*\*  
## t -676.85 73.29 -9.235 6.51e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4760 on 35 degrees of freedom  
## Multiple R-squared: 0.709, Adjusted R-squared: 0.7007   
## F-statistic: 85.29 on 1 and 35 DF, p-value: 6.512e-11

# Pull out coeff of the transform model   
#INtercept  
B0\_Ice = summary(IceMod)$coefficients[1,1] # Intercept  
   
#Slope  
B1\_Ice = summary(IceMod)$coefficients[2,1] # Slope   
  
# Solve for the curve with math   
curve((B0\_Ice+B1\_Ice\*x)^(1/5), add = TRUE) # Tke 5th root of each side, jsut solved



## STOR 455 Class 6 Outliers and Points of Influence

# message=FALSE, warning=FALSE suppress warnings and messages from appearing in knitted files  
  
library(readr)  
library(Stat2Data)

**Single Quantitative Predictor Model** Notation:  
- Y = Response variable - X = Predictor variable

Assume (for now) that both Y and X are quantitative variables.

**Simple Linear Model** X = Single quantitative predictor Y = Quantitative response

Find a line that best summarizes the trend in the data.

Y = Bo + B1X + E Response = intercept + Slope\*Predictor + Random Error

**Simple Linear Model- Conditions** **Model:** 1.Linearity: The means for Y vary as a linear function of X. **Error:** 2.Zero Mean: The distribution of the errors is centered at zero. 3.Constant variance: The variance for Y is the same at each X. (Homoscedasticity) 4.Independence: No relationships among errors. 5.Normality: Residuals are normally distributed (sometimes) At each X, the Y’s follow a normal distribution.

*Look at* What potent do these points have to influence our model?

**Types of “Unusual” Points in SLM** - Two Types - **Outlier:** A data point that is far from the regression line. –Points really above or below the regression line – Doesn’t always have a lot of influence on the model – Could be big enough that it has influence, but mostly depends on the value of the predictor – Data points that are closer to the edges of the predictor value (high or low) have a higher chance of having inlfuence in our model - **Influential point:** A data point that has a large effect on the regression fit. – can come from many things

**Detecting Unusual Cases - Overview** 1. Compute residuals “raw”, standardized, studentized 2. Plots of residuals (or std. residuals) Boxplot, scatterplot, normal plot 3. Leverage Unusual values for the predictors 4. Cook’s distance Cases with large influence

*This notebook covers the first two from above* - leverage = the potiential for a certain value to have infleuence on the model - Cook’s distance combines a lot of the things to do some calculations for us

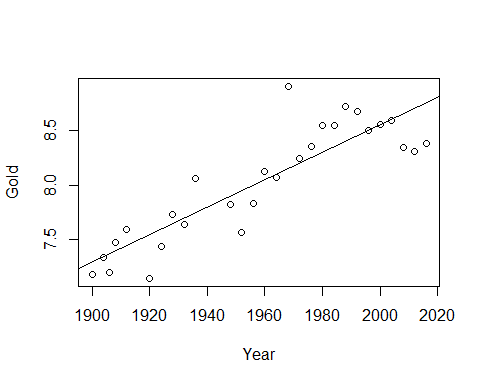
**Raw Residual** ei = yi - yhati *How can we tell if a residual is unusually large?*  CONTEXT! Example: Y = GPA ei = 2.6 is very large Y = SAT ei = 2.6 is very small

**Example: Men’s Olympic Long Jump**

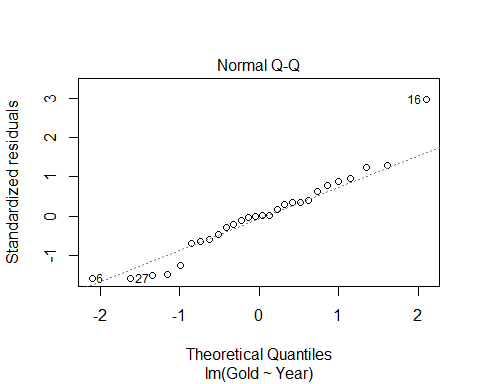
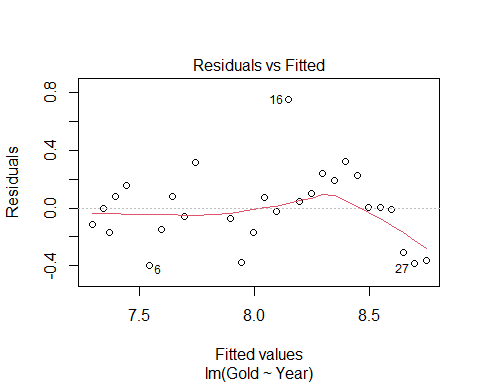
data("LongJumpOlympics2016")  
head(LongJumpOlympics2016)

## Year Gold  
## 1 1900 7.185  
## 2 1904 7.340  
## 3 1906 7.200  
## 4 1908 7.480  
## 5 1912 7.600  
## 6 1920 7.150

plot(Gold~Year, data=LongJumpOlympics2016) # Predict longjump distance by year   
GoldModel = lm(Gold~Year, data=LongJumpOlympics2016)  
abline(GoldModel) # Draw the line we made onto the plot



plot(GoldModel, 1:2) # To see the residual models, we see that there is one really big outlier



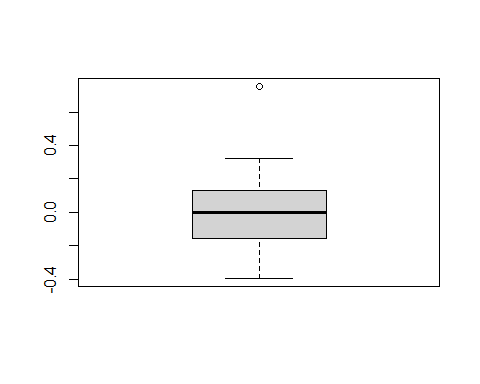
# Linearity looks like an issue because of the fitted plot. The red line goes down; the prediction curves   
# Point 16 is our outlier   
# R tells us which row this is because it looks like an ourlier to R   
# Looking at the normal QQ Plot, normal might be an issue   
# Constant variance doesn't appear to be an issue   
summary(GoldModel)

##   
## Call:  
## lm(formula = Gold ~ Year, data = LongJumpOlympics2016)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.39610 -0.15495 -0.00137 0.11606 0.75349   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -16.470194 2.666282 -6.177 1.56e-06 \*\*\*  
## Year 0.012508 0.001361 9.191 1.19e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2595 on 26 degrees of freedom  
## Multiple R-squared: 0.7646, Adjusted R-squared: 0.7556   
## F-statistic: 84.47 on 1 and 26 DF, p-value: 1.192e-09

# Look at estimate of the year: For every 1 year increase than that's how many meters we think the winning long distance jump is going to increase as well   
# Every 4 years it looks like its increasing by 5 cms

**What if we wanted to see what the data looked like without the outlier?**

boxplot(GoldModel$residuals)



# Outliers are more than 1.5 IQR’s beyond the Quartiles  
# Will give idea of how different that one value is from the others   
  
# Wee see there is an outlier, but how much of an outlier is this?   
# LOOK AT STANDARDIZED VALUES OF RESIDUALS INSTEAD  
  
max(GoldModel$residuals)

## [1] 0.7534932

which.max(GoldModel$residuals)

## 16   
## 16

**Standardized Residuals** - HOW TO TELL HOW MUCH OF AN OUTLIER THIS IS - ROughly equal to the acutal - predicted/stdeve; basically a zscore, but not exactly

* Fact: If X has mean mu and std. dev, then (𝑋−𝜇)/𝜎 has mean 0 and std. dev.=1.
* For residuals: mean=0 and std. dev. of errors
* Standardized Residuals about equals (yi - yhat)/stad dev of the population errors
* Look for values beyond +/-2 (mild) or beyond +/-3
* Once you have fit mymodel=lm(Y~X)
* Use: rstandard(mymodel)

*notes* - It will give us a thing centered at zero +/- unites; and that’s how many std they are away from teh average - THink about this as: - Once you are +/- 2 std away and its normally dist, then you’re in teh outer 5% of the data, so that’s starting to eb an outlier - IF you’re +/-3 away, then you’re into the .05 of the data and outliers, so it’s pretty extreme

*Below* - If we look at the standaized residual, it’s 2.96. so it’s 2.96 std above the line, which is an outlier

rstandard(GoldModel) # Put the model in

## 1 2 3 4 5 6   
## -0.45846037 -0.02447194 -0.69927805 0.34245659 0.62395356 -1.58872766   
## 7 8 9 10 11 12   
## -0.60354515 0.33349306 -0.22282237 1.24041457 -0.28038843 -1.47783119   
## 13 14 15 16 17 18   
## -0.65303672 0.28865616 -0.10391829 2.96025507 0.17095826 0.40756162   
## 19 20 21 22 23 24   
## 0.96124413 0.76569814 1.28443310 0.89029407 0.01303494 0.01295953   
## 25 26 27 28   
## -0.02754601 -1.24791995 -1.58338033 -1.51206192

# Will show standardized residuals   
# Taht's great when we have a small dataset, but if its not small, then don't use the above code   
  
which.max(GoldModel$residuals)

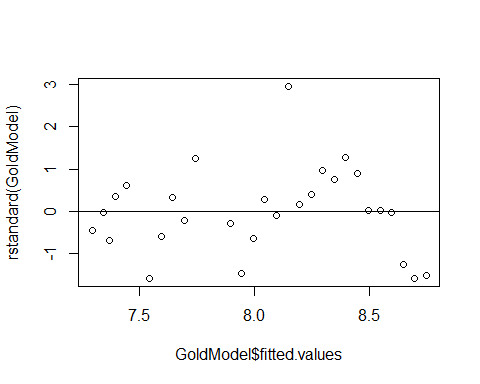
## 16   
## 16

rstandard(GoldModel)[16] # This will target the key point we ar elooking at

## 16   
## 2.960255

**We now know it’s an outlier, BUT WE DONT KNOW IF IT HAS ANY INFLUENCE ON TEH MODEL YET** - Can plot the rstand of the residuals by the fitted values - The plot is going to look identical to the others, other htan the axies, so but it’s a different measure of scale - How much infleunce is that really having?

plot(rstandard(GoldModel)~GoldModel$fitted.values)  
abline(0,0)

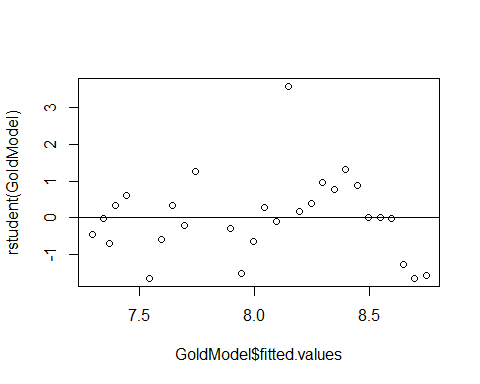
 **THING ABOUT THE STUDNTIZED RESIDUALS TO SEE HOW MUCH INFLUENCE SOMETHING HAS** - Standard = Outliers - Student = Influence (uses a different standard deviation)

**Studentized Residual** - Takes the single data out of the dataset, make a new regression line and get a new std of resid and seeing how far away the new line is from the outlier point in terms of the new standard residuals - If we take out an oulter, the std of the residuals is going to go down because we are removing an extreme case, so now its going to take more std to get to the outlier, and its going to give us a value that is bigger - When we see a studentized residual that is larger thant the stadnard residual, then it tells us that by removing this point, we are reallying changing the varibaility of the model and condenseing it more. - uses a different standard deviation than standariza

* **Concern:** An unusual value may exert great influence on the fit
* Its residual might be underestimated because the model “moves” a lot to fit it and/or
* The standard error of regression may be inflated due to the outlier error
* **Studentize:** Fit the model without that case, then find new 𝑦−𝑦hat and 𝜎\_𝜀 (of the population) to standardize. (R does this for every point)

**Influence** The effect of a single data point on the regression line depends on: - how well it matches the “trend” of the rest of the points - how “unusual” is its predictor value

plot(rstudent(GoldModel)~GoldModel$fitted.values)  
abline(0,0)



rstudent(GoldModel)[16]

## 16   
## 3.565083

# When we took the outlier from teh thing, its more away, which is bigger   
# There is some kind of influence here, but is it really noticable influence?

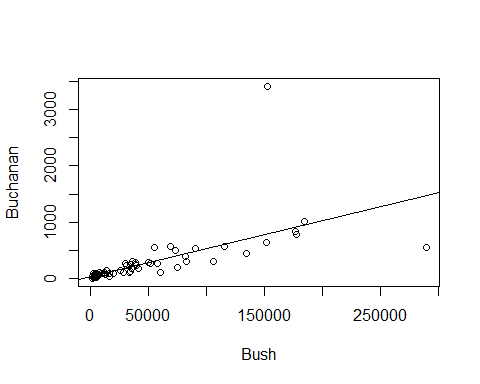
# plot(IceModel3) # From the previous notes   
  
# max(rstandard(IceModel3))  
# max(rstudent(IceModel3))  
# When we look at the values, themore different they are the mode influence they have in the model   
# The more close they are, then they have less influence on the model   
# No real bounds on what is a big or little influence on set number ot look at

**Dataset: PalmBeach** - County vote counts in Florida (n=67) for George Bush and Pat Buchanan in 2000. - Model: Use Bush votes to predict Buchanan votes.

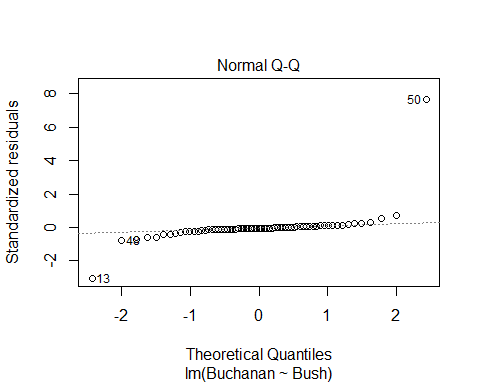
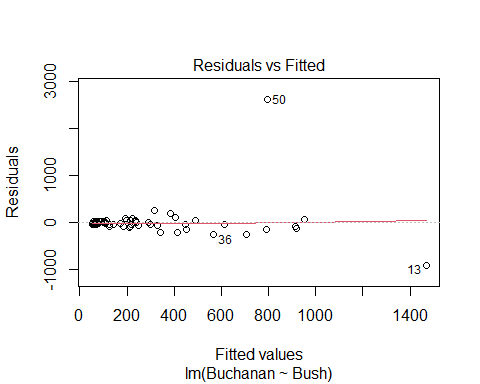
data(PalmBeach)  
head(PalmBeach)

## County Buchanan Bush  
## 1 ALACHUA 262 34062  
## 2 BAKER 73 5610  
## 3 BAY 248 38637  
## 4 BRADFORD 65 5413  
## 5 BREVARD 570 115185  
## 6 BROWARD 789 177279

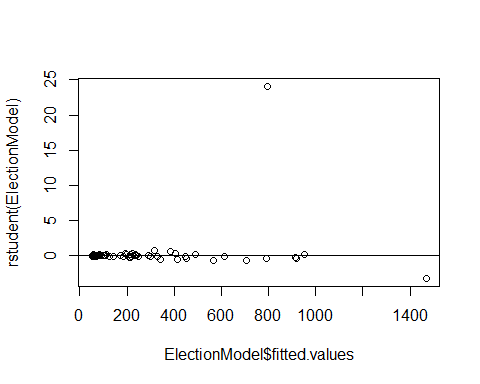
ElectionModel = lm(Buchanan~Bush, data=PalmBeach)  
# Bush = republican   
# Buchana - the other person   
  
plot(Buchanan~Bush, data=PalmBeach) # Look at the residuals and check the conditions   
abline(ElectionModel)

 **Example: Palm Beach Butterfly Ballot** - Palm Beach is the outlier

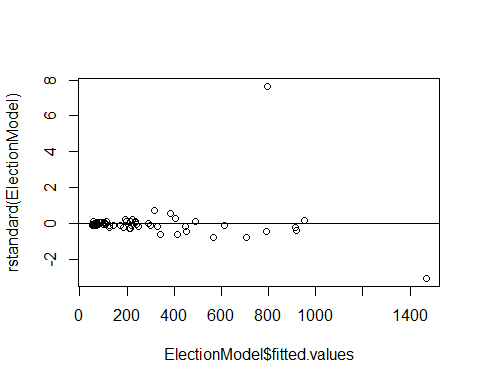
plot(ElectionModel, 1:2)



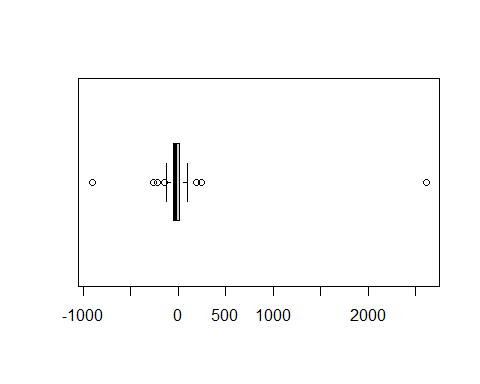
plot(rstudent(ElectionModel)~ElectionModel$fitted.values)  
abline(0,0)



plot(rstandard(ElectionModel)~ElectionModel$fitted.values)  
abline(0,0)



boxplot(ElectionModel$residuals, horizontal=TRUE) # Look at the outliers, there looks like ther eare a lot of outliers, but it's okay we are only looking at one of them



**What to do with an extreme residual?** - Try a transformation (Loging works really well, log(Data)) - Redo the analysis with the point omitted

*Below, we redo with the point omitted*

newdata = subset(PalmBeach, County!="PALM BEACH")  
  
ElectionModel\_noPB = lm(Buchanan~Bush, data=newdata)  
  
summary(ElectionModel)

##   
## Call:  
## lm(formula = Buchanan ~ Bush, data = PalmBeach)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -907.50 -46.10 -29.19 12.26 2610.19   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.529e+01 5.448e+01 0.831 0.409   
## Bush 4.917e-03 7.644e-04 6.432 1.73e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 353.9 on 65 degrees of freedom  
## Multiple R-squared: 0.3889, Adjusted R-squared: 0.3795   
## F-statistic: 41.37 on 1 and 65 DF, p-value: 1.727e-08

summary(ElectionModel\_noPB)

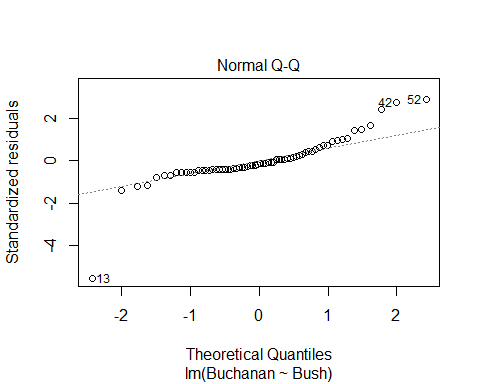
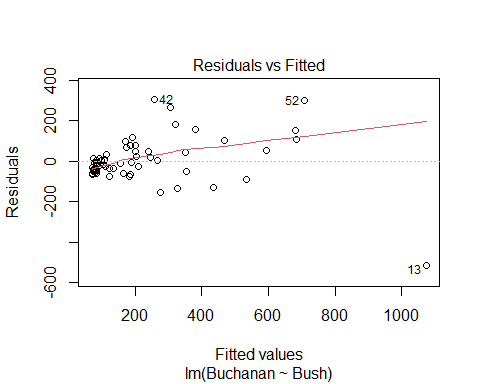
##   
## Call:  
## lm(formula = Buchanan ~ Bush, data = newdata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -512.43 -47.97 -17.09 41.78 305.45   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.557e+01 1.733e+01 3.784 0.000343 \*\*\*  
## Bush 3.482e-03 2.501e-04 13.923 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 112.5 on 64 degrees of freedom  
## Multiple R-squared: 0.7518, Adjusted R-squared: 0.7479   
## F-statistic: 193.8 on 1 and 64 DF, p-value: < 2.2e-16

# Compare teh summary of old vs new model   
# We see that the new model residuals std error is really high; it s std value   
# The origanl value is a lot smaller for std

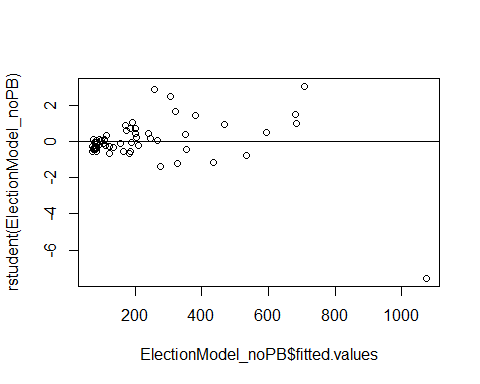
**Model with/without Palm Beach** - If compare the rstudent and rstandard of the with palm beach, we see that there is a really big jump between tehse numbers; tells us we have a value that is taking the whole regression and dragging it - So all our other predictiosn are going to be higher becuase of this vlaue

* If we look at the intercept and slope
* the slope is more dramatic when we take out the point, with the intercept goes down
* the residual standarad errors are very different between teh two
* the linearirity is really good without the palm beach
* the normal residuals look pretty bad when you take out palm beach

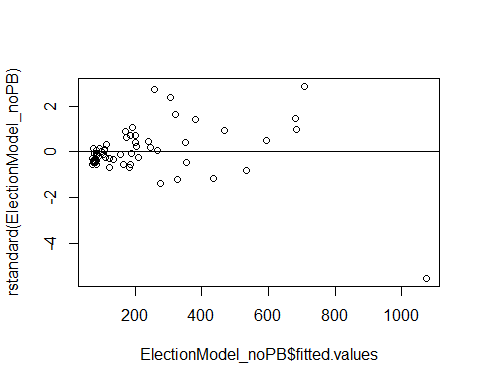
plot(ElectionModel\_noPB, 1:2)



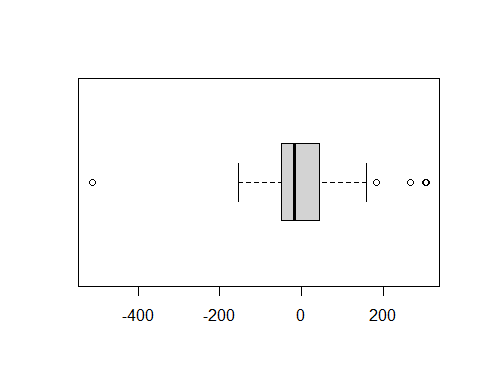
plot(rstudent(ElectionModel\_noPB)~ElectionModel\_noPB$fitted.values)  
abline(0,0)



plot(rstandard(ElectionModel\_noPB)~ElectionModel\_noPB$fitted.values)  
abline(0,0)



boxplot(ElectionModel\_noPB$residuals, horizontal=TRUE)



**BE careful with the student and standard** - look at how differen tbetween student and standard, you cna’t just know based on the one value

## STOR 455 Class 7 Outliers and Points of Influence Again

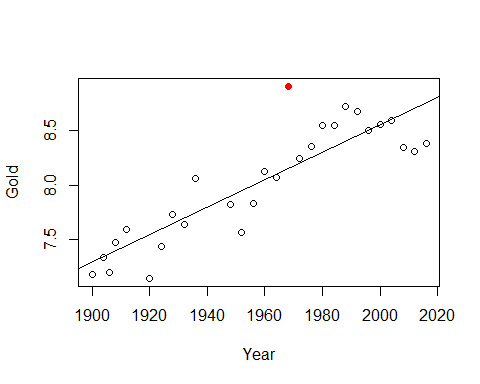
library(Stat2Data)  
  
data("LongJumpOlympics2016")  
data("PalmBeach")

**Types of “Unusual” Points in SLM** -**Outlier:** A data point that is far from the regression line. -**Influential point:** A data point that has a large effect on the regression fit.

**Detecting Unusual Cases - Overview** 1. Compute residuals - “raw”, standardized, studentized 2. Plots of residuals (or std. residuals) - Boxplot, scatterplot, normal plot 3. Leverage - Unusual values for the predictors 4. Cook’s distance - Cases with large influence

*Below:* We are making the same model as last time, in class 06

GoldModel=lm(Gold~Year, data=LongJumpOlympics2016)  
  
plot(Gold~Year, data=LongJumpOlympics2016)  
abline(GoldModel)  
points(LongJumpOlympics2016$Year[16], LongJumpOlympics2016$Gold[16], col="red", pch=16)



# Points just stands out a certain point   
# We need to give it the x and y coordinates, the the color and style   
  
summary(GoldModel) # The 0.2595 meaters is the standard error of the residual

##   
## Call:  
## lm(formula = Gold ~ Year, data = LongJumpOlympics2016)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.39610 -0.15495 -0.00137 0.11606 0.75349   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -16.470194 2.666282 -6.177 1.56e-06 \*\*\*  
## Year 0.012508 0.001361 9.191 1.19e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2595 on 26 degrees of freedom  
## Multiple R-squared: 0.7646, Adjusted R-squared: 0.7556   
## F-statistic: 84.47 on 1 and 26 DF, p-value: 1.192e-09

**Standardized Residuals** - Roughly equal to zscore - For residuals: mean=0 and std. dev. ≈𝜎 ̂\_𝜀 - Standardized Residual about = (yi-yhat)/std - Look for values: beyond +/-2 (mild) or beyond +/-3 - **Definition:** The standardized residuals are: – std. resi = ((yi-yhat)/(stdsqrt(1-hi))); where hi = leverage

*SQRT 1=leverae* - It wont tell us that it has influence, but it will tell us the potential for influence -

StanResidEst = GoldModel$resid/summary(GoldModel)$sigma # Pulls out the standard error of the residuals   
  
StanResidEst - rstandard(GoldModel) # What this leverage is

## 1 2 3 4 5   
## 0.0310213900 0.0014925005 0.0404484002 -0.0187751023 -0.0306759692   
## 6 7 8 9 10   
## 0.0624609303 0.0211917539 -0.0104655385 0.0062653144 -0.0314031236   
## 11 12 13 14 15   
## 0.0055095767 0.0275829381 0.0118383564 -0.0052073979 0.0019121259   
## 16 17 18 19 20   
## -0.0568637884 -0.0034991146 -0.0090387925 -0.0233983067 -0.0206451799   
## 21 22 23 24 25   
## -0.0385895552 -0.0299054544 -0.0004902075 -0.0005455954 0.0012967412   
## 26 27 28   
## 0.0655674849 0.0926377694 0.0982533934

**What is this leverage?** - Depending on where the value is, depending ont eh x coord, determines the leverage - Things that follow the mean along the predictor because it doesn’t have a lot of leverage - If it’s twoards the right or left of end the range of values, then even a small, very little values could have a lot fo influence on the model - Think of leverage as a seesaw. – if you have a long seesaw on one side, but short on the other side, it doesn’t matter who is on the short side, they won’t have as much leverage. – If you have a long seesaw on the other side, then if it’s an adult or child may impact the stronger amount of leverage. Think of if it was an adult of child by how far up or down the y axis it goes. If it’s an adult, it’s oging to be a bigger different from the axis, but if it’s a child it’s going to be smaller and have less influence. - **Bottom line** A lower weight, if its further away fromt eh balance point, can have more effect on the model than a higher weight that is closer to the balance point – **The balance point is the mean**

StanResid = GoldModel$resid/(summary(GoldModel)$sigma \* sqrt(1 - hatvalues(GoldModel)))  
  
StanResid - rstandard(GoldModel)

## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26   
## 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0   
## 27 28   
## 0 0

**Studentized Residuals** - If we took that point out of the model, how does that impact the variability of the residuals? - Definition: The studentized residuals are: stud.resi = *same as standardized, but without the one point* - If th epoint had a lot of influence on the varibaility, then the studentized value would be much bigger than the standarized value

* **The more different the standardized vs studentized values are, the more influence that point has on the model**

**Typical Leverage** - **For one point:** - For a simple linear model: hi = (1/n)+(((xi-xbar)2)/sum(xi-xbar)2) - **For all leverage in a model:** - sum(leverage) = sum(1/n) + ((sum(xi-xbar)2)/sum(xi-xbar)2) = 1+1 = 2 - The sum of the leverages of all of the points = 2 – this is usefulbecause we can think of what a typical leverage value is – If all teh leverages are 2, then 2/n = give you what the mean leverage is - *Look for:* – leverage > 2(2/n) – leverage hi > 3(2/n)

*Notes* - How different is this point from the average predictive value in compraision to all those other values - Xi - xbar = predictor where the xi = a specific predictor value, and xbar = the mean of the values - Xi-xbar squared = in a horizontal way, how different is that point regardless if above or below mean - The sum(xi-xbar)^2 = how different is this point proportionally to the rest of the data - scared accoriding the the sample of the data with the 1/n - One point will haev mor einfluence if the data is small - values that are double or triple the average leverage are the ones we have to worry about because they may have influence in the model – **We are still talking about potenital here because we are only looking at the predictors** –Because we are only looking at years in this model, vs the entire context of the data –It tells us the data points we should look at more closely

# average leverage : 2/28   
# IF poitns are bigger than this, might have a lot of influence on the model   
  
2\*(2/28)

## [1] 0.1428571

3\*(2/28)

## [1] 0.2142857

# Above are just the cut off points for influence, we don't really care too much about them   
  
# The below code will show the ordered data because of the way the data is organized in the dataset   
# We can see how the leverages are going to be similar there where the points that are small, very different for the mean, they wil start with a higher leverage, and by the time we get the middle values, its going down to a leverage of nearly zero,   
# Get further waya from average year, we will get higher leverages reported   
  
hatvalues(GoldModel)

## 1 2 3 4 5 6 7   
## 0.13075010 0.11825693 0.11234035 0.10664378 0.09591065 0.07708445 0.06899139   
## 8 9 10 11 12 13 14   
## 0.06177835 0.05544533 0.04999234 0.03891349 0.03698058 0.03592770 0.03575484   
## 15 16 17 18 19 20 21   
## 0.03646200 0.03804918 0.04051638 0.04386361 0.04809086 0.05319813 0.05918543   
## 22 23 24 25 26 27 28   
## 0.06605275 0.07380009 0.08242745 0.09193483 0.10232224 0.11358967 0.12573712

# Iif you look at the 16th value, it has a leverage of 0.038; based onteh creitera foor what tis a big lleverage value, this point doesn't appear to be high leverage   
# the 16the point has to be a REALLY big outlier for it to have influence, because its potential to have influence is low

**Wanted to look at the leverages better** - Sort them to have a better idea of where the potenitals are to gave influence - Sort the hat values - sort will by default go from acending order/increasing order - mostly interested in the biggest leverages - Want to change it to decreasing = TRUE so we see the biggest leverages first instead of the smallest ones first

2\*(2/28)

## [1] 0.1428571

3\*(2/28)

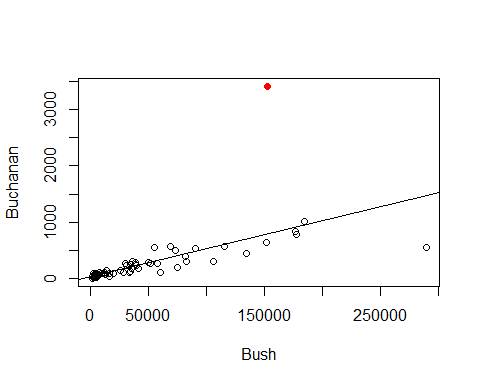
## [1] 0.2142857

sort(hatvalues(GoldModel), decreasing = TRUE)

## 1 28 2 27 3 4 26   
## 0.13075010 0.12573712 0.11825693 0.11358967 0.11234035 0.10664378 0.10232224   
## 5 25 24 6 23 7 22   
## 0.09591065 0.09193483 0.08242745 0.07708445 0.07380009 0.06899139 0.06605275   
## 8 21 9 20 10 19 18   
## 0.06177835 0.05918543 0.05544533 0.05319813 0.04999234 0.04809086 0.04386361   
## 17 11 16 12 15 13 14   
## 0.04051638 0.03891349 0.03804918 0.03698058 0.03646200 0.03592770 0.03575484

# The one with the biggests index is 0.1307, even the smallest and biggest predicotr value aresnt so far from teh mean predictor compared to the other values that hey have potiental for influence.   
# They would need to be big outliers to have influence   
# If we had more datapoints, then we could also just look at some of the values   
# The head function can be added to the sorting hatvalues

ElectionModel=lm(Buchanan~Bush,data=PalmBeach)  
plot(Buchanan~Bush,data=PalmBeach)  
abline(ElectionModel)  
points(PalmBeach$Bush[50], PalmBeach$Buchanan[50], col="red", pch=16)



# Red is palm beach

#Average leverage = 2/67  
  
#Potiential for influence   
2\*(2/67)

## [1] 0.05970149

3\*(2/67)

## [1] 0.08955224

head(sort(hatvalues(ElectionModel), decreasing=TRUE), n=10)

## 13 52 6 29 50 16 48   
## 0.29747301 0.10761608 0.09859725 0.09820784 0.07085197 0.07007421 0.05365982   
## 5 36 53   
## 0.03899504 0.03331541 0.02511932

# We see that point 13 probably has potiential leverage   
  
# We want to see the first 6 with the highest potienital for influence so we can look at them closer  
# We want all columns, so just leave it with the ,  
# This tells us teh counties in the order of teh output, not the leverages though   
# Want to add the leverages to it though   
PalmBeach[c(6,13,16,29,50,52),]

## County Buchanan Bush  
## 6 BROWARD 789 177279  
## 13 DADE 561 289456  
## 16 DUVAL 650 152082  
## 29 HILLSBOROUGH 836 176967  
## 50 PALM BEACH 3407 152846  
## 52 PINELLAS 1010 184312

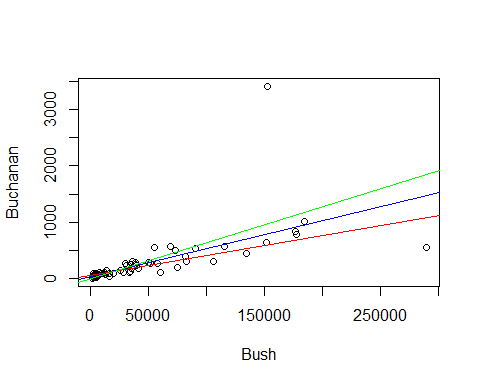
# Defines new variable inside of the dataframe  
PalmBeach$Leverage = hatvalues(ElectionModel)  
# This is putting the leverage into the dataframe   
  
# Now run this again: PalmBeach[c(6,13,16,29,50,52),]  
PalmBeach[c(6,13,16,29,50,52),]

## County Buchanan Bush Leverage  
## 6 BROWARD 789 177279 0.09859725  
## 13 DADE 561 289456 0.29747301  
## 16 DUVAL 650 152082 0.07007421  
## 29 HILLSBOROUGH 836 176967 0.09820784  
## 50 PALM BEACH 3407 152846 0.07085197  
## 52 PINELLAS 1010 184312 0.10761608

# This will now include the palmbeach levearge

*Want to visualize the impact dave county has vs palm beach* - Dave has a smaller outlier, it’s leverage is high so it could havea similar impact on the model

plot(Buchanan~Bush,data=PalmBeach)  
abline(ElectionModel, col="blue")  
# The above is the model we have now   
  
# Below, make new model with no palmbeach   
NoPalmBeach=subset(PalmBeach,County!="PALM BEACH")  
ElectionModel\_noPB=lm(Buchanan~Bush,data=NoPalmBeach)  
abline(ElectionModel\_noPB, col="red")  
  
# Below, make a new model with no dade  
NoDade = subset(PalmBeach,County!="DADE")  
ElectionModel\_noD=lm(Buchanan~Bush,data=NoDade)  
abline(ElectionModel\_noD, col="green")



# Then we compare the lines that we have written to see how the slope is different.   
# Red = no palm beach   
# Green = no dade   
# Blue = with both   
# The change in slope appears the same, so it tells me that the impact these datapoint shave on our model are about similar   
# Even though one is a bigger outlier, it has a lower potenital to have infleunce,   
# The other is smaller outlier, but has a high leverage, it has abotu the same typ eof influence

*NOtes* - FOr high leverage, doesnt need to be a big outlier to have an impact on the model - for high outlier, it doesnt have to have that much leverage - We are looking at cook’s distance, which combines teh two above things and quantify the value instead of the other vlaues - Cook’s will tell you if the pointhas influence on the model - Brings in stanard resid; the leverage of the point; and how many predictors there are - Cooks’ distance is acombien of leverage, student, adn standard

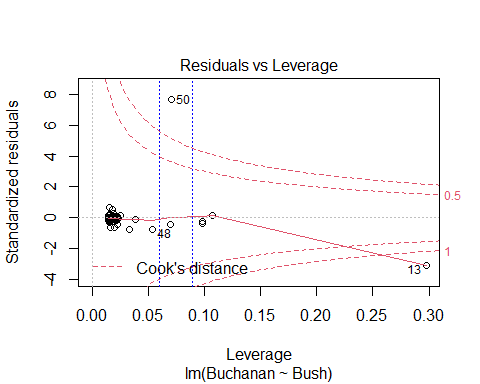
**Cook’s Distance** - How much would the fit change if one data value were omitted? - Di increases with either poor fit (std.resi) and high leverage (hi). 1. Compare to other Di’s. 2. Study any case with Di > 0.5; worry if Di > 1.0.

*Cook’s Di:* = (((std.resi)^2)/(k+1))\*(hi/(1-hi))

head(sort(cooks.distance(ElectionModel), decreasing=TRUE), n=5)

## 50 13 48 36 6   
## 2.231935359 1.981365681 0.016228381 0.009781056 0.007928585

# Two ar emuch bigger adn the rest of smaller, those two points rae the main that are havign influence on our model   
# 50 = =palm beach   
# 13 = Dade   
# They're not exactly the same, but have about teh same impact ont eh model we are making   
  
#We've seen Cook's distance between with teh plots for the model   
plot(ElectionModel,5)  
# Will show you what point shave leverage   
# LOOK AT WHAT THE THING IS   
# HOw big of an outlier is this value   
# Mode values have standardiex values within +/- 2  
# Th etwo that have big differences are point 50 and 13  
# Point 50 = std score of 8   
# Point 13, Dade, has a -3.0 std value; but it is high leverage   
# Points low leverage = closer to the left side of the graph   
# Further right = different data, more leverage   
  
# O.5  
# Gives us a cook's distance of 0.5  
# OUtler line: 1 = cook's distance of 1   
# Above these lines, we can claim that by putting the leverage and output together it's having an influence on the model   
# POint 50, has small leverage, but a big outlier that it's goig outside of the curve and having influence on the model   
# Dade county has a huge leverage, but it's not an outlier, but it's enough to put it outside of the curve and have an high impact on teh model   
  
# Think about: WHY DO THESE POINTS HAVE IMPACT ON THE MDOEL>   
# DO WE WANT THEM TO IMPACT THE MODEL?   
# sometimes might not want to inlcude some things int eh model, but it's good to look at it anyways so you know what is going on   
abline(v = 4/67, col="blue", lty=3)  
abline(v = 6/67, col="blue", lty=3)



## STOR 455 Class 8 Inference for Regression Slope

library(readr)  
library(Stat2Data)  
library(metRology)  
  
DistanceHome <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/DistanceHome.csv")  
Domestic=subset(DistanceHome,Distance<250)  
  
source("https://raw.githubusercontent.com/JA-McLean/STOR455/master/scripts/CIPIPlot.R")   
# Function that is not built into the thing   
# The source will allow you to call the function as you want

*NOtes* - If the data is packed around the line, then we can predict that the data will be by the line

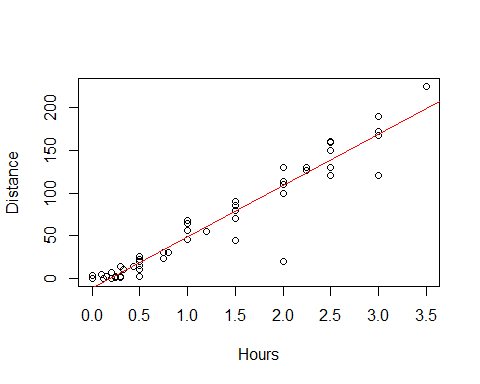
**Inference for Slope and Intercept** - Find a confidence interval of plausible values for the parameter. - Test a hypothesis about a possible value for the parameter

**Bootstrap Distribution** - THink of poopulation - Estimate the distribution and variability (SE) of β ̂\_𝑖 from the bootstraps - WE ar elooking atht the distnace data, - THink of the population and there is some population slope here - If I knew everyone’s distance from home, this would be the realsiton they have; we know the sample but not the population - It would be nice if we could take a lot of samples ad keep calculating that regression line (then you could see teh distribution) - Think of seeds that fall from a tree; from each seed that falls from teh tree we get a slope from that line. - We cant keep taking from teh ample, and we need to find where the actual population vlaue is

* Hypothesis what the population might look like based on this sample
* Maybe the population is just a lot of copies of this sample
* If we take samples form teh bootstrap poulation, what variability is there?
* We are making many copies of them
* We just have out 54 key people with our data, taking one out and seeing where they fall on the regression line, putting htem back and then pulling someone else – If you do this a bunch of times, sometimes you ll get a sample that is like the population or not; youll get a range of values

**Simple Linear Regression** *See below*

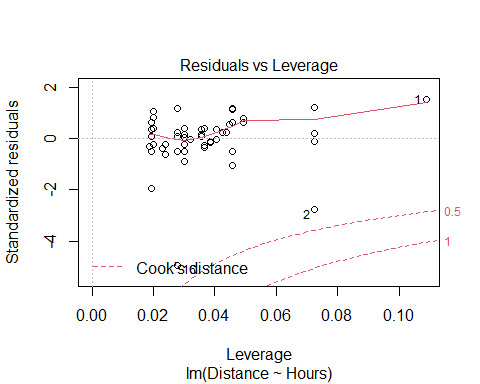
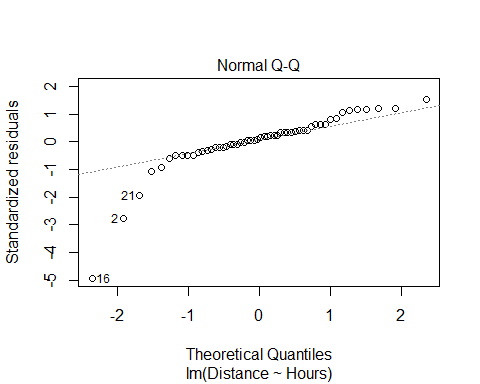
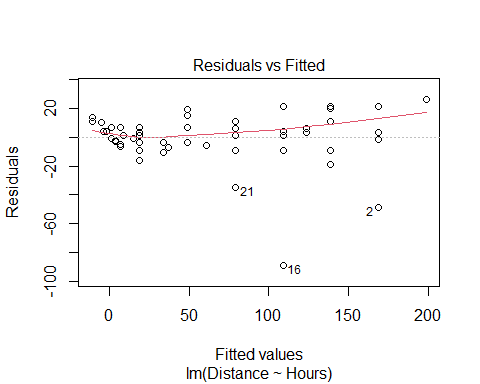
moddist=lm(Distance~Hours, data=Domestic)  
# Hours per distnace; predict distance form home based on hours it takes you to get home   
  
plot(Distance~Hours, data=Domestic)  
abline(moddist, col="red")



summary(moddist)

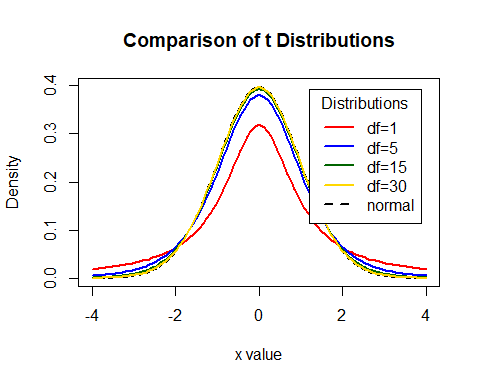
##   
## Call:  
## lm(formula = Distance ~ Hours, data = Domestic)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -88.892 -4.680 2.172 7.082 26.141   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -11.063 4.056 -2.727 0.00868 \*\*   
## Hours 59.977 2.484 24.144 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.26 on 52 degrees of freedom  
## Multiple R-squared: 0.9181, Adjusted R-squared: 0.9165   
## F-statistic: 582.9 on 1 and 52 DF, p-value: < 2.2e-16

# Slope = 59; for every 1 hour people ar eaway from home, then they are about 60 miles more away from home   
# We want to think about; How close is this to the population value?   
# What claims can I make about the population?  
  
plot(moddist, c(1, 2, 5))



# THis whole chunk of code creates a linear model, plots it against the data, and checks the conditions.   
# Pretty linear relaitonship overall: hours away from home, distance increase   
# Residuals have a curve that s alittle ; overall it doesnt seem very bad   
# normal QQ plot, ther eis a problem where one side has a skew   
  
#COok's ditance   
# High leverage, far right   
# High influence, up/down

# Display the Student's t distributions with various  
# degrees of freedom and compare to the normal distribution  
  
x <- seq(-4, 4, length=100)  
hx <- dnorm(x)  
  
degf <- c(1, 5, 15, 30)  
colors <- c("red", "blue", "darkgreen", "gold", "black")  
labels <- c("df=1", "df=5", "df=15", "df=30", "normal")  
  
plot(x, hx, type="l", lty=2, xlab="x value",  
 ylab="Density", main="Comparison of t Distributions")  
  
for (i in 1:4){  
 lines(x, dt(x,degf[i]), lwd=2, col=colors[i])  
}  
  
legend("topright", inset=.05, title="Distributions",  
 labels, lwd=2, lty=c(1, 1, 1, 1, 2), col=colors)



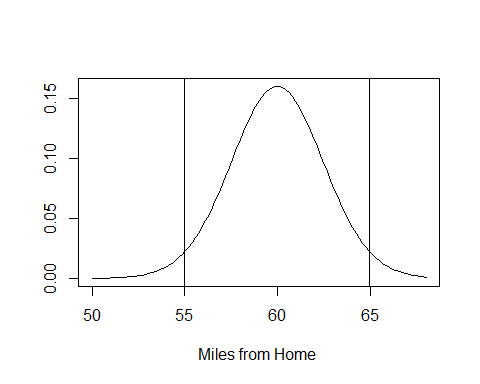
# What the tdist looks like   
#df comes down to sample size and other factors (predicotrs and modeltype)   
# Bigger sample, more degree of freedom   
# For 30 df, a sample of 32, this is pretty much a normal curve   
# As it gests smaller, we are pressing down and pressing out the sides of the graph   
#We are triyng to make this based on what the standard error of our sample is

**CI for Slope or Intercept** - HOw R is doing the math - BUild a distribution where the cente ris where the sample value is adnd we are making a curve on top of that and how far in each direction do we need to go were the middle 95% of the area under the curve is met? Bihar +/- tstar*StandardError or the Bihat Y = Bo + B1X+E - t* comes from a t-distribution with n-2 d.f and depends on the level of confidence - For 1−𝛼 level confidence, use qt(1-α/2,df)in R - **e.g. for 95% confidence and 52 df, qt(0.975,52)** - tstar = tells us how many STDERRORS in each direction we need to go

qt(0.975, 52)

## [1] 2.006647

# T = t distribution   
# area under curve to teh left of the point we want = first arguement; 0.975 = 95% confi int   
# df = 2nd argument; sampel size minus 2   
# Result = 2.00664 ish   
#If i want to see the confidence interval, you have to start at your sample of slope  
#qt gives you tstar  
  
curve(  
 dt.scaled(  
 x,   
 52,  
 mean = summary(moddist)$coef[2,1],  
 sd = summary(moddist)$coef[2,2]  
 ),   
 from = 50, to = 68,  
 xlab = "Miles from Home ",  
 ylab = " "  
 )  
  
  
abline(  
 v=c(  
 qt.scaled(  
 0.025,   
 52,   
 mean = summary(moddist)$coef[2,1],   
 sd = summary(moddist)$coef[2,2]  
 ),  
 qt.scaled(  
 0.975,   
 52,   
 mean = summary(moddist)$coef[2,1],   
 sd = summary(moddist)$coef[2,2]  
 )  
 )  
 )



# IF you want to see the confidence interval   
summary(moddist)$coef[2,1]-qt(0.975, 52)\*summary(moddist)$coef[2,2] #LOwer bound for confidence interval

## [1] 54.99264

summary(moddist)$coef[2,1]+qt(0.975, 52)\*summary(moddist)$coef[2,2] #Upper bound for confidence interval

## [1] 64.96233

#We are predicting that with 95% confidence the solution is between the 54.99-64.96

If we think tha the population might not be normal, then we probably want to do a bootstrap method

**How to find a confidence interval?** -confint(mymodel,level =0.XX) and adjust for the confidence level.

# HOW TO FIND CONFIDENCE INTERVAL; default 95% confidence   
# generally the intercept is not very useful to think about   
# We are predicting taht it could be close to zero   
# WE are mostly looking at the coeffs   
# The hours are about teh same as above   
confint(moddist, level=0.95)

## 2.5 % 97.5 %  
## (Intercept) -19.20166 -2.92358  
## Hours 54.99264 64.96233

**Accuracy of Predictions** Example: It takes a student 2.25 hours to drive from home. How many miles do we predict that thy are away from home? How accurate is that prediction? - Want to make a prediction for a specific case - Wnted regalr prediction, just plug the 2.25 into the regression line - It matterse what you are rtrying to predict - all people or the specific person’s distance from home? - There is a difference; one person has ore variability (Say they’re biking) - If we have the ditribution, ontop of that is some normal curve; most of the people are close to thtat, but they trail off a bit

**Two Forms of Intervals for Regression** 1. Confidence Interval for μY (mean Y) Where is the “true” line for that x? or Where is the average Y for all with that x? 2. Prediction Interval for Individual Y Where are most Y’s for that x?

\_\_CI for μY when X=x\*\_\_ - Predicting in general SSX = ∑▒〖(𝑥\_𝑖 − 𝑥 ̅)〗^2 yhat +/- tstar*standerror*sqrt((1/n)+((xstar-xbar)^2)/SSX)

\_\_Prediction Interval for Individual Y’s when X=x\*\_\_ - predicting for one person yhat +/- tstar*standerror*sqrt(1+ (1/n)+((xstar-xbar)^2)/SSX) Just add 1 in the sqrt

\_\_CI and PI via R when X=x\*\_\_

newx=data.frame(Hours=2.25) # Creat e a new person   
head(newx)

## Hours  
## 1 2.25

predict.lm(moddist, newx, interval="confidence") # Predict mean for all people who are 2.25 away from hoe

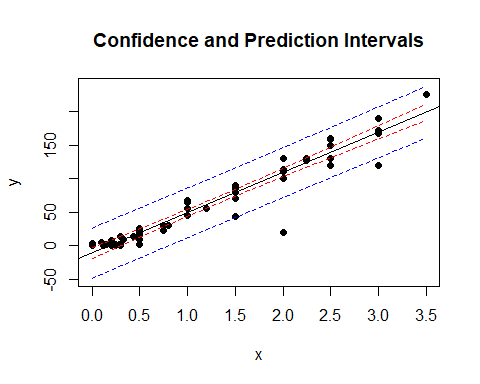
## fit lwr upr  
## 1 123.8867 116.9764 130.797

predict.lm(moddist, newx, interval="prediction") # Distance home for one specific person

## fit lwr upr  
## 1 123.8867 86.59458 161.1789

# Both gives us a fitted value, its a point in the regression lie   
# THis si what we would get if we plugged and chugged   
# One person is going to be a wider range because we want to make sure we get teh one person

CIPIPlot(Domestic$Hours, Domestic$Distance) # Visualize different between confidence and prediction



# calculates   
# For every possible point in teh data, or for the  
# What would be the confidence interveral for that value and what would be the prediction interval

* The red lines are the confidence interval
* if we are trying to predict the mean vlaue for people’s distance away from home based on tehse hours, we will predict the mean is somewhere between the red lines and its tight by the regression line with 95% confidence Th eblue line = much wider
* there’s a lot more variability there
* much wider

## STOR 455 Class 8 Inference for Regression Slope

library(readr)  
library(Stat2Data)  
library(metRology)  
  
DistanceHome <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/DistanceHome.csv")  
Domestic=subset(DistanceHome,Distance<250)  
  
source("https://raw.githubusercontent.com/JA-McLean/STOR455/master/scripts/CIPIPlot.R")   
# Function that is not built into the thing   
# The source will allow you to call the function as you want

*NOtes* - If the data is packed around the line, then we can predict that the data will be by the line

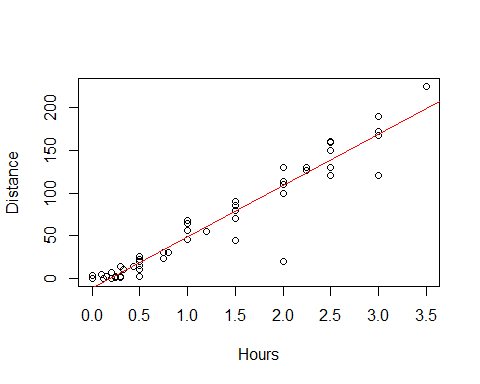
**Inference for Slope and Intercept** - Find a confidence interval of plausible values for the parameter. - Test a hypothesis about a possible value for the parameter

**Bootstrap Distribution** - THink of poopulation - Estimate the distribution and variability (SE) of β ̂\_𝑖 from the bootstraps - WE ar elooking atht the distnace data, - THink of the population and there is some population slope here - If I knew everyone’s distance from home, this would be the realsiton they have; we know the sample but not the population - It would be nice if we could take a lot of samples ad keep calculating that regression line (then you could see teh distribution) - Think of seeds that fall from a tree; from each seed that falls from teh tree we get a slope from that line. - We cant keep taking from teh ample, and we need to find where the actual population vlaue is

* Hypothesis what the population might look like based on this sample
* Maybe the population is just a lot of copies of this sample
* If we take samples form teh bootstrap poulation, what variability is there?
* We are making many copies of them
* We just have out 54 key people with our data, taking one out and seeing where they fall on the regression line, putting htem back and then pulling someone else – If you do this a bunch of times, sometimes you ll get a sample that is like the population or not; youll get a range of values

**Simple Linear Regression** *See below*

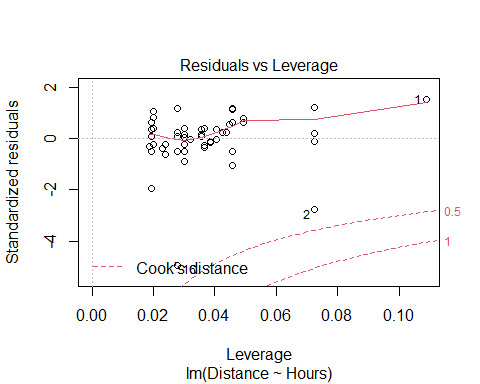
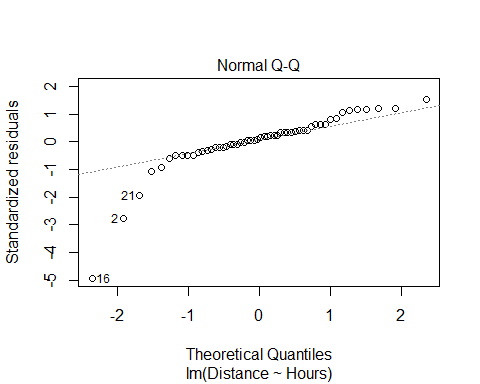
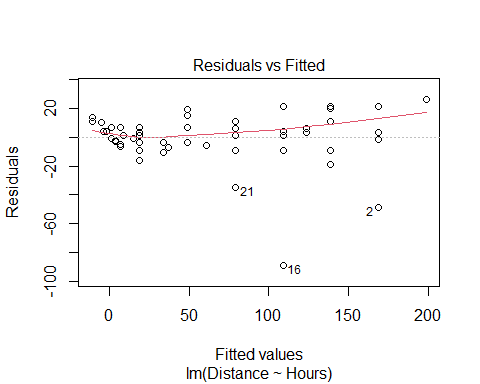
moddist=lm(Distance~Hours, data=Domestic)  
# Hours per distnace; predict distance form home based on hours it takes you to get home   
  
plot(Distance~Hours, data=Domestic)  
abline(moddist, col="red")



summary(moddist)

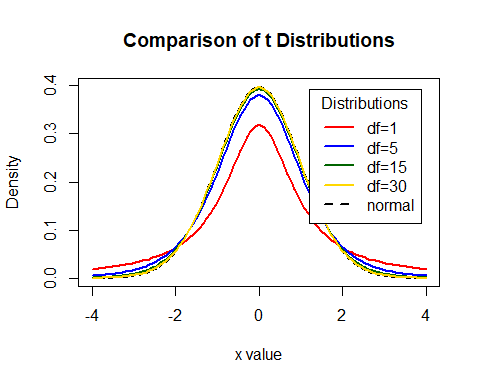
##   
## Call:  
## lm(formula = Distance ~ Hours, data = Domestic)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -88.892 -4.680 2.172 7.082 26.141   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -11.063 4.056 -2.727 0.00868 \*\*   
## Hours 59.977 2.484 24.144 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.26 on 52 degrees of freedom  
## Multiple R-squared: 0.9181, Adjusted R-squared: 0.9165   
## F-statistic: 582.9 on 1 and 52 DF, p-value: < 2.2e-16

# Slope = 59; for every 1 hour people ar eaway from home, then they are about 60 miles more away from home   
# We want to think about; How close is this to the population value?   
# What claims can I make about the population?  
  
plot(moddist, c(1, 2, 5))



# THis whole chunk of code creates a linear model, plots it against the data, and checks the conditions.   
# Pretty linear relaitonship overall: hours away from home, distance increase   
# Residuals have a curve that s alittle ; overall it doesnt seem very bad   
# normal QQ plot, ther eis a problem where one side has a skew   
  
#COok's ditance   
# High leverage, far right   
# High influence, up/down

# Display the Student's t distributions with various  
# degrees of freedom and compare to the normal distribution  
  
x <- seq(-4, 4, length=100)  
hx <- dnorm(x)  
  
degf <- c(1, 5, 15, 30)  
colors <- c("red", "blue", "darkgreen", "gold", "black")  
labels <- c("df=1", "df=5", "df=15", "df=30", "normal")  
  
plot(x, hx, type="l", lty=2, xlab="x value",  
 ylab="Density", main="Comparison of t Distributions")  
  
for (i in 1:4){  
 lines(x, dt(x,degf[i]), lwd=2, col=colors[i])  
}  
  
legend("topright", inset=.05, title="Distributions",  
 labels, lwd=2, lty=c(1, 1, 1, 1, 2), col=colors)



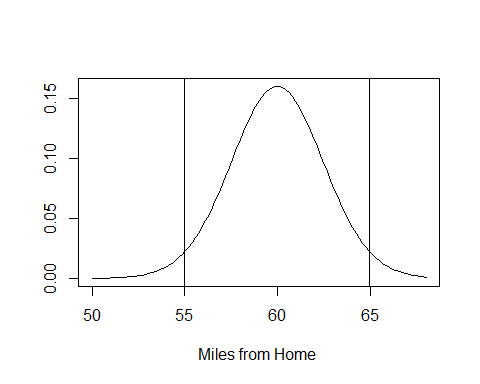
# What the tdist looks like   
#df comes down to sample size and other factors (predicotrs and modeltype)   
# Bigger sample, more degree of freedom   
# For 30 df, a sample of 32, this is pretty much a normal curve   
# As it gests smaller, we are pressing down and pressing out the sides of the graph   
#We are triyng to make this based on what the standard error of our sample is

**CI for Slope or Intercept** - HOw R is doing the math - BUild a distribution where the cente ris where the sample value is adnd we are making a curve on top of that and how far in each direction do we need to go were the middle 95% of the area under the curve is met? Bihar +/- tstar*StandardError or the Bihat Y = Bo + B1X+E - t* comes from a t-distribution with n-2 d.f and depends on the level of confidence - For 1−𝛼 level confidence, use qt(1-α/2,df)in R - **e.g. for 95% confidence and 52 df, qt(0.975,52)** - tstar = tells us how many STDERRORS in each direction we need to go

qt(0.975, 52)

## [1] 2.006647

# T = t distribution   
# area under curve to teh left of the point we want = first arguement; 0.975 = 95% confi int   
# df = 2nd argument; sampel size minus 2   
# Result = 2.00664 ish   
#If i want to see the confidence interval, you have to start at your sample of slope  
#qt gives you tstar  
  
curve(  
 dt.scaled(  
 x,   
 52,  
 mean = summary(moddist)$coef[2,1],  
 sd = summary(moddist)$coef[2,2]  
 ),   
 from = 50, to = 68,  
 xlab = "Miles from Home ",  
 ylab = " "  
 )  
  
  
abline(  
 v=c(  
 qt.scaled(  
 0.025,   
 52,   
 mean = summary(moddist)$coef[2,1],   
 sd = summary(moddist)$coef[2,2]  
 ),  
 qt.scaled(  
 0.975,   
 52,   
 mean = summary(moddist)$coef[2,1],   
 sd = summary(moddist)$coef[2,2]  
 )  
 )  
 )



# IF you want to see the confidence interval   
summary(moddist)$coef[2,1]-qt(0.975, 52)\*summary(moddist)$coef[2,2] #LOwer bound for confidence interval

## [1] 54.99264

summary(moddist)$coef[2,1]+qt(0.975, 52)\*summary(moddist)$coef[2,2] #Upper bound for confidence interval

## [1] 64.96233

#We are predicting that with 95% confidence the solution is between the 54.99-64.96

If we think tha the population might not be normal, then we probably want to do a bootstrap method

**How to find a confidence interval?** -confint(mymodel,level =0.XX) and adjust for the confidence level.

# HOW TO FIND CONFIDENCE INTERVAL; default 95% confidence   
# generally the intercept is not very useful to think about   
# We are predicting taht it could be close to zero   
# WE are mostly looking at the coeffs   
# The hours are about teh same as above   
confint(moddist, level=0.95)

## 2.5 % 97.5 %  
## (Intercept) -19.20166 -2.92358  
## Hours 54.99264 64.96233

**Accuracy of Predictions** Example: It takes a student 2.25 hours to drive from home. How many miles do we predict that thy are away from home? How accurate is that prediction? - Want to make a prediction for a specific case - Wnted regalr prediction, just plug the 2.25 into the regression line - It matterse what you are rtrying to predict - all people or the specific person’s distance from home? - There is a difference; one person has ore variability (Say they’re biking) - If we have the ditribution, ontop of that is some normal curve; most of the people are close to thtat, but they trail off a bit

**Two Forms of Intervals for Regression** 1. Confidence Interval for μY (mean Y) Where is the “true” line for that x? or Where is the average Y for all with that x? 2. Prediction Interval for Individual Y Where are most Y’s for that x?

\_\_CI for μY when X=x\*\_\_ - Predicting in general SSX = ∑▒〖(𝑥\_𝑖 − 𝑥 ̅)〗^2 yhat +/- tstar*standerror*sqrt((1/n)+((xstar-xbar)^2)/SSX)

\_\_Prediction Interval for Individual Y’s when X=x\*\_\_ - predicting for one person yhat +/- tstar*standerror*sqrt(1+ (1/n)+((xstar-xbar)^2)/SSX) Just add 1 in the sqrt

\_\_CI and PI via R when X=x\*\_\_

newx=data.frame(Hours=2.25) # Creat e a new person   
head(newx)

## Hours  
## 1 2.25

predict.lm(moddist, newx, interval="confidence") # Predict mean for all people who are 2.25 away from hoe

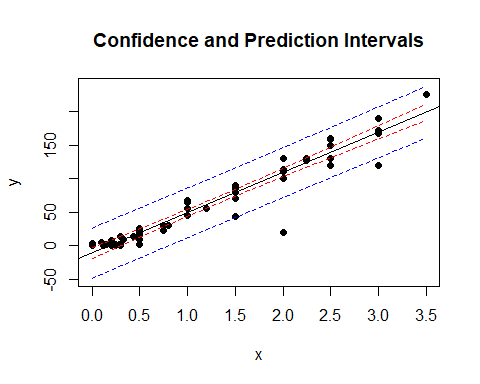
## fit lwr upr  
## 1 123.8867 116.9764 130.797

predict.lm(moddist, newx, interval="prediction") # Distance home for one specific person

## fit lwr upr  
## 1 123.8867 86.59458 161.1789

# Both gives us a fitted value, its a point in the regression lie   
# THis si what we would get if we plugged and chugged   
# One person is going to be a wider range because we want to make sure we get teh one person

CIPIPlot(Domestic$Hours, Domestic$Distance) # Visualize different between confidence and prediction



# calculates   
# For every possible point in teh data, or for the  
# What would be the confidence interveral for that value and what would be the prediction interval

* The red lines are the confidence interval
* if we are trying to predict the mean vlaue for people’s distance away from home based on tehse hours, we will predict the mean is somewhere between the red lines and its tight by the regression line with 95% confidence Th eblue line = much wider
* there’s a lot more variability there
* much wider

## STOR 455 Class 9 Partitioning Variability - ANOVA

library(readr)  
library(Stat2Data)  
  
DistanceHome <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/DistanceHome.csv")  
  
Domestic = subset(DistanceHome, Distance<250)

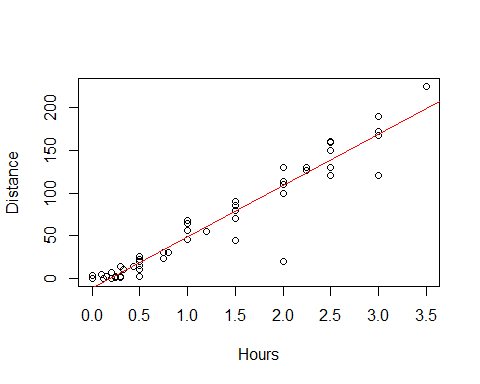
*NOtes* - Evidence to say if there is some kind of relationship exists? - How we center teh distributios is a little different

**T-test for Slope**  Ho: B1 = 0 (for Y = Bo + Error) Ha: B1 != 0 (for Y = Bo + B1X + Error)

t.s. = Bhat1/SEofBhat1 - Find p-value using a t-distribution and n-2 d.f. -p-value is small then Reject Ho

**How to find a P-value?** - Statistically significant evidence suggests ( p-value < 2.22\*10-16) that there is a relationship between the hours that a student spend travelling to campus and their distance from campus.

plot(Distance~Hours, data=Domestic) # Predict distance home, based onhow many hours it takes to get there   
moddist = lm(Distance~Hours, data=Domestic)  
abline(moddist, col="red") # Draw linear model on top of plot



summary(moddist) # Take summary of plot model

##   
## Call:  
## lm(formula = Distance ~ Hours, data = Domestic)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -88.892 -4.680 2.172 7.082 26.141   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -11.063 4.056 -2.727 0.00868 \*\*   
## Hours 59.977 2.484 24.144 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.26 on 52 degrees of freedom  
## Multiple R-squared: 0.9181, Adjusted R-squared: 0.9165   
## F-statistic: 582.9 on 1 and 52 DF, p-value: < 2.2e-16

# In coef table,   
# Does a hypothesis test for us   
# COmpares two models to eachother   
# The linear model to the constant model, where the model is just the mean   
# If the model is just a horizontal line, the it's that the slope would be zero  
# slope zero = no cheged based on distance from home   
# Alternative: AS hours change, the ditsnce from home changes   
# How likely is it that we get a sampel slope like we did, what the chance we would get what we did   
# The test stat is the tstat/Se = how many std err are we from teh null; how unlikely is it that we get the thing that we did   
  
# Slope: 59.99  
# STd err: 2.48 = Tightly bound   
# tvalue = 24. = if our null is 0, then we are 24 std errs from the null hyp of 0; thats unlikely to happen by chance   
# Pvalue: How unlikely we are to get a sample that we did, if the null is true; and we have a really low pvalue for this, which means that it is not likely we would get the null hypothesis

**Finding Correlation in R** -For data in two variables:

cor(Domestic$Distance, Domestic$Hours)

## [1] 0.9581758

**Finding Correlation in R** For all variables in a dataframe:

data(Houses)  
head(Houses)

## Price Size Lot  
## 1 212000 4148 25264  
## 2 230000 2501 11891  
## 3 339000 4374 25351  
## 4 289000 2398 22215  
## 5 160000 2536 9234  
## 6 85000 2368 13329

# Can put in whole dataframes   
# This will give you a corr matrix; a table with a ll possible combination snad the correlations for those variables   
  
cor(Houses)

## Price Size Lot  
## Price 1.0000000 0.6848219 0.7157072  
## Size 0.6848219 1.0000000 0.7668722  
## Lot 0.7157072 0.7668722 1.0000000

**Finding Correlation in R** Watch out – variables must be numeric! May need to choose numeric columns:

data(Cereal)  
head(Cereal)

## Cereal Calories Sugar Fiber  
## 1 Common Sense Oat Bran 100 6 3  
## 2 Product 19 100 3 1  
## 3 All Bran Xtra Fiber 50 0 14  
## 4 Just Right 140 9 2  
## 5 Original Oat Bran 70 5 10  
## 6 Heartwise 90 5 6

#cor(Cereal) <- this doesn't work because it has character vectors in it   
  
cor(Cereal[c(2:4)]) # Tells R Which columns we want to take from the thing

## Calories Sugar Fiber  
## Calories 1.0000000 0.5154008 -0.7150123  
## Sugar 0.5154008 1.0000000 -0.5025772  
## Fiber -0.7150123 -0.5025772 1.0000000

**Test for a Linear Relationship via Correlation** - Let p (rho) denote the population correlation

Ho: p = 0 Ha: p =!= 0

t = ((r\*sqrt(n-2))/sqrt(1-r^2)) - Compare to a t-distribution with n-2 d.f.

**Correlation t-test in R** *SEe below:*

cor.test(Domestic$Distance, Domestic$Hours) # If null is true (if 0 correlation) how likely is it that we have a sample with this strong correlation? That's what this will tell you

##   
## Pearson's product-moment correlation  
##   
## data: Domestic$Distance and Domestic$Hours  
## t = 24.144, df = 52, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.9286787 0.9756277  
## sample estimates:  
## cor   
## 0.9581758

# OUtputs: Correlation bt the two things   
# What the cof int is for the population based here   
# t = 24.144 =how many stand dev from the null we are (Same value as before for test for slope)  
# Tells us really low pvalue, which means that its ar really low chance we could get this result by chance   
  
# All the results are going to be the same if it's simple linear regression   
# Gives us F test stat and p value

**ANOVA for Regression** Data = Model + ERror Total variation in response, Y = variation explained by MODEL + Unexplained variation in RESIDUALS

Key question: Does the MODEL explain a “significant” amount of the TOTAL variability?

**Partitioning Variability - SLM** Y = Bo + B1X + E SSTotal = SSModel +SSE

**NOtes** - ANOVA = analaysis of variances - we cant explain all variability, but we can try -*Total variations in reponse:* Is how far away each of our points are from the mean value - *Variation explained by model* total variability = full distance from teh mean down to the point; by fitting teh line to it, most of the variability is explained by the variation ; - *Resisudals* Distance that is left over from teh points - WE want a sig amount of variability in the model to be explained by the regression line vs a horizontal line - We really only care about teh sum of squares - The f test stat = how likely this variability is explained by model compared to what is left over in the error term if the null model were true that there is no relationship between tehse things **Can look at this by using ANOVA**

anova(moddist)

## Analysis of Variance Table  
##   
## Response: Distance  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Hours 1 194417 194417 582.93 < 2.2e-16 \*\*\*  
## Residuals 52 17343 334   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# We have teh SSModel and SSError terms; does not give SSTotal, but if we add teh two, we could get that   
# What it is is that the SS in the HOurs row, is for the model, so if we had the line going back; how far away for each point the line is for that value away from teh mean of that value; how much variability is explained by the model and all squared and summed together   
# Residual SS; looks at the similar thing, but how far each point is from the line squared and summing them totgether   
  
# Think about that proportionally; how different are these two? Scale them by df.  
# MSE = comparing models for the vriabiilty and i'm taking the SSTotal/df. Hours df = 1, and residual df = 52 (which is 54-2)  
# F test stat = MSModel/MSError; that value is big it says that a lot of variability is explained by the model; when that value is small, it's saying not so much varibility is being explained by the model   
  
# Big and small are relative, depends on sample size   
# F test stat says its unlikely we would get this result by chance

**ANOVA Test for Regression** - Basic idea: Find two estimators of 2 - Model: SSModel/1 = MSModel ERror = variance of error = SSE/n-2 = MSE - We want to compare the model and the error

t.s. = MSModel/MSE - COmpare to F-dist with I and n-2 d.f

* ANOVA Test for regression Ho: B1 = 0 (ybar) Ha: B1 != 0 (yhat)
* Same null and alternative as slope test; asusming there is no relationship bt predictor and response
* so the slope of the model is 0 and the mean = the model that we use
* Alternative = some nonzero slope better descrbes this model and how likely we would get this kind of model if the null were true

Source, d.f, Sum of Squares, Mean Swuare, F, Pvalue Model, 1, SSModel, SSModel/1, MSModel/MSE, F(1,n-2) REsidual, n-2, SSE, SSE/n-2, See above got F and PValue TOtal, n-1, SSTOtal - in R use 1-pf(Fstat,1,n-2)

**What is r2?** r2 = proportion of total variability in the response (Y) that is “explained” by the model. 𝑟^2=𝑆𝑆𝑀𝑜𝑑𝑒𝑙/𝑆𝑆𝑇𝑜𝑡𝑎𝑙=1−𝑆𝑆𝐸/𝑆𝑆𝑇𝑜𝑡𝑎𝑙 - The amount of variability explaiend by the model out of the total variability - If we look at a plot ; the total variability (how far away each point is away from teh mean line), then our error term; varibaility explained by model = mean down tot the line - high values of rsquares = most of the variability in teh response is explained by the predictors a - low values = oppsitie

**Visualizing r2 for a SLM** Basic Idea: How much “better” does the least squares line do than a “prediction” that doesn’t use X at all? - Using NO predictor: 𝑦 ̂=𝑦 ̄

* Least Squares Line: 𝑦 ̂=𝛽 ̂\_𝑜+𝛽 ̂\_1 𝑥
* the cor coof being squared
* if corr coef is always bt -1 and 1; r^2 has to be bt 0 and 1
* big difference; r-sqqare we can look at multipel predictors at once

**Why is it called r2?** - Def: The correlation, r, measures the strength of linear association between two quantitative variables. -1< r <1 to 0 < r2 < 1 - 0 = Explains no variability - 1 = Explains all variability **Simple Linear Regression - R**

summary(moddist)

##   
## Call:  
## lm(formula = Distance ~ Hours, data = Domestic)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -88.892 -4.680 2.172 7.082 26.141   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -11.063 4.056 -2.727 0.00868 \*\*   
## Hours 59.977 2.484 24.144 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.26 on 52 degrees of freedom  
## Multiple R-squared: 0.9181, Adjusted R-squared: 0.9165   
## F-statistic: 582.9 on 1 and 52 DF, p-value: < 2.2e-16

**Which test is best?** -**T-test for slope:** Ho: B1=0 Ha: B1!=0

𝑡=𝑏\_1/(𝑆𝐸\_(𝑏\_1 ) )

Compare to t n-2 - How well related is this predictor with this response after taking into account the whole rest of the model, all other predicotrs - Not vacume - look at all in one room

* **ANOVA for regression:** Ho: B1=0 Ha: B1!=0

𝐹=𝑀𝑆𝑀𝑜𝑑𝑒𝑙/𝑀𝑆𝐸

Compare to F1,n-2 - Bigger picutre - Have 1 predictor, but once we have more, it will let us see if there is any relation bt any of these predictors and repsonse here - one test to see if there is any relation anywhere - useful for more tests

* **T-test for correlation:** Ho: p =0 Ha: p !=0

𝑡=(𝑟√(𝑛−2))/√(1−𝑟^2 )

Compare to t n-2 (tn-2)2 = F1,n-2

* Focuse on a predictor ans response
* how are they related in a vaccume ingnoring everyhting else

**We have 3 different tests for when we get to multiple regression**

## STOR 455 Quiz 1 Solutions

LEGO is a type of building toy created and made by the Lego Group, a company in Denmark. “Lego Bricks” are colorful plastic building blocks that can be joined together easily and are the most popular building toy in the world. Since the 1950s, the Lego Group has released thousands of sets with a variety of themes, including space, robots, pirates, trains, Vikings, castle, dinosaurs, undersea exploration, and wild west. Over the years, Lego has licensed themes from numerous cartoon and film franchises and even some from video games. These include Batman, Indiana Jones, Pirates of the Caribbean, Harry Potter, Star Wars, and Minecraft. For this quiz you will focus on the relationship between the number of pieces in a LEGO set and the set’s selling price on Amazon.

library(readr)

lego = read\_csv('https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/lego.csv')

**Quiz A**

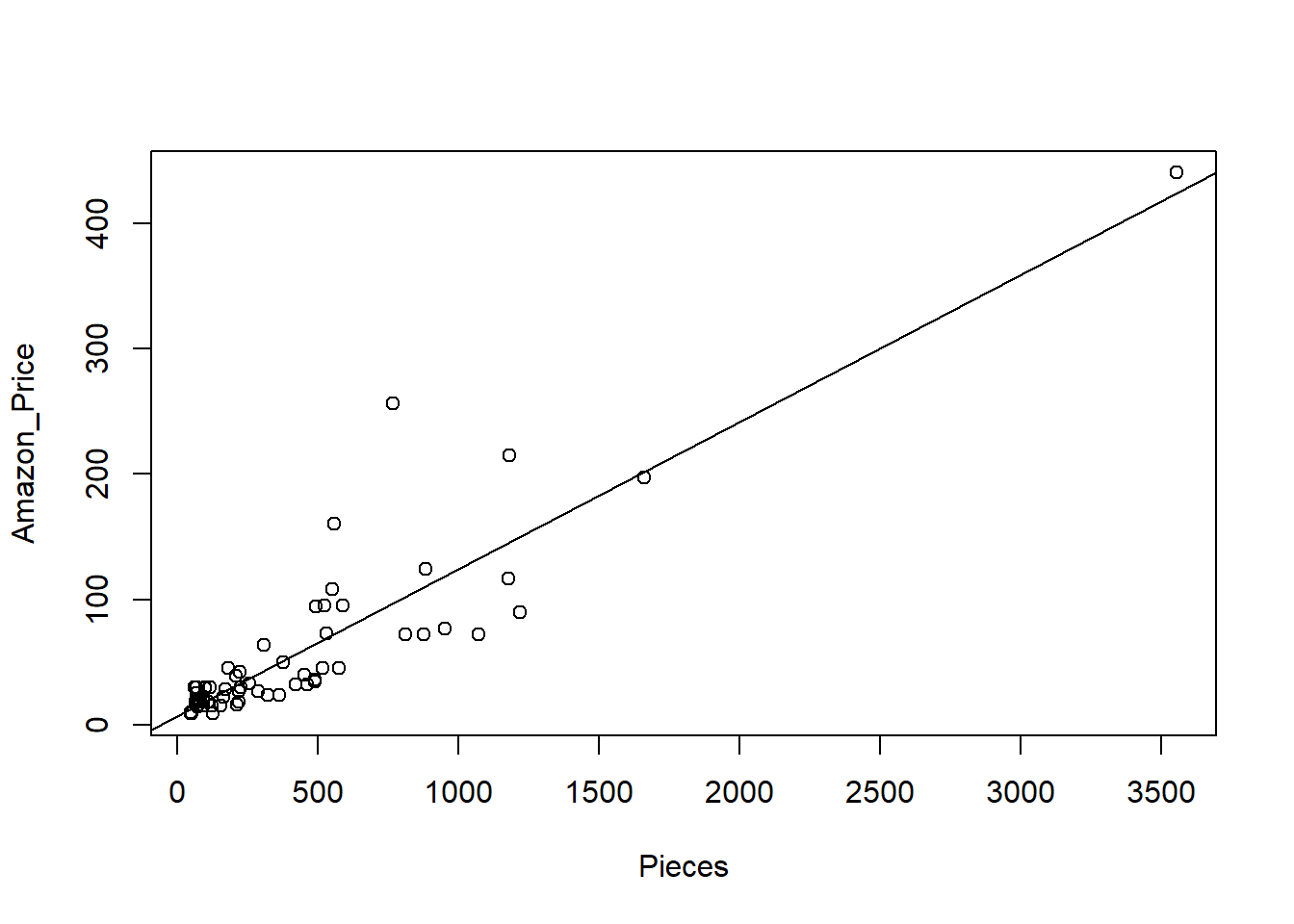
1. Construct a dataframe that contains LEGO sets that only have the *Theme* of *NINJAGO*. Use this dataframe to construct a linear model that predicts the *Amazon\_Price* based on the number of *Pieces* in the LEGO set. Plot the *Amazon\_Price* and *Pieces* data from this new dataframe, as well as your linear model, on one plot. You do not need to check the conditions for a linear model nor consider models with transformations. **5 pts**

ninjago = subset(lego, Theme == 'NINJAGO')

plot(Amazon\_Price ~ Pieces, data = ninjago)

ninjago\_mod = lm(Amazon\_Price ~ Pieces, data = ninjago)

abline(ninjago\_mod)



1. For each increase in 20 lego pieces in a *NINJAGO* LEGO set, how much does your model predict will be the increase in *Amazon\_Price*? **1.5 pts**

20 \* summary(ninjago\_mod)$coef[2,1]

## [1] 2.345026

1. Which NINJAGO LEGO set (by *Set\_Name*) has the largest influence on the regression model? How do you know this? Would this be considered an influential **1.5 pts**

max(cooks.distance(ninjago\_mod))

## [1] 0.3822737

which.max(cooks.distance(ninjago\_mod))

## 65

## 65

ninjago$Set\_Name[65]

## [1] "NINJAGO City Docks"

The NINJAGO City Docks set has the largest Cook’s Distance, therefore has the most influential on the model. Since this Cook’s Distance is less than 0.5, we would not consider this LEGO set to be influential.

1. What does your model predict, with 90% confidence, will be the *Amazon\_Price* for a *NINJAGO* LEGO set with 742 *Pieces*? Your answer should be an interval of values, and not one number. **2pts**

new\_set = data.frame(Pieces = 742)

predict.lm(ninjago\_mod, new\_set, interval="prediction", level=0.90)

## fit lwr upr

## 1 94.19487 40.97408 147.4157

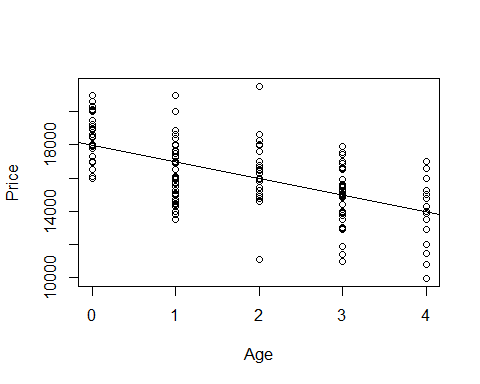
# Unit 3: Multiple Regression

## STOR 455 Class 11 R Assessing Multiple Linear Regression Models

library(readr)  
library(Stat2Data)  
  
data("Houses")  
  
source("https://raw.githubusercontent.com/JA-McLean/STOR455/master/scripts/anova455.R")  
  
# Or you can download the R script from Sakai  
# Save the script in the same folder as the notebook  
# comment out the above sourced script from github  
# Run code below  
#source("anova455.R)

# install if needed  
# dplyr package used for sample\_n() function  
  
library(dplyr)  
  
UsedCars <- read\_csv("UsedCars.csv")  
  
# selects same random sample each time  
set.seed(09132021)  
  
# random sample of 200 CamrySE  
CamrySE = sample\_n(subset(UsedCars, Model=="CamrySE"), 200)  
# Gets a random sample of a subset of the full dataset, where the model is camrySE, the second arguement is just how many it's taking

CamrySE$Age = 2017 - CamrySE$Year  
CamrySE\_Model = lm(Price~Age, data=CamrySE) # model  
plot(Price~Age, data=CamrySE) # Plot model   
abline(CamrySE\_Model)

 Look above; from here we have to think about the linear conditions - There is one situation were we need to look at a different way of the conditoins - When we have predictors wiht lots of values for the response for it ;we dont care about teh normalitiy to the residuals overall, we care about the normiality of the residuals for each response - In other words, for all the cars that are zero years old, are those residuals normally distirbuted for one, two, three and four are they all normally distributed? - Not a lot of years ar ehere, but we can try to see if we haev that normal distributio or are these some probailemns betwen the years

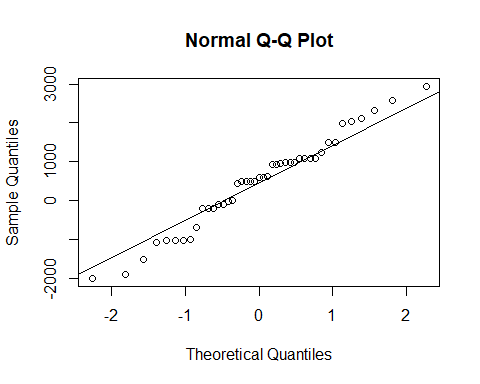
*How to lok at athat* - Can split it up by the ages of teh car

**Below:** - Looking at the camry’s by age

# Splits residuals by Age  
  
Resid\_by\_age = split(CamrySE\_Model$residuals, CamrySE$Age) # SUPER USEFUL FUNCTION  
Resid\_by\_age

## $`0`  
## 8 13 22 26 29 34   
## 1486.908341 -1022.091659 -204.091659 1095.908341 982.908341 -13.091659   
## 35 39 46 47 49 55   
## -1019.091659 1487.908341 2942.908341 591.908341 942.908341 2330.908341   
## 56 58 59 60 61 69   
## -1914.091659 937.908341 2043.908341 589.908341 -2008.091659 1983.908341   
## 71 75 80 85 88 89   
## -107.091659 987.908341 617.908341 2119.908341 1087.908341 -8.091659   
## 96 105 115 117 119 121   
## -204.091659 1095.908341 -1018.091659 430.908341 -1012.091659 495.908341   
## 123 128 129 137 143 149   
## 987.908341 495.908341 -1518.091659 -113.091659 1095.908341 924.908341   
## 150 159 165 171 175 193   
## -204.091659 2583.908341 -1077.091659 -704.091659 1237.908341 495.908341   
## 198   
## 495.908341   
##   
## $`1`  
## 2 5 16 18 24 25   
## -2036.580152 -1997.580152 -3506.580152 -1077.580152 -2722.580152 -1999.580152   
## 27 31 33 38 45 53   
## -891.580152 -26.580152 -1284.580152 1383.419848 257.419848 -942.580152   
## 57 62 63 65 67 68   
## -1100.580152 -3003.580152 -1514.580152 1868.419848 -2391.580152 -1342.580152   
## 77 78 79 81 83 84   
## -3175.580152 495.419848 -1493.580152 1384.419848 -992.580152 229.419848   
## 90 92 93 94 95 103   
## -3006.580152 593.419848 333.419848 -942.580152 -2609.580152 998.419848   
## 106 107 113 122 124 127   
## 1896.419848 3.419848 498.419848 293.419848 -1342.580152 673.419848   
## 130 135 138 139 140 141   
## -3210.580152 1003.419848 -2500.580152 -1996.580152 -96.580152 1007.419848   
## 144 148 152 155 157 161   
## -1091.580152 1612.419848 -492.580152 -2310.580152 1383.419848 -3006.580152   
## 164 167 170 172 174 180   
## -3506.580152 -2992.580152 508.419848 -1742.580152 -1891.580152 3003.419848   
## 185 190 191 196 199   
## -492.580152 2998.419848 599.419848 -1342.580152 4008.419848   
##   
## $`2`  
## 4 6 10 14 28 32   
## 2012.93135 2304.93135 2673.93135 530.93135 21.93135 -700.06865   
## 36 44 50 51 52 64   
## 997.93135 -976.06865 1643.93135 -976.06865 5548.93135 1003.93135   
## 70 74 76 97 99 101   
## 678.93135 477.93135 -67.06865 1014.93135 18.93135 1022.93135   
## 104 108 112 131 132 136   
## 323.93135 -189.06865 -549.06865 768.93135 11.93135 -978.06865   
## 142 145 146 156 160 162   
## -1094.06865 273.93135 2004.93135 1614.93135 1005.93135 -4898.06865   
## 166 169 188 189 192   
## -981.06865 -1377.06865 1018.93135 -1177.06865 2018.93135   
##   
## $`3`  
## 1 3 7 9 12 17   
## 289.442861 -1018.557139 -967.557139 -2060.557139 2448.442861 -960.557139   
## 19 20 21 23 37 43   
## -150.557139 -535.557139 30.442861 -521.557139 1939.442861 2019.442861   
## 48 54 66 72 73 86   
## 1534.442861 -1980.557139 260.442861 -1963.557139 37.442861 638.442861   
## 87 98 100 109 110 111   
## 1918.442861 539.442861 -1063.557139 499.442861 524.442861 1536.442861   
## 114 118 120 125 126 133   
## 651.442861 1530.442861 2939.442861 971.442861 -1970.557139 619.442861   
## 134 151 153 158 168 173   
## -3060.557139 -960.557139 -1461.557139 -3565.557139 -965.557139 4.442861   
## 176 177 178 179 181 182   
## 129.442861 1656.442861 -3975.557139 -1238.557139 -69.557139 2039.442861   
## 184 186 187 194 197 200   
## 2563.442861 2034.442861 1925.442861 36.442861 939.442861 1590.442861   
##   
## $`4`  
## 11 15 30 40 41 42   
## -450.04563 349.95437 50.95437 -3152.04563 1049.95437 2049.95437   
## 82 91 102 116 147 154   
## -1950.04563 -3988.04563 2649.95437 -2458.04563 3045.95437 849.95437   
## 163 183 195   
## -1057.04563 -70.04563 1304.95437

# Different lists, of the string 0; all residuals for 2017 models   
# Then 1, 2, and 3, etc old cars   
# Year year each year gets its own string   
  
# If want to look at normaility, then look at the qqnorm plot of it   
# Need to do for each thing individually   
  
# This is for the zero year old cars   
qqnorm(Resid\_by\_age$'0')  
qqline(Resid\_by\_age$'0')



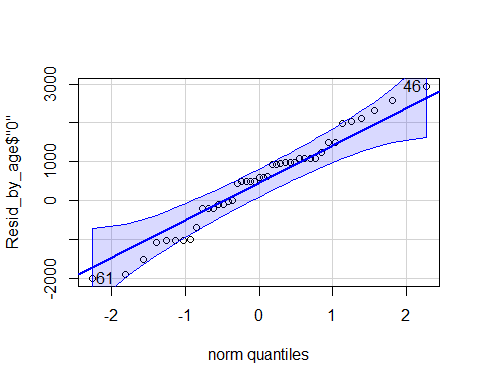
# install if needed  
# car package used for qqPlot()  
  
library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

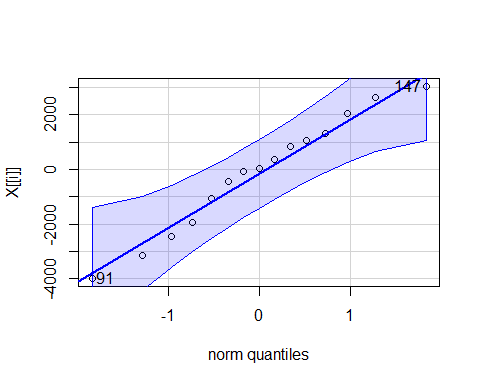
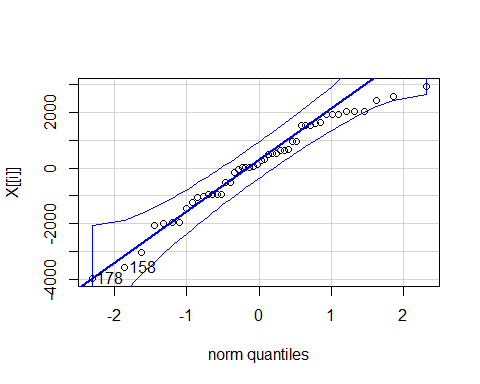
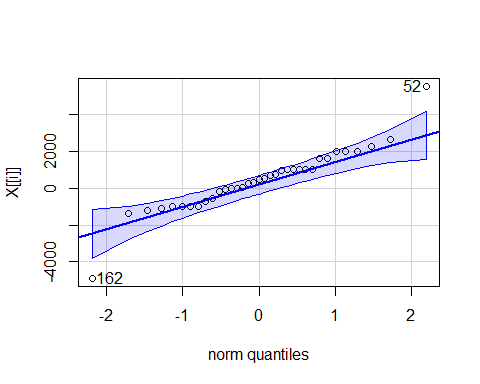
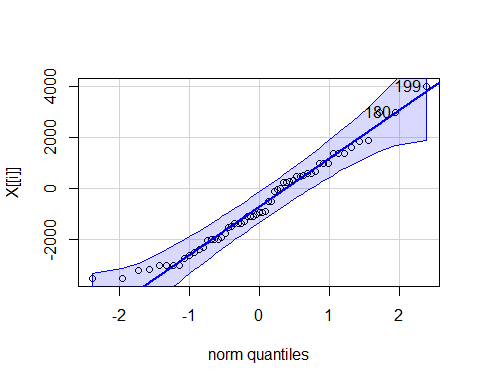
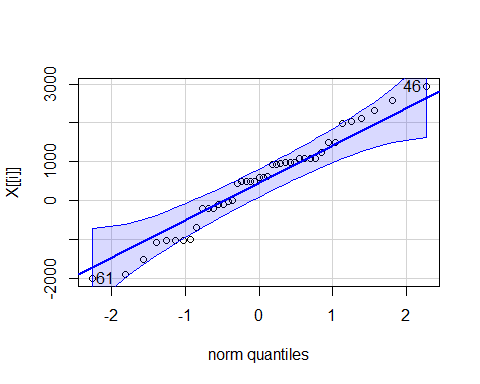
qqPlot(Resid\_by\_age$'0') # THIS IS A SUPER USEFUL FUNCTION



## 61 46   
## 17 9

# Shows teh QQNorm and QQLine plot if norma distributioned   
# Gives a band of what we would expect to see if this data of this size was normally distributed   
# The varibaility we would expect from a thing of this size   
# It looks pretty good in this interval   
# Want to use this for all the ages of the cars

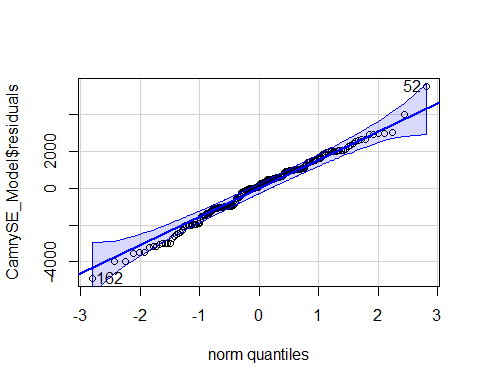
sapply(Resid\_by\_age, qqPlot)



## 0 1 2 3 4  
## 61 17 59 30 39 8  
## 46 9 54 11 34 11

# Applies the qqPlot function to all the residual\_by\_age   
# This shows us that it's normal enough for each age, so it's good to go

qqPlot(CamrySE\_Model$residuals)



## [1] 52 162

# Narrow band because bigger dataset   
# Age 4 is wider ebcause there is less data   
# Less data = wider band, more data = smaller band (Same idea as confidence intervals)

**Three different types of tests** **T-test for Slope** - Ho: B1 =0 - Ha: B2 != 0

* t = b1/SEofb1
* Compare to t(subof n-2)
* Do we have evidence to say the relationship bt the predict and resposne has a non horizontal relationship/non zero slope?

**ANOVA for regression** - Ho: B0 = 0 - Ha: B1 != 0

* F = MSModel/MSE
* Compare to F1, n-2
* (tn-2)^2 = F1, n-2
* For simple linear models, did the same as the t test; tried to look at ho wmuch variability is being explained by odel compared to a null model (so a horizontal line)

**T-test for correlation** - Ho: p = 0 - Ha: p != 0

* t = ((r\*sqrt(n-2))/sqrt(1-r^2))
* Compare to tn-2
* Do we have evidence to say there is some kind of correlation, +/- for the two variables?

**Simple Linear Regression Model** - Y = B0 + B1X + Error - Where Error follows N(0,stdof error) and independent (normal and independent) *What if we have more than one potential predictor?* - We got teh same values if we did the above three different tests on a simple linear regression model, so its not super helpful when you’re doing a simple linear regression model

**Multiple Regression Model** - Y = B0 +B1X1 + B2X2 +….+BkXk + Error - where error is assumed to follow a normal and independent - THis is in many dimentions with more predictors

Data?  
We need n data cases, each with values for Y and all of the predictors X1,…,Xk.

**R - Correlation Matrix**

head(Houses)

## Price Size Lot  
## 1 212000 4148 25264  
## 2 230000 2501 11891  
## 3 339000 4374 25351  
## 4 289000 2398 22215  
## 5 160000 2536 9234  
## 6 85000 2368 13329

# Just a small dataset of 20 houses   
  
cor(Houses)

## Price Size Lot  
## Price 1.0000000 0.6848219 0.7157072  
## Size 0.6848219 1.0000000 0.7668722  
## Lot 0.7157072 0.7668722 1.0000000

# We can do this because they are all quanatitive variables   
# all have fairly strong, postiive correlations of eachother   
# Are tehse big enough to take this claim tot he population?

*Look at a test between them*

**t-test for Correlation**

#cor.test(Houses) # This doesnt work because there's too many   
  
# Correlation looks at 2 vars once and looks at the relationship bt two vars   
  
cor.test(Houses$Size, Houses$Price)

##   
## Pearson's product-moment correlation  
##   
## data: Houses$Size and Houses$Price  
## t = 3.9871, df = 18, p-value = 0.0008643  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.3476582 0.8651583  
## sample estimates:  
## cor   
## 0.6848219

# Ho: No correlation between teh two in teh population   
# Ha: There is a correlation between teh two in the population   
# Want to see how likely it is that we would get the smapel line in the population   
# Output tells us the correlation bt the two, the t test of 3.9 (this tells us that it's pretty unlikely by chance); pvalue of 0.000008 - that is the probability that we would get a sample like this or one as extreme as this if there was no realtionship in teh popualtion; so this is pretty unlikely this would happen by chance   
  
# Can do the same thing above for the other realtionships as well   
cor.test(Houses$Lot, Houses$Price)

##   
## Pearson's product-moment correlation  
##   
## data: Houses$Lot and Houses$Price  
## t = 4.3478, df = 18, p-value = 0.0003878  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.3998134 0.8796343  
## sample estimates:  
## cor   
## 0.7157072

cor.test(Houses$Size,Houses$Lot)

##   
## Pearson's product-moment correlation  
##   
## data: Houses$Size and Houses$Lot  
## t = 5.0694, df = 18, p-value = 7.991e-05  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.4909631 0.9029654  
## sample estimates:  
## cor   
## 0.7668722

# BOth have teh same conclusion and hypothesis that   
# Ho: No relationshio in the popualtion   
# Ha: Some realtionship in population   
# Both have small pvalues   
# Both have strong evidence to say there is a realtionship between these things in teh population

**t-test for Correlation** - Ho: p = 0 - Ha: p != 0

* t = ((r\*sqrt(n-2))/sqrt(1-r^2))
* Find p-value with t n-2

Use this to: 1. Identify potential good predictors of Y. 2. Look for relationships among predictors.

**Prediction Equation** - where the coefficients are chosen to minimize: SSE = sum(y-yhat)^2

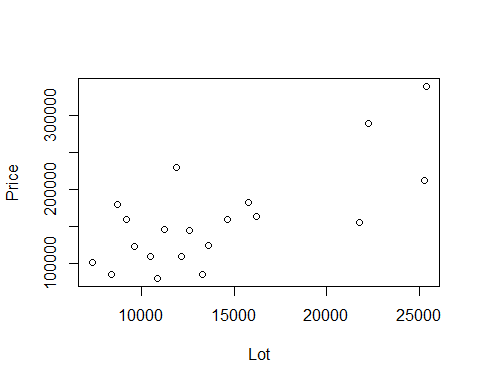
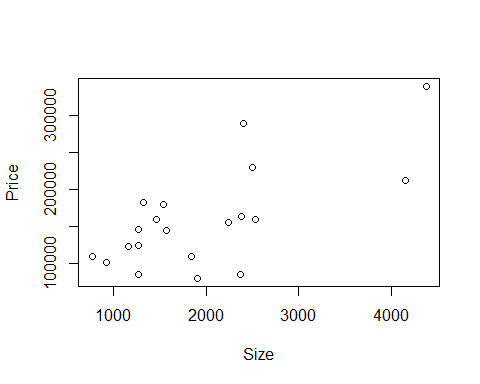
To fit a multiple regression in R: model=lm(y~pred1+pred2+pred3,data= ) - Other tests keep in mind the model that we are workign with, the correlation test always lookas only at 2, while the other tests look at more than 2 variables at once

**R Regression: Individual T-tests** - Look at the P value of the predictors (Size and lot in this example)

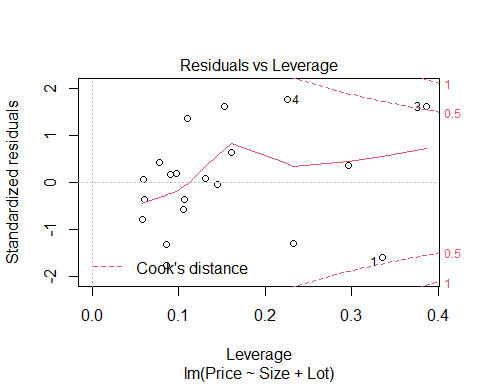
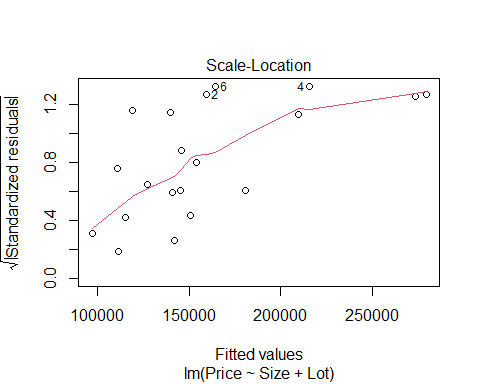
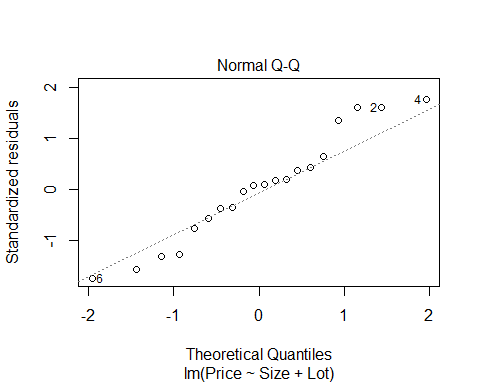
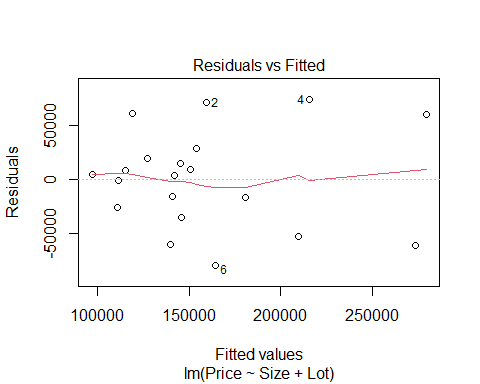
HouseModel=lm(Price~Size+Lot,data=Houses) # Multiple linear regression model   
summary(HouseModel)

##   
## Call:  
## lm(formula = Price ~ Size + Lot, data = Houses)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -79532 -28464 3713 21450 73507   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 34121.649 29716.458 1.148 0.2668   
## Size 23.232 17.700 1.313 0.2068   
## Lot 5.657 3.075 1.839 0.0834 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 47400 on 17 degrees of freedom  
## Multiple R-squared: 0.5571, Adjusted R-squared: 0.505   
## F-statistic: 10.69 on 2 and 17 DF, p-value: 0.000985

# The summary output:   
# The test we looked at before for individual slopes are still there, just an extra line is there for teh extra predictor   
# The resuls are not what we would expect; where invidually when we compare size with teh price of the house and the lot with t price of the house, we had small pvalues saying we would have a really low chance of getting this sample if there was no realtion   
# Same tests here that the coeff of size is vs teh alternatieve it's not, we get pvalues athat are higher and possibily not sign at this rate   
# There is something goeing on here, which is called multicolinearity (See next class's notes)  
# The bottome line gives the anova tests, does something different   
  
plot(Price~Size+Lot,data=Houses)



# It's not giving us one plot we can look at and think about lineariry and constant variance to the full modle   
# Its giving us two different plots, because this plot would be given in 3 dimentions   
# We have aplot by size by price and lot size by price   
# Hard to see the realtionship here with this visual format   
# We know teh realtionship with each redcitor in teh response could be useful when looking at t transformations in multiple regression   
  
plot(HouseModel)



# We see that it's not all conditions are really met   
# Same idea as simple lienar regression because for simple linear regression, just compare residuals to the fitted values or looking at teh normaility of the residuals   
# Here, no matter how many predictors we have, we still haev how far off are we and what are the residuals, so we can still look at the residuals and fitted plots   
# Does teh red line look roughly horizontal or some defined curve? IT's a small dataset, so we expect some kind of variability, that's looks pretty good. there's not some clear curve over the whole model   
# Normaility, can judge the same as simple lienar regression   
# the residuals by fitted, we can still look at this roughly the same, but leverage is a little different now  
# For simple linear models it was how far are we fro teh mean predictors, for mutkpel, we could be far from some predictors but close to others  
#Cook's distance, look at it the same as simple linear models outside bounds, teh have influence

We are saying that a price in house is about 34121.649 + 23.23size + 5.657lot **R - Multiple Regression** (𝑃𝑟𝑖𝑐𝑒) ̂=34121.6+23.232∙𝑆𝑖𝑧𝑒+5.657∙𝐿𝑜𝑡 - We are summing each test has a slope of zero - the t test stat = how many SE we are from zero - about the same, but different dfs

**Multicolinearity** is when size is a good predictor of lot already, so when we look at things in the future; if lots of predictors are correlated with eachother, then they may be explaineng a similar amount of variability

**T-test for Slope** -Note: We now have multiple “slopes” to test - Ho: B1 = 0 - Ha: Bi != 0

* t.s. = Bi/SEofBi
* All given in R with a p-value
* Compare to t n-k-1
* **lose 1 d.f. for each coefficient**
* Reject Ho if The ith predictor is useful in this model

**Coefficient of Multiple Determination** - 𝑅^2=𝑆𝑆𝑀𝑜𝑑𝑒𝑙/𝑆𝑆𝑇𝑜𝑡𝑎𝑙 - Now interpreted as the % of variability in the response variable (Y) that is “explained” by a linear combination of these predictors. *NOTes* - Variability in response being explained by the predictior - for Simple linear regression this was just correlation squared - now ti’s different because we haev 2 predictorsl its not just looing at each predictor and the response its lookig at the variability in teh resposne based on all teh predictors in the model - may explain overlaping responses in the mdoel

* Look at teh adjustesd R squared as 0.55, it says 55% of the variability in the price of the house is explained by the size of the house and the lot size of the house; that leaves an extray 44% that is not explained by these things by the data we have; so there may be other variables that explain that varibaility and we just don thaev those variablies

**t-test for Correlation vs. t-test for Slope** - **t-test for correlation:** Assesses the linear association between two variables by themselves. – in a vaccum; compare two things, ignore world - **t-test for slope:** Assesses the linear association after accounting for the other predictors in the model. – accounts for other predicotrs

**Partitioning Variability** - Y = B0+B1X1+…+BkXk + Error - SSTotal = SSModel + SSE - SSModel = Total explained by the regression - SSE = Error after regression - SSTotal = Total variability in Y

* About the same thing, but just in different dimentions
* we have amodel that predicts that data dn teh error around that model that we haev
* can look at teh SS the same as before and this case, we jsut condense teh SSModel not just from 1 predicotr, but from multiple predictors we are looking at
* ANOVA wise, the idea of accounting for that avariability is about the same

**ANOVA test for Overall Fit** - Ho: B1 = B2 = …= Bk = 0 (weak model) - Ha: Some Bi != 0 (Effective model)

Source, d.f, Sum of Squares (SS), MeanAquare, t.s., P-value Model, k, SSModel, SSModelk, MSModel/MSE, Fk, n-k-1 Residual, n-k-1, SSE, SSE/(n-k-1), t.s. is the same, p value is the same Total, n-1, SSTotal

* Only difference than before is that there can be multiple predictors
* Still haev the other things, just accounts for other htings in the model
* still trying to figure out how good is the mdoel doing
* WE ARE ASSUMING NOTHING IN THE MODEL IS USEFUL TO US
* We are assuming that the coeffe of the predictors are zero **ANOVA Hypo Test**
* Ho: There is no point of using the model; all coeffs are zero
* Ha: Assume at least one is non-zero
* Use to see if the predictors are useful
* will talk more about this when it comes to errors later
* if we haev 10 preictors in the mode, we dont wat to test each of teh predicotrs individuals because tehres agood chance we will amek an error an dhow a type 1 error
* if we do any overal test, it will give us an way to see if anything is useful to us
* good to use for nestsed tests
* comapres the mdoel with all predictors to a null modle where teh coeff are both 0
* the same process can be used to compare nested models together
* think: Model with price by size and lot, compared to a model of just price by size; does adding this extra variabile explain a sig amount more of the variability?

**R - Regression ANOVA** - **Important note:** R shows a “sequential” sum of squares in the ANOVA table, i.e. how much new variability is explained as each predictor is added. Add the components to find the SSModel.

anova(HouseModel)

## Analysis of Variance Table  
##   
## Response: Price  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Size 1 4.0447e+10 4.0447e+10 18.0018 0.0005485 \*\*\*  
## Lot 1 7.6013e+09 7.6013e+09 3.3831 0.0833990 .   
## Residuals 17 3.8196e+10 2.2468e+09   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# ANOVA by default will looka at ta sequnetial SS  
# This iwll loo at the first line of the model that has price predicted by size and it says is teh slope of that mdoel nonzero? we have pvalue of 0.0005,   
# Next line says compares a model with asize and lot in it to a model of just size; if we add lot to our odel, does it explain more varibaility? In this case, no.   
# Use ANOVA455 if you DONT want it to look at it sequentially

**A “Local” ANOVA Function** - To find ANOVA for a multiple regression model that is NOT split sequentially for each predictor…

anova455(HouseModel)

## ANOVA Table  
## Model: Price ~ Size + Lot   
##   
## Df Sum Sq Mean Sq F value P(>F)   
## Model 2 4.8048e+10 2.4024e+10 10.693 0.000985 \*\*\*  
## Error 17 3.8196e+10 2.2468e+09   
## Total 19 8.6244e+10   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# It condesnes the model and looks at it instead of line by linea nd adding it one at a time, its looking at teh overal model   
# It condesnes all teh SS together rather than the SS of each individual predictor   
  
# We see that its going in twith teh assumption that the coeffes for lot and size are 0 and ttrying to see if we have evidence to say at least one of them is non zero   
# Have as amll pvalue that at least one is nonzero, but it seems that it contradicts what we have before   
# Where we lookeda t teh summary, then both the pvalues were big, so it's a bit contradictory because of teh multiocolinearity that we talk about next class

**Example: Houses** 1. Test #1: Compute and test the correlation between Size and Lot in Houses

cor.test(HousesLot) t = 5.0694, df = 18, p-value = 7.991e-05

1. Test #2: Compute and test the coefficient of Size in a multiple regression model (along with Lot) to predict Price. (Estimate Std Error t value Pr(>|t|) Intercept) 34121.649 29716.458 1.148 0.2668 Size 23.232 17.700 1.313 0.2068 Lot 5.657 3.075 1.839 0.0834

F-statistic: 10.69 on 2 and 17 DF, p-value: 0.000985

## STOR 455 Class 12 R [Correlated Predictors & Model Selection Methods](https://sakai.unc.edu/portal/site/ff98023c-6e12-47a7-acba-0c12abe4203b/tool/f03494dc-48e2-44b2-8904-0aa5ba69b16a#Class 12)

library(readr)  
library(Stat2Data)  
library(car)  
  
data("Houses")  
  
StateSAT <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/StateSAT.csv")  
  
source("https://raw.githubusercontent.com/JA-McLean/STOR455/master/scripts/anova455.R")

head(Houses)

## Price Size Lot  
## 1 212000 4148 25264  
## 2 230000 2501 11891  
## 3 339000 4374 25351  
## 4 289000 2398 22215  
## 5 160000 2536 9234  
## 6 85000 2368 13329

cor(Houses)

## Price Size Lot  
## Price 1.0000000 0.6848219 0.7157072  
## Size 0.6848219 1.0000000 0.7668722  
## Lot 0.7157072 0.7668722 1.0000000

HouseModel=lm(Price~Size+Lot,data=Houses)  
# Linear model that predicts jprice by size and lot of the house.  
  
summary(HouseModel)

##   
## Call:  
## lm(formula = Price ~ Size + Lot, data = Houses)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -79532 -28464 3713 21450 73507   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 34121.649 29716.458 1.148 0.2668   
## Size 23.232 17.700 1.313 0.2068   
## Lot 5.657 3.075 1.839 0.0834 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 47400 on 17 degrees of freedom  
## Multiple R-squared: 0.5571, Adjusted R-squared: 0.505   
## F-statistic: 10.69 on 2 and 17 DF, p-value: 0.000985

# Tests teh coef size and lot are equal to zero; that they do not have a realtionship with price   
# Alternative: that at least one of those is nonzero   
# Very low pvalue, very unlikly that we would get this sample if the nuill was true; we have evide nce to say that at least one of these has a non zero slope   
  
# Whtat is the FTest stats and what it is useful?   
# The pvcaalue is really what we want  
# The f tests stat, if its big or small depends on the sample size and the number of predictors;   
# GFOr a small sample of small predictors, 10 = big number   
# Large sample wiht lots of predictors, 10 = small number   
# Mostly focus on the pvalue and how to interpret that   
  
# Looking at the coeff table   
# The Ho: Coeff of size = 0  
# Ha: Coef size != 0   
 # Same thing for lot   
# Each ahas an indivual test sthat ehre is probably not evidence of a relationshipo   
# The two appear to be contradictory there   
# When you're writng out the hypotehsis, you can say in words what he ahsa said or you can write it our mathmatically

#cor.test(Houses)  
  
cor.test(Houses$Size, Houses$Price)

##   
## Pearson's product-moment correlation  
##   
## data: Houses$Size and Houses$Price  
## t = 3.9871, df = 18, p-value = 0.0008643  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.3476582 0.8651583  
## sample estimates:  
## cor   
## 0.6848219

# jfThe relationship bt price and size are with a low pvalue, 0.0008; havbe evidence of a relationship here that is non zero correlation   
# the same test with the other; there is no evidence of ra realtionship   
cor.test(Houses$Lot, Houses$Price)

##   
## Pearson's product-moment correlation  
##   
## data: Houses$Lot and Houses$Price  
## t = 4.3478, df = 18, p-value = 0.0003878  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.3998134 0.8796343  
## sample estimates:  
## cor   
## 0.7157072

cor.test(Houses$Size,Houses$Lot)

##   
## Pearson's product-moment correlation  
##   
## data: Houses$Size and Houses$Lot  
## t = 5.0694, df = 18, p-value = 7.991e-05  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.4909631 0.9029654  
## sample estimates:  
## cor   
## 0.7668722

# We get contradcitory results because of multicollinearity   
# there is an issue wher ethe variance is being inflated when we do these tests   
# IF we loook at the relationship between teh predictors (Lot and size) we see that eh correaltion is really high 0.7686 ish (Not exaclty) it's a signifigant realtionship   
# This is driving the conficting results because too much is being explained by the two thing s  
# It's not inherantly bad, it's not telling ust htat there is no realtionshipship, it's just saying that we dont haev evidence to say ther eis s asignfigiant realtionship   
# If we have a lto of predictors that are highly correlationed you might not want to use them all on our model   
# If these predcitors are explaining the same thing, then why include both? IT sjust going to cause problems   
# THis can cause overfitting problems

*Simple models are idea, than overaly complicated ones*

**Multicollinearity**

* What is it? – When two or more predictors are strongly associated with each other.
* Why is it a problem? –Individual coefficients and t-tests can be deceptive and unreliable.

*NOPtes* - Makes the tests deceptive and we need to know that there is multicoloinarity going on - More its unrealiable tests if there are multicollinearity - It makes it harder to interpret but it means that our model acna be simpler than what we have - so it really means, jsut change you rmodel a little

**Effects of Multicollinearity** - If predictors are highly correlated among themselves: 1. The regression coefficients and tests can be extremely variable and difficult to interpret individually. 2. One variable alone might work as well as many.

anova(HouseModel)

## Analysis of Variance Table  
##   
## Response: Price  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Size 1 4.0447e+10 4.0447e+10 18.0018 0.0005485 \*\*\*  
## Lot 1 7.6013e+09 7.6013e+09 3.3831 0.0833990 .   
## Residuals 17 3.8196e+10 2.2468e+09   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova455(HouseModel)

## ANOVA Table  
## Model: Price ~ Size + Lot   
##   
## Df Sum Sq Mean Sq F value P(>F)   
## Model 2 4.8048e+10 2.4024e+10 10.693 0.000985 \*\*\*  
## Error 17 3.8196e+10 2.2468e+09   
## Total 19 8.6244e+10   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

*NOtes* -0 WE can see where the correlation is bt lot and size - One way we can test is to see how closely correlated things are – THis is fine whwen you’re jsut looking at two things; if we have a lot of predictors then teh correlation between things cna be hard to use as a measure ebcause there are more things to look at - Solution: Build a new model for each predcitor where th remaining predcitors are the predictors of that model - In this case, we could build a model for the size of a house and use the rest of the predcitors as predictoyrs (Would just be lot in this case) and do teh same hting for lot and make a model where size is the predictor for that

*See below*

mod=lm(Size~Lot, data=Houses)  
summary(mod)

##   
## Call:  
## lm(formula = Size ~ Lot, data = Houses)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -872.42 -591.71 -47.96 397.03 1214.17   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 91.50286 395.12226 0.232 0.819   
## Lot 0.13324 0.02628 5.069 7.99e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 631.2 on 18 degrees of freedom  
## Multiple R-squared: 0.5881, Adjusted R-squared: 0.5652   
## F-statistic: 25.7 on 1 and 18 DF, p-value: 7.991e-05

# We want ot look at the multiple r sqaured; it says that almost 59% is beign predictoed by how big the lot is   
# Thats just the correlation squared,. the .77 squared; when we get the bigger models, we are going to have to do more to calcualte that   
# We see how much variability is explained there by the two predicotrs   
  
# can use this to se ehow much the variance is being inflated   
# The variance that is being calcualted that is for each paredictor is not done in isolation; its taking into account the other predictors; more multicollinearity will increase the variance   
summary(mod)$r.squared

## [1] 0.588093

# how to pull out the multiple r-squiared from the model

**How do we detect multicollinearity?** 1. Look at a correlation matrix of the predictors.

round(cor(Houses), 2)

## Price Size Lot  
## Price 1.00 0.68 0.72  
## Size 0.68 1.00 0.77  
## Lot 0.72 0.77 1.00

1. Compute the Variance Inflation Factor (VIF).

* (Beware if VIF > 5)
* where Ri2 is for predicting Xi with the other predictors.
* 𝑉𝐼𝐹 > 5 or 𝑅𝑖2 >80%

# How to account for the inflated variance in places with possible multicollinearity   
VIF = 1/(1-summary(mod)$r.squared)  
VIF

## [1] 2.427732

# If VIF is 5 or more, then ther emight be a lot of multicollinearity going on  
# This would mean the adjusted r sqaured would be above 80 or more   
# We are saying that the variance is being aadjusted by a factor of 2.42  
# We get the 2.42 by the VIF   
  
# If we look at the summary of the housemod   
# the variance of size and the stderror = 17.7, when we are doing a hypothesis test for the slope of size, then we are caclauting a t stest stat - the estimate for slope/Stderror;   
 # 23.2/17.9 = 1.313 which is the tvalue   
# that's where that tvalue is coming from   
# We could pull it our better with a summary funciton, but we're not going to   
# So this outcome is about 1.31 stdar devations away if we didnt have a relation between teh things if there was no realtion   
# The variance that we used in this calvcualtion, because of the multicollinearity is being inflated by this facotr   
# This is teh variance inflation factor and we are calculting the stadard error   
# Std = sqrt(variance)   
  
#Go down to sqrtt(VIF) code

**Finding VIF with R** 1. 1. Brute force. Fit a model to predict Xi using the other predictors and find 𝑅𝑖2. - Compute: 𝑉𝐼𝐹=1/(1−𝑅𝑖2) - Example: Find VIF for Size when using Lot to predict Size

summary(HouseModel)

##   
## Call:  
## lm(formula = Price ~ Size + Lot, data = Houses)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -79532 -28464 3713 21450 73507   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 34121.649 29716.458 1.148 0.2668   
## Size 23.232 17.700 1.313 0.2068   
## Lot 5.657 3.075 1.839 0.0834 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 47400 on 17 degrees of freedom  
## Multiple R-squared: 0.5571, Adjusted R-squared: 0.505   
## F-statistic: 10.69 on 2 and 17 DF, p-value: 0.000985

sqrt(VIF) # This is how much that variance is being inflated

## [1] 1.558118

# Look at the summary of the houses model   
summary(HouseModel)$coeff[2,2]/sqrt(VIF)

## [1] 11.36012

# sqrt(VIF) = how much its being inflated   
# If no correlation., then we would haev a variance of 11.1; so now if were to divide the slope by this value instead   
  
summary(HouseModel)$coeff[2,1]/summary(HouseModel)$coeff[2,2]/sqrt(VIF) # This is what we would get for the slope if we didn't have the multicollinearity

## [1] 0.8423846

# So multicollienarity has a really big impact on the model   
  
# We dont have to do the math above in practice, it's really just to learn how and why the infaltion is affecting teh table

**Finding VIF with R** 2. 2. Install car package - use vif( )function or use VIF.R script from Sakai *See below for installing car package and using the vif function*

vif(HouseModel)

## Size Lot   
## 2.427732 2.427732

# This is how to do what you did above, but really short form   
# This is what you would use to look at the inflation   
# This will be the same when you are lookign at t athing with 2 predictors   
# Ity will be different when you ahev mutliple predictors   
  
# Even though it changes our result in the data from a sig to a non sig realtionship; multilcolinearity wise, it snot a huge realtionshio  
# mostly because its samll dataset

**What to do if you’ve got Multicollinearity?** 1.Choose a better set of predictors 2.Eliminate some of the redundant predictors to leave a more independent set. 3.Combine predictors into a scale. 4.“Ignore” the individual coefficients and tests.

Note: Predictions aren’t necessarily worse if some predictors are related – it’s just conclusions about individual terms that might be confused.

**NOTES** - Looking at a bigger dataset when it’s not so straightforward to know if there is multicollinearity or not

**Example: State SAT Scores** Source: Statistical Sleuth, Case 12.1 pg. 339  
Response Variable:  
SAT =Average combined SAT Score Potential Predictors:  
Takers = % taking the exam Income = median family income ($100’s) Years = avg. years of study (SS, NS, HU) Public = % public school Expend = spend per student ($100’s) Rank = median class rank of takers

head(StateSAT)

## # A tibble: 6 x 8  
## State SAT Takers Income Years Public Expend Rank  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Iowa 1088 3 326 16.8 87.8 25.6 89.7  
## 2 SouthDakota 1075 2 264 16.1 86.2 20.0 90.6  
## 3 NorthDakota 1068 3 317 16.6 88.3 20.6 89.8  
## 4 Kansas 1045 5 338 16.3 83.9 27.1 86.3  
## 5 Nebraska 1045 5 293 17.2 83.6 21.0 88.5  
## 6 Montana 1033 8 263 15.9 93.7 29.5 86.4

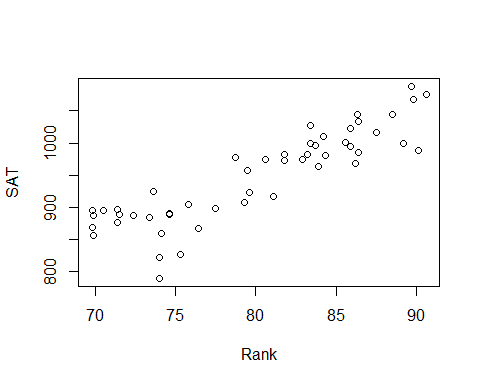
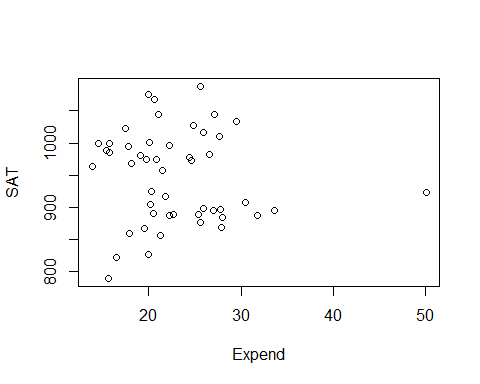
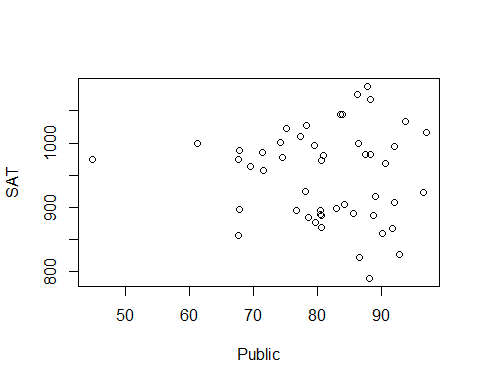
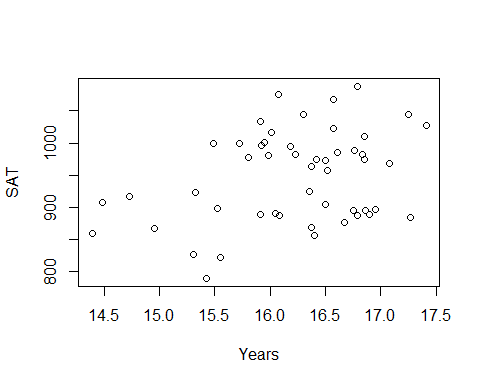
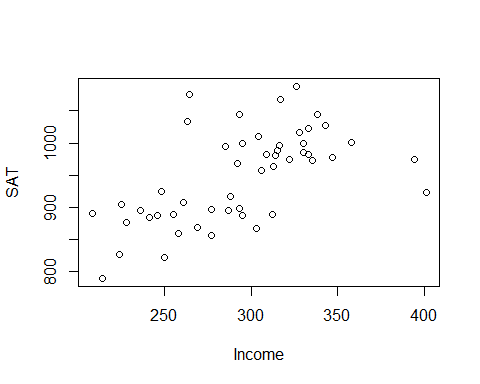
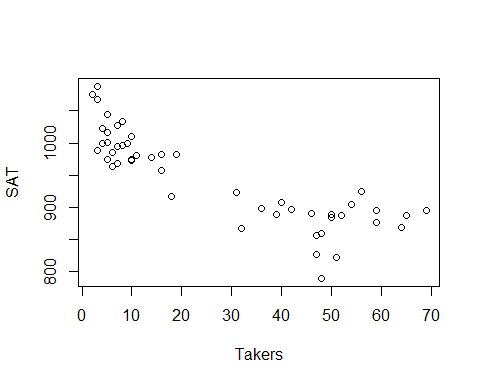
# Some states might have lower SAT takers ebcause the ACT mgiht be better

**Example: Predicting State SAT** Data: StateSAT  
Response: SAT Possible Predictors: Takers Income Years Public Expend Rank

Find the “best” model for GPA using some or all of these predictors.

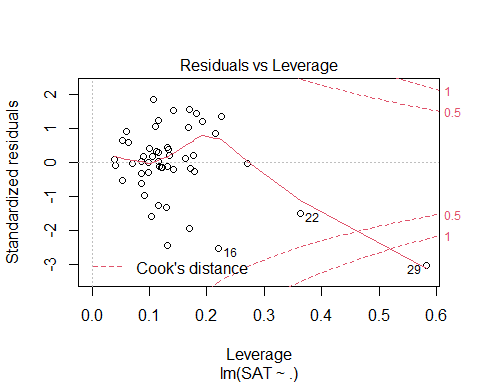
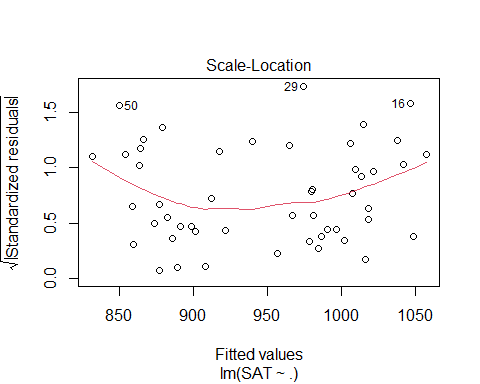
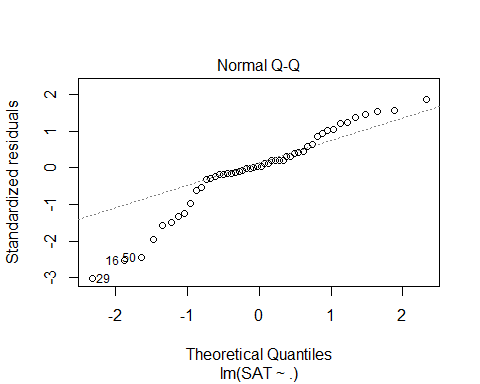
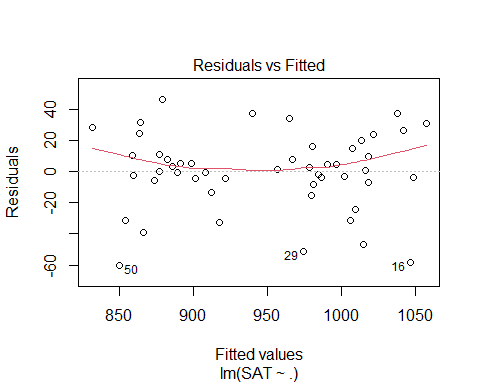
*What determines “best”?*

plot(SAT~., StateSAT[,2:8])



# Want ot predict the average SAT per state   
# The ~ will take all the remaining columns

SAT\_Model = lm(SAT~., data=StateSAT[,2:8])  
plot(SAT\_Model)



summary(SAT\_Model)

##   
## Call:  
## lm(formula = SAT ~ ., data = StateSAT[, 2:8])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -60.046 -6.768 0.972 13.947 46.332   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -94.659109 211.509584 -0.448 0.656731   
## Takers -0.480080 0.693711 -0.692 0.492628   
## Income -0.008195 0.152358 -0.054 0.957353   
## Years 22.610082 6.314577 3.581 0.000866 \*\*\*  
## Public -0.464152 0.579104 -0.802 0.427249   
## Expend 2.212005 0.845972 2.615 0.012263 \*   
## Rank 8.476217 2.107807 4.021 0.000230 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 26.34 on 43 degrees of freedom  
## Multiple R-squared: 0.8787, Adjusted R-squared: 0.8618   
## F-statistic: 51.91 on 6 and 43 DF, p-value: < 2.2e-16

# We see some weird things are happening where there ar ea lot of NAs happenign here   
#It's because we have categorical variables and r is trying to get those results in number form of us   
# We want to use only the nnumeric things   
# That is what this model does above   
  
# The bottom line does a n anova test witht eh assumption that the slope for the columsn are all zero and there is no relation bt sat scores vs the alternative that a tleast one is non zero in thei model   
# We Think base don this model at least one is a good predictor   
# IF we look at the individual tests for slope, we see where 3 of them have low pvalues (YEars, Expend, and Rank) Where the p value is low; this model those seem to have a strong relationship with SAT score   
# These results might be decieving if there are multicollinearity   
# there is most likely multicollinearity here because its all interfomration fro one state   
# There is probably some lurking variables there that are in the background making these things highly correlated to each other

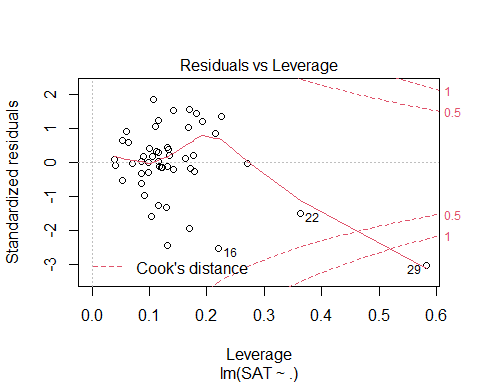
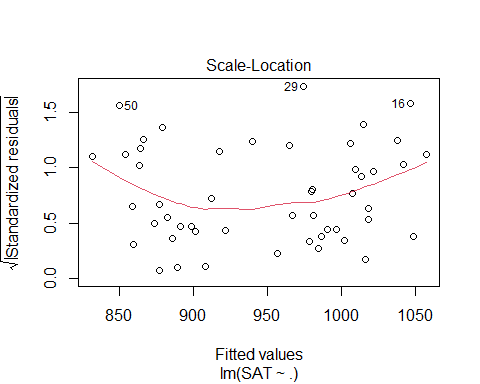
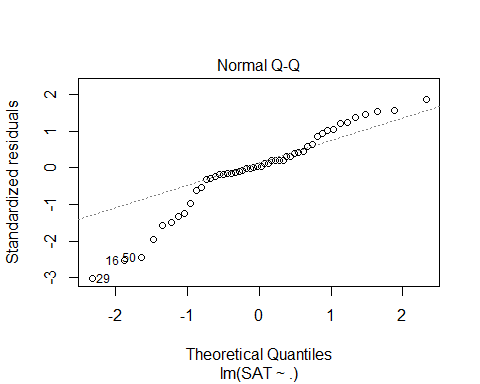
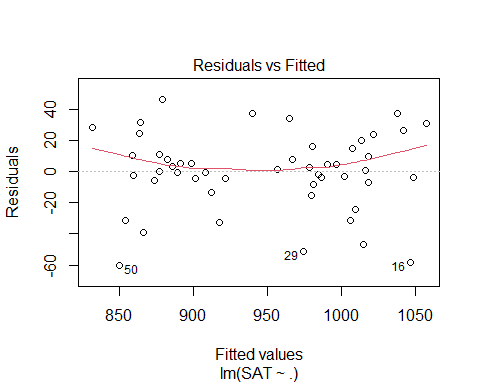
vif(SAT\_Model)

## Takers Income Years Public Expend Rank   
## 16.478636 3.128848 1.379408 2.288398 1.907995 13.347394

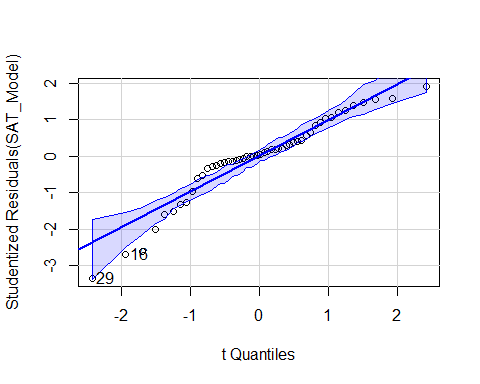
# looking at the VIF for this model, we see that Takers and Rank have really high inflation rates, which means they should probably be expcluded from the final regression model   
# It teslls us wif we make a mdoel with takers as the repsonse, we will get a really high r squared value; almost all the other varibaility in takers is being expained by the other varianceles   
# we probably dont need takers in teh model if everythign else is already doing that for us   
# We could probaly not need rank either as well for the same reason   
# Or maybe just rank or just takers is all we need to predcit.  
# It gives us some informatoin, but the biggest thing is we are skipping a good ifrst step   
# Does it meet the lienar model conditions?

*Does it meet the linear model conditions?* - Look at residual analysis of the data - Does this data meet the criteria, and if it doesn’t, where are these problems occuring? - Becaufore we had 1 variable and wanted to do a transfomration, we could jsut ransforma the predictor/response - Now we have ore predictors and some might have lienar issues while others don’t

plot(SAT\_Model)



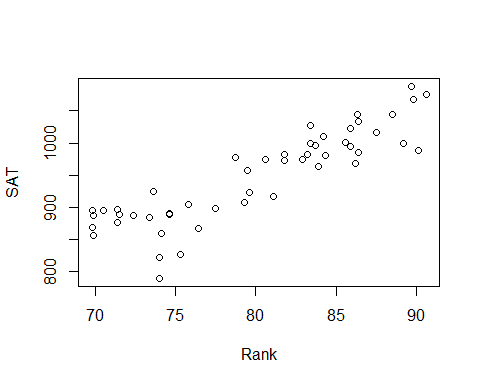
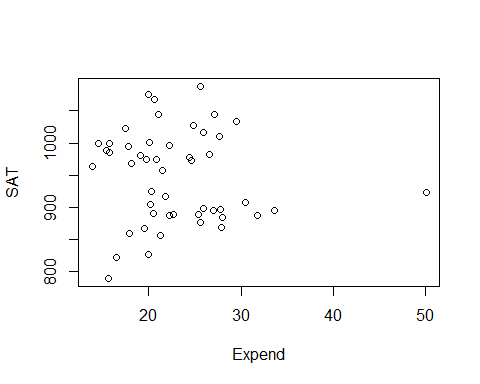
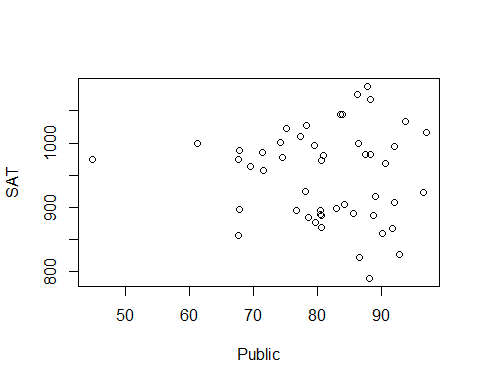
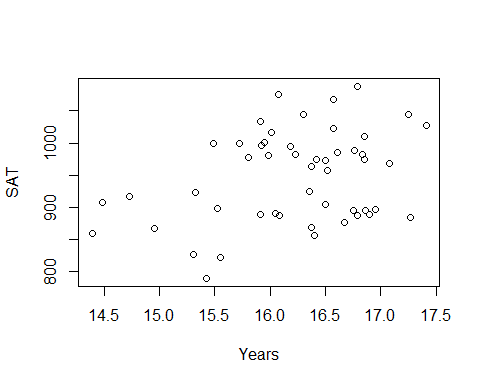
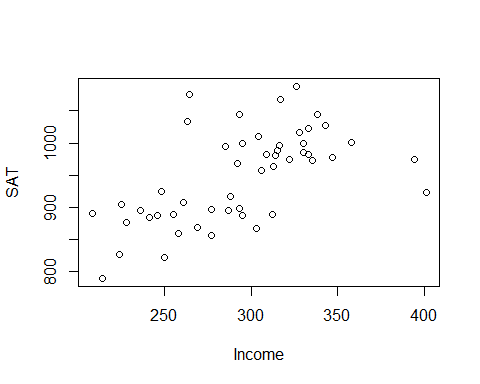
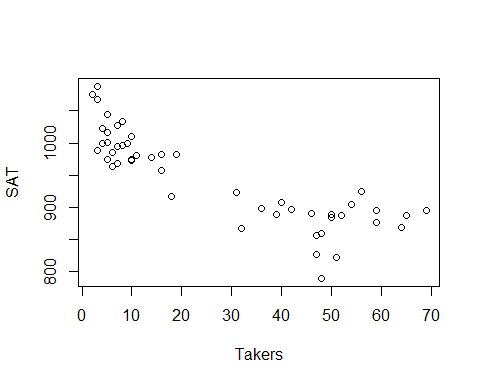
# We can look at these plots to see if the ocnditions are met   
# The model doesn't appear to fit a line well   
# Transfomration might be useful   
# Or, dont include all the variables in teh model   
# Maybe tranfomraiotn or may less predicotrs   
qqPlot(SAT\_Model)



## [1] 16 29

# normalitiy is a big issue   
  
# COnstancce variance: Not a big issue   
# Leverage/Cook's distance - super wonky looking, probably wouldn't trust it right now   
# One data point has high influence in model, should probably look at state 29 because of teh cook's plot

# WE could also look at which varibales are probkematic for us  
plot(SAT~., data = StateSAT[,2:8])



# Tehse plots will help us see wehre things might be a good fit for choosing a predictor   
# Will also help show where we might want to look towards what has outliers   
# first is SAT by takers - there is a curve we could use; does a transformation help this? not so much , but some others might   
# Takers seems a pretty good predicotrorl; would haev to worry about lower states   
  
# Income, doesn't look like a good predicotr, but it is appear to have a conneciton somewhere; but it's not the best; not super linear and not as clear as other s  
# YEars: nothing jumps out, but it's hard to see a pattern, there is something there   
# Public, its hard to say what is going on, that is one that is different than the rest, this is probably the state 29 that has high influence; this variable is probably messing up our data   
# Expend; same issue with one state is apearing to spend more omoeny than the rest of the state; one point has a lot of influence   
# Rank: This is pretty definded realtionship; not a line, but appears to be a good varibale here   
# Guessing: The high VIF bt Takers and RAnk; they appear to expain similar amount of varibility withteh SAT scores

* We have all tehse ariables; how do we make the best model?
* We could go on teh r-squared alone, then it’s pretty good; but you should check the condiotns and that makes it a sus model
* Different ransfomraitons could make the model better and make the model better
* need to see the realtionships to see if there are different combos that will give a better linear conditoins and realtionship between teh model predictors

**Predictor Selection Methods** Choosing an effective set of predictors - All subsets (All combinations of predicotrs in the model and compare all to each other; there is a certain point in which you cant really do this on a compauter because its really hard on a compauter) - Backward elimination - Forward selection - Stepwise regression

## STOR 455 - Class 13 - R Model Section Methods

library(readr)  
library(car)  
library(corrplot) #Install first if needed  
library(leaps) #Install first if needed  
  
StateSAT <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/StateSAT.csv")  
  
source("https://raw.githubusercontent.com/JA-McLean/STOR455/master/scripts/ShowSubsets.R")

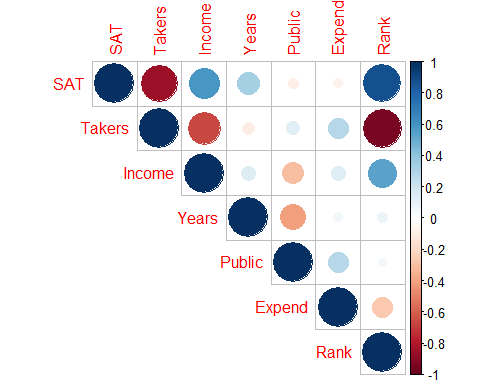
head(StateSAT)

## # A tibble: 6 x 8  
## State SAT Takers Income Years Public Expend Rank  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Iowa 1088 3 326 16.8 87.8 25.6 89.7  
## 2 SouthDakota 1075 2 264 16.1 86.2 20.0 90.6  
## 3 NorthDakota 1068 3 317 16.6 88.3 20.6 89.8  
## 4 Kansas 1045 5 338 16.3 83.9 27.1 86.3  
## 5 Nebraska 1045 5 293 17.2 83.6 21.0 88.5  
## 6 Montana 1033 8 263 15.9 93.7 29.5 86.4

# want to keep in mind what teh corerlation between things are to see what may be useful for a good model   
  
cor(StateSAT[c(2:8)])

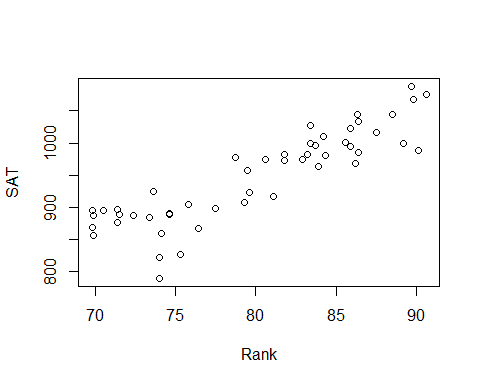
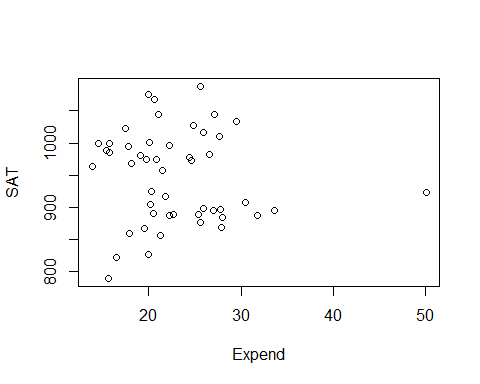
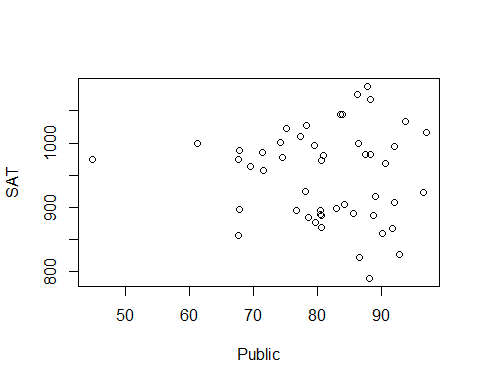
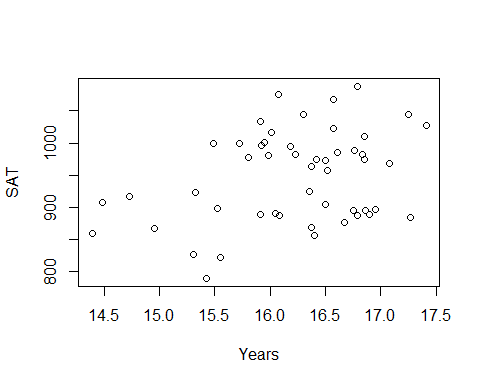
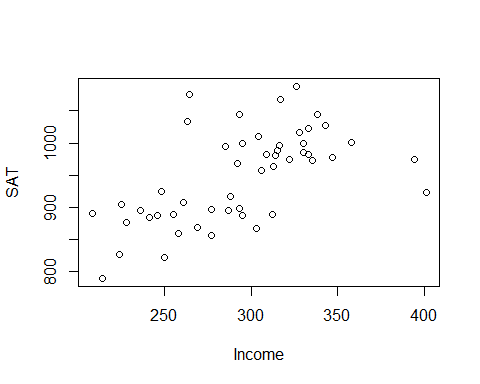
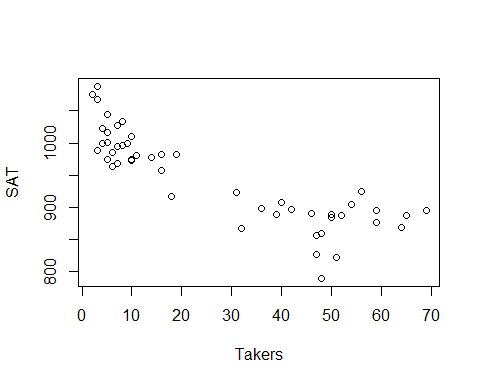
## SAT Takers Income Years Public Expend  
## SAT 1.00000000 -0.8578100 0.5844666 0.33096886 -0.08035688 -0.06287764  
## Takers -0.85780996 1.0000000 -0.6619351 -0.10154350 0.12355625 0.28363041  
## Income 0.58446657 -0.6619351 1.0000000 0.13476231 -0.30656703 0.13151942  
## Years 0.33096886 -0.1015435 0.1347623 1.00000000 -0.41711822 0.05982861  
## Public -0.08035688 0.1235563 -0.3065670 -0.41711822 1.00000000 0.28459116  
## Expend -0.06287764 0.2836304 0.1315194 0.05982861 0.28459116 1.00000000  
## Rank 0.87990910 -0.9428331 0.5326999 0.07022360 0.05062355 -0.26496897  
## Rank  
## SAT 0.87990910  
## Takers -0.94283311  
## Income 0.53269989  
## Years 0.07022360  
## Public 0.05062355  
## Expend -0.26496897  
## Rank 1.00000000

# This makes a correlation matrix that will tell us the correlation between everything in the dataste   
# only owrks for numeric data   
# Not super easy to read   
# Takers has a negative correlation   
# Rank has a strong postive correlation   
# Income and years and other corerlation   
# Doens't tel lme if there is a linera realtionship; its assuming linear relation   
  
corrplot(cor(StateSAT[c(2:8)]), type="upper")



# Helps to visualize the matrix better than other things   
# A nicer visual of the correlation matrix   
# Dark blue = strong correlation   
# Darker and bigger circle = stronger positive or negative correlation   
# Type = "upper" just gives us the upper part of it, it avoids duplicate infomraiotn   
# could also tell where we could have multicllinearity   
# Takers may have multicolinearity from income and rank   
# INcome andr ank will have the same prediction power as takers   
# We can see rank, income and takers have high correlation, so we proabbly dont need all three of those int eh same model because they might explain similar things

plot(SAT~., data=StateSAT[c(2:8)])

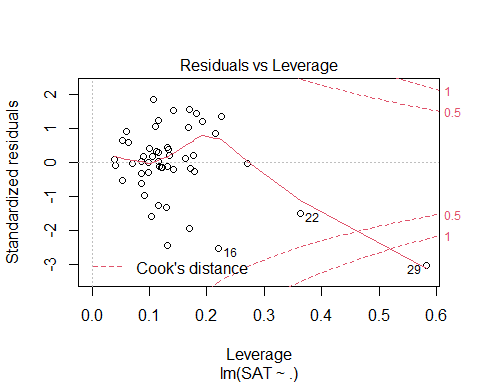
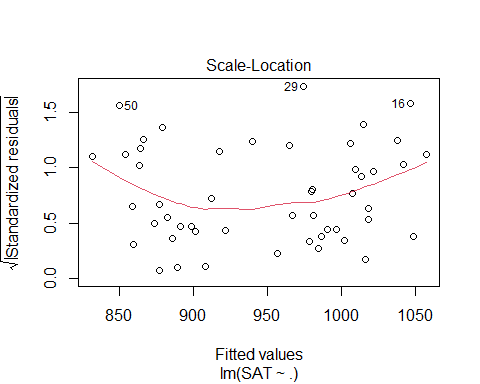
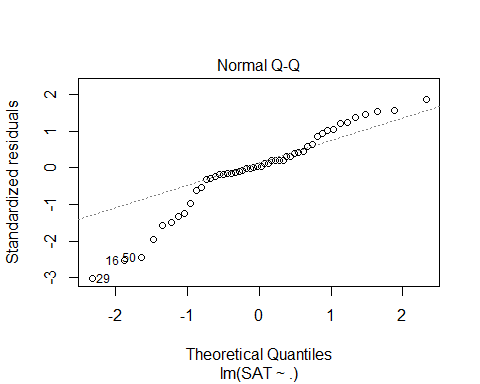
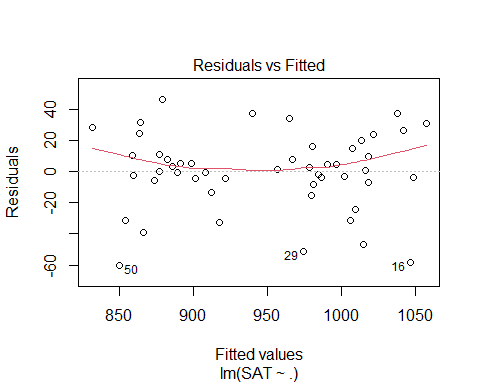


# Plot the data against each of the predictors int eh dataframe   
# Excludes state (Because we would have to factor state, and that would be a lot of information to process)  
  
# Rank adn takers have a recise pattern wtih SAT scores; its appears to have a curved realtionship there   
# Might not have a good linear realtion model conditions, but we can transforms them and work with them   
# Public and Expend = there is one state that is really different htan teh otehrs and thats causing some issues, so we might not want ot use that because it might impact the model in ways we dont wnat

modSAT1 = lm(SAT~., data=StateSAT[c(2:8)])  
# Make a linear model with all the variables   
summary(modSAT1)

##   
## Call:  
## lm(formula = SAT ~ ., data = StateSAT[c(2:8)])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -60.046 -6.768 0.972 13.947 46.332   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -94.659109 211.509584 -0.448 0.656731   
## Takers -0.480080 0.693711 -0.692 0.492628   
## Income -0.008195 0.152358 -0.054 0.957353   
## Years 22.610082 6.314577 3.581 0.000866 \*\*\*  
## Public -0.464152 0.579104 -0.802 0.427249   
## Expend 2.212005 0.845972 2.615 0.012263 \*   
## Rank 8.476217 2.107807 4.021 0.000230 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 26.34 on 43 degrees of freedom  
## Multiple R-squared: 0.8787, Adjusted R-squared: 0.8618   
## F-statistic: 51.91 on 6 and 43 DF, p-value: < 2.2e-16

# Pvalue of kiw, so we can say that some of these we can sue   
# Rank, expend and years have low pvalues; but we could have icorrect infomraiton because of multicollinarity   
# Rank has a similar issue, but it's small pvalue, so it might be a better predictor model   
# Some have high pvalues even though the correlation looked okay  
  
# R squared is the precentage of sat scores that are explained by teh model; this is hgih, but teh conditoins are really met, so we cant use that as a relaibale model   
  
plot(modSAT1) # Too look at residuals



# Lineariry isnt super good   
# normial is really bad, the tail has an issue   
# Residual plot has one state that has really different values than other things  
vif(modSAT1) # To see if there is any inflation of variance

## Takers Income Years Public Expend Rank   
## 16.478636 3.128848 1.379408 2.288398 1.907995 13.347394

**Criteria to Compare Models?** 1. Look for large R2 - But R2 is always best for the model with all predictors - R squared will never go down because if you add something, you’re not explaining less variability you can only explain that much or more; - Just because it’s high rsquared, deosnt mean they are signifigiant

1. Look for large adjusted R2

* Helps factor in the number of predictors in the model
* Adj r squared formuals: -𝑅\_𝑎𝑑𝑗2=1−(𝜎̂2\_𝜀2)/(𝑆\_𝑌^2 )
* 𝑅^2=𝑆𝑆𝑀𝑜𝑑𝑒𝑙/𝑆𝑆𝑇𝑜𝑡𝑎𝑙 =1−𝑆𝑆𝐸/𝑆𝑆𝑇𝑜𝑡𝑎𝑙
* 𝑅\_𝑎𝑑𝑗^2=1−(𝑆𝑆𝐸⁄((𝑛−𝑘−1)))/(𝑆𝑆𝑇𝑜𝑡𝑎𝑙⁄((𝑛−1))) =1−(𝜎̂\_𝜀2)/(𝑠\_𝑌2 )
* (adjusts for the number of predictors in the model)
* THis penalizes teh r squared based ont eh predictors that we have
* it tells us that we know we will have an increased rsquared with extra predictors, so we need a certain amoutn explained to increase teh rsquared

1. Look at individual t-tests

* Might be susceptible to multicollinearity problems
* There could be decent variables, but we aren’t seeing the full story

**How to Choose Models to Compare?** 1. Method #1: **All Subsets!** - Consider all possible combinations of predictors. - How many are there? - Pool of k predictors then 2𝑘−1 subsets - *Advantage:* Find the best model for your criteria - *Disadvantage:* LOTS of computation

*NOtes* - All subsets: - Can look at all subsets or 1 predictors, 2, 3, 4, 5, etc. - We can make a lot of predictors.  
- Can get out of hand quickly if you have a lot of variables - Catgegorical variables make this message because when you factor it you get a variable for the category

all = regsubsets(SAT~., data = StateSAT[c(2:8)], nbest = 2, nvmax = 6)  
# nbest will tell you the two best models with 6, 5, 4, 3, 2, and 1 predictor   
# nvmax will say only look at models with up to 6 predicotrs here; so it is like an upper bound; its not applicable here, but if we had a bigger selection it would be needed   
summary(all)

## Subset selection object  
## Call: regsubsets.formula(SAT ~ ., data = StateSAT[c(2:8)], nbest = 2,   
## nvmax = 6)  
## 6 Variables (and intercept)  
## Forced in Forced out  
## Takers FALSE FALSE  
## Income FALSE FALSE  
## Years FALSE FALSE  
## Public FALSE FALSE  
## Expend FALSE FALSE  
## Rank FALSE FALSE  
## 2 subsets of each size up to 6  
## Selection Algorithm: exhaustive  
## Takers Income Years Public Expend Rank  
## 1 ( 1 ) " " " " " " " " " " "\*"   
## 1 ( 2 ) "\*" " " " " " " " " " "   
## 2 ( 1 ) " " " " "\*" " " " " "\*"   
## 2 ( 2 ) " " " " " " " " "\*" "\*"   
## 3 ( 1 ) " " " " "\*" " " "\*" "\*"   
## 3 ( 2 ) " " "\*" "\*" " " " " "\*"   
## 4 ( 1 ) " " " " "\*" "\*" "\*" "\*"   
## 4 ( 2 ) "\*" " " "\*" " " "\*" "\*"   
## 5 ( 1 ) "\*" " " "\*" "\*" "\*" "\*"   
## 5 ( 2 ) " " "\*" "\*" "\*" "\*" "\*"   
## 6 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*"

#ISsue: THis doesn't compare the models between eachother

# IMPORTANT  
ShowSubsets(all)

## Takers Income Years Public Expend Rank Rsq adjRsq Cp  
## 1 ( 1 ) \* 77.42 76.95 34.03  
## 1 ( 2 ) \* 73.58 73.03 47.64  
## 2 ( 1 ) \* \* 84.71 84.05 10.22  
## 2 ( 2 ) \* \* 80.54 79.71 24.97  
## 3 ( 1 ) \* \* \* 87.11 86.27 3.69  
## 3 ( 2 ) \* \* \* 85.84 84.91 8.21  
## 4 ( 1 ) \* \* \* \* 87.71 86.61 3.58  
## 4 ( 2 ) \* \* \* \* 87.67 86.57 3.72  
## 5 ( 1 ) \* \* \* \* \* 87.87 86.49 5.00  
## 5 ( 2 ) \* \* \* \* \* 87.73 86.34 5.48  
## 6 ( 1 ) \* \* \* \* \* \* 87.87 86.18 7.00

# this iwll give you more infomraiton   
# For each model, what's teh rsquared, the adj rsquared adn teh mallo cp  
  
# We want a small Mallo Cp  
# The first line with rank, it says 77% the stuff is explained, but its' not taking into accoun the otehr variables

**Mallow’s Cp** - Note: R2, Adjusted R2, SSE, all depend only on the predictors in the model being evaluated – NOT the other potential predictors in the pool. - Mallow’s Cp: When evaluating a subset of m predictors from a larger set of k predictors, - m = # predictors in the reduced model - 𝐶\_𝑝=(𝑆𝑆𝐸\_𝑚)/(𝑀𝑆𝐸\_𝑘 )+2(𝑚+1)−𝑛 *notes* - The amount of var explained with reduced model (What we are just using) compared with teh full model with all of the possible predictors in it (The entire model) - What fraction of the model is explained - It penalizes bigger models - If we look at the full model, it gives us the SSE fule/MSE + the left over (the 2(m-1) etc. - Mallow cp = number of predictors + 1 - If numbers are lower than that number, then thats a useful model - So 2 predictor model, look for a Cp of 3 or less

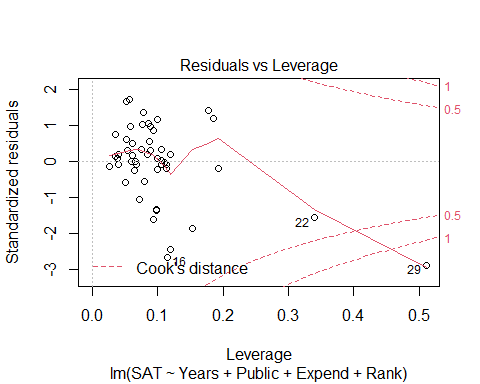
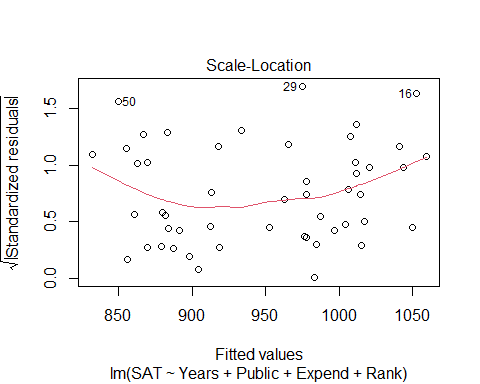
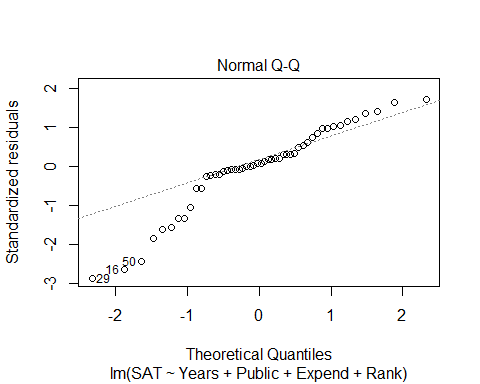
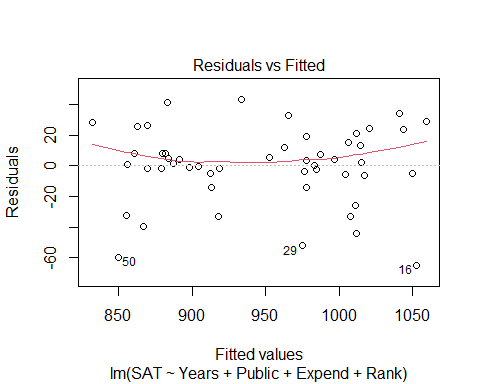
**Notes on Cp** - Cp depends on the larger pool of predictors as well as the set being considered. - For full model Cp = k+1 - For a “good” set of predictor, Cp should be small. - Like Adj R2, Cp weighs both the effectiveness of the model (SSEm) and the # of predictors (m).

**Predictor Selection Methods** - Think, consult, graph… but if that fails, then: 1. All subsets 2. Backward elimination 3. Forward selection 4. Stepwise regression

modSAT3 = lm(SAT~Years+Public+Expend+Rank, data=StateSAT) # this is lowest mallow Cp from best subsets above   
summary(modSAT3)

##   
## Call:  
## lm(formula = SAT ~ Years + Public + Expend + Rank, data = StateSAT)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -64.931 -5.471 1.932 14.980 43.280   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -204.5982 117.6871 -1.738 0.088963 .   
## Years 21.8905 6.0372 3.626 0.000731 \*\*\*  
## Public -0.6638 0.4500 -1.475 0.147154   
## Expend 2.2416 0.6782 3.305 0.001868 \*\*   
## Rank 10.0032 0.6033 16.581 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 25.93 on 45 degrees of freedom  
## Multiple R-squared: 0.8771, Adjusted R-squared: 0.8661   
## F-statistic: 80.25 on 4 and 45 DF, p-value: < 2.2e-16

plot(modSAT3)



vif(modSAT3)

## Years Public Expend Rank   
## 1.301929 1.426831 1.266145 1.129034

# Look at sum; it's sig because we know allsubsets   
# Public has a higher pvalue, but thats because of multicollinearity; they were all highly correlated; public is being inflated a bit   
# We can see that ints not inflated too much because teh VIF is amll; maybe Public just isnt that good   
# Problem: The residual anaysis, we still have nonlinearitiy; if we too things taht din't haev lienar relation with teh response, then we are going to have problems   
# We need to try and make tehse lienar realtions work first, then put it in the model selction process.

**Backward Elimination** 1. Start with the full model (all predictors) 2. Calculate if the model would be “better” by removing each of the predictor individually 3. Find the “least significant” predictor 4. Does removing the predictor create a “better” model? - No, then Keep the predictor & stop - Yes, then Delete the predictor and go back to step 2 with the reduced model.

* *Advantages:* Removes “worst” predictors early Relatively few models to consider Leaves only “important” predictors
* *Disadvantages:* Most complicated models first Individual t-tests may be unstable Susceptible to multicollinearity

summary(modSAT1)

##   
## Call:  
## lm(formula = SAT ~ ., data = StateSAT[c(2:8)])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -60.046 -6.768 0.972 13.947 46.332   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -94.659109 211.509584 -0.448 0.656731   
## Takers -0.480080 0.693711 -0.692 0.492628   
## Income -0.008195 0.152358 -0.054 0.957353   
## Years 22.610082 6.314577 3.581 0.000866 \*\*\*  
## Public -0.464152 0.579104 -0.802 0.427249   
## Expend 2.212005 0.845972 2.615 0.012263 \*   
## Rank 8.476217 2.107807 4.021 0.000230 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 26.34 on 43 degrees of freedom  
## Multiple R-squared: 0.8787, Adjusted R-squared: 0.8618   
## F-statistic: 51.91 on 6 and 43 DF, p-value: < 2.2e-16

# See that we would want t amodel with no income in it because it's the worse predictor   
  
#This is what backwards elimiation is doing, but step by step  
  
modSAT2.1 = lm(SAT~Takers+Years+Public+Expend+Rank, data=StateSAT)  
summary(modSAT2.1)

##   
## Call:  
## lm(formula = SAT ~ Takers + Years + Public + Expend + Rank, data = StateSAT)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -59.890 -6.637 0.975 13.872 46.261   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -100.4737 179.7256 -0.559 0.578969   
## Takers -0.4621 0.6007 -0.769 0.445883   
## Years 22.6688 6.1486 3.687 0.000620 \*\*\*  
## Public -0.4523 0.5291 -0.855 0.397344   
## Expend 2.1859 0.6851 3.190 0.002620 \*\*   
## Rank 8.4964 2.0505 4.144 0.000153 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 26.04 on 44 degrees of freedom  
## Multiple R-squared: 0.8787, Adjusted R-squared: 0.8649   
## F-statistic: 63.74 on 5 and 44 DF, p-value: < 2.2e-16

# We look at the summary of the new model and then choose the next worse predictor that we want to get rid of   
  
modSAT2.2 = lm(SAT~Years+Public+Expend+Rank, data=StateSAT)  
# This is the new model without takers, because takers probably wasn't signfigant   
summary(modSAT2.2)

##   
## Call:  
## lm(formula = SAT ~ Years + Public + Expend + Rank, data = StateSAT)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -64.931 -5.471 1.932 14.980 43.280   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -204.5982 117.6871 -1.738 0.088963 .   
## Years 21.8905 6.0372 3.626 0.000731 \*\*\*  
## Public -0.6638 0.4500 -1.475 0.147154   
## Expend 2.2416 0.6782 3.305 0.001868 \*\*   
## Rank 10.0032 0.6033 16.581 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 25.93 on 45 degrees of freedom  
## Multiple R-squared: 0.8771, Adjusted R-squared: 0.8661   
## F-statistic: 80.25 on 4 and 45 DF, p-value: < 2.2e-16

# We look at the summary of the new model ad tehn choose the next worse predictor that we want to get rid of   
  
modSAT2.3 = lm(SAT~Years+Expend+Rank, data=StateSAT)  
# This is the new model without takers and public, because public probably wasnt signifigant either   
summary(modSAT2.3)

##   
## Call:  
## lm(formula = SAT ~ Years + Expend + Rank, data = StateSAT)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -64.802 -6.798 2.169 17.525 49.706   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -303.7243 97.8415 -3.104 0.00326 \*\*   
## Years 26.0952 5.3894 4.842 1.49e-05 \*\*\*  
## Expend 1.8609 0.6351 2.930 0.00526 \*\*   
## Rank 9.8258 0.5987 16.412 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 26.25 on 46 degrees of freedom  
## Multiple R-squared: 0.8711, Adjusted R-squared: 0.8627   
## F-statistic: 103.6 on 3 and 46 DF, p-value: < 2.2e-16

**How to do backwards elimination in R**

# Fit the full model  
Full=lm(SAT~Takers+Income+Years+Public+Expend+Rank, data=StateSAT)  
# Find the MSE for the full model  
  
MSE=(summary(Full)$sigma)^2  
# Backward: use the step( ) command starting with the full model  
#MSE = variance of the residuals   
  
step(Full,scale=MSE) # this is the step back so it can step by the mallow cp, so it will get teh model with the smallest mallo cp

## Start: AIC=7  
## SAT ~ Takers + Income + Years + Public + Expend + Rank  
##   
## Df Sum of Sq RSS Cp  
## - Income 1 2.0 29844 5.0029  
## - Takers 1 332.4 30175 5.4789  
## - Public 1 445.8 30288 5.6424  
## <none> 29842 7.0000  
## - Expend 1 4744.9 34587 11.8369  
## - Years 1 8897.8 38740 17.8208  
## - Rank 1 11223.0 41065 21.1712  
##   
## Step: AIC=5  
## SAT ~ Takers + Years + Public + Expend + Rank  
##   
## Df Sum of Sq RSS Cp  
## - Takers 1 401.3 30246 3.5812  
## - Public 1 495.5 30340 3.7169  
## <none> 29844 5.0029  
## - Expend 1 6904.4 36749 12.9515  
## - Years 1 9219.7 39064 16.2876  
## - Rank 1 11645.9 41490 19.7836  
##   
## Step: AIC=3.58  
## SAT ~ Years + Public + Expend + Rank  
##   
## Df Sum of Sq RSS Cp  
## <none> 30246 3.5812  
## - Public 1 1462 31708 3.6884  
## - Expend 1 7343 37589 12.1618  
## - Years 1 8837 39083 14.3141  
## - Rank 1 184786 215032 267.8394

##   
## Call:  
## lm(formula = SAT ~ Years + Public + Expend + Rank, data = StateSAT)  
##   
## Coefficients:  
## (Intercept) Years Public Expend Rank   
## -204.5982 21.8905 -0.6638 2.2416 10.0032

#R uses Cp (AIC) to pick next model  
# Builds model with all predictors; if we removed any predictors, tehn what would the model be if weremove: none = 7; if we remove income, takers, or public then it would get better, but the expend, years, and rank would be bad to get rid of   
# It will take teh worse predictor and get rid of it   
# the best model will be at the bottom   
# This can take a lot of screen, so you can add "trace = FALSE" to the end, which will just give you the last output

**Forward Selection** 1. Start with the best single predictor 2. Is that predictor significant? Yes, then Include predictor in the model No, then Don’t include predictor & stop 3. Find the “most significant” new predictor from among those NOT in the model. Return to step 2.

* *Advantages:* Uses smaller models early (parsimony) Less susceptible to multicollinearity Shows “most important” predictors
* *Disadvantages:* Need to consider more models Predictor entered early may become redundant later
* Continue until adding something is no longer useful
* Want to start with no predictors in the model

# Start with a model with NO predictors  
none=lm(SAT~1,data=StateSAT)  
  
 #Specify the direction  
step(none,scope=list(upper=Full),scale=MSE, direction= "forward")# Full is the full model, you have to tell R what the end point is, it wouldn't have an end point if you didn't include that

## Start: AIC=306.48  
## SAT ~ 1  
##   
## Df Sum of Sq RSS Cp  
## + Rank 1 190471 55539 34.027  
## + Takers 1 181024 64987 47.639  
## + Income 1 84038 161973 187.388  
## + Years 1 26948 219063 269.648  
## + Public 1 1589 244422 306.189  
## <none> 246011 306.478  
## + Expend 1 973 245038 307.076  
##   
## Step: AIC=34.03  
## SAT ~ Rank  
##   
## Df Sum of Sq RSS Cp  
## + Years 1 17913.6 37626 10.215  
## + Expend 1 7671.0 47868 24.974  
## + Income 1 4601.1 50938 29.397  
## + Public 1 3847.7 51692 30.483  
## + Takers 1 1761.8 53778 33.488  
## <none> 55539 34.027  
##   
## Step: AIC=10.22  
## SAT ~ Rank + Years  
##   
## Df Sum of Sq RSS Cp  
## + Expend 1 5917.6 31708 3.6884  
## + Income 1 2782.4 34843 8.2059  
## <none> 37626 10.2152  
## + Takers 1 778.7 36847 11.0931  
## + Public 1 37.0 37589 12.1618  
##   
## Step: AIC=3.69  
## SAT ~ Rank + Years + Expend  
##   
## Df Sum of Sq RSS Cp  
## + Public 1 1462.46 30246 3.5812  
## <none> 31708 3.6884  
## + Takers 1 1368.28 30340 3.7169  
## + Income 1 848.47 30860 4.4659  
##   
## Step: AIC=3.58  
## SAT ~ Rank + Years + Expend + Public  
##   
## Df Sum of Sq RSS Cp  
## <none> 30246 3.5812  
## + Takers 1 401.32 29844 5.0029  
## + Income 1 70.95 30175 5.4789

##   
## Call:  
## lm(formula = SAT ~ Rank + Years + Expend + Public, data = StateSAT)  
##   
## Coefficients:  
## (Intercept) Rank Years Expend Public   
## -204.5982 10.0032 21.8905 2.2416 -0.6638

# Shows you what will happen to the mallow cp if you add a certian predictor to it   
# Computationally, it is a little heavy because it has a lot to look at   
# Sometimes though, the first predictor isnt good once you reach the end

step(none, scope=list(upper=Full), scale=MSE, direction="forward", trace=FALSE) # This is how you get the forward selection, but just the end solution

##   
## Call:  
## lm(formula = SAT ~ Rank + Years + Expend + Public, data = StateSAT)  
##   
## Coefficients:  
## (Intercept) Rank Years Expend Public   
## -204.5982 10.0032 21.8905 2.2416 -0.6638

**Stepwise Regression** - Basic idea: Alternate forward selection and backward elimination 1. Use forward selection to choose a new predictor and check its significance. 2. Use backward elimination to see if predictors already in the model can be dropped.

* What would happen if you add or substract certain things and how would that impact eth mallow cp

# Start with a model with NO predictors  
none=lm(SAT~1,data=StateSAT)  
  
 # Don’t specify a direction  
step(none,scope=list(upper=Full),scale=MSE)

## Start: AIC=306.48  
## SAT ~ 1  
##   
## Df Sum of Sq RSS Cp  
## + Rank 1 190471 55539 34.027  
## + Takers 1 181024 64987 47.639  
## + Income 1 84038 161973 187.388  
## + Years 1 26948 219063 269.648  
## + Public 1 1589 244422 306.189  
## <none> 246011 306.478  
## + Expend 1 973 245038 307.076  
##   
## Step: AIC=34.03  
## SAT ~ Rank  
##   
## Df Sum of Sq RSS Cp  
## + Years 1 17914 37626 10.215  
## + Expend 1 7671 47868 24.974  
## + Income 1 4601 50938 29.397  
## + Public 1 3848 51692 30.483  
## + Takers 1 1762 53778 33.488  
## <none> 55539 34.027  
## - Rank 1 190471 246011 306.478  
##   
## Step: AIC=10.22  
## SAT ~ Rank + Years  
##   
## Df Sum of Sq RSS Cp  
## + Expend 1 5918 31708 3.6884  
## + Income 1 2782 34843 8.2059  
## <none> 37626 10.2152  
## + Takers 1 779 36847 11.0931  
## + Public 1 37 37589 12.1618  
## - Years 1 17914 55539 34.0268  
## - Rank 1 181437 219063 269.6479  
##   
## Step: AIC=3.69  
## SAT ~ Rank + Years + Expend  
##   
## Df Sum of Sq RSS Cp  
## + Public 1 1462 30246 3.5812  
## <none> 31708 3.6884  
## + Takers 1 1368 30340 3.7169  
## + Income 1 848 30860 4.4659  
## - Expend 1 5918 37626 10.2152  
## - Years 1 16160 47868 24.9737  
## - Rank 1 185667 217375 269.2161  
##   
## Step: AIC=3.58  
## SAT ~ Rank + Years + Expend + Public  
##   
## Df Sum of Sq RSS Cp  
## <none> 30246 3.5812  
## - Public 1 1462 31708 3.6884  
## + Takers 1 401 29844 5.0029  
## + Income 1 71 30175 5.4789  
## - Expend 1 7343 37589 12.1618  
## - Years 1 8837 39083 14.3141  
## - Rank 1 184786 215032 267.8394

##   
## Call:  
## lm(formula = SAT ~ Rank + Years + Expend + Public, data = StateSAT)  
##   
## Coefficients:  
## (Intercept) Rank Years Expend Public   
## -204.5982 10.0032 21.8905 2.2416 -0.6638

# In this case we end up with the same case, but this isn't always the case   
# you might end up with different things

**Missing Values** - Warning! If data are missing for any of the predictors in the pool, R’s “Stepwise” and “Best Subsets” procedures will eliminate the data case from all\* models. - Thus, running the model for the selected subset of predictors alone may produce different results than within the stepwise or best subsets procedures. - \*R’s step( ) sometimes gives an error.

## STOR 455 Homework #2

40 points - Due Wednesday 9/15 at 5:00pm

**Situation:** Suppose that you are interested in purchasing a used car. How much should you expect to pay? Obviously the price will depend on the type of car you get (the model) and how much it’s been used. For this assignment you will investigate how the price might depend on the age and mileage.

**Data Source:** To get a sample of cars, begin with the UsedCars CSV file. The data was acquired by scraping TrueCar.com for used car listings on 9/24/2017 and contains more than 1.2 million used cars. For this assignment you will choose a car *Model* for which there are at least 100 of that model listed for sale in a state of your choice (that is not North Carolina). After constructing a subset of the UsedCars data under these conditions, check to make sure that there is a reasonable amount of variability in the years and mileages for your car (ie, all of your cars are not from the same year). The model that you choose should have cars ranging over at least 5 years. You should add a variable called *Age* which is 2017-year (since the data was scraped in 2017).

**Directions:** The code below should walk you through the process of selecting data from a particular model and state of your choice. Each of the following two R chunks begin with {r, eval=FALSE}. eval=FALSE makes these chunks not run when I knit the file. Before you knit these chunks, you should revert them to {r}.

library(readr)

# This line will only run if the UsedCars.csv is stored in the same directory as this notebook!

UsedCars <- read\_csv("UsedCars.csv")

# Delete the \*\* below and enter the two letter abbreviation for the state of your choice.

StateOfMyChoice = "NY"

# Delete the \*\* below and enter the model that you chose from the Enough\_Cars data.

ModelOfMyChoice = "Civic"

# Takes a subset of your model car from your state

MyCars = subset(UsedCars, Model==ModelOfMyChoice & State==StateOfMyChoice)

# Add a new variable for the age of the cars.

MyCars$Age = 2017 - MyCars$Year

**MODEL #1: Use Age as a predictor for Price**

1. Calculate the least squares regression line that best fits your data. Interpret (in context) what the slope estimate tells you about prices and ages of your used car model. Explain why the sign (positive/negative) makes sense.

**0.5 point** code for model  
**0.5 point** explanation that the slope describes the change in price of the car as it ages 1 year. For my data, for each one year a Civic ages, I predict that the Price of the car will decrease by $1312.35

modq1 = lm(Price~Age, data=MyCars)

summary(modq1)$coef[2,1]

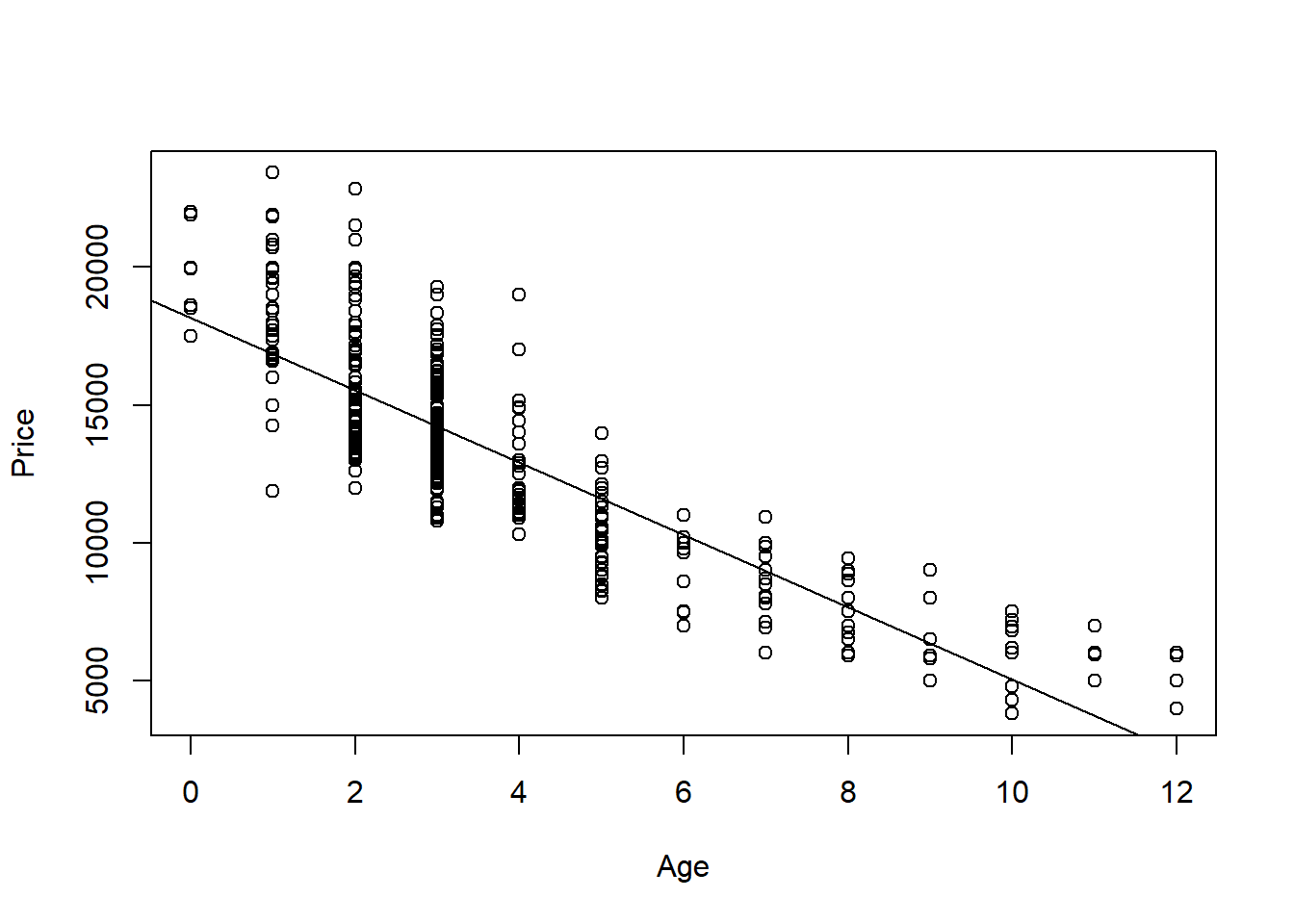
## [1] -1312.345

1. Produce a scatterplot of the relationship with the regression line on it.

**0.5 point** code for plot  
**0.5 point** abline

plot(Price~Age, data=MyCars)

abline(modq1)



1. Produce appropriate residual plots and comment on how well your data appear to fit the conditions for a linear model. Don’t worry about doing transformations at this point if there are problems with the conditions.

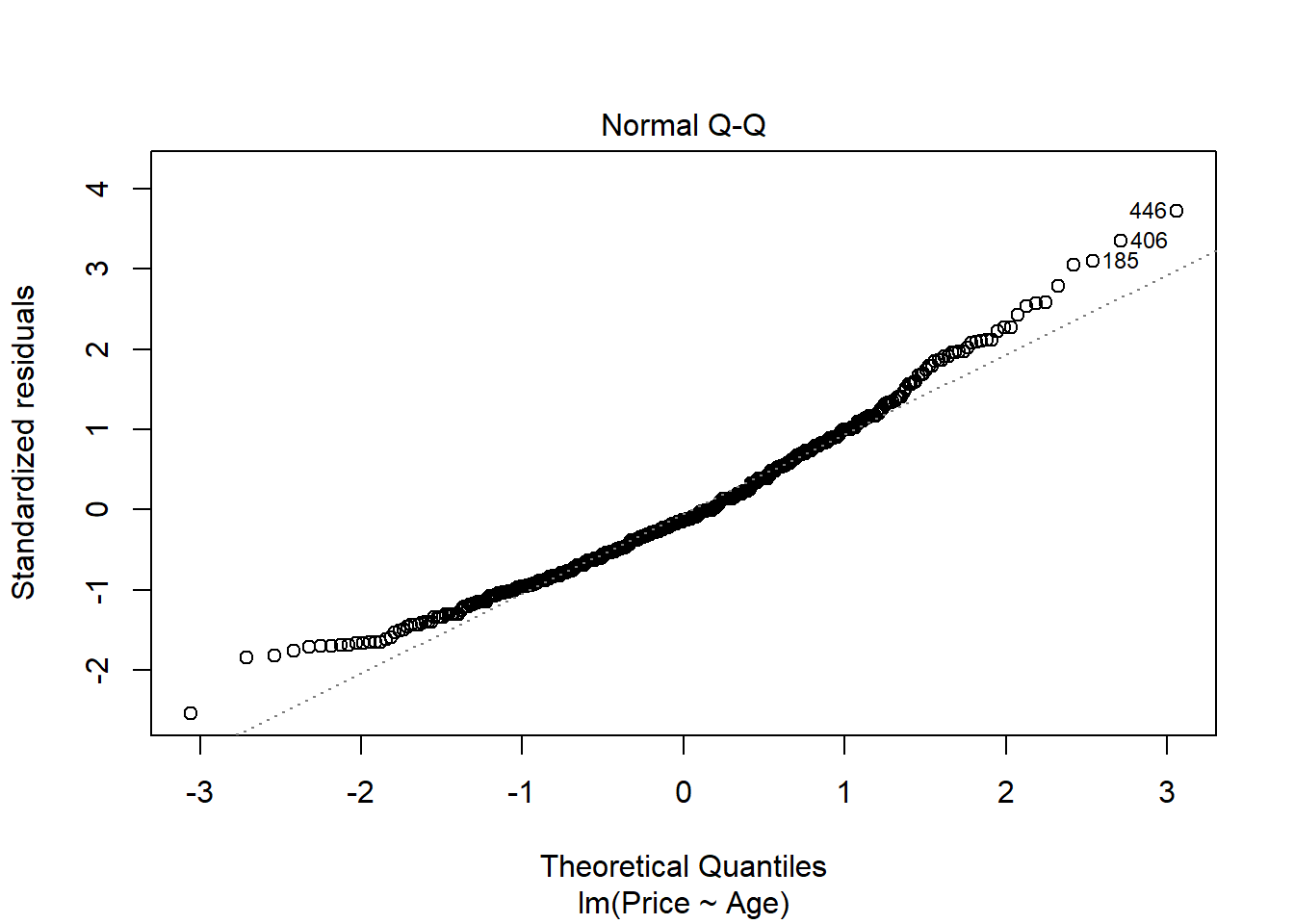
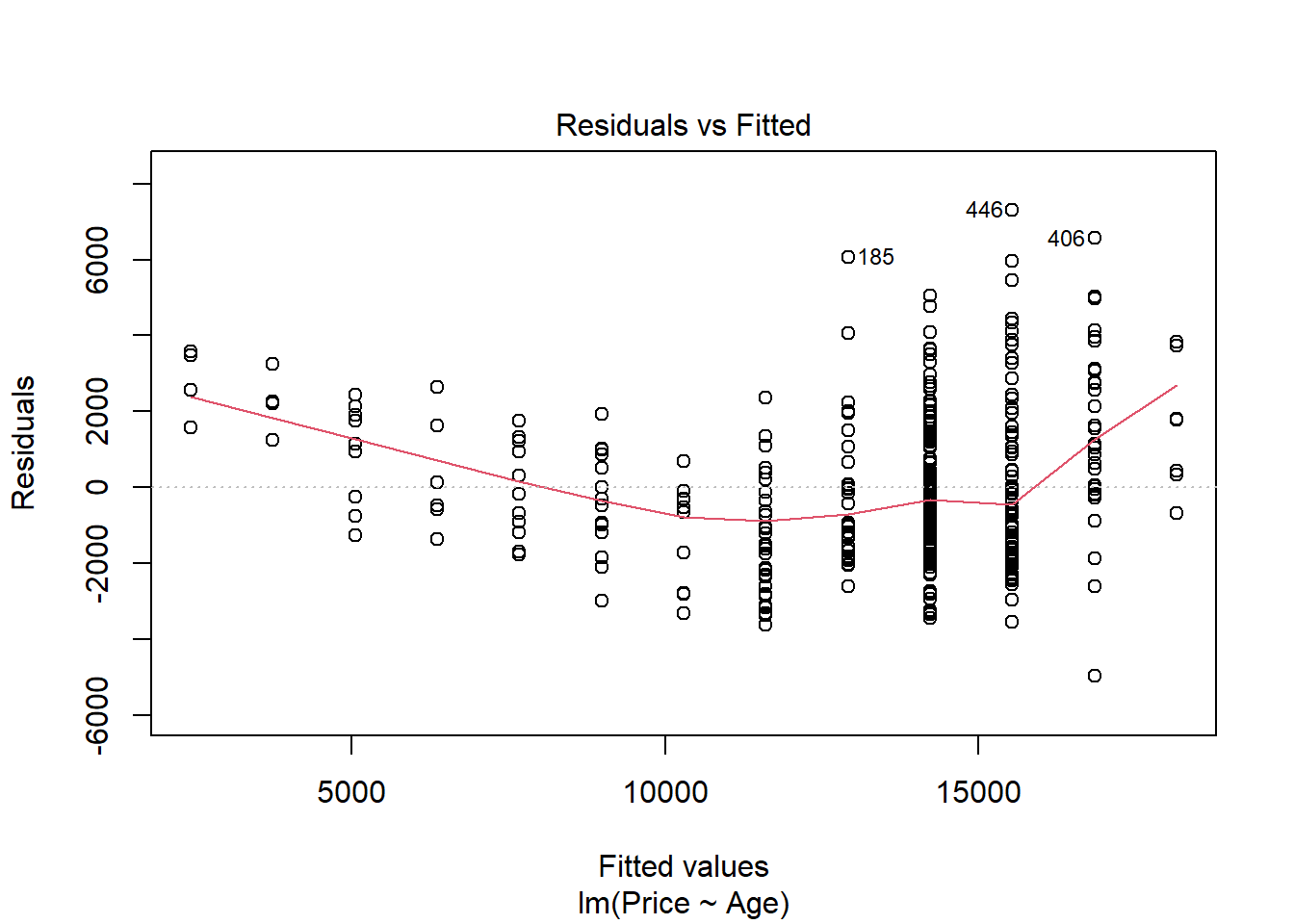
**0.5 points** residuals vs fitted plot  
**0.5 points** qqnorm (or histogram) for Normality of residuals  
**1.5 points** discussion of conditions (linearity, constant variance, and normality of residuals) You can give 0.5 pts each. They can describe the conditions without explicitly using these terms.

Note 1: For linearity, they should have some discussion if the line seems to describe the data, using either the scatter plot or residual vs fitted plot. For constant variance they should discuss if the variability (vertical distances) from the line seems to follow any pattern as value of the predictor changes. For normality of the residuals, they should note the adherence of the residuals (or not) to the qqline, or bell curve shape or skew in a histogram. As each student will have a different plot, any reasonable assertions of the conditions being met (or not) supported by an argument is fine for full credit.

Note 2: plot and discussion of the additional condition, normality of residuals for each value of the predictor, is not required for this assignment (although realistically likely needed for this data). They should split the residuals by Age (or using some other method) and comment on how for each predictor value, the residuals do, or do not seem to be distributed approximately normal.

Note 3: they may use the plot(model) to produce all of the plots at once, or separately produce each of the plots with different lines of code.

plot(modq1, 1:2)

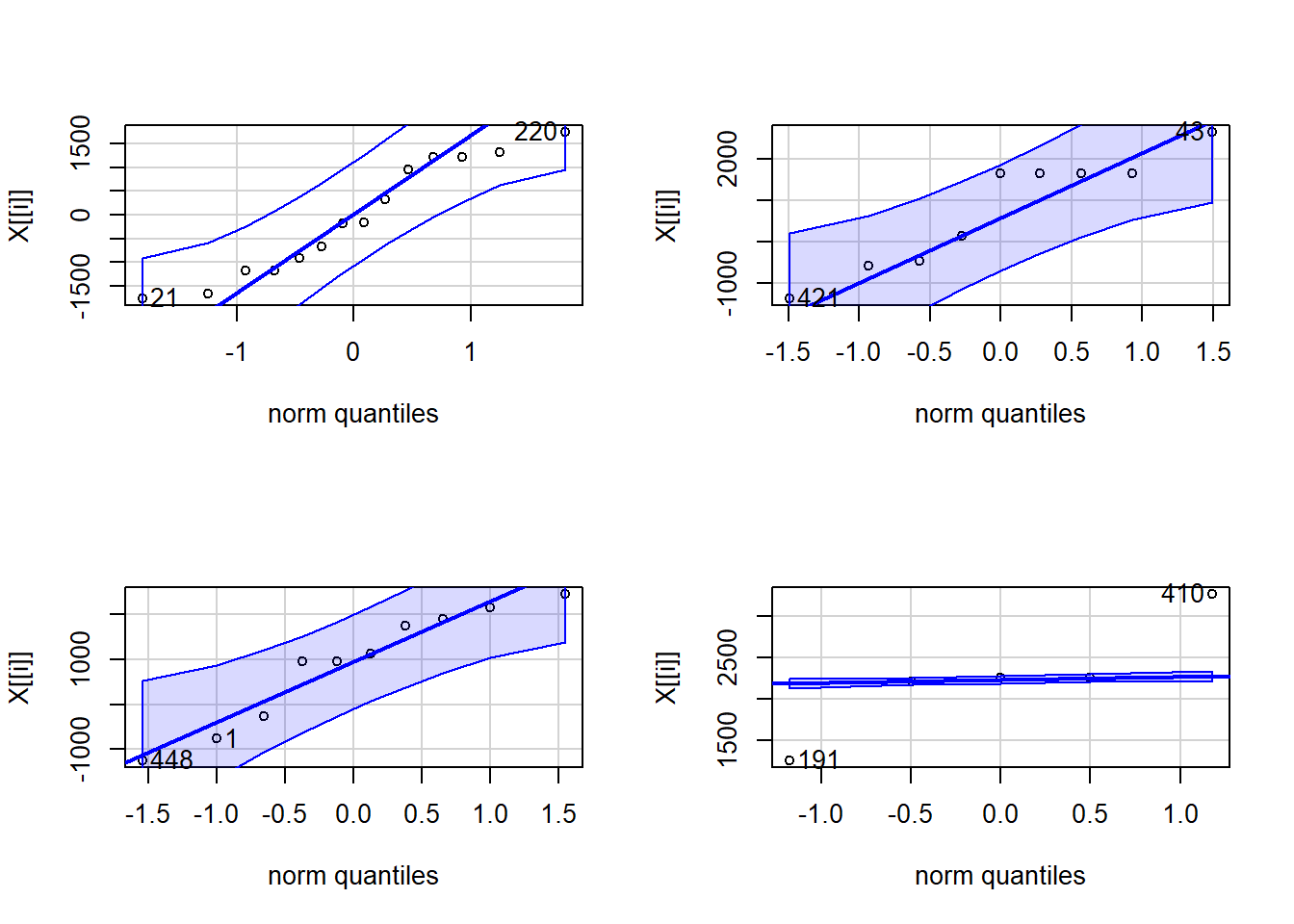
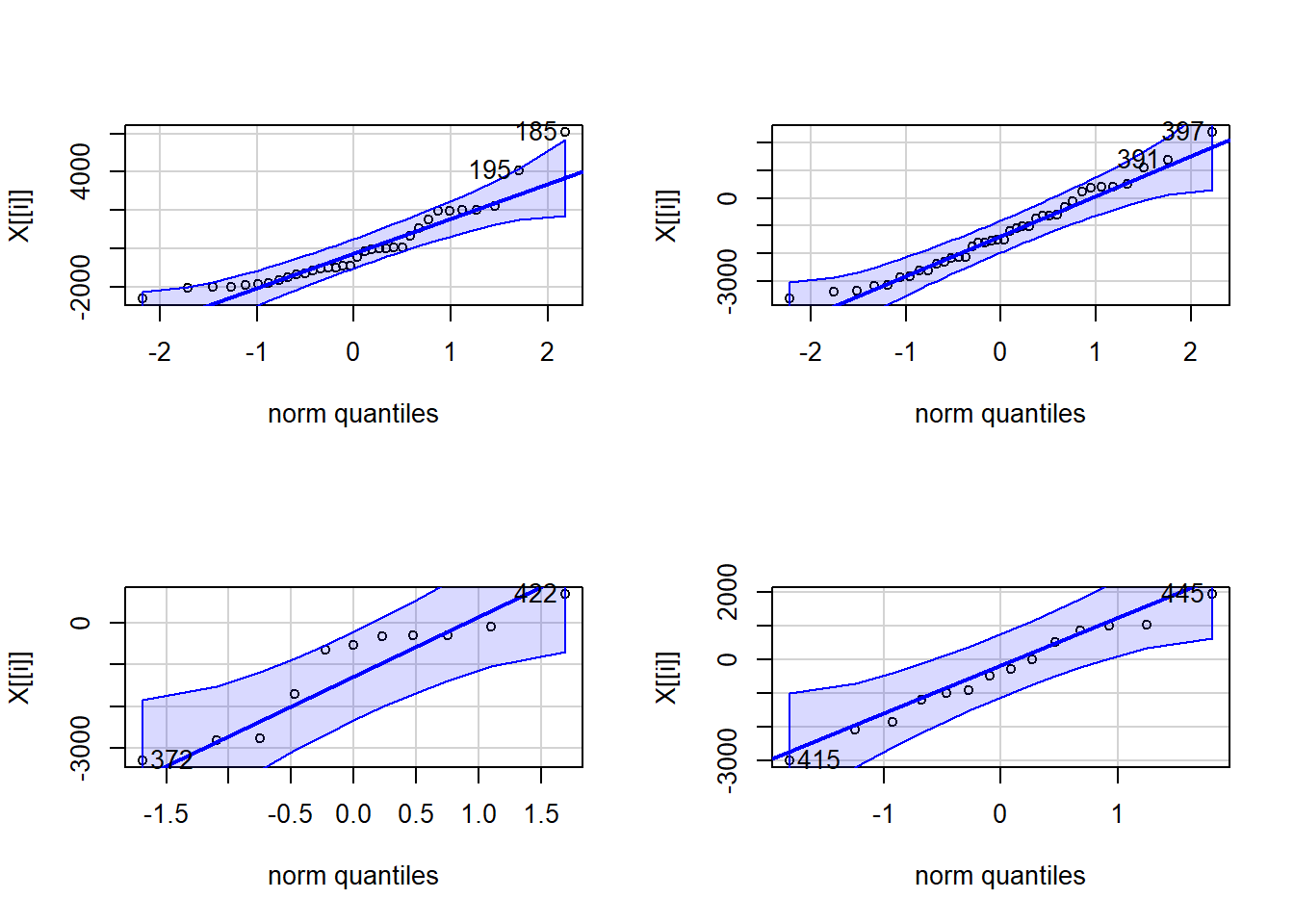
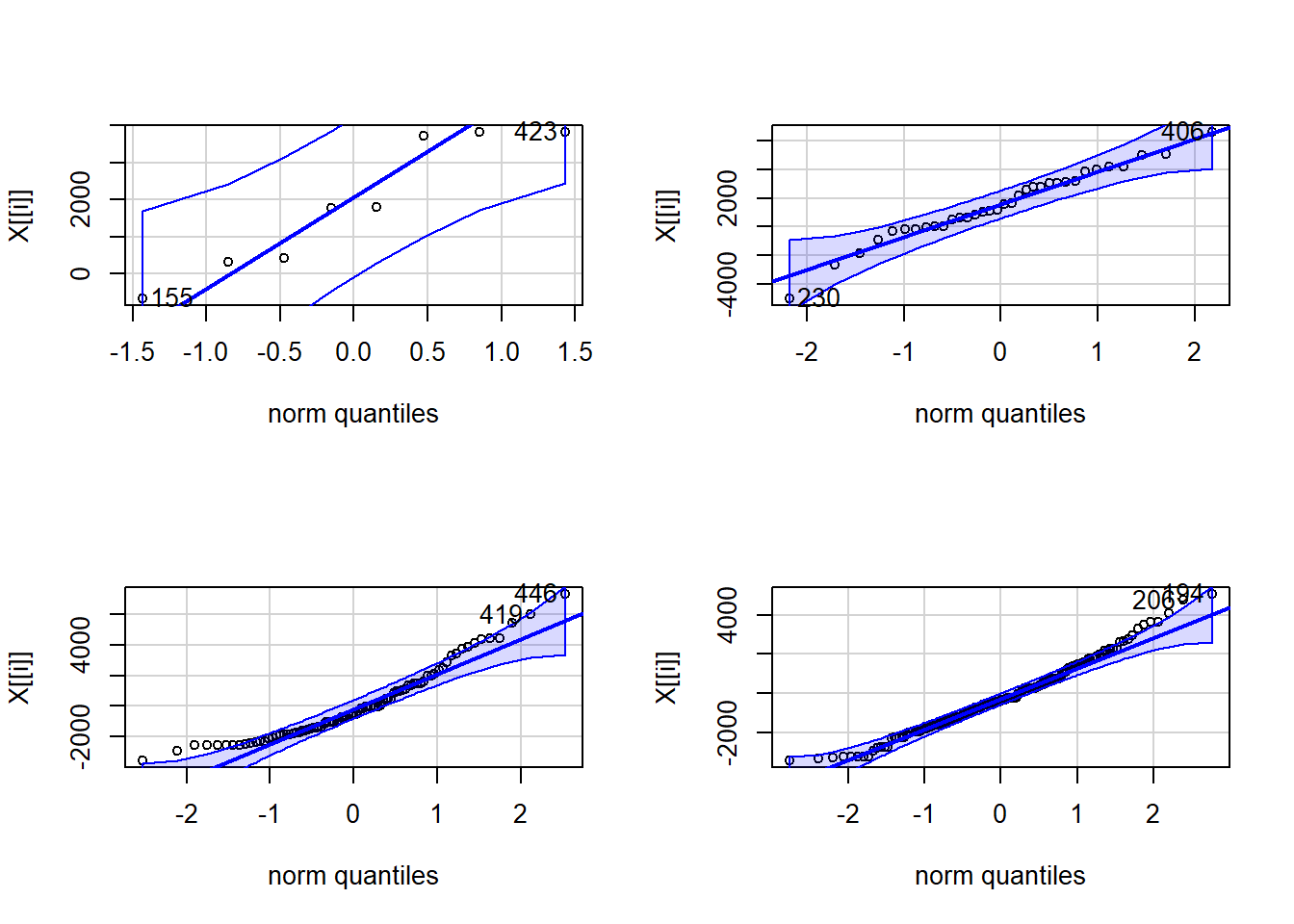


# Not required for this assignment

library(car)

par(mfrow=c(2,2)) #combined the plots for better output

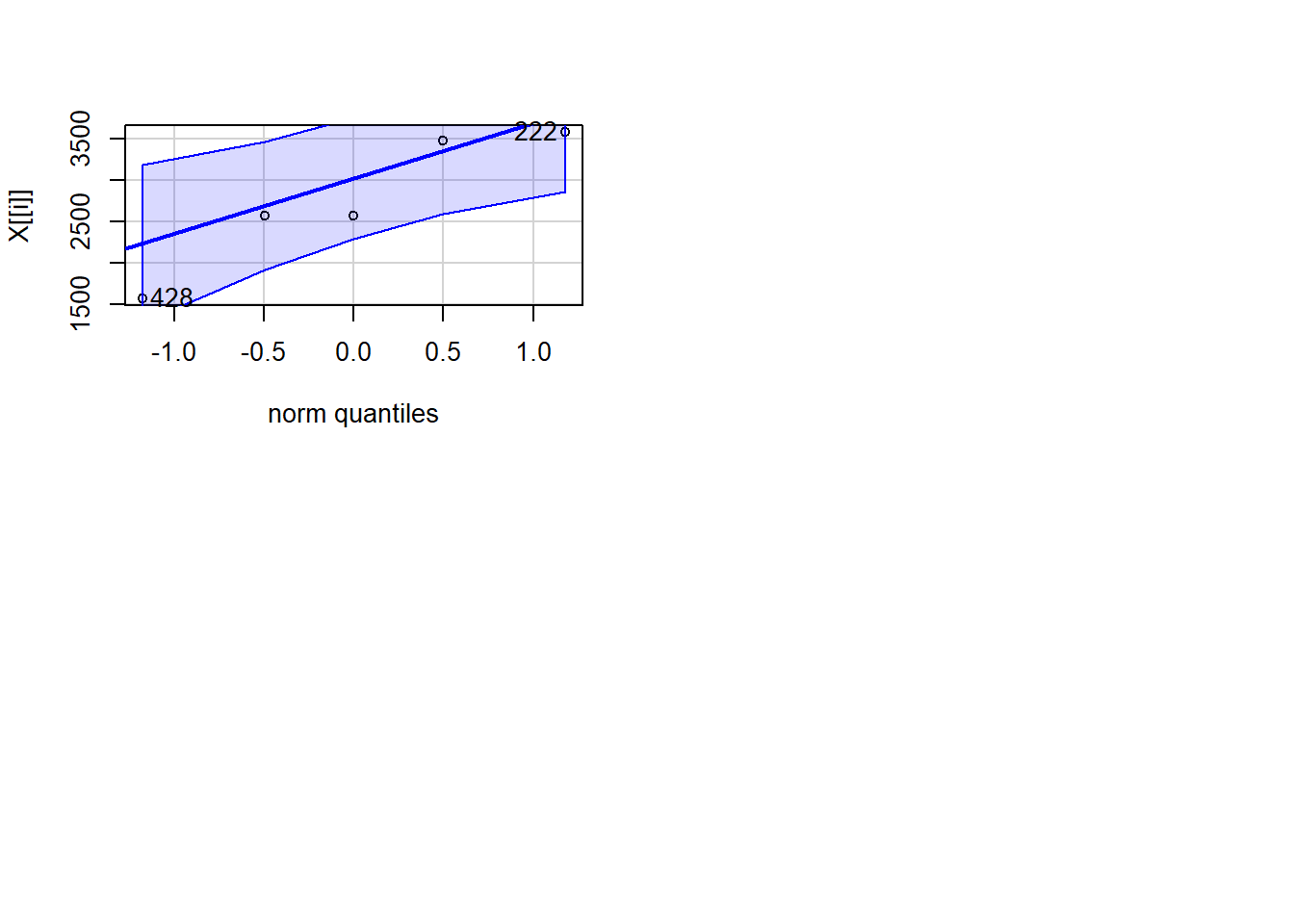
sapply(split(modq1$residuals, MyCars$Age), qqPlot)



## 0 1 2 3 4 5 6 7 8 9 10 11 12

## 155 2 18 87 86 21 33 9 13 9 9 10 4 5

## 423 7 31 80 88 22 32 11 14 3 2 1 2 3



1. Find the car in your sample with the largest residual (in magnitude - positive or negative). For that car, find its standardized and studentized residual. Based on these residuals, could this value be considered influential?

**0.5 points** largest absolute residual  
**0.5 points** standarized residual  
**0.5 points** studentized residual  
**0.5 points** discussion of influence. This will likely be based on the similarity or difference in the standardized and studentized residual. If they are similar, that point likely has little influence. They should not earn points for saying the point is an outlier, with a standardized/studentized residual over 2 or 3, and is hence influential. This is not always true. For my values, the standardized and studentized values are similar for each index, so it does not seem they that have much influence on the model. They may also find the cooks distance and compare the value to 0.5 or 1 to determine if it has a large influence.

which.max(abs(modq1$resid))

## 446

## 446

rstandard(modq1)[446]

## 446

## 3.729341

rstudent(modq1)[446]

## 446

## 3.784635

cooks.distance(modq1)[446]

## 446

## 0.02335437

1. Determine the leverages for the cars with the ten largest absolute residuals. What do these leverage values say about the potential for each of these ten cars to be influential on your model?

**1.0 point** Determine leverages  
**1.0 point** Discuss potential influence from leverage. Should compare to 2(2/sample size) and 3(2/sample). If the leverage of their value is more than double or triple the average leverage, then the point has higher potential to influence the model.

Note: Leverage measures potential for influence. This alone does not determine if the points actually have influence. They do not need to use the manner that I did to find the indices and check the leverages.

# I assume that there are better ways to do this...

# Extracts the indices for the 10 largest absolute residuals

top\_resid\_indices = sort(

abs(

modq1$resid

),

decreasing=TRUE,

index.return=TRUE

)$ix[1:10]

# Find the leverages for the points with the 10 largest absolute residuals

hatvalues(modq1)[top\_resid\_indices]

## 446 406 185 419 208 194

## 0.003347170 0.005085658 0.002276505 0.003347170 0.003347170 0.002410786

## 186 230 386 206

## 0.005085658 0.005085658 0.005085658 0.002410786

# Compares leverages to 2(2/n)

# Since no leverages are above this 2(2/n), the statement is false for all 10 points

# This make the line below sum to 0.

# This implies that these 10 points have low potential to influence the model

sum(hatvalues(modq1)[top\_resid\_indices] > (2 \* 2/dim(MyCars)[1]))

## [1] 0

1. Determine the Cook’s distances for the cars with the ten largest absolute residuals. What do these Cook’s distance values say about the influence of each of these ten cars on your model?

**0.5 point** Determine Cook’s distances  
**0.5 point** Discuss influence based on comparing Cook’s Distance values to 0.5 or 1.0

cooks.distance(modq1)[top\_resid\_indices]

## 446 406 185 419 208 194

## 0.023354365 0.028754590 0.010959237 0.015632738 0.013006155 0.008060807

## 186 230 386 206

## 0.016869283 0.016503209 0.016442434 0.007136331

# Since no Cook's distances are greater than 0.5, this sums to 0 (all FALSE values)

sum(cooks.distance(modq1)[top\_resid\_indices] > 0.5)

## [1] 0

1. Compute and interpret in context a 90% confidence interval for the slope of your regression line.

**0.5 point** code for interval  
**1.0 point** explanation - With 90% confidence the prices of this model car decrease between $1377.09 and $1247.60 as the age increases by one year

confint(modq1, level=.90)

## 5 % 95 %

## (Intercept) 17889.589 18454.264

## Age -1377.091 -1247.599

1. Test the strength of the linear relationship between your variables using each of the three methods (test for correlation, test for slope, ANOVA for regression). Include hypotheses for each test and your conclusions in the context of the problem.

**0.5 points** cor.test() code  
**0.5 points** cor.test() hypotheses and conclusion; Null: *ρ*

= 0, Alternative *ρ* ≠ 0. They can write this in notation or in words. Conclusion should note the p-value from the code, either explicitly stating its value or noting that it is very small, and then reject the null or fail to support the alternative either with those words explicitly or inferred.  
**0.5 points** slope test code (likely just summary())  
**0.5 points** slope test hypotheses and conclusion; Same as cor.test, but with *β* instead of *ρ*

**0.5 points** anova() code (or anova455() or from summary())  
**0.5 points** anova hypotheses and conclusion; same as slope test

**Note:** They should not mention significance or p-values in their hypotheses. The hypotheses have nothing to do with these things. The p-values will help support or fail to support the hypotheses. Give 0 points for any of the three hypotheses that do this. The hypotheses state the the parameter is equal to 0 vs it is not equal to 0.

cor.test(MyCars$Age, MyCars$Price)

##

## Pearson's product-moment correlation

##

## data: MyCars$Age and MyCars$Price

## t = -33.409, df = 446, p-value < 2.2e-16

## alternative hypothesis: true correlation is not equal to 0

## 95 percent confidence interval:

## -0.8698098 -0.8165831

## sample estimates:

## cor

## -0.8452806

summary(modq1)

##

## Call:

## lm(formula = Price ~ Age, data = MyCars)

##

## Residuals:

## Min 1Q Median 3Q Max

## -4971.6 -1407.0 -250.2 1212.6 7302.8

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 18171.93 171.29 106.09 <2e-16 \*\*\*

## Age -1312.35 39.28 -33.41 <2e-16 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 1961 on 446 degrees of freedom

## Multiple R-squared: 0.7145, Adjusted R-squared: 0.7139

## F-statistic: 1116 on 1 and 446 DF, p-value: < 2.2e-16

anova(modq1)

|  |
| --- |
|  |

|  | **Df**  **<int>** | **Sum Sq**  **<dbl>** | **Mean Sq**  **<dbl>** | **F value**  **<dbl>** | **Pr(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- |
| Age | 1 | 4294333596 | 4294333596 | 1116.167 | 1.778551e-123 |
| Residuals | 446 | 1715936839 | 3847392 | NA | NA |

2 rows

# May also use anova455

source("https://raw.githubusercontent.com/JA-McLean/STOR455/master/scripts/anova455.R")

anova455(modq1)

|  |
| --- |
|  |

|  | **Df**  **<dbl>** | **Sum Sq**  **<dbl>** | **Mean Sq**  **<dbl>** | **F value**  **<dbl>** | **P(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- |
| Model | 1 | 4294333596 | 4294333596 | 1116.167 | 0 |
| Error | 446 | 1715936839 | 3847392 | NA | NA |
| Total | 447 | 6010270436 | NA | NA | NA |

3 rows

1. Suppose that you are interested in purchasing a car of this model that is four years old (in 2017). Determine each of the following: 90% confidence interval for the mean price at this age and 90% prediction interval for the price of an individual car at this age. Write sentences that carefully interpret each of the intervals (in terms of car prices).

**1.0 points** - Construct dataframe with single car of age 4. Give 0 points if they instead take a subset from their sample of all cars that are of age 3.  
**1.0 points** - code for confidence interval.  
**1.0 points** - code for predcition interval.  
**0.5 points** - sentence interpreting the confidence interval, such as with 90% confidence I predict that the mean price of all 4 year old Civics sold in NY is between $12768.29 and $13076.8  
**0.5 points** - sentence interpreting the prediction interval, such as with 90% confidence I predict that the price of a 4 year old Civic sold in NY is between $9685.81 and $16159.28

Note: They don’t need sentences with this exact wording, but it should be clear that the confidence interval is a prediction for the mean price of all cars like this, while the prediction interval is predicting the price of one specific car. It should also be clear that there is 90% confidence in the process.

single\_car = data.frame(Age = 4)

predict.lm( modq1, single\_car, interval = "confidence", level = 0.90)

## fit lwr upr

## 1 12922.55 12768.29 13076.8

predict.lm( modq1, single\_car, interval = "prediction", level = 0.90)

## fit lwr upr

## 1 12922.55 9685.81 16159.28

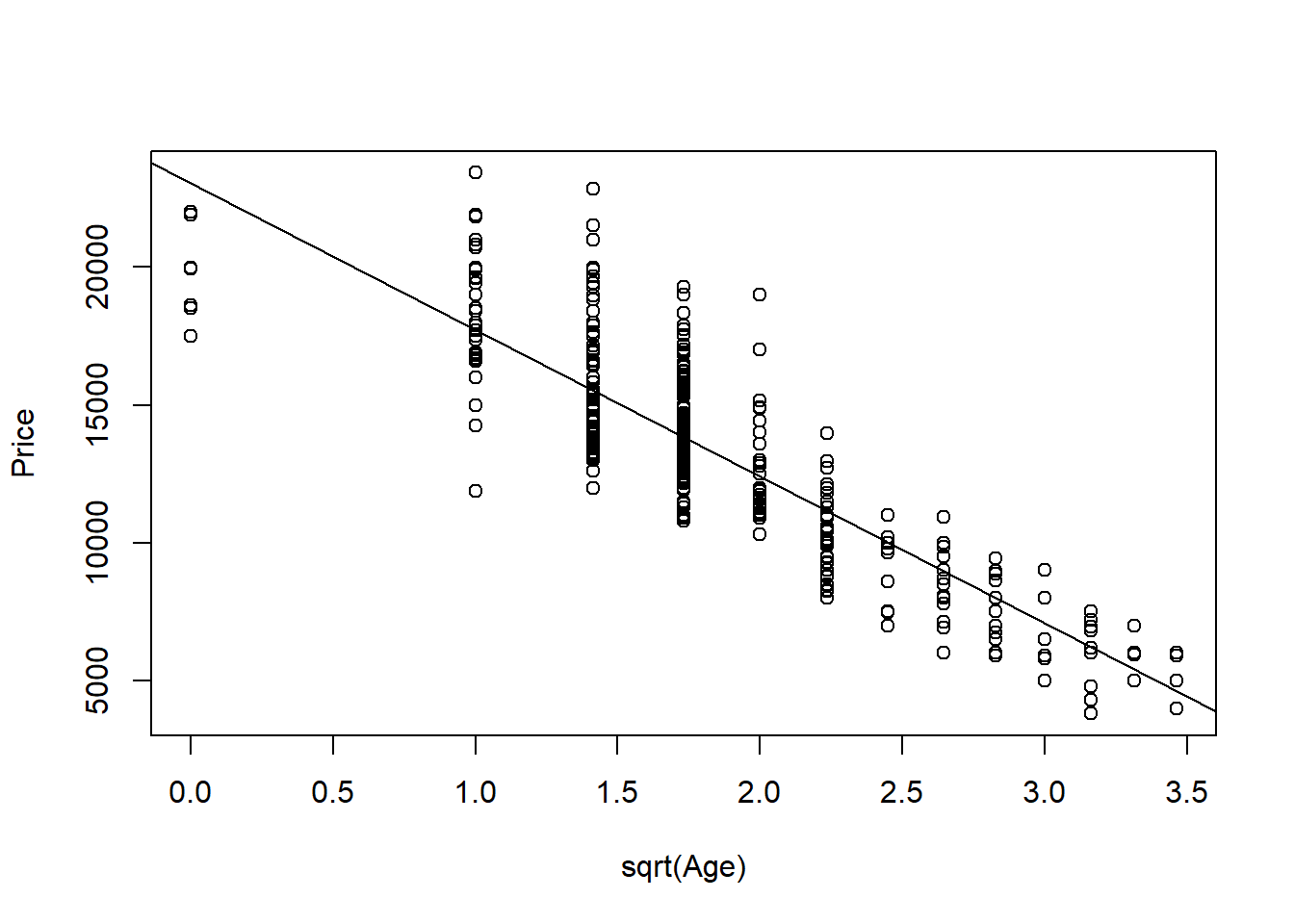
1. Experiment with some transformations to attempt to find one that seems to do a better job of satisfying the linear model conditions. Include the summary output for fitting that model and a scatterplot of the original data with this new model (which is likely a curve on the original data). Explain why you think that this transformation does or does not improve satisfying the linear model conditions.

**2 points** - transformation of some kind that tries to improve the model. It’s possible that for some students, no transformation is needed, but they should still show the attempt to improve the model.  
**1 points** - Discussion of how the transformed model improves at least one of the conditions for a linear model. Or, if no better transformed model is found, a discussion of how the transformation did not improve linear model conditions.  
**1 points** - Plot of the best model on the orginal axes.

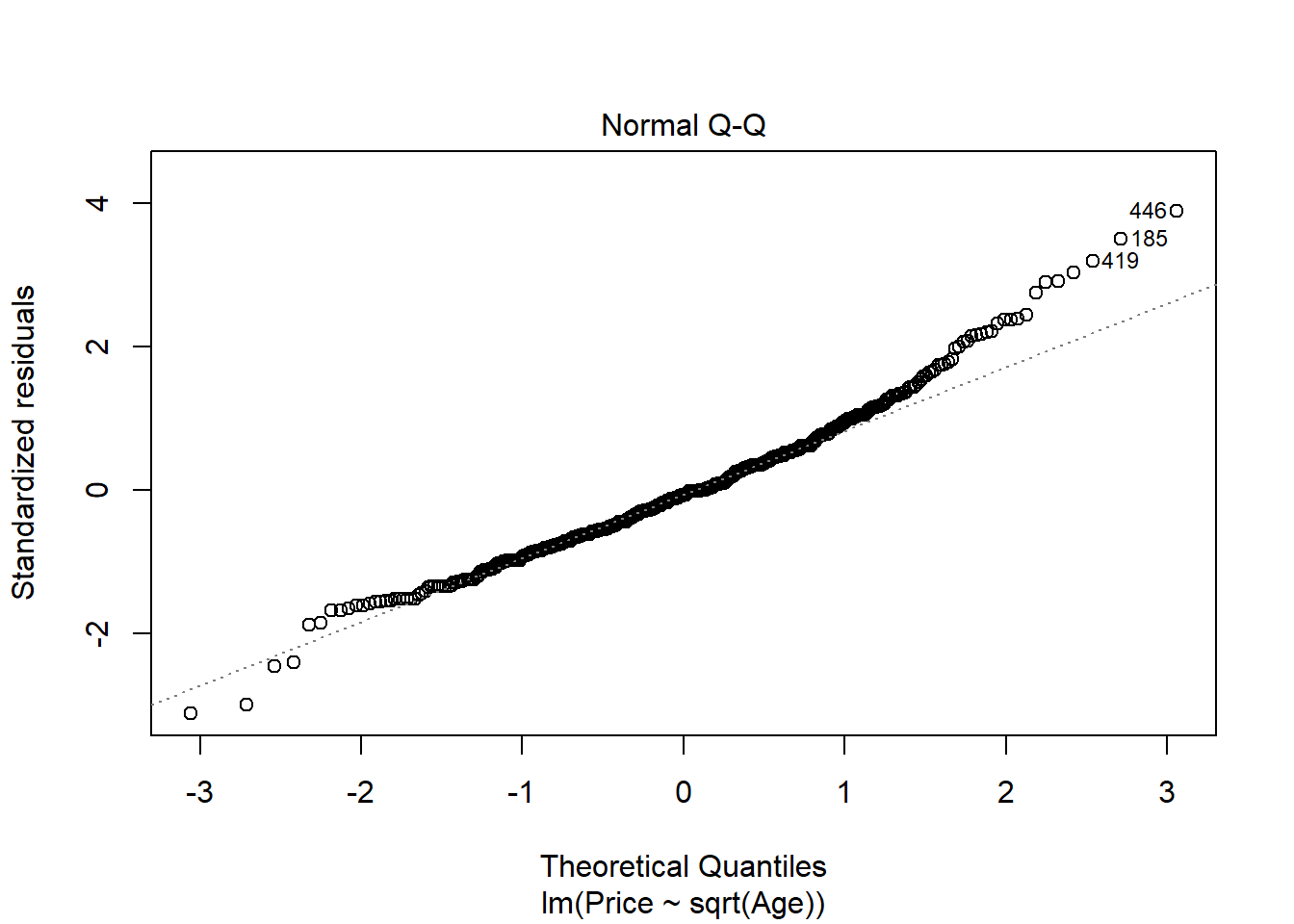
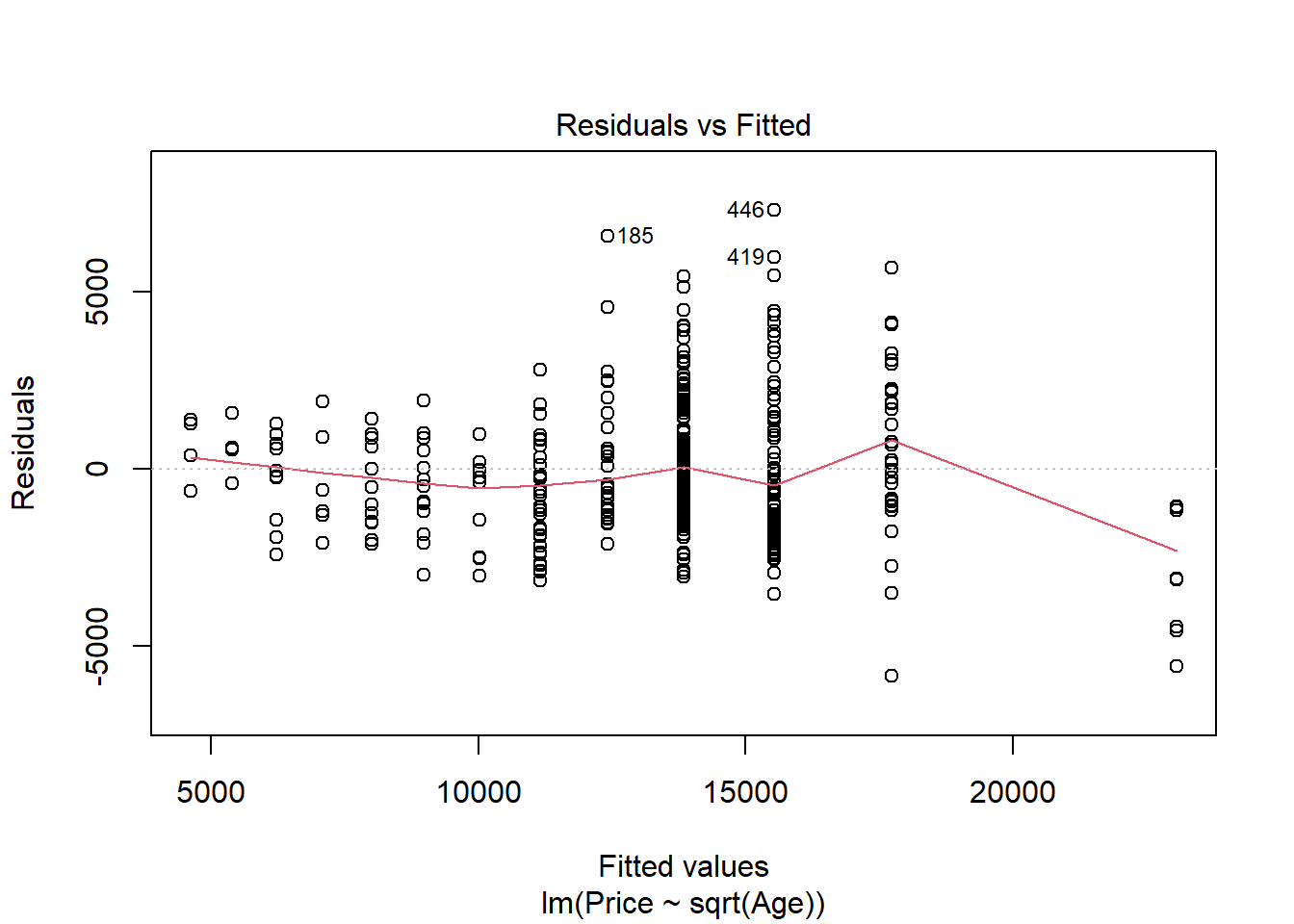
modq9.1 = lm(Price~sqrt(Age), data=MyCars)

plot(Price~sqrt(Age), data=MyCars)

abline(modq9.1)



plot(modq9.1, 1:2)



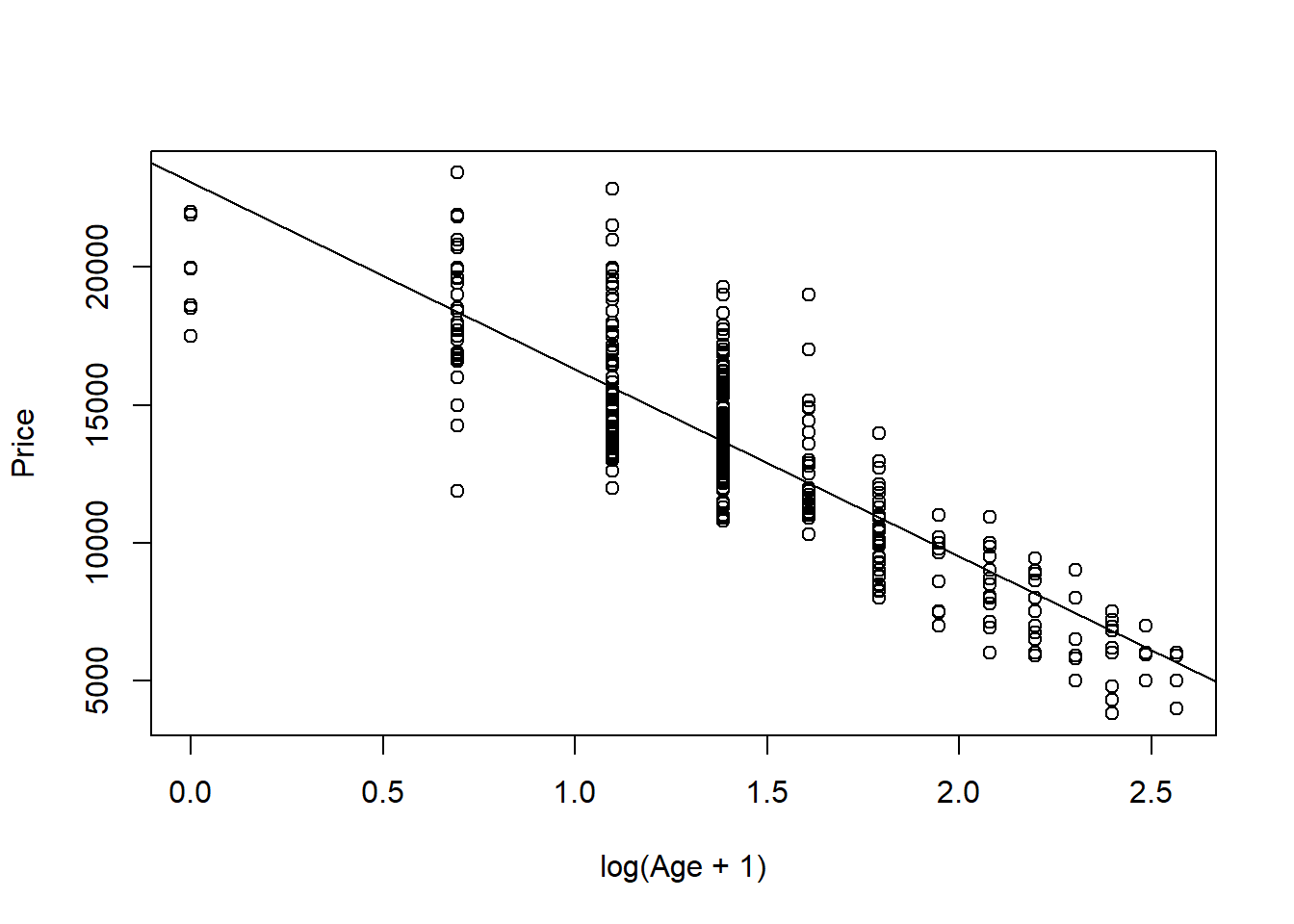
# Since there are 0s in the data, I added 1 to Age so I could use log

# I didn't do this in class, but some students may do this.

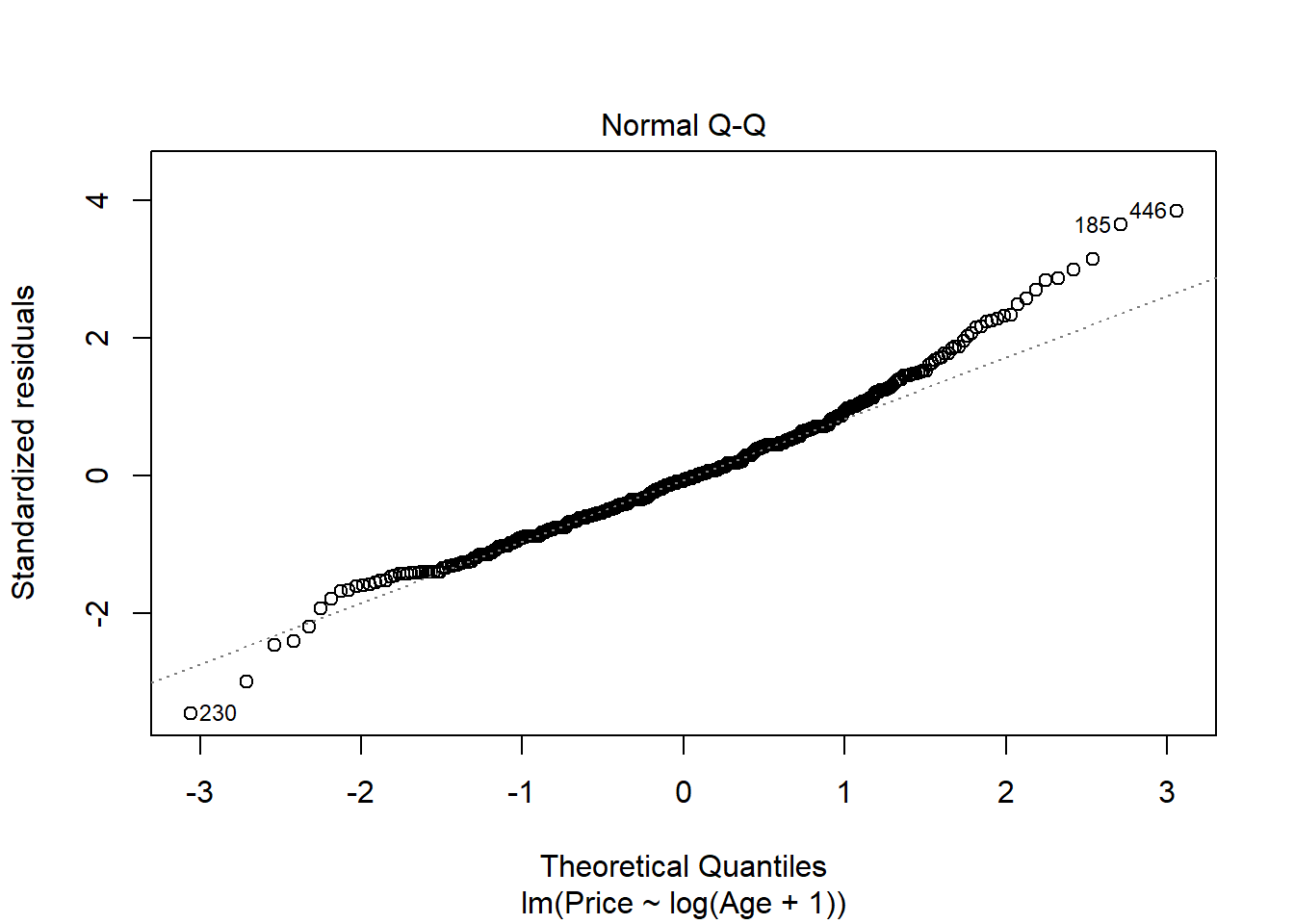
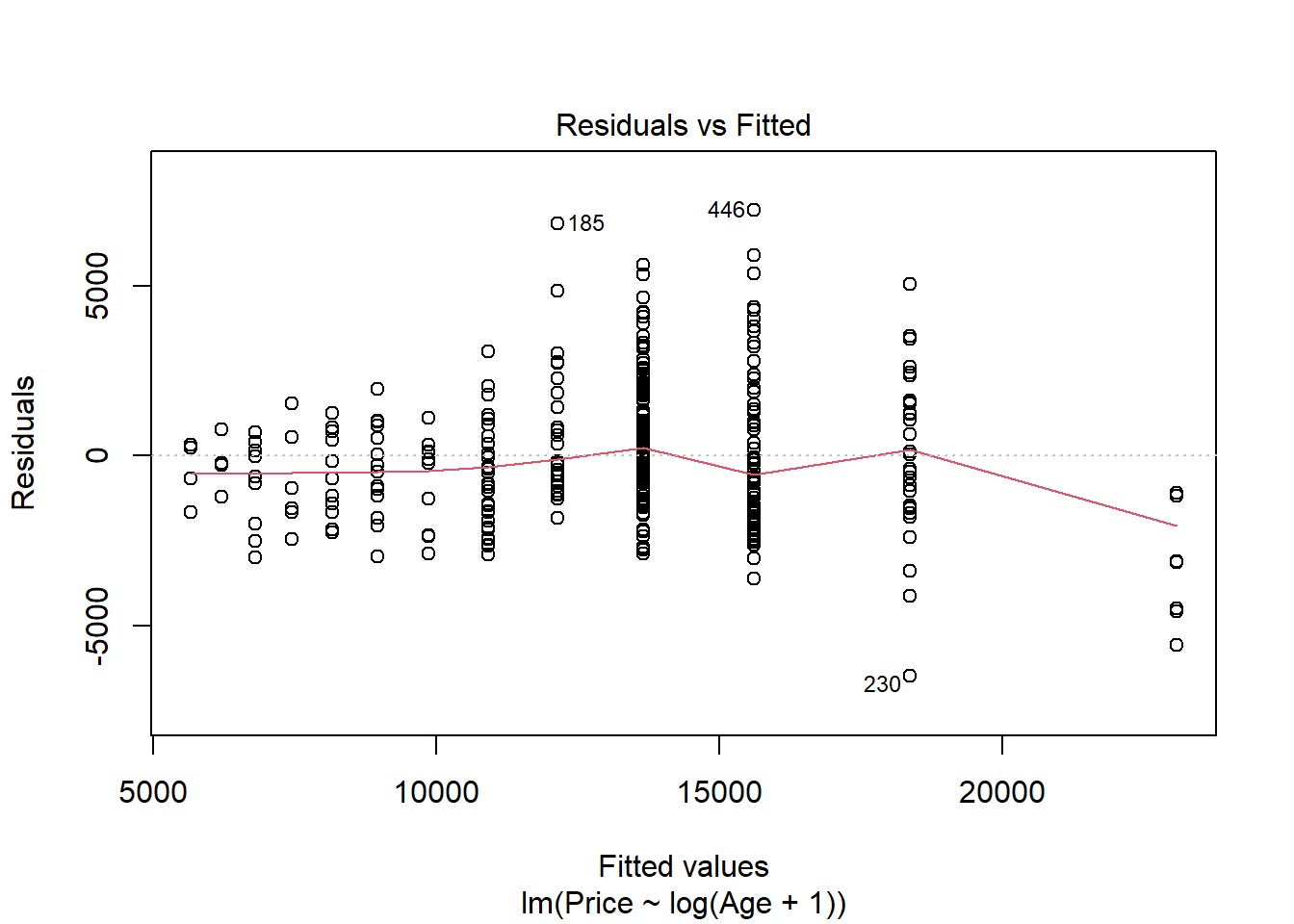
modq9.2 = lm(Price~log(Age+1), data=MyCars)

plot(Price~log(Age+1), data=MyCars)

abline(modq9.2)



plot(modq9.2, 1:2)

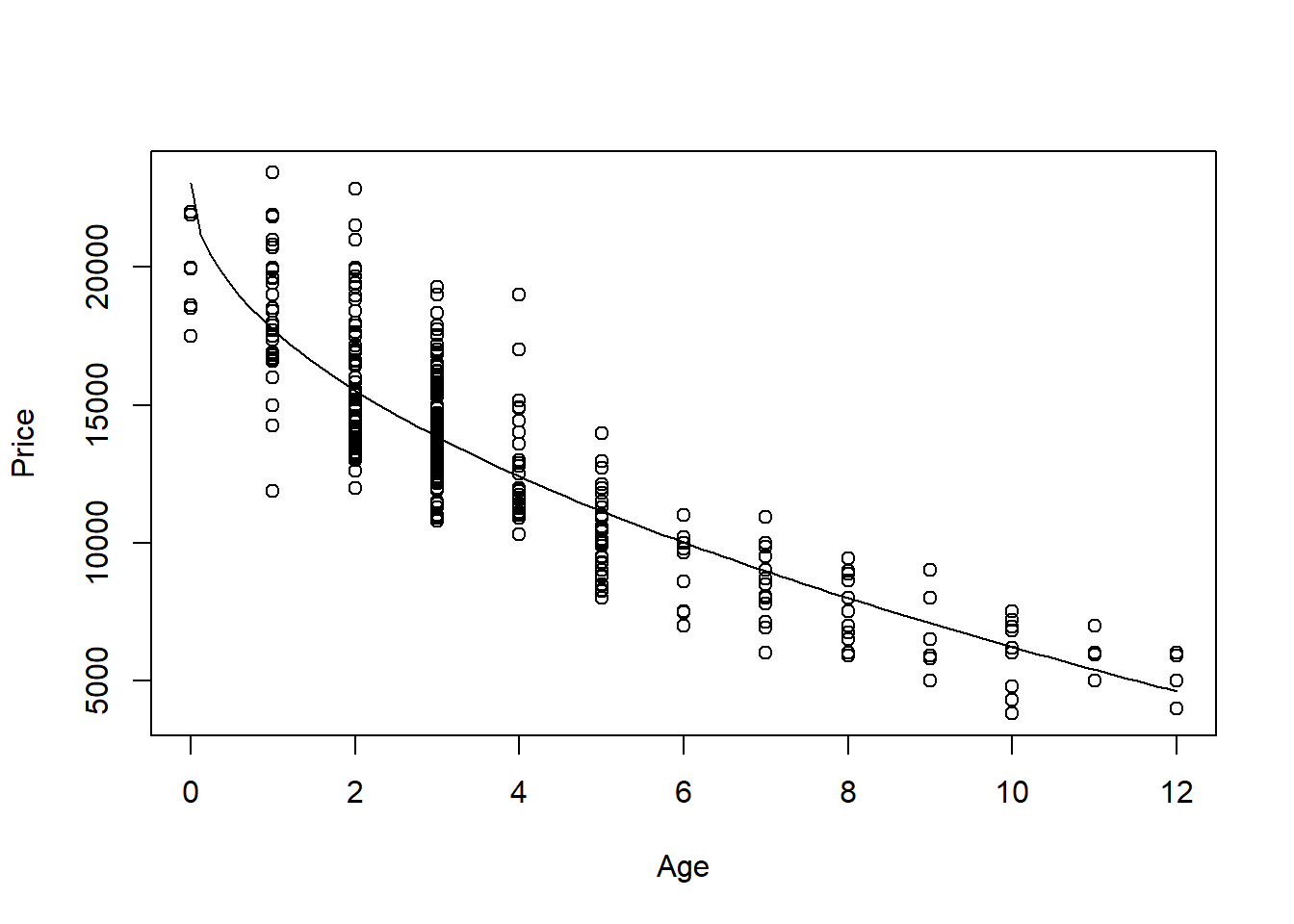


plot(Price~Age, data=MyCars)

B0 = summary(modq9.1)$coef[1,1]

B1 = summary(modq9.1)$coef[2,1]

curve(B0 + B1\*sqrt(x), add=TRUE)



1. According to your transformed model, is there an age at which the car should be free? If so, find this age and comment on what the “free car” phenomenon says about the appropriateness of your model.

**1.0 points** - Find (or approximate) at what age the car has a price of zero.  
**1.0 points** - Brief description of the meaning here. This could relate to the price that the car is worth $0 being well outside of the range of ages in the sample, which leads to extrapolation. The model may fit the given data, but not fit data for a wider range of cars. Or they may say somethng along the lines that you could sell the car up until this age, after that you’ll need to pay someone to take it away. Anything reasonable is fine.

Note: They will likely approximate this by manually solving for Price=0 in their linear model. Other methods such as approximating from the plot (shown below) or use of specific functions are fine as well. It’s possible that their models will never have a car worth $0, which is fine for full credit as well.

# For modq9.1 Car price is $0 at almost 19 years old

((-1\*B0)/B1)^2

## [1] 18.74863

1. Again suppose that you are interested in purchasing a car of this model that is four years old (in 2017). Determine each of the following using your model constructed in question 9: 90% confidence interval for the mean price at this age and 90% prediction interval for the price of an individual car at this age. Write sentences that carefully interpret each of the intervals (in terms of car prices).

**1.0 points** - code for confidence interval.  
**1.0 points** - code for predcition interval.  
**0.5 points** - sentence interpreting the confidence interval, such as with 90% confidence I predict that the mean price of all 4 year old Civic sold in NY is between $12256.62 and 12562.88  
**0.5 points** - sentence interpreting the prediction interval, such as with 90% confidence I predict that the price of a 4 year old Civic sold in NY is between $9303.123 and 15516.37

Note: They don’t need sentences with this exact wording, but it should be clear that the confidence interval is a prediction for the mean price of all cars like this, while the prediction interval is predicting the price of one specific car. It should also be clear that there is 90% confidence in the process. For my transformed model, I only transformed the predictor. This left my intervals below in dollars. If students used a transformation that transformed the response, then the intervals below will be in terms of transformed dollars. They must “untransform” the intervals to be back in dollars to earn full credit.

predict.lm( modq9.1, single\_car, interval = "confidence", level = 0.90)

## fit lwr upr

## 1 12409.75 12256.62 12562.88

predict.lm( modq9.1, single\_car, interval = "prediction", level = 0.90)

## fit lwr upr

## 1 12409.75 9303.123 15516.37

**MODEL #2: Use Mileage as a predictor for Price**

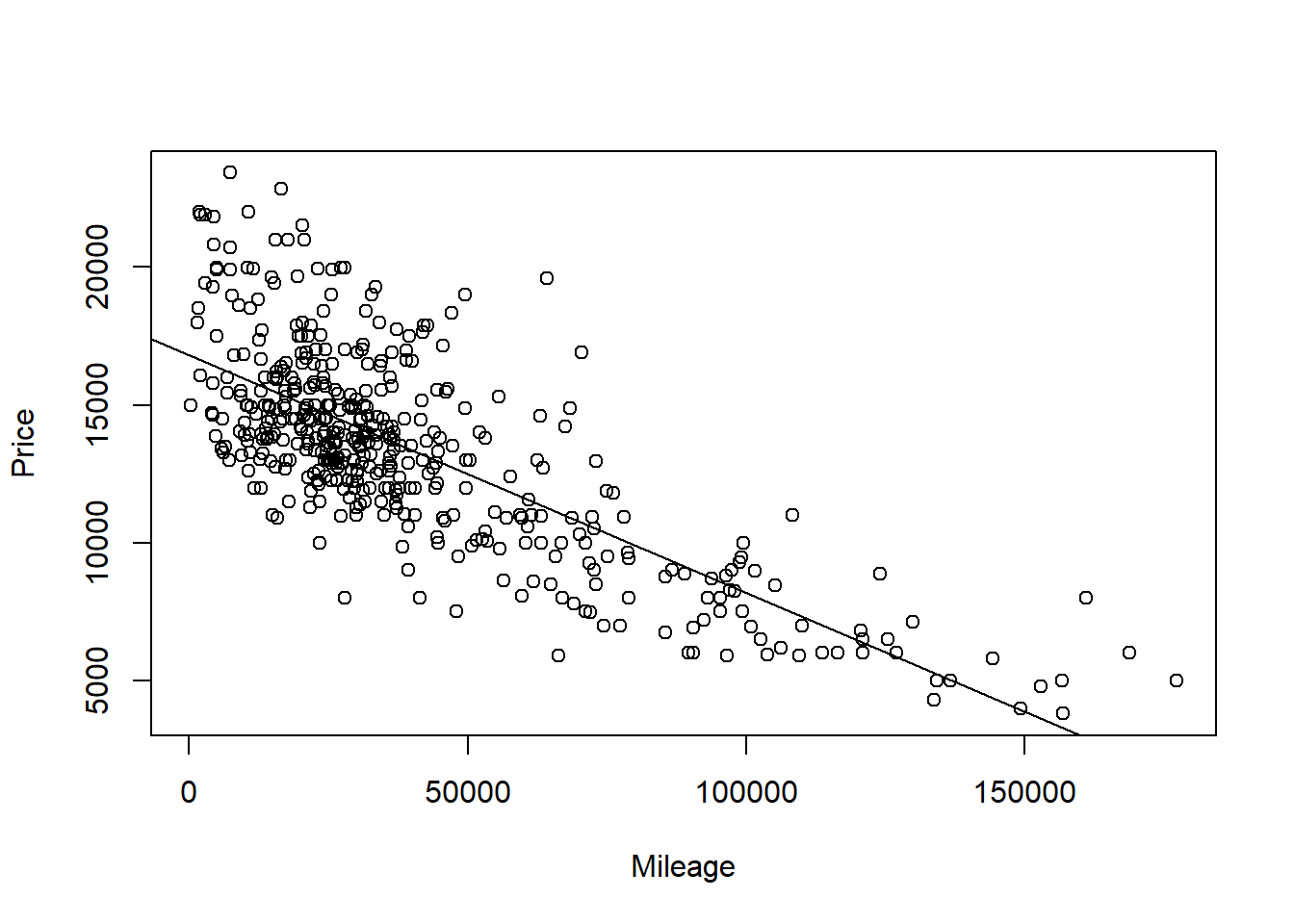
1. Calculate the least squares regression line that best fits your data (with *Mileage* now as the predictor) and produce a scatterplot of the relationship with the regression line on it.

**0.5 point** code for model  
**0.5 point** code for plot  
**0.5 point** abline

modq12 = lm(Price~Mileage, data=MyCars)

plot(Price~Mileage, data=MyCars)

abline(modq12)



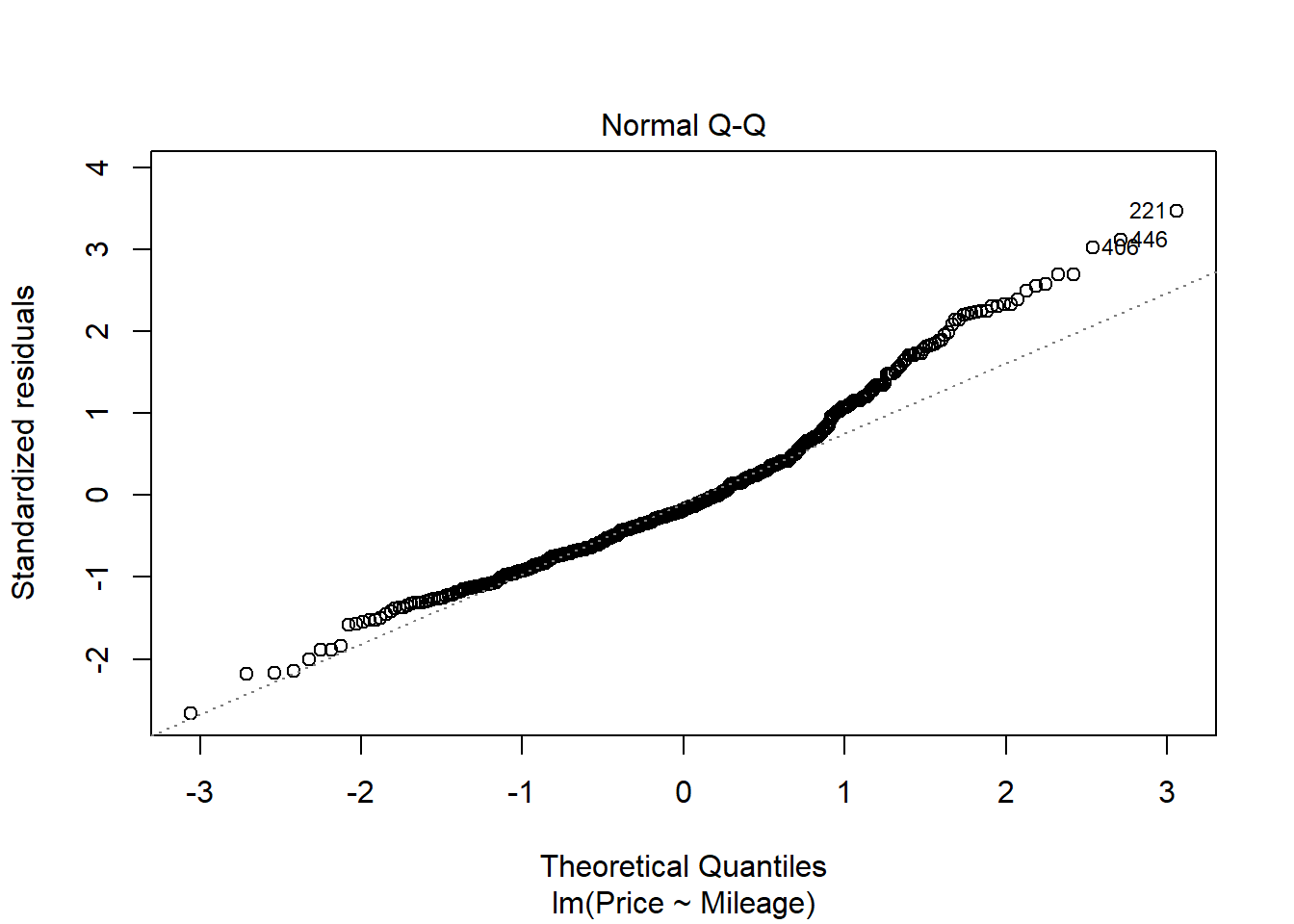
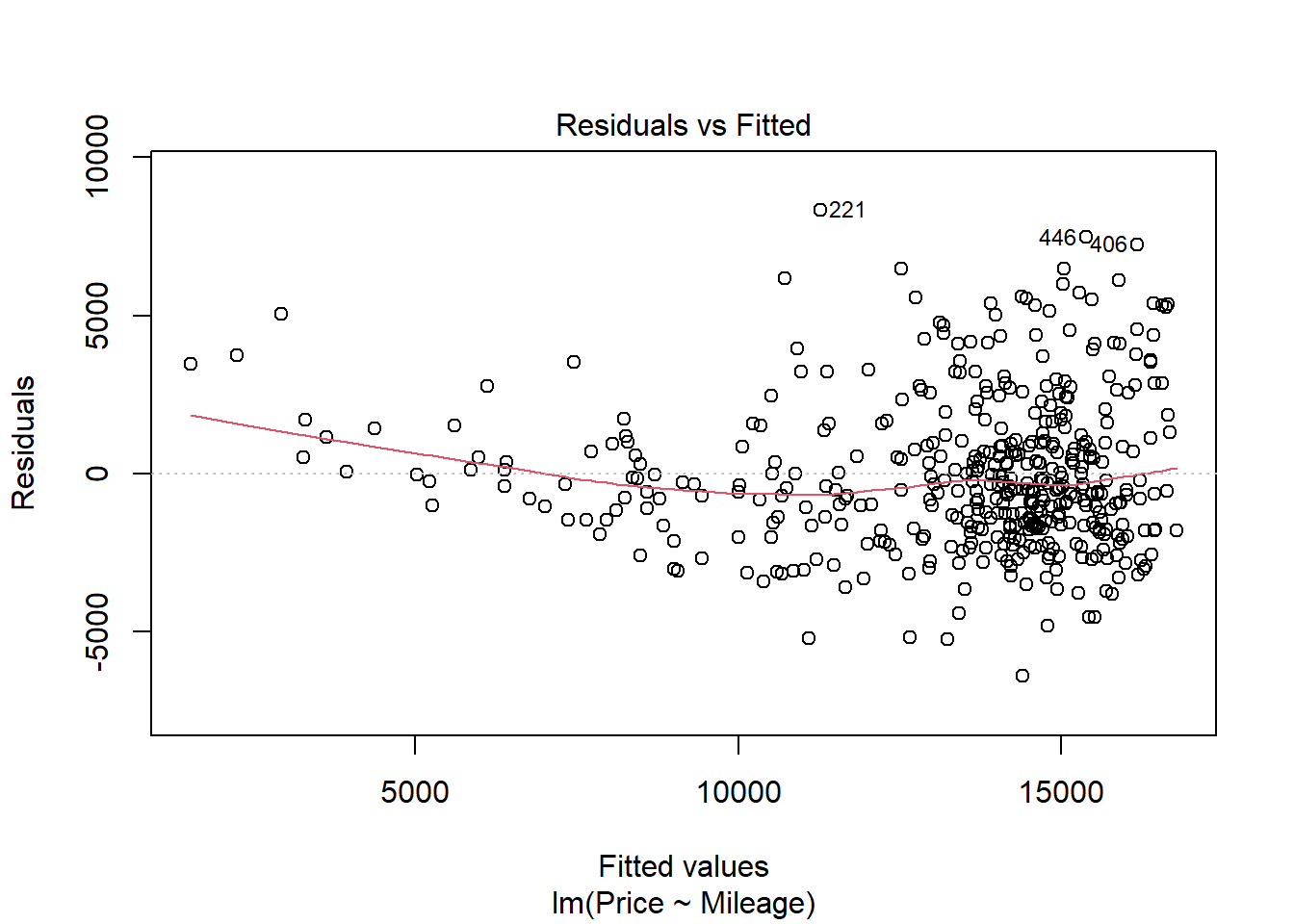
1. Produce appropriate residual plots and comment on how well your data appear to fit the conditions for a simple linear model. Don’t worry about doing transformations at this point if there are problems with the conditions.

**0.5 points** residuals vs fitted plot  
**0.5 points** qqnorm (or histogram) for Normality of residuals  
**1.5 points** discussion of conditions (linearity, constant variance, and normality of residuals) You can give 0.5 pts each. They can describe the conditions without explicitly using these terms.

Note 1: For linearity, they should have some discussion if the line seems to describe the data, using either the scatter plot or residual vs fitted plot. For constant variance they should discuss if the variability (vertical distances) from the line seems to follow any pattern as value of the predictor changes. For normality of the residuals, they should note the adherence of the residuals (or not) to the qqline, or bell curve shape or skew in a histogram. As each student will have a different plot, any reasonable assertions of the conditions being met (or not) supported by an argument is fine for full credit.

Note 2: they may use the plot(model) to produce all of the plots at once, or separately produce each of the plots with different lines of code.

plot(modq12, 1:2)



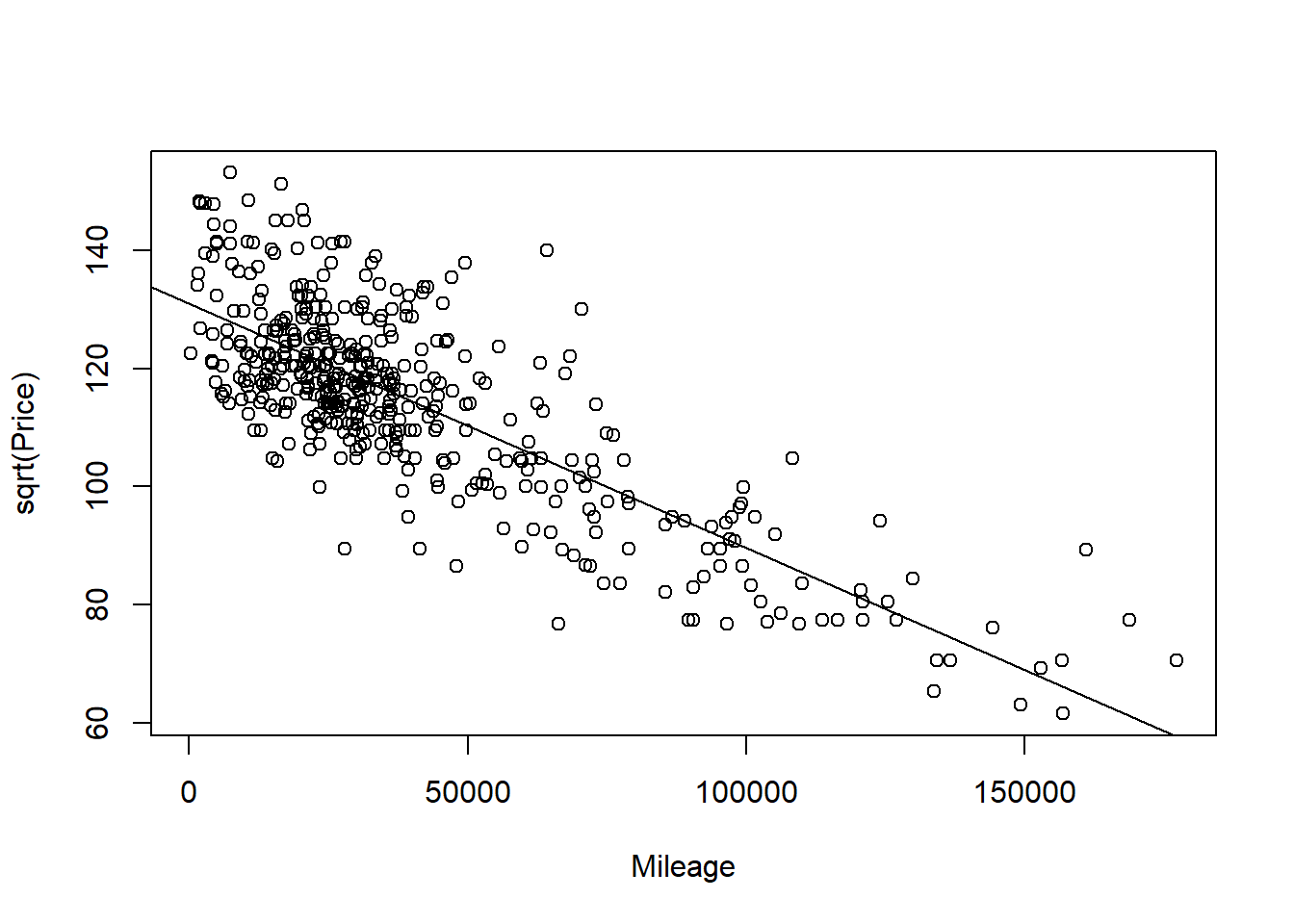
1. Experiment with some transformations to attempt to find one that seems to do a better job of satisfying the linear model conditions. Include the summary output for fitting that model and a scatterplot of the original data with this new model (which is likely a curve on the original data). Explain why you think that this transformation does or does not improve satisfying the linear model conditions.

**2 points** - transformation of some kind that tries to improve the model. It’s possible that for some students, no transformation is needed, but they should still show the attempt to improve the model.  
**1 points** - Discussion of how the transformed model improves at least one of the conditions for a linear model. Or, if no better transformed model is found, a discussion of how the transformation did not improve linear model conditions.  
**1 points** - Plot of the best model on the orginal axes.

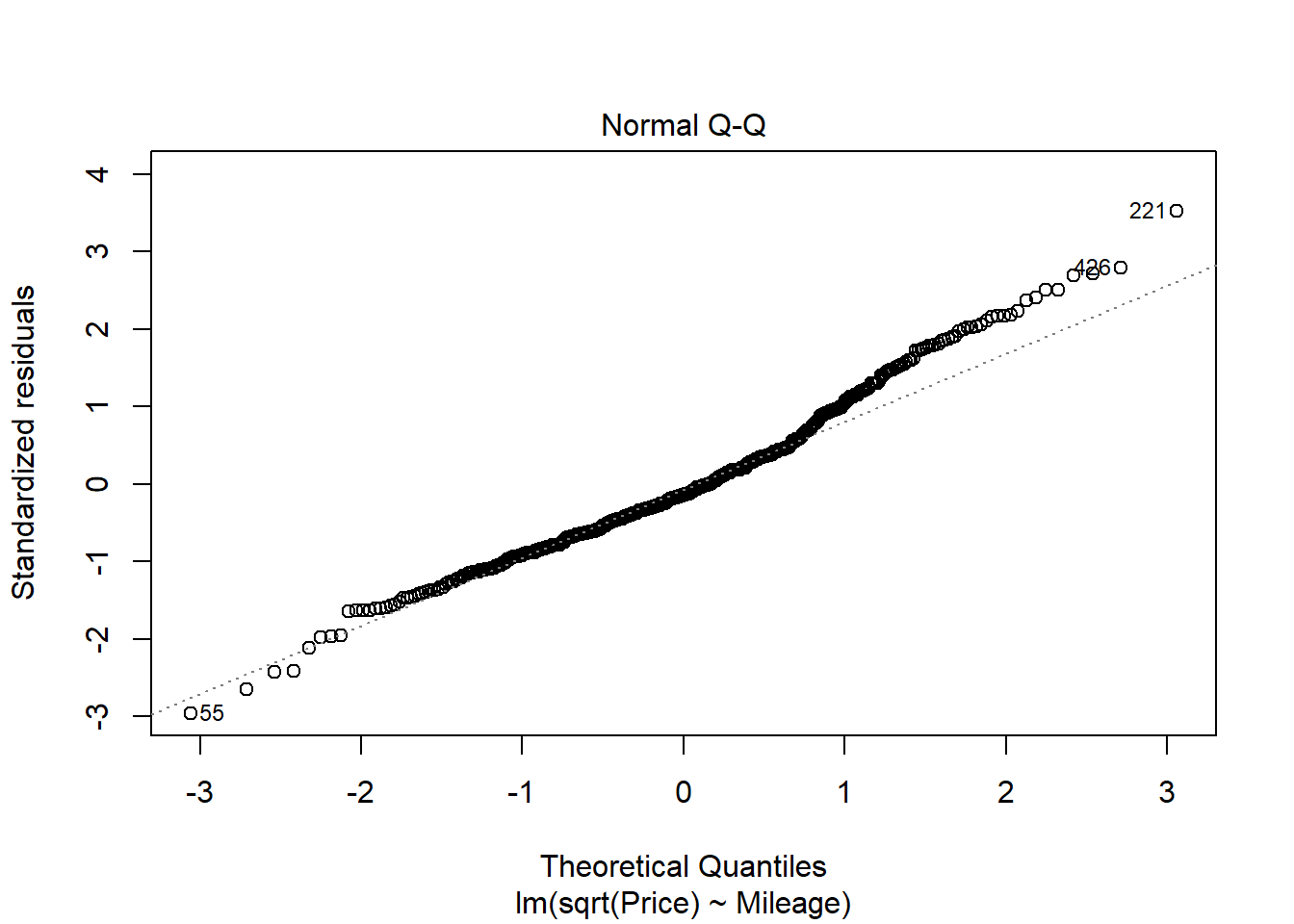
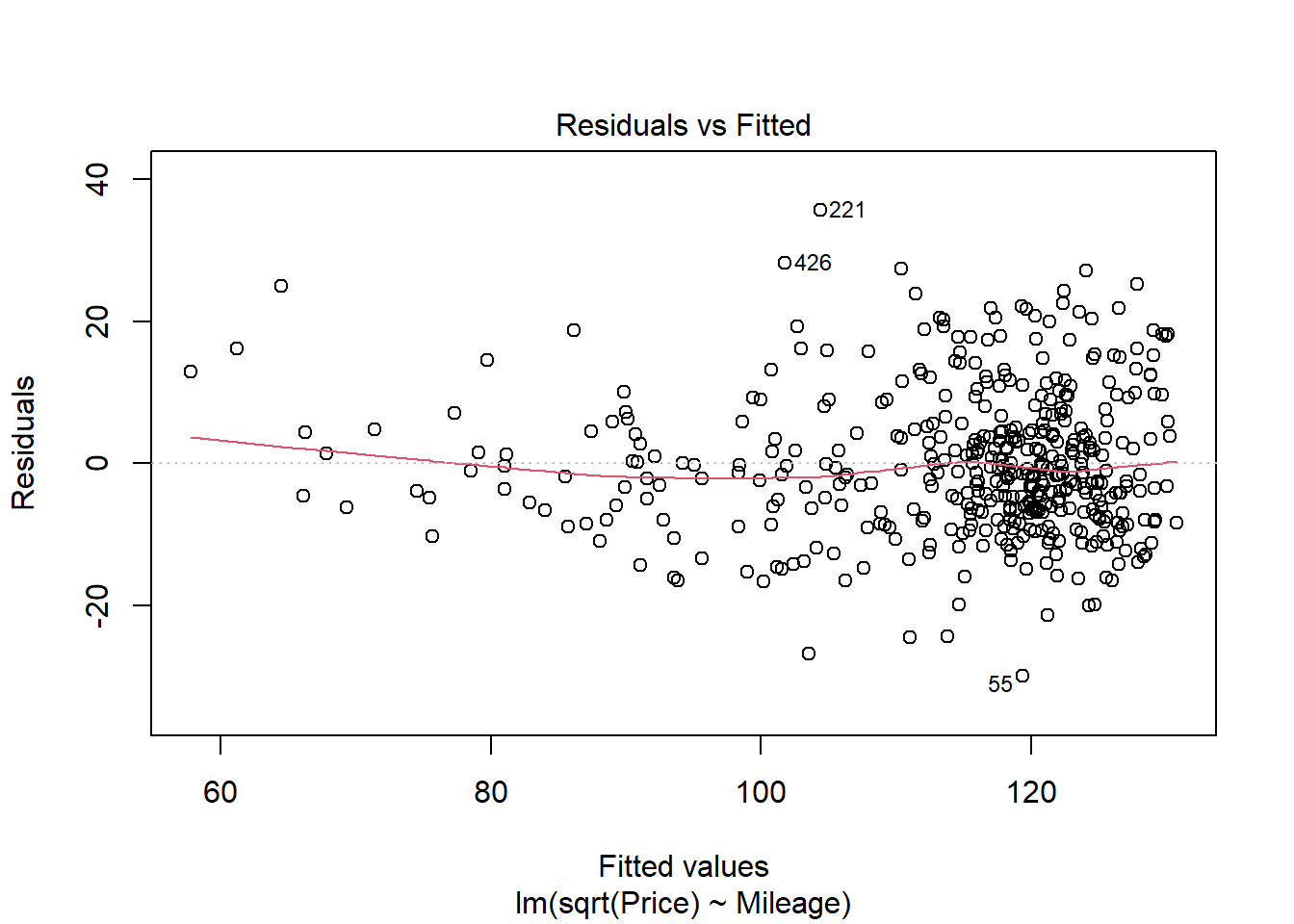
modq14.1 = lm(sqrt(Price)~Mileage, data=MyCars)

plot(sqrt(Price)~Mileage, data=MyCars)

abline(modq14.1)



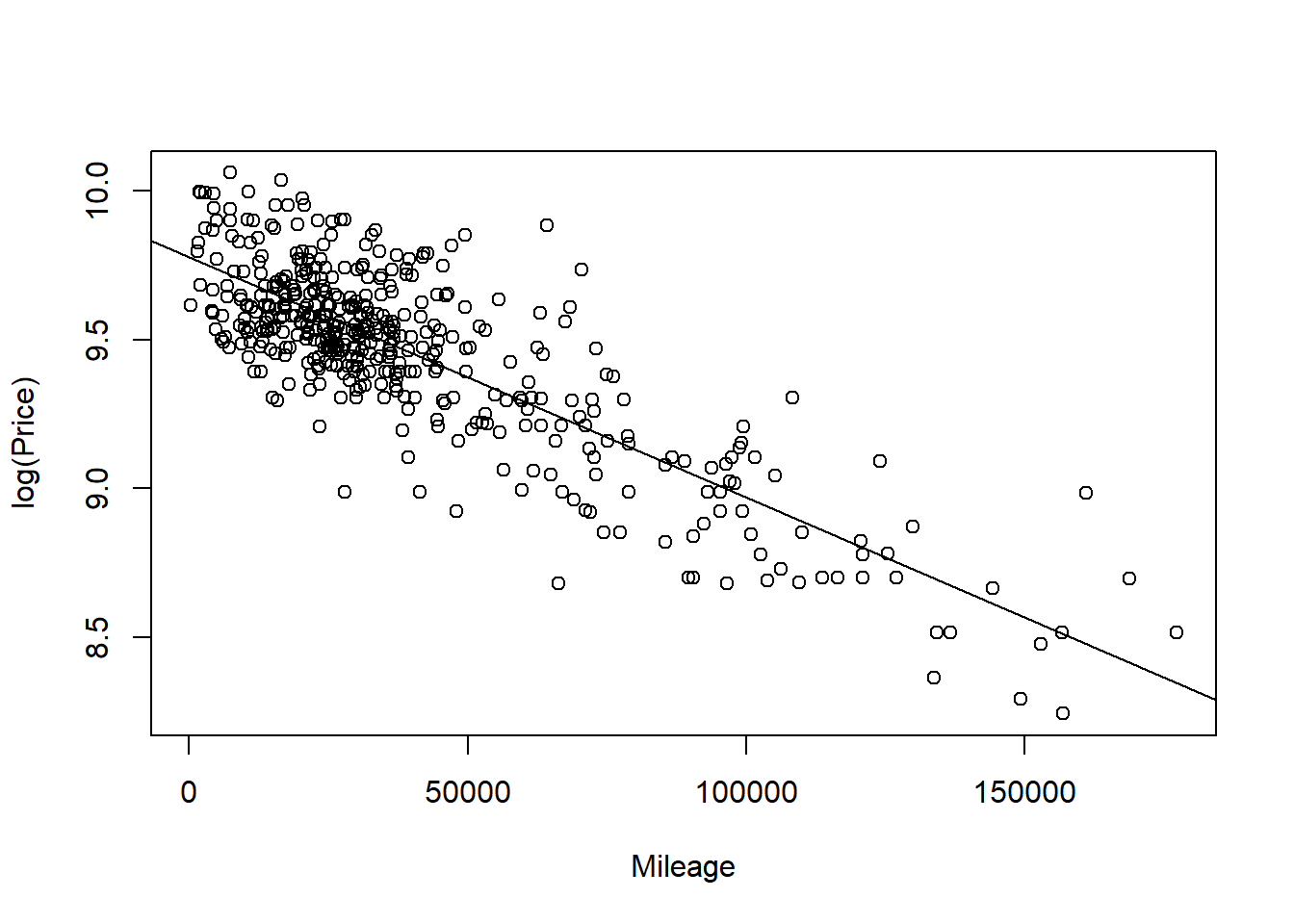
plot(modq14.1, 1:2)



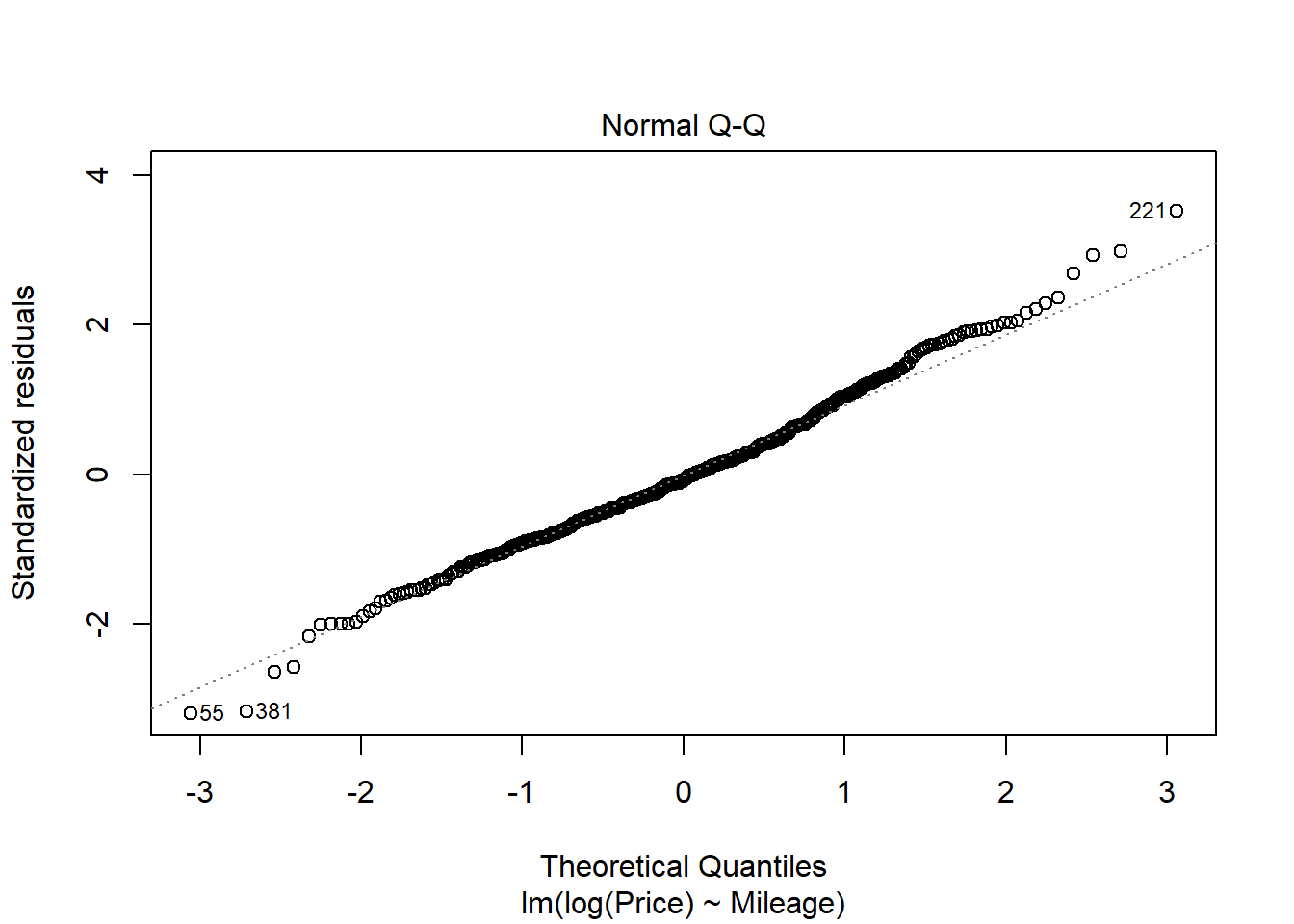
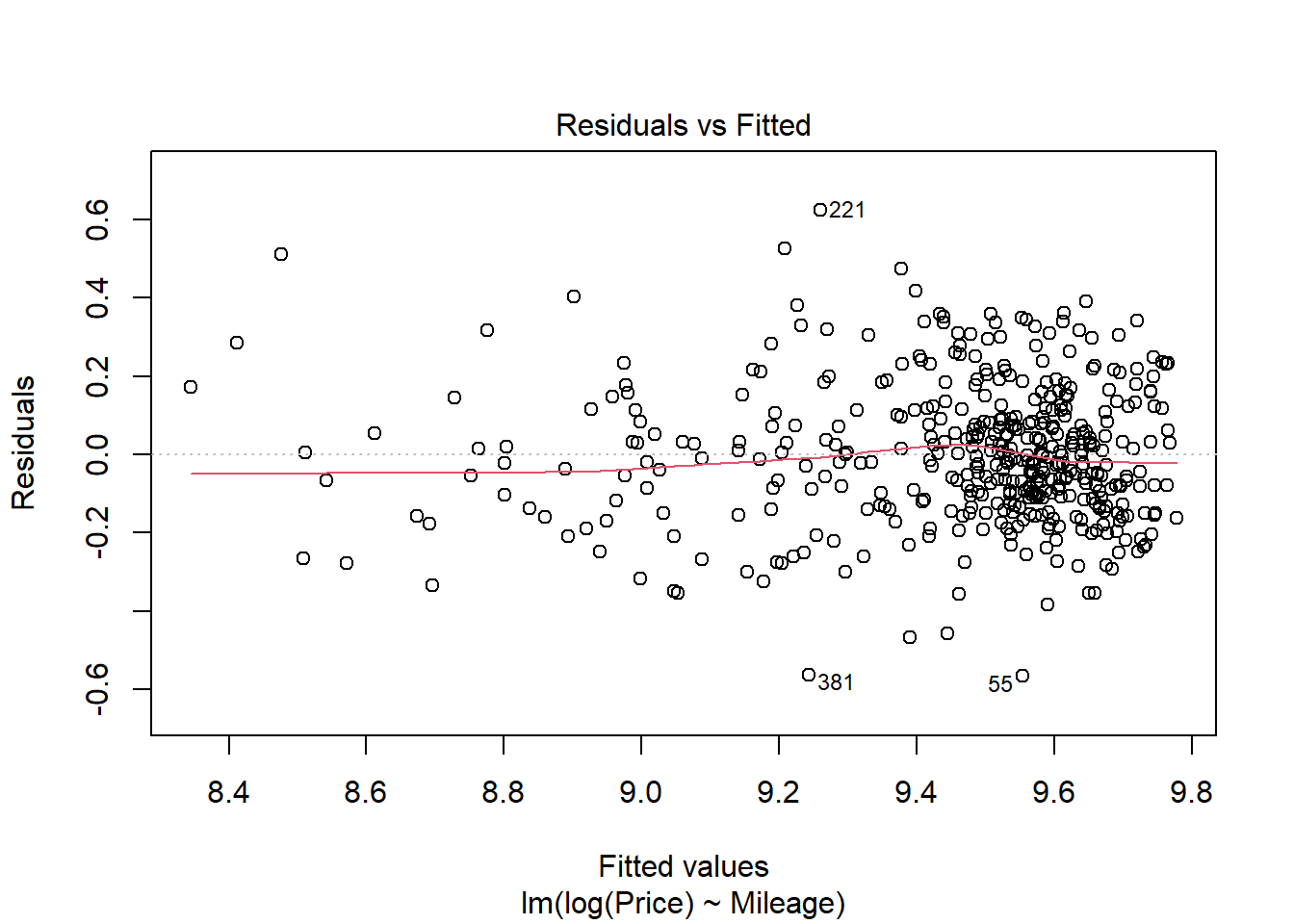
modq14.2 = lm(log(Price)~Mileage, data=MyCars)

plot(log(Price)~Mileage, data=MyCars)

abline(modq14.2)



plot(modq14.2, 1:2)

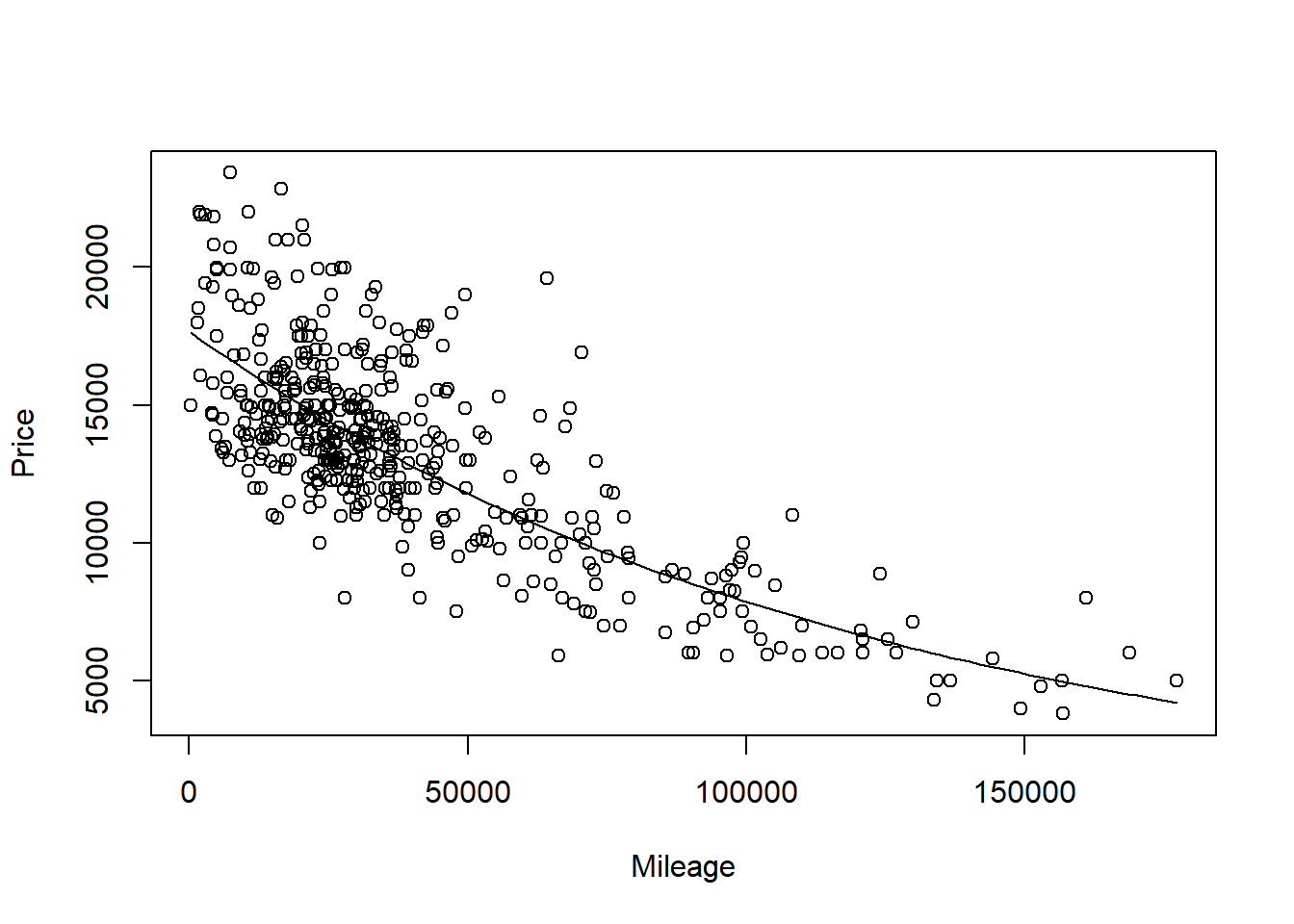


plot(Price~Mileage, data=MyCars)

B0\_q15 = summary(modq14.2)$coef[1,1]

B1\_q15 = summary(modq14.2)$coef[2,1]

curve(exp(B0\_q15 + B1\_q15\*x), add=TRUE)



1. How do the models, using either *Age* or *Mileage* as the predictor compare? Does one of the models seem “better” or do they seem similar in their ability to predict *Price*? Explain.

**1 points** - Any reasonable comparison is fine for full credit. Students may comment on things such as the p-value of signifcance tests, Rsq, etc…

**MODEL #3: Again use Age as a predictor for Price, but now for new data**

1. Select a new sample from the UsedCar dataset using the same *Model* car that was used in the previous sections, but now from cars for sale in North Carolina. You can mimic the code used above to select this new sample.

**1 points** - THey can set this up differently, but should have the same model car, a different state, and a new variable *Age*.

StateOfMyChoice2 = "NC"

ModelOfMyChoice = "Civic"

# Takes a subset of your model car from your state

MyCars2 = subset(UsedCars, Model==ModelOfMyChoice & State==StateOfMyChoice2)

# Add a new variable for the age of the cars.

MyCars2$Age = 2017 - MyCars2$Year

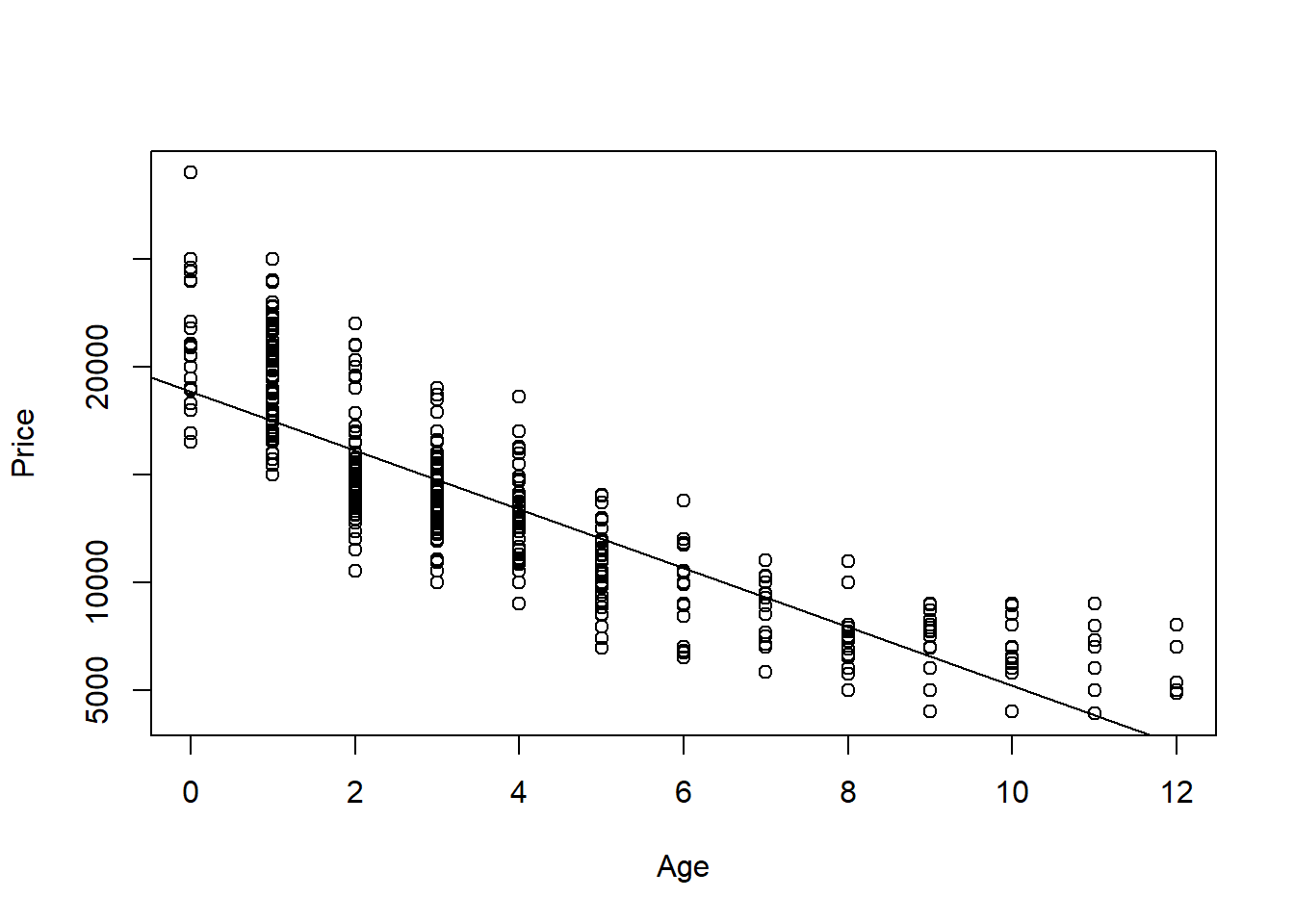
1. Calculate the least squares regression line that best fits your new data and produce a scatterplot of the relationship with the regression line on it.

**0.5 point** code for model  
**0.5 point** code for plot  
**0.5 point** abline

modq17 = lm(Price~Age, data=MyCars2)

plot(Price~Age, data=MyCars2)

abline(modq17)



1. How does the relationship between *Price* and *Age* for this new data compare to the regression model constructed in the first section? Does it appear that the relationship between *Age* and *Price* for your *Model* of car is similar or different for the data from your two states? Explain.

**1 points** - Any reasonable comparison is fine for full credit. They may compare scatterplots, summary intercepts and slope, etc. My data for NY may not show quite the same curvature, as for NC cars, but the overall trend of a curve better fitting the data than a line holds the same.

1. Again suppose that you are interested in purchasing a car of this model that is four years old (in 2017) from North Carolina. How useful do you think that your model will be? What are some possible cons of using this model?

**0.5 points** - Any reasonable discussion is fine for full credit.

In an earlier question I predicted with 90% confidence I predict that the price of a 4 year old Civic sold in NY is between $9303.123 and 15516.37. This gives me some idea of how much a car will cost, but is still a wide interval. For students that chose models with fewer of that car in the data, their intervals will be much wider. It would be more useful to have a narrower interval, that would require more data, or possibly a multiple regression model to explain more of the variability in price.

## Homework #3 Individual Part

40 points - Due 9/29 at 5:00pm   
Directions: You will be randomly assigned to a group of three to four students for this   
assignment. Parts 1 & 2 should be turned in individually to Gradescope by each student in   
your group. Parts 3, 4, & 5 should be submitted by as a group to Gradescope. There are   
separate places to submit the individual and group portions of the assignment.   
Situation: Can we predict the selling price of a house in Ames, Iowa based on recorded   
features of the house? That is your task for this assignment. Each group will have a dataset   
with information on forty potential predictors and the selling price (in $1,000’s) for a   
sample of homes. The data set for your group is in AmesTrain??.csv (where ?? corresponds   
to your group number) and can be found in the AmesTrain zipped file under class 14 in   
Sakai. A separate file identifies the variables in the Ames Housing data and explains some of   
the coding.   
Part 1. Build an initial “basic” model   
Your basic model can use any of the quantitative variables in the dataset but should NOT   
use the categorical variables, transformations, or interactions (we’ll discuss these in class   
soon) – those will come in a later assignment. Use your data to select a set of predictors to   
include in your model. Keep track of the process you use and decisions you make to arrive   
at an initial set of predictors. Your report should include a summary of this process. You   
don’t need to show all the output for every model you consider, but you should give a clear   
description of the path you took and the criteria that you used to compare competing   
models. Also, use at least two model selection methods to find a model (e.g. don’t just check   
all subsets, although it will work well here, this method will fail in future assignments).   
In addition to the commentary on model selection, include the following information for   
this initial choice of a model: the summary() output for your model, comments on which (if   
any) of the predictors in the model are not significant at a 5% level, and comments on what   
the VIF values tell you about the individual predictors in your model.   
Do not consider the Order variable (that is just an observation number) as one of your   
predictors. Avoid predictors that are exactly related. For example, if   
GroundSF=FirstSF+SecondSF you will likely get trouble if you try to put all three in the   
same model.   
setwd <- "C:/Users/adeve/Desktop"   
houses <- read.csv("AmesTrain24.csv")   
source("https://raw.githubusercontent.com/JA-  
McLean/STOR455/master/scripts/ShowSubsets.R")   
  
num <- lapply(houses, is.numeric)   
names(houses[num=="TRUE"])

## [1] "Order" "Price" "LotFrontage" "LotArea"   
## [5] "Quality" "Condition" "YearBuilt" "YearRemodel"   
## [9] "BasementFinSF" "BasementUnFinSF" "BasementSF" "FirstSF"   
## [13] "SecondSF" "GroundSF" "BasementFBath" "BasementHBath"   
## [17] "FullBath" "HalfBath" "Bedroom" "TotalRooms"   
## [21] "Fireplaces" "GarageCars" "GarageSF" "WoodDeckSF"   
## [25] "OpenPorchSF" "EnclosedPorchSF" "ScreenPorchSF"   
Method 1: All Subsets First, I ran various subset combinations and noted their MallowCP   
and R-squared values. I did this until I found the lowest Mallow CP value for the highest R-  
Squared value.   
library(readr)   
library(car)   
## Loading required package: carData   
library(corrplot)   
## corrplot 0.90 loaded   
library(leaps)   
  
all <- regsubsets(Price~Fireplaces+GarageSF+GroundSF, houses, nbest=4)   
  
summary(all)   
## Subset selection object   
## Call: regsubsets.formula(Price ~ Fireplaces + GarageSF + GroundSF,   
## houses, nbest = 4)   
## 3 Variables (and intercept)   
## Forced in Forced out   
## Fireplaces FALSE FALSE   
## GarageSF FALSE FALSE   
## GroundSF FALSE FALSE   
## 4 subsets of each size up to 3   
## Selection Algorithm: exhaustive   
## Fireplaces GarageSF GroundSF   
## 1 ( 1 ) " " " " "\*"   
## 1 ( 2 ) " " "\*" " "   
## 1 ( 3 ) "\*" " " " "   
## 2 ( 1 ) " " "\*" "\*"   
## 2 ( 2 ) "\*" "\*" " "   
## 2 ( 3 ) "\*" " " "\*"   
## 3 ( 1 ) "\*" "\*" "\*"   
ShowSubsets(all)   
## Fireplaces GarageSF GroundSF Rsq adjRsq Cp   
## 1 ( 1 ) \* 47.24 47.15 254.24   
## 1 ( 2 ) \* 42.72 42.62 327.04

STOR 455 Homework #3 Group Part   
40 points - Due 9/29 at 5:00pm   
Group Members: Alyssa Warnock, Holden Mclaughlin, Clayton Brockmann, Alex Huml   
Directions: You will be randomly assigned to a group of three to four students for this   
assignment. Parts 1 & 2 should be turned in individually to Gradescope by each student in   
your group. Parts 3, 4, & 5 should be submitted by as a group to Gradescope. There are   
separate places to submit the individual and group portions of the assignment.   
Situation: Can we predict the selling price of a house in Ames, Iowa based on recorded   
features of the house? That is your task for this assignment. Each group will have a dataset   
with information on forty potential predictors and the selling price (in $1,000’s) for a   
sample of homes. The data set for your group is in AmesTrain??.csv (where ?? corresponds   
to your group number) and can be found in the AmesTrain zipped file under class 14 in   
Sakai. A separate file identifies the variables in the Ames Housing data and explains some of   
the coding.   
library(readr)   
library(Stat2Data)   
library(tidyverse)   
## -- Attaching packages --------------------------------------- tidyverse   
1.3.1 --   
## v ggplot2 3.3.5 v dplyr 1.0.7   
## v tibble 3.1.3 v stringr 1.4.0   
## v tidyr 1.1.3 v forcats 0.5.1   
## v purrr 0.3.4   
## -- Conflicts ------------------------------------------   
tidyverse\_conflicts() --   
## x dplyr::filter() masks stats::filter()   
## x dplyr::lag() masks stats::lag()   
library(ggplot2)   
library(car)   
## Loading required package: carData   
##   
## Attaching package: 'car'   
## The following object is masked from 'package:dplyr':   
##   
## recode

## The following object is masked from 'package:purrr':   
##   
## some   
library(corrplot)   
## corrplot 0.90 loaded   
library(leaps)   
  
source("https://raw.githubusercontent.com/JA-  
McLean/STOR455/master/scripts/ShowSubsets.R")   
  
amesHouseData <- read\_csv("AmesTrain24.csv")   
## Rows: 600 Columns: 42   
## -- Column specification --------------------------------------------------  
------   
## Delimiter: ","   
## chr (15): LotConfig, HouseStyle, ExteriorQ, ExteriorC, Foundation,   
BasementH...   
## dbl (27): Order, Price, LotFrontage, LotArea, Quality, Condition,   
YearBuilt,...   
##   
## i Use `spec()` to retrieve the full column specification for this data.   
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this   
message.   
head(amesHouseData)   
## # A tibble: 6 x 42   
## Order Price LotFrontage LotArea LotConfig HouseStyle Quality Condition   
## <dbl> <dbl> <dbl> <dbl> <chr> <chr> <dbl> <dbl>   
## 1 1318 128 35 6300 Inside 2.5Unf 6 6   
## 2 527 241 65 8125 Inside 2Story 7 5   
## 3 1100 315 80 9200 Inside 2Story 8 5   
## 4 761 116 60 10320 Inside 2Story 6 7   
## 5 1244 136 75 9464 Corner 1Story 6 7   
## 6 1120 247 87 10037 Corner 1Story 8 5   
## # ... with 34 more variables: YearBuilt <dbl>, YearRemodel <dbl>,   
## # ExteriorQ <chr>, ExteriorC <chr>, Foundation <chr>, BasementHt <chr>,   
## # BasementC <chr>, BasementFin <chr>, BasementFinSF <dbl>,   
## # BasementUnFinSF <dbl>, BasementSF <dbl>, Heating <chr>, HeatingQC   
<chr>,   
## # CentralAir <chr>, FirstSF <dbl>, SecondSF <dbl>, GroundSF <dbl>,   
## # BasementFBath <dbl>, BasementHBath <dbl>, FullBath <dbl>, HalfBath   
<dbl>,   
## # Bedroom <dbl>, KitchenQ <chr>, TotalRooms <dbl>, Fireplaces <dbl>, ...

# Unit 4: Multiple Regression II

## STOR 455 - Class 17 – Comparing two regression lines

What do we do with categorical variables? **Example: Pulse Rates** - Response Variable:  
- Y =Active pulse - Predictors:  
- X1 = Resting pulse - X2 = Sex (0=M, 1=F) - Datat that looks at heart rate and how we can use sex as a categorical indicator varianble and how we should consider it if we add it to our model - If we just use sex –If we wanted to see the difference bt the binary groups, then we could do a two sample t test

**Categorical Predictor** Example:  
Response = Y = Active pulse Predictor = X = Sex  
- *Are active pulse rates related to sex? “Usual” procedure?* - Two-sample t-test (difference in means) – How different are the two means vs what I would expect to see from them? – A few ways you can do difference in means test – How unlikely would it be that they are this different by chance in our smaples

library(readr)  
  
Pulse <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/Pulse.csv")  
head(Pulse)

## # A tibble: 6 x 7  
## Active Rest Smoke Sex Exercise Hgt Wgt  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 97 78 0 1 1 63 119  
## 2 82 68 1 0 3 70 225  
## 3 88 62 0 0 3 72 175  
## 4 106 74 0 0 3 72 170  
## 5 78 63 0 1 3 67 125  
## 6 109 65 0 0 3 74 188

**(using pooled variances), Two-sample T-test for Means** Ho: mu1 = mu2 Ha: mu1 != mu2

where, (pooled standard deviation): 𝑠\_𝑝=√(((𝑛\_1−1) 𝑠\_1^2+〖(𝑛\_2−1)𝑠〗\_2^2)/(𝑛\_1+𝑛\_2−2))

𝑡=(𝑦̅\_1−𝑦̅\_2)/(𝑠\_𝑝 √(1/𝑛\_1 +1/𝑛\_2 )) Compare to t with 𝑛\_1+𝑛\_2−2 d.f.

**R - Two-sample T-test** - pvalue of 0.004853 is significant

* We want to see if active heart rates show any difference
* Will add var.equal=TRUE
* By default the dif in mean test assumes that the means are euqla bt the groups and wants to see how unlikely you get this result by chance
* ytou could also make the assumption that they come from teh same popuilation, so the means and spread should be equal as well
* We want withthe assumption that the mean active heart rates are equal by sex;
* Ho: Mean active heart rates = by sex
* Ha: No equal
* Evidence there is some difference by sex

t.test(Active~Sex, var.equal=TRUE, data=Pulse)

##   
## Two Sample t-test  
##   
## data: Active by Sex  
## t = -2.8329, df = 373, p-value = 0.004863  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## -8.992040 -1.623594  
## sample estimates:  
## mean in group 0 mean in group 1   
## 84.60753 89.91534

**“Dummy” Predictors** - We can code a categorical predictor as (0,1) - How should this be interpreted in a regression? - Example: Y = Active pulse, where 0 = male and 1 = female

**For summary of modelP** Table, estimate, std, error, tvalue, P Intercept, mean for males, 1.330, 63.607, pvalue sex, difference for females, 1.874, 2.833, 0.00486 residual standard error….

**t.test(active~Sex, var.equal=TRUE, data = Pulse)** <data:Active> by Sex t = -2.8329, df = 373, pvalue = 0.004863 <- this is the t-test for significant difference

sample estimates: mean in group 0 mean in group 1 84.60753 (mean for males), 89.91534 (mean for female)

modelP=lm(Active~Sex, data=Pulse) #Active hr predicted by sex   
# In teh data,t eh intercept = mean for the meales ebc its where sex = 0   
# The mean for 0 = 84 = the male active heart rate   
# The mean for 1 = 89 = female active heart rate, if   
# If out linear model is just an intercept + sex\*slope, and sex is a 1 or a 0, then it ends up being intercept + 0 or intercept + slope   
# Getting together = female active heart rate   
summary(modelP)

##   
## Call:  
## lm(formula = Active ~ Sex, data = Pulse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -36.915 -12.761 -1.915 9.392 69.392   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 84.608 1.330 63.607 < 2e-16 \*\*\*  
## Sex 5.308 1.874 2.833 0.00486 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.14 on 373 degrees of freedom  
## Multiple R-squared: 0.02106, Adjusted R-squared: 0.01844   
## F-statistic: 8.025 on 1 and 373 DF, p-value: 0.004863

# looking at t tests  
# same pvalue, dif sig digets   
# Can use a linear model like we did d fiference in means test in teh past

**Quantitative + Indicator Predictors** Example: Y = Active pulse rate X1 = Resting pulse rate X2 = Sex (0,1) (𝐴𝑐𝑡𝑖𝑣𝑒)̂=8.016+1.165𝑅𝑒𝑠𝑡+2.326𝑆𝑒𝑥

*How do we interpret the coefficient of sex?* Ho: B2 = 0 Ha: B2 != 0

With pvalue of 0.116 (because B2 is Sex), There is not evidence to reject the null hypothesis and suggest that B2 != 0

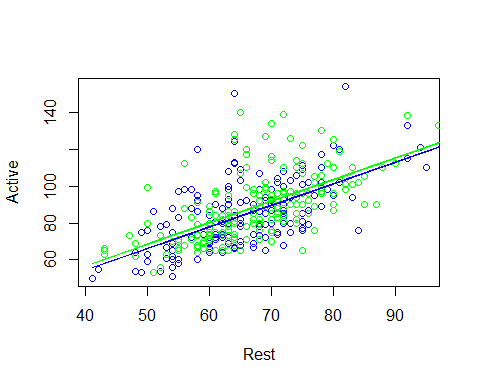
modelP2=lm(Active~Rest+Sex, data=Pulse) # model with quanti and categorical variables  
summary(modelP2)

##   
## Call:  
## lm(formula = Active ~ Rest + Sex, data = Pulse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.706 -9.396 -2.742 6.787 67.434   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.01600 5.04661 1.588 0.113   
## Rest 1.16484 0.07511 15.508 <2e-16 \*\*\*  
## Sex 2.32642 1.47471 1.578 0.116   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.16 on 372 degrees of freedom  
## Multiple R-squared: 0.4055, Adjusted R-squared: 0.4023   
## F-statistic: 126.8 on 2 and 372 DF, p-value: < 2.2e-16

# Ho: Rest = 0   
# Ha: Rest != 0   
# Small pvalue = reject Ho  
  
# Ho: Sex = 0   
# Ha: Sex != 0   
# Higher pvalue = fail to reject Ho

**Same slope, different intercepts**

plot(Active~Rest, col="blue", data=subset(Pulse,Sex==0))  
  
points(Active~Rest, col="green", data=subset(Pulse,Sex==1))  
# plots the points on the graph   
  
# Below shows where we got the things from   
B\_Int = summary(modelP2)$coef[1,1]  
B\_Rest = summary(modelP2)$coef[2,1]  
B\_Sex = summary(modelP2)$coef[3,1]  
  
# Plots the line of males and females separately   
lines(  
 B\_Int + B\_Rest \* Rest ~ Rest,   
 col = "blue",   
 data = Pulse  
 )  
  
lines(  
 (B\_Int + B\_Sex) + B\_Rest \* Rest ~ Rest,   
 col = "green",   
 data = Pulse  
 )



# THis shows that it is forcing us to assume that there is the same slope per sex

**Fit Models to Subsets** (𝐴𝑐𝑡𝑖𝑣𝑒)̂=9.440+1.1432 𝑅𝑒𝑠𝑡 (Males)

Males=subset(Pulse, Sex==0)  
modelPM=lm(Active~Rest, data=Males)  
summary(modelPM)

##   
## Call:  
## lm(formula = Active ~ Rest, data = Males)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -29.468 -9.426 -2.462 8.109 67.396   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.4399 7.4324 1.27 0.206   
## Rest 1.1432 0.1119 10.21 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.09 on 184 degrees of freedom  
## Multiple R-squared: 0.3618, Adjusted R-squared: 0.3583   
## F-statistic: 104.3 on 1 and 184 DF, p-value: < 2.2e-16

# Intercept 9.43 and a slope of 1.1432 for the male model

**Fit Models to Subsets** (𝐴𝑐𝑡𝑖𝑣𝑒)̂=9.153+1.1823 𝑅𝑒𝑠𝑡 (Females)

Females=subset(Pulse, Sex==1)  
modelPF=lm(Active~Rest, data=Females)  
summary(modelPF)

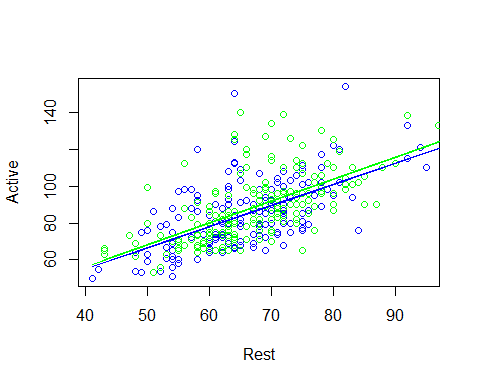
##   
## Call:  
## lm(formula = Active ~ Rest, data = Females)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.822 -9.088 -3.177 6.010 54.000   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.1527 7.0198 1.304 0.194   
## Rest 1.1823 0.1016 11.633 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.26 on 187 degrees of freedom  
## Multiple R-squared: 0.4198, Adjusted R-squared: 0.4167   
## F-statistic: 135.3 on 1 and 187 DF, p-value: < 2.2e-16

# Intercept of 9.1527 witha slope of 1.1823 for the female model

We see some difference between the output. Is this a signifigant difference or would I just expect to see this by chance?

**Plotting the lines** - Are these lines “significantly” different?

plot(Active~Rest, col="blue", data=subset(Pulse,Sex==0))  
  
points(Active~Rest, col="green", data=subset(Pulse,Sex==1))  
# The above code puts the dots on the graph   
  
# the below code puts the line of the models with male and female on the graph  
  
lines(  
 summary(modelPM)$coef[1,1] + summary(modelPM)$coef[2,1] \* Rest ~ Rest,   
 col = "blue",   
 data = Pulse  
 )  
  
lines(  
 summary(modelPF)$coef[1,1] + summary(modelPF)$coef[2,1] \* Rest ~ Rest,   
 col = "green",   
 data=Pulse  
 )



# Now we are working with 2 different models instead of one   
  
# The slopes are slightly different, and we are allowing for the different rate of change; we want to do this with the entire dataset.

Y = B0 + B1X1 + B2X2 + B3X1\*X2+Error - We want to create a line that by changing the value of one indivator variable, we can change what teh intercept of the prediction is, as well as what the slope of that model is - The interaction term does this - When sex = 1 what should slope be and when sex = 0 what should slope be? - When sex = 0, then the some things go away, but when sex = 1, the B2 Term will effect the intercept of the model and B3 will affect the slope of the model

**Comparing Two Regression Lines (with a multiple regression)** - Example:  
- Y=Active pulse - X1= Resting pulse - X2= Sex(0,1)

𝑌=𝛽\_𝑜+𝛽\_1 𝑋\_1+𝛽\_2 𝑋\_2+𝛽\_3 𝑋\_1 𝑋\_2+𝜀 Y = Intercept + Quantitative + Indicator + Interaction

**Quantitative + Indicator +Interaction** (𝐴𝑐𝑡𝑖𝑣𝑒)̂=9.440+1.1432𝑅𝑒𝑠𝑡−0.287𝑆𝑒𝑥+0.039𝑅𝑒𝑠𝑡∗𝑆𝑒𝑥 - How does this relate to the two lines? - Substitute Sex=0 and Sex=1

# Interaction terms  
# ANOVA Assumptions   
# Ho: All Bi = 0   
# Ha: At least one Bi != 0   
modelPint=lm(Active~Rest+Sex+Rest\*Sex, data=Pulse)  
# BEcause rest is sig, it doesn't appear that the intercept change is useful for us to do   
# interaction bt rest and sex has a high pvalue, tells us we might not need the interaction term   
summary(modelPint)

##   
## Call:  
## lm(formula = Active ~ Rest + Sex + Rest \* Sex, data = Pulse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.822 -9.251 -2.893 6.784 67.396   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.43987 7.47902 1.262 0.208   
## Rest 1.14319 0.11264 10.149 <2e-16 \*\*\*  
## Sex -0.28717 10.22830 -0.028 0.978   
## Rest:Sex 0.03907 0.15130 0.258 0.796   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.17 on 371 degrees of freedom  
## Multiple R-squared: 0.4056, Adjusted R-squared: 0.4008   
## F-statistic: 84.37 on 3 and 371 DF, p-value: < 2.2e-16

*How can we make a sig test for this?* - When we dont appear to have evidence for sig difference in slope or intercept? - Are there different lines to predict what we want? - If there are, then the last two coeff would be 0

**Tests to Compare Two Regression Lines** Y = Bo + B1X1 + B2X2 + B3X1X2 + Error

* **Different Slope** - T test – Ho: B3 = 0 – Ha: B3 != 0
* **Different Intercept** - T test – Ho: B2 = 0 – Ha: B2 != 0
* **Different lines** - See Multiple Regression Model section below – Ho: B2=B3=0 – Ha: B2 != 0 or B3 != 0

**Multiple regression model** - **Testing one term at a time:** – T-test – Ho: B1 = 0 – Ha: B1 != 0

* **Testing all terms at once** – ANOVA – Ho: B2=B3=0 – Ha: Some Bi != 0

*Is there anything in between?*

**Nested Models** - **Definition:** If all of the predictors in Model A are also in a bigger Model B, we say that Model A is nested in Model B.

* Example: 𝐴𝑐𝑡𝑖𝑣𝑒=𝛽\_0+𝛽\_1 𝑅𝑒𝑠𝑡+ 𝜀 is nested in 𝐴𝑐𝑡𝑖𝑣𝑒=𝛽\_0+𝛽\_1 𝑅𝑒𝑠𝑡+\_2 𝑆𝑒𝑥+\_3 𝑅𝑒𝑠𝑡∗𝑆𝑒𝑥+𝜀
* *Test for Nested Models:*
* Do we really need the extra terms in Model B?
* i.e. How much do they “add” to Model A?

# Something in between   
# ANOVA = all coef are zero vs at least one is nonzero - we compare to null mode, how much do we explain vs a null model?   
# We can sub a different model to the null model   
# Caveat: The sub model has to have a nested subset of what we are working with in the bigger model   
  
modelP\_Reduced = lm(Active~Rest, data=Pulse)  
  
# NEsted F Test   
# Do we need all these things in the model?   
# IF we add things tothe model we may get a smaller mallow CP, but is it signifigant improvement? This tells you   
anova(modelP\_Reduced, modelPint) # Nested test

## Analysis of Variance Table  
##   
## Model 1: Active ~ Rest  
## Model 2: Active ~ Rest + Sex + Rest \* Sex  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 373 75050   
## 2 371 74538 2 512.14 1.2746 0.2808

# First line = test compare model to null model, and its sig   
# THen build a model with rest and sex in it and comparing the model before to this model   
# To see if addingteh sex predictor increases teh varaibility signifigant;y (It is not in this explame)  
# Third line compares rest, sex, and interaction to just the model with rest and sex in it  
# Tells you if we are explaining an extra amount of the varibility by adding the interactiont erm (This tells you you are not explaining a good amount extra)   
# Only showed up this way ebcause fo the order you put it in ANOVA, if you change the order, you change the order it analyzes things and it might change what it says   
  
# F test stat is calc similiarlly, the dif is when we think about sum of squares; its not how much teh SS in this model it's how much they are in this model taking away what is in the reduced model   
# How much variability is being explained by these extra things, not the whole model as a whole

anova(modelPint)

## Analysis of Variance Table  
##   
## Response: Active  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Rest 1 50342 50342 250.5710 <2e-16 \*\*\*  
## Sex 1 499 499 2.4824 0.1160   
## Rest:Sex 1 13 13 0.0667 0.7964   
## Residuals 371 74538 201   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Nested F-test** - Basic idea: 1. Find how much “extra” variability is explained when the “new” terms being tested are added. 2. Divide by the number of new terms to get a mean square for the new part of the model. 3. Divide this mean square by the MSE for the “full” model to get an F-statistic. 4. Compare to an F-distribution to find a p-value.

* Test: Ho: Bi = 0 for a “subset” of predictors Ha: Bi != 0 for some predictors in the subset
* 𝐹=(((𝑆𝑆𝑀𝑜𝑑𝑒𝑙\_𝐹𝑢𝑙𝑙−𝑆𝑆𝑀𝑜𝑑𝑒𝑙\_𝑅𝑒𝑑𝑢𝑐𝑒𝑑))⁄(# 𝑝𝑟𝑒𝑑𝑖𝑐𝑡𝑜𝑟𝑠))/𝑀𝑆𝐸𝐹𝑢𝑙𝑙
* F = ((Explained by full model - explained by reduced model)/#predictors tested in Ho)/MSEFullthat is based on the full model
* Compare to F-distribution

**Nested F-test** - 𝐴𝑐𝑡𝑖𝑣𝑒 =𝛽\_0+𝛽1𝑅𝑒𝑠𝑡+\_2 𝑆𝑒𝑥+ 3𝑅𝑒𝑠𝑡𝑆𝑒𝑥 +𝜀 - H0: β2=β3=0 - Ha: Some βi != 0 - Compare mean square for the “extra” variability to the mean square error for the full model.

modelPint2 = lm(Active~Sex+Rest+Sex\*Rest, data = Pulse)  
anova(modelPint2)

## Analysis of Variance Table  
##   
## Response: Active  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Sex 1 2641 2641 13.1454 0.0003285 \*\*\*  
## Rest 1 48200 48200 239.9081 < 2.2e-16 \*\*\*  
## Sex:Rest 1 13 13 0.0667 0.7963721   
## Residuals 371 74538 201   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# First one being done is sex a sig pred of active heart rate?   
# Yes it is, but it looks different than the other tests because they were doing soemthing different   
# This test is after considering teh variability explained in teh active heart rat eby a resting heart ratel is sex sig after that?   
# We have a big p value so its not,   
#If we dont take into account the varability in account by resting heart rate first, then sex alone is useful   
# ADding rest and then sex adn rest together has a pvalue of 2.2 to the -16, to its a sig amount explained   
# The last row is teh same because we are still comparing teh same test as we are before   
# A mdoel wtih all 3 vs a model withjust sex and rest

modelPint3 = lm(Active~Rest\*Sex, data = Pulse)  
summary(modelPint3)

##   
## Call:  
## lm(formula = Active ~ Rest \* Sex, data = Pulse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.822 -9.251 -2.893 6.784 67.396   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.43987 7.47902 1.262 0.208   
## Rest 1.14319 0.11264 10.149 <2e-16 \*\*\*  
## Sex -0.28717 10.22830 -0.028 0.978   
## Rest:Sex 0.03907 0.15130 0.258 0.796   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.17 on 371 degrees of freedom  
## Multiple R-squared: 0.4056, Adjusted R-squared: 0.4008   
## F-statistic: 84.37 on 3 and 371 DF, p-value: < 2.2e-16

# Even though its jsut an interaction term, R assumed we wanted each indidivual terms as well   
# YOu dont want an interaction without each of the terms in teh model before that   
# R knew that I should do that, so it gave it to you

ANOVA Tests - We dont want ot just put one function into it becase thatis depend on the order of the predictors of the model - ONly lets us test one predictor at a time - Adds one layer at a time

We want to compare a model with all 3 predictors with just 1 predictor (just restin gheart rate)

modelP\_Reduced = lm(Active~Rest, data = Pulse)  
summary(modelP\_Reduced)

##   
## Call:  
## lm(formula = Active ~ Rest, data = Pulse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -31.675 -9.142 -2.725 7.062 66.309   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.15295 5.05592 1.613 0.108   
## Rest 1.18029 0.07462 15.818 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.18 on 373 degrees of freedom  
## Multiple R-squared: 0.4015, Adjusted R-squared: 0.3999   
## F-statistic: 250.2 on 1 and 373 DF, p-value: < 2.2e-16

* Then we wnat to do the ANOVA on the reduced and others
* To see how much variability is being explaine dby adding those new terms to it

anova(modelP\_Reduced, modelPint)

## Analysis of Variance Table  
##   
## Model 1: Active ~ Rest  
## Model 2: Active ~ Rest + Sex + Rest \* Sex  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 373 75050   
## 2 371 74538 2 512.14 1.2746 0.2808

# This tells us the df is 2, whih tells us tehre are 2 variables different between tehse two models   
# We get an idea of teh variability explained bt these two models and see tha tthe SS (the 512.14) = the amount of varabiltiy explained by adding teh sex and interaction term to the model ; that is a small amount in the long run and that is why the f test stat is small and the p value is high   
# This lets you test both those terms together   
  
# BEnefit: We are jsut comparing two extra terms; if there were 10 terms different, then we could just do tests of each 10 terms individually, but tehn we'll run into more error issues   
# When we do it all at once, we get a big pitcure if there are any differences and if we need to we can investigate further to see wehre teh differences are

Nested tests tell you if you’re looking at noise or if you’re looking at something signifignat

## STOR 455 - Class Coding categorial variables

library(readr)  
library(leaps)  
  
Pulse <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/Pulse.csv")  
source("https://raw.githubusercontent.com/JA-McLean/STOR455/master/scripts/ShowSubsets.R")  
  
head(Pulse)

## # A tibble: 6 x 7  
## Active Rest Smoke Sex Exercise Hgt Wgt  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 97 78 0 1 1 63 119  
## 2 82 68 1 0 3 70 225  
## 3 88 62 0 0 3 72 175  
## 4 106 74 0 0 3 72 170  
## 5 78 63 0 1 3 67 125  
## 6 109 65 0 0 3 74 188

**Nested F-test** 𝐴𝑐𝑡𝑖𝑣𝑒 =𝛽\_0+𝛽1𝑅𝑒𝑠𝑡+\_2 𝑆𝑒𝑥+ 3𝑅𝑒𝑠𝑡𝑆𝑒𝑥 +𝜀 H0: β2=β3=0 Ha: Some βi≠0

Compare mean square for the “extra” variability to the mean square error for the full model.

anova(modelP\_Reduced, modelPint) Analysis of Variance Table

Model 1: Active ~ Rest Model 2: Active ~ Rest + Sex + Rest \* Sex Res.Df RSS Df Sum of Sq F Pr(>F) 1 373 75050  
2 371 74538 2 512.14 1.2746 0.2808

**More than Two Categories** Example: (Active pulse)

* Exercise: – 1 = Slight – 2 = Moderate – 3 = Lots

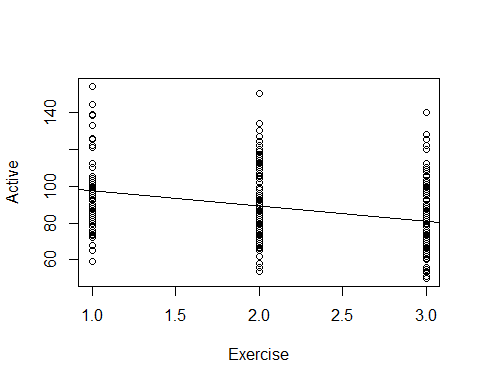
-Try a model to predict Y=Active pulse rates using X=Exercise. How should the coefficients be interpreted?

\_Predicting Active with Exercise\_\_

modelEX = lm(Active ~ Exercise, data=Pulse) # Predict active heart rate by exercise rate   
summary(modelEX)

##   
## Call:  
## lm(formula = Active ~ Exercise, data = Pulse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -38.613 -12.879 -1.613 9.121 60.754   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 105.979 2.878 36.829 < 2e-16 \*\*\*  
## Exercise -8.367 1.224 -6.834 3.37e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17.28 on 373 degrees of freedom  
## Multiple R-squared: 0.1113, Adjusted R-squared: 0.1089   
## F-statistic: 46.71 on 1 and 373 DF, p-value: 3.372e-11

plot(Active ~ Exercise, data=Pulse)  
abline(modelEX)



# We are saying there is some change between exervise levels   
# Does exercising a moderate and a lot amount have the same impact as exercising a small and mdoerate amount?   
# We are summing that you exercising changes is constant regardless of group

* 105 = exercise rate of 0, but we don thave one
* exercise 1 = whatever the intercept is plus the slope because we are just going over on eunit and intercept is negative; r=poreict would be 105-8.37 = 89 prediction show the same distance bteween groups We dont want to make that asusmption here

Tkae more care with things that are not binary; we need to foroce varibales to be binary

**Active Pulse vs. Exercise Categories**

tapply(Pulse$Active, Pulse$Exercise, mean)

## 1 2 3   
## 96.24242 90.41290 80.29221

# Slpit groups by exercise levels   
# WE want to know what the average is for each thing   
#Is the “slope” from 1 to 2 the same as from 2 to 3?  
#Note: Using Exercise as a quantitative predictor forces the “slopes” to be the same.  
  
# The oringial model is telling me that there is no change between the mean heart rates based on exercise level; this is telling me that there is a change.  
# WE dont know if it's a significant change or not yet.   
  
#It's ordnal, the exercise levels

\_-Dummy Indicators for Multiple Categories\_\_ For a categorical predictor with k levels, we use k-1 dummy indicators. - X1 = 1 if group #1, 0 if otherwise - Xk-1 = 1 if graph is k-1, 0 if otherwise

*Below: R Trick: (To create indicator variables)* What happens to Group #k?

*Predicting Active Using Slight and Moderate Exercise Indicators* Call: lm(formula = Active ~ Slight + Moderate, data = Pulse)

Coefficients: Estimate Std. Error t value Pr(>|t|)  
(Intercept) 80.292 1.392 57.670 < 2e-16 Slight 15.950 2.542 6.275 9.74e-10  
Moderate 10.121 1.966 5.148 4.27e-07

Multiple R-squared: 0.1144, Adjusted R-squared: 0.1096 F-statistic: 24.02 on 2 and 372 DF, p-value: 1.541e-10

Pulse$Moderate=(Pulse$Exercise==2)\*1 # Be careful! this is 2!  
# This says that if it is 2, it will be true  
Pulse$Slight=(Pulse$Exercise==1)\*1  
# this says if it is 1, then it will be true  
  
# Multiplying it by 1 will treat the trues and falses as 1 and 0   
# We only need to do this for all but 1, because if all false, then it's whwatever is left over  
  
modelEX2 = lm(Active ~ Slight + Moderate, data = Pulse)  
summary(modelEX2)

##   
## Call:  
## lm(formula = Active ~ Slight + Moderate, data = Pulse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -37.242 -12.413 -1.292 8.647 59.708   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 80.292 1.392 57.670 < 2e-16 \*\*\*  
## Slight 15.950 2.542 6.275 9.74e-10 \*\*\*  
## Moderate 10.121 1.966 5.148 4.27e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17.28 on 372 degrees of freedom  
## Multiple R-squared: 0.1144, Adjusted R-squared: 0.1096   
## F-statistic: 24.02 on 2 and 372 DF, p-value: 1.541e-10

#small pvalue; so we do have some evi that at least one of the coef are not zero   
# other predictors look good   
# The rsquared, only 11% is explained, so it's not that its not explaining, buyt alone it's probably not best by itself   
  
# Look at thow th emodel is set up, we dont see exercise a lot   
# The intercept = those who exercise a lot   
# For those who exercise a lot, we predict their active heart rate is 80.29  
# IF you look a th em eanthe mean value = the same active heart rate   
# People who exercise a slight aount; then slight would be 1 and moderate would be zero ; then we would get intercept of 96 for slight   
# Doing it this way, we dont have to assume that the change is consistent among the levels of our categorical variables   
  
# WE dont need an extra variable, and if we include it, then we will probably get NA values

**Handling Categorical Predictors in R** - If a predictor in lm( ) has “text” values, R will automatically create indicators for all but one category. - Using factor( )around a quantitative predictor in lm( )creates the indicators. - If you let R decide, then R will decide which one to elave out and you might not know which one it stalking about - R Treats categorical varibales this way - If the categories were Slight, mdoerate and high, then R would factor it right - IF we want to use a numeric value as a category, then use factors

modelEX3=lm(Active~factor(Exercise),data=Pulse)  
summary(modelEX3)

##   
## Call:  
## lm(formula = Active ~ factor(Exercise), data = Pulse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -37.242 -12.413 -1.292 8.647 59.708   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 96.242 2.127 45.253 < 2e-16 \*\*\*  
## factor(Exercise)2 -5.830 2.539 -2.296 0.0223 \*   
## factor(Exercise)3 -15.950 2.542 -6.275 9.74e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17.28 on 372 degrees of freedom  
## Multiple R-squared: 0.1144, Adjusted R-squared: 0.1096   
## F-statistic: 24.02 on 2 and 372 DF, p-value: 1.541e-10

# Looks a little differen than before, because we have a different reference category   
# IT chose to leave out the people who exercise a slight amount   
# Intercept = slight amount average   
# Intercept + Eecise 2 = moderate maount   
# 96-15 = high amount   
  
# No reason we can't include more, so look below for more inclusions

**Multiple Categories in Regression** - With indicator variables for categories we can include quantitative and categorical predictors in the same model

modelEX4=lm(Active~Rest+factor(Exercise),data=Pulse)  
summary(modelEX4)

##   
## Call:  
## lm(formula = Active ~ Rest + factor(Exercise), data = Pulse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.653 -9.206 -2.629 7.231 65.073   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.25869 6.70517 1.381 0.168   
## Rest 1.15698 0.08611 13.436 <2e-16 \*\*\*  
## factor(Exercise)2 1.62128 2.15805 0.751 0.453   
## factor(Exercise)3 -0.51883 2.38266 -0.218 0.828   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.19 on 371 degrees of freedom  
## Multiple R-squared: 0.4043, Adjusted R-squared: 0.3995   
## F-statistic: 83.92 on 3 and 371 DF, p-value: < 2.2e-16

# We looked at the lines f coef table for exericse; maybe not useful due to pvalue   
# Need to do to nested test value because if one is small pvalue adn the other is big we dont want to use one level fothe categorical varible we want one or all   
# Unless we look at if exercise a lot has effect on heart rate; we just want to know if you exercise a lot or you dont; then just look at one category   
# IN general we wnast ot keep all of the categories   
  
# Could do a nested test to see if exercise is s auseful predictor in the odel   
mod = lm(Active~Rest, data=Pulse)  
anova(mod, modelEX4)

## Analysis of Variance Table  
##   
## Model 1: Active ~ Rest  
## Model 2: Active ~ Rest + factor(Exercise)  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 373 75050   
## 2 371 74699 2 350.9 0.8714 0.4192

# Careful here not comparing two predictors to one; its compare with 3 to 1   
# Cator excerise will give 3 var because it has 3 levels   
# Test will do is do a test is the coef of exercise factor 2 = to 0 and the coef of exercise factor 3 = 0 vs the alternative that at least one of them is nonzero?  
# We get a big pvalue; we dont have evidence that adding the exercise terms are improving the model   
# They are not a sig improvement   
  
# We then run into the same issue with binary cate variables that there is some relation between teh resting and active heart rate, but does that change for those who exercise a slight moderate and a lot?   
# Maybe the resting is not so different, but the active heart rates might be differnt?   
  
# THis model is assuming there is a same splot and same realtionship bt active and rest for all exercise levels   
# We are just changing the intercept

**Multiple Categories in Regression with Interactions** - With indicator variables for categories we can include quantitative, categorical, and interaction predictors in the same model

modelEX4int=lm(Active~Rest+factor(Exercise)+Rest\*factor(Exercise),data=Pulse)  
# Adds the interaction term   
# THis will add ac ouple of terms in here, but ti will tell you if the itneraction ebtween things is sig or not   
  
summary(modelEX4int)

##   
## Call:  
## lm(formula = Active ~ Rest + factor(Exercise) + Rest \* factor(Exercise),   
## data = Pulse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.420 -9.609 -2.467 7.008 64.374   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7.5850 13.1326 -0.578 0.5639   
## Rest 1.3810 0.1731 7.977 1.91e-14 \*\*\*  
## factor(Exercise)2 28.6009 16.4065 1.743 0.0821 .   
## factor(Exercise)3 16.5284 15.7150 1.052 0.2936   
## Rest:factor(Exercise)2 -0.3715 0.2240 -1.659 0.0980 .   
## Rest:factor(Exercise)3 -0.2273 0.2216 -1.026 0.3056   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.18 on 369 degrees of freedom  
## Multiple R-squared: 0.4087, Adjusted R-squared: 0.4007   
## F-statistic: 51.01 on 5 and 369 DF, p-value: < 2.2e-16

anova(lm(Active~Rest, data=Pulse), modelEX4int)

## Analysis of Variance Table  
##   
## Model 1: Active ~ Rest  
## Model 2: Active ~ Rest + factor(Exercise) + Rest \* factor(Exercise)  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 373 75050   
## 2 369 74146 4 903.87 1.1246 0.3446

# A line can show the differnce btw active and resting heart rate   
# Exercise elvel 1 which is the level not including this model, we have an intercept of -7.58 and a slope of 1.38 - that st he realtionship   
# For exercise level two , there would be 2 adj to the mopdel; the intercept is going to be the value for our intercept +28, because the intercept is going to change a bit and our resting relationship is going to be the 1.38 slope - .37  
# These are our adjustments   
# Looks like a drastic change   
# factor exercise 3, people who exercise a lto - the intercept will change by this amount and the slope will change by the .22; these are pretty differen tlines   
# If we plotted them we would see there is a big difference   
# We could do some tests to see if they are sid dif.

**Model Selection with Categorical and Interaction Predictors** - Use each of the four model selection methods discussed in class (AllSubsets, Backwards, Forwards, and Stepwise) and compare the processes and outcomes for the predictor pool: Rest, Exercise, Hgt, Wgt, Rest & Exercise, Hgt & Exercise, and Wgt & Exercise - They dont all treat them in teh same way

**All subsets**

library(leaps)  
all = regsubsets(Active~   
 Rest+  
 factor(Exercise)+  
 Rest\*factor(Exercise)+  
 Hgt\*factor(Exercise)+  
 Wgt\*factor(Exercise),   
 data = Pulse, nvmax = 11)  
  
ShowSubsets(all)

## Rest factor(Exercise)2 factor(Exercise)3 Hgt Wgt  
## 1 ( 1 ) \*   
## 2 ( 1 ) \* \*   
## 3 ( 1 ) \* \* \*  
## 4 ( 1 ) \* \* \* \*  
## 5 ( 1 ) \* \* \* \*  
## 6 ( 1 ) \* \* \* \*  
## 7 ( 1 ) \* \* \* \* \*  
## 8 ( 1 ) \* \* \* \*  
## 9 ( 1 ) \* \* \* \*  
## 10 ( 1 ) \* \* \* \*  
## 11 ( 1 ) \* \* \* \* \*  
## Rest:factor(Exercise)2 Rest:factor(Exercise)3 factor(Exercise)2:Hgt  
## 1 ( 1 )   
## 2 ( 1 )   
## 3 ( 1 )   
## 4 ( 1 )   
## 5 ( 1 )   
## 6 ( 1 ) \*  
## 7 ( 1 ) \*   
## 8 ( 1 ) \* \*  
## 9 ( 1 ) \* \* \*  
## 10 ( 1 ) \* \* \*  
## 11 ( 1 ) \* \* \*  
## factor(Exercise)3:Hgt factor(Exercise)2:Wgt factor(Exercise)3:Wgt  
## 1 ( 1 )   
## 2 ( 1 )   
## 3 ( 1 )   
## 4 ( 1 )   
## 5 ( 1 ) \*   
## 6 ( 1 ) \*   
## 7 ( 1 ) \*   
## 8 ( 1 ) \* \*   
## 9 ( 1 ) \* \*   
## 10 ( 1 ) \* \* \*  
## 11 ( 1 ) \* \* \*  
## Rsq adjRsq Cp  
## 1 ( 1 ) 40.15 39.99 14.57  
## 2 ( 1 ) 40.43 40.11 14.77  
## 3 ( 1 ) 41.70 41.23 8.57  
## 4 ( 1 ) 42.01 41.38 8.58  
## 5 ( 1 ) 42.84 42.07 5.20  
## 6 ( 1 ) 43.02 42.09 6.04  
## 7 ( 1 ) 43.29 42.20 6.35  
## 8 ( 1 ) 43.52 42.28 6.87  
## 9 ( 1 ) 43.58 42.19 8.45  
## 10 ( 1 ) 43.65 42.10 10.03  
## 11 ( 1 ) 43.65 41.94 12.00

# Scroll, over to see where the lowest mallow cp is   
  
ShowSubsets(all)[5,] # Best mallow Cp

## Rest factor(Exercise)2 factor(Exercise)3 Hgt Wgt  
## 5 ( 1 ) \* \* \* \*  
## Rest:factor(Exercise)2 Rest:factor(Exercise)3 factor(Exercise)2:Hgt  
## 5 ( 1 )   
## factor(Exercise)3:Hgt factor(Exercise)2:Wgt factor(Exercise)3:Wgt  
## 5 ( 1 ) \*   
## Rsq adjRsq Cp  
## 5 ( 1 ) 42.84 42.07 5.2

# This is not idea because it isnt taking all levels of the varibaile; it might include an interaction term, but i t might not include the indivudal values; which is bad

Full = lm(Active~Rest+Hgt+Wgt+Wgt\*factor(Exercise)+Rest\*factor(Exercise)+ Hgt\*factor(Exercise), data = Pulse)  
# Fullmodel with all predictors we want   
  
none = lm(Active~1, data = Pulse)  
# Model with non   
  
MSE = (summary(Full)$sigma)^2  
# Pull out MSE  
  
# Sets up the process

**Backwards Selection**

step(Full, sclae=MSE)

## Start: AIC=1988.5  
## Active ~ Rest + Hgt + Wgt + Wgt \* factor(Exercise) + Rest \* factor(Exercise) +   
## Hgt \* factor(Exercise)  
##   
## Df Sum of Sq RSS AIC  
## - Wgt:factor(Exercise) 2 368.83 71026 1986.5  
## - Rest:factor(Exercise) 2 388.85 71046 1986.6  
## - Hgt:factor(Exercise) 2 600.51 71258 1987.7  
## <none> 70657 1988.5  
##   
## Step: AIC=1986.45  
## Active ~ Rest + Hgt + Wgt + factor(Exercise) + Rest:factor(Exercise) +   
## Hgt:factor(Exercise)  
##   
## Df Sum of Sq RSS AIC  
## - Rest:factor(Exercise) 2 414.99 71441 1984.6  
## <none> 71026 1986.5  
## - Hgt:factor(Exercise) 2 1233.28 72259 1988.9  
## - Wgt 1 1606.43 72632 1992.8  
##   
## Step: AIC=1984.64  
## Active ~ Rest + Hgt + Wgt + factor(Exercise) + Hgt:factor(Exercise)  
##   
## Df Sum of Sq RSS AIC  
## <none> 71441 1984.6  
## - Hgt:factor(Exercise) 2 1270 72711 1987.2  
## - Wgt 1 1683 73123 1991.4  
## - Rest 1 34858 106298 2131.7

##   
## Call:  
## lm(formula = Active ~ Rest + Hgt + Wgt + factor(Exercise) + Hgt:factor(Exercise),   
## data = Pulse)  
##   
## Coefficients:  
## (Intercept) Rest Hgt   
## 84.97301 1.13968 -1.33728   
## Wgt factor(Exercise)2 factor(Exercise)3   
## 0.10212 -4.19657 -70.52397   
## Hgt:factor(Exercise)2 Hgt:factor(Exercise)3   
## 0.09612 1.02785

# Not saying we could take out hgiehg, says we would haev to remove the interaction term as well   
# Removing weight is possible because th te interaction is gone   
# Takes itno account the restrictions for the model

**forward Method**

step(none, scope=list(upper = Full), sclae = MSE, direction = "forward")

## Start: AIC=2181.6  
## Active ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + Rest 1 50342 75050 1991.1  
## + factor(Exercise) 2 14342 111050 2140.1  
## + Hgt 1 3238 122154 2173.8  
## <none> 125392 2181.6  
## + Wgt 1 397 124995 2182.4  
##   
## Step: AIC=1991.12  
## Active ~ Rest  
##   
## Df Sum of Sq RSS AIC  
## <none> 75050 1991.1  
## + Hgt 1 350.00 74700 1991.4  
## + Wgt 1 148.18 74902 1992.4  
## + factor(Exercise) 2 350.90 74699 1993.4

##   
## Call:  
## lm(formula = Active ~ Rest, data = Pulse)  
##   
## Coefficients:  
## (Intercept) Rest   
## 8.153 1.180

# starts with none, puts in rest, adn doesnt igve option to add interactions  
# Can only add interaciton if the two thigns were in it   
# Tells you to use just rest

**Stepwise**

step(none, scope = list(upper=Full), scale=MSE)

## Start: AIC=271.2  
## Active ~ 1  
##   
## Df Sum of Sq RSS Cp  
## + Rest 1 50342 75050 14.568  
## + factor(Exercise) 2 14342 111050 201.516  
## + Hgt 1 3238 122154 256.563  
## + Wgt 1 397 124995 271.162  
## <none> 125392 271.200  
##   
## Step: AIC=14.57  
## Active ~ Rest  
##   
## Df Sum of Sq RSS Cp  
## <none> 75050 14.568  
## + Hgt 1 350 74700 14.770  
## + Wgt 1 148 74902 15.806  
## + factor(Exercise) 2 351 74699 16.765  
## - Rest 1 50342 125392 271.200

##   
## Call:  
## lm(formula = Active ~ Rest, data = Pulse)  
##   
## Coefficients:  
## (Intercept) Rest   
## 8.153 1.180

# Tells you about the same thing, with only rest   
# Stepwise and forward are very different based on what they do   
  
# Backwards eleminiation goes backwards, least compuational, but you might have a bigger model thatn you need   
# Forward start with nothign and risk a too small method   
# Stepwise is noramlly between, but in this case it was like forward   
  
# We like thes other methods becuase they treat the intearciton terms differently.  
  
#I would say if there are a lot of interaction terms, then you should probably use backwards selection

## STOR 455 - Class 19 – Testing a subset of predictors

library(readr)  
library(leaps)  
  
Pulse <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/Pulse.csv")  
StateSAT <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/StateSAT.csv")  
  
source("https://raw.githubusercontent.com/JA-McLean/STOR455/master/scripts/ShowSubsets.R")  
source("https://raw.githubusercontent.com/JA-McLean/STOR455/master/scripts/anova455.R")

\*Is there a sig dif bt a model with these extra predictors compared to something smaller? **Comparing Two Regression Lines (with a multiple regression)** - dhould reg be considered the same? - The interaction terms - can interact slope and intercept depending on values

**Multiple regression model** - We had anova, but is there someone inbetween? - That’s th enested test! - Instead of comparing to anull, we compare to a subset of the model - and that subset is the base point - ANOVA looks at how much more is explained to a horizontal line - NOw a nested test is comparing the model to a different model

**Nested Models** - Definition: If all of the predictors in Model A are also in a bigger Model B, we say that Model A is nested in Model B. - Example: 𝐴𝑐𝑡𝑖𝑣𝑒=𝛽\_0+𝛽\_1 𝑅𝑒𝑠𝑡+ 𝜀 is nested in - 𝐴𝑐𝑡𝑖𝑣𝑒=𝛽\_0+𝛽\_1 𝑅𝑒𝑠𝑡+\_2 𝑆𝑒𝑥+\_3 𝑅𝑒𝑠𝑡∗𝑆𝑒𝑥+𝜀 - Test for Nested Models: - Do we really need the extra terms in Model B? - i.e. How much do they “add” to Model A?

**Nested F-test** - Want to see how much variability is explained by adding these new values Basic idea: 1. Find how much “extra” variability is explained when the “new” terms being tested are added. 2. Divide by the number of new terms to get a mean square for the new part of the model. 3. Divide this mean square by the MSE for the “full” model to get an F-statistic. 4. Compare to an F-distribution to find a p-value.

**Nested F-test** Test: Ho: Bi=0 for a “subset” of predictors Ha: Bi != 0 for some predictors in the subset - F = ((SSModelFull - SSModelReduced)/# Predictors)/MSEFull - F = ((Explained by Full model - Explained by reduced model)/predictors tested in Ho)/ based on full model - Compared to a f distribution

**Nested F-test** 𝐴𝑐𝑡𝑖𝑣𝑒 =𝛽\_0+𝛽1𝑅𝑒𝑠𝑡+B\_2 𝑆𝑒𝑥+ B3𝑅𝑒𝑠𝑡𝑆𝑒𝑥 +𝜀 H0: β2=β3=0 Ha: Some βi≠0 -Compare mean square for the “extra” variability to the mean square error for the full model.

**Nested F-test Code Example**

modelPint=lm(Active~Rest+Sex+Rest\*Sex, data=Pulse) # Total model;   
# Predict active heart rate by resting rate, sex and the interaction bt rest and sex   
# including the interaction term makes sure that we don't assume that rest and sex have the same slope and intercept   
summary(modelPint)

##   
## Call:  
## lm(formula = Active ~ Rest + Sex + Rest \* Sex, data = Pulse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.822 -9.251 -2.893 6.784 67.396   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.43987 7.47902 1.262 0.208   
## Rest 1.14319 0.11264 10.149 <2e-16 \*\*\*  
## Sex -0.28717 10.22830 -0.028 0.978   
## Rest:Sex 0.03907 0.15130 0.258 0.796   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.17 on 371 degrees of freedom  
## Multiple R-squared: 0.4056, Adjusted R-squared: 0.4008   
## F-statistic: 84.37 on 3 and 371 DF, p-value: < 2.2e-16

# If we want to test to see if adding sex to see if the slope adn itnercept are different, we want to comapre to one without sex and the interaction   
modelP\_Reduced = lm(Active~Rest, data=Pulse)  
  
# This compares the two models to tell us if the interaction and sex term are significant in our model   
anova(modelP\_Reduced, modelPint)

## Analysis of Variance Table  
##   
## Model 1: Active ~ Rest  
## Model 2: Active ~ Rest + Sex + Rest \* Sex  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 373 75050   
## 2 371 74538 2 512.14 1.2746 0.2808

# How much extra variability is expalined? Its the difference int eh sum of squares; if that's a big differene, a higher SSqures is better? Yes

# This tells us, individually, if the predictors are significant in our model  
anova455(modelPint)

## ANOVA Table  
## Model: Active ~ Rest + Sex + Rest \* Sex   
##   
## Df Sum Sq Mean Sq F value P(>F)   
## Model 3 50854 16951.5 84.373 < 2.2e-16 \*\*\*  
## Error 371 74538 200.9   
## Total 374 125392   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova455(modelP\_Reduced)

## ANOVA Table  
## Model: Active ~ Rest   
##   
## Df Sum Sq Mean Sq F value P(>F)   
## Model 1 50342 50342 250.2 < 2.2e-16 \*\*\*  
## Error 373 75050 201   
## Total 374 125392   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# The resting term has a sig relationship   
# The itneraciton model, its at 50854 and the other model is 50342  
# Subbing these gives us the additional variability exampled;  
  
# That's SSDif that is below

SS\_diff = anova455(modelPint)[1,2] - anova455(modelP\_Reduced)[1,2]  
SS\_diff # The additional variability explained

## [1] 512.1413

# that's where the 512 is coming from in teh table above, the difference in teh sum of squares  
  
MS\_diff = SS\_diff/(anova455(modelPint)[1,1] - anova455(modelP\_Reduced)[1,1])  
MS\_diff # Means squared difference

## [1] 256.0706

# Divide SS\_dif by the difference in predictors of the model   
# WE want to see what the differences are in teh df   
# 3 - 1 = 2   
  
F\_diff = MS\_diff/anova455(modelPint)[2,3]  
F\_diff # The F value difference

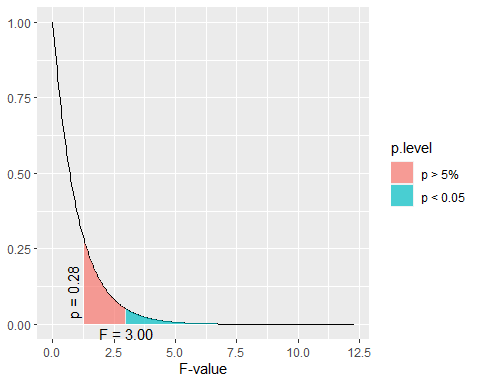
## [1] 1.274553

library(sjPlot)

## Warning: package 'sjPlot' was built under R version 4.1.2

## #refugeeswelcome

dist\_f(f = F\_diff,   
 deg.f1 = anova455(modelPint)[1,1] - anova455(modelP\_Reduced)[1,1],   
 deg.f2 = anova455(modelPint)[2,2],  
 )



# The area under the curve is a pvalue   
# Plots teh f distribution to see graphically how extreme it is   
# We need to tell it the difference of the predictors and the df of the error term (that's what the deg.f1 and f2 are)  
# This graph will vary depending onthe degrees of freedom   
# WE see that the 1.28 is around the p value of 0.28;   
# we would expect ot seet his variation about 28% of the time if there was no useful ness of adding things into the model   
# WE need an f test stat up to 3 to show sig results   
# This tells us that its not beneficial to add these terms to our model bcuase we don't see a stat sig dif bet the two models (using sex to predcit active heart rate)

**Example: State SAT Scores** Source: Statistical Sleuth, Case 12.1 pg. 339  
Response Variable:  
SAT =Average combined SAT Score Potential Predictors:  
Takers = % taking the exam Income = median family income ($100’s) Years = avg. years of study (SS, NS, HU) Public = % public school Expend = spend per student ($100’s) Rank = median class rank of takers

SATModel = lm(SAT~., data=StateSAT[,2:8])  
summary(SATModel)

##   
## Call:  
## lm(formula = SAT ~ ., data = StateSAT[, 2:8])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -60.046 -6.768 0.972 13.947 46.332   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -94.659109 211.509584 -0.448 0.656731   
## Takers -0.480080 0.693711 -0.692 0.492628   
## Income -0.008195 0.152358 -0.054 0.957353   
## Years 22.610082 6.314577 3.581 0.000866 \*\*\*  
## Public -0.464152 0.579104 -0.802 0.427249   
## Expend 2.212005 0.845972 2.615 0.012263 \*   
## Rank 8.476217 2.107807 4.021 0.000230 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 26.34 on 43 degrees of freedom  
## Multiple R-squared: 0.8787, Adjusted R-squared: 0.8618   
## F-statistic: 51.91 on 6 and 43 DF, p-value: < 2.2e-16

# IF we think about polynomial regression we can make a good model with it   
# We have a few good predicotrs here

**R: Best Subsets for StateSAT**

all = regsubsets(SAT~., data=StateSAT[,2:8])  
ShowSubsets(all)

## Takers Income Years Public Expend Rank Rsq adjRsq Cp  
## 1 ( 1 ) \* 77.42 76.95 34.03  
## 2 ( 1 ) \* \* 84.71 84.05 10.22  
## 3 ( 1 ) \* \* \* 87.11 86.27 3.69  
## 4 ( 1 ) \* \* \* \* 87.71 86.61 3.58  
## 5 ( 1 ) \* \* \* \* \* 87.87 86.49 5.00  
## 6 ( 1 ) \* \* \* \* \* \* 87.87 86.18 7.00

# Tells us the best model is

SATModel1 = lm(SAT ~ Years + Expend + Rank, data = StateSAT)  
SATModel2 = lm(SAT ~ Years + Public + Expend + Rank, data = StateSAT)  
summary(SATModel1)

##   
## Call:  
## lm(formula = SAT ~ Years + Expend + Rank, data = StateSAT)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -64.802 -6.798 2.169 17.525 49.706   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -303.7243 97.8415 -3.104 0.00326 \*\*   
## Years 26.0952 5.3894 4.842 1.49e-05 \*\*\*  
## Expend 1.8609 0.6351 2.930 0.00526 \*\*   
## Rank 9.8258 0.5987 16.412 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 26.25 on 46 degrees of freedom  
## Multiple R-squared: 0.8711, Adjusted R-squared: 0.8627   
## F-statistic: 103.6 on 3 and 46 DF, p-value: < 2.2e-16

summary(SATModel2)

##   
## Call:  
## lm(formula = SAT ~ Years + Public + Expend + Rank, data = StateSAT)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -64.931 -5.471 1.932 14.980 43.280   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -204.5982 117.6871 -1.738 0.088963 .   
## Years 21.8905 6.0372 3.626 0.000731 \*\*\*  
## Public -0.6638 0.4500 -1.475 0.147154   
## Expend 2.2416 0.6782 3.305 0.001868 \*\*   
## Rank 10.0032 0.6033 16.581 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 25.93 on 45 degrees of freedom  
## Multiple R-squared: 0.8771, Adjusted R-squared: 0.8661   
## F-statistic: 80.25 on 4 and 45 DF, p-value: < 2.2e-16

# Null: Added coefficients for the added predictors are equal to zero   
# Alternative: At least 1 is nonzero   
# tehre is only 1 added predictor; we are testing that public = 0 vs the alternative that it is nonzero

# Nested test on the things   
# This tells us the same pvaule resut  
# Doing a nested test for the difference with one term in our model is the same as doing those individual tests for slope   
##IMPORTANT ABOVE  
# The below anova is less useful with one term at a time, but it's pretty useful if judgeing multiple terms at a time   
anova(SATModel1, SATModel2)

## Analysis of Variance Table  
##   
## Model 1: SAT ~ Years + Expend + Rank  
## Model 2: SAT ~ Years + Public + Expend + Rank  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 46 31708   
## 2 45 30246 1 1462.5 2.1759 0.1472

**Model Selection with Categorical and Interaction Predictors** Use each of the four model selection methods discussed in class (AllSubsets, Backwards, Forwards, and Stepwise) and compare the processes and outcomes for the predictor pool: Rest, Exercise, Hgt, Wgt, Rest & Exercise, Hgt & Exercise, and Wgt & Exercise

* WE saw in teh past that the regsubsets method wasn’t the best becuase it included things that weren’t as useful
* it picked a chose levels of things when we wanted all of the levels or none of the levels; and it also liked to pick and choose certain interaction terms, some of which were not included in the model
* If you want to include an interaction term, you have to have both terms already in the model

# THis is setting things up  
Full=lm(Active~Rest+Hgt+Wgt+factor(Exercise)+Rest\*factor(Exercise)+ Hgt\*factor(Exercise) + Wgt\*factor(Exercise), data=Pulse)  
none=lm(Active~1,data=Pulse)  
MSE=(summary(Full)$sigma)^2

#Backwards selection  
back\_mod = step(Full,scale=MSE, trace=FALSE)  
back\_mod

##   
## Call:  
## lm(formula = Active ~ Rest + Hgt + Wgt + factor(Exercise) + Hgt:factor(Exercise),   
## data = Pulse)  
##   
## Coefficients:  
## (Intercept) Rest Hgt   
## 84.97301 1.13968 -1.33728   
## Wgt factor(Exercise)2 factor(Exercise)3   
## 0.10212 -4.19657 -70.52397   
## Hgt:factor(Exercise)2 Hgt:factor(Exercise)3   
## 0.09612 1.02785

# Forward selection  
forward\_mod = step(none,scope=list(upper=Full), scale=MSE,direction="forward", trace=FALSE)  
forward\_mod

##   
## Call:  
## lm(formula = Active ~ Rest, data = Pulse)  
##   
## Coefficients:  
## (Intercept) Rest   
## 8.153 1.180

# Stepwise selection  
step\_mod = step(none,scope=list(upper=Full),scale=MSE, trace=FALSE)  
step\_mod

##   
## Call:  
## lm(formula = Active ~ Rest, data = Pulse)  
##   
## Coefficients:  
## (Intercept) Rest   
## 8.153 1.180

# Comaring the nested backwards selection model to the stepwise selection method  
# IF we look at the nested test values of these   
# Ho: At there is no difference between the models   
# Ha: At least one variable is non zero   
# Do we have sig evidence that at least one of these predictors coefficient is non zero?   
anova(back\_mod, step\_mod)

## Analysis of Variance Table  
##   
## Model 1: Active ~ Rest + Hgt + Wgt + factor(Exercise) + Hgt:factor(Exercise)  
## Model 2: Active ~ Rest  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 367 71441   
## 2 373 75050 -6 -3608.9 3.0899 0.005788 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# There are 6 predictors different from the two   
# The factor exercise has 2 additional dummy variables and the interactio nhas 2 addiitonal variables   
# is the coeff for these 6 extra terms equal to zero or evidence that non zero   
# Small pvalue, evidence that at least 1 is non zero   
# mallow Cps may not fit for this model, we might have a lower mallow Cp for rest, it slooks like its a sig imporvement ot add these different criteria to it   
  
#It is an addiitonal tool to build a bigger model

## STOR 455 - Class 20 – new predictors from old

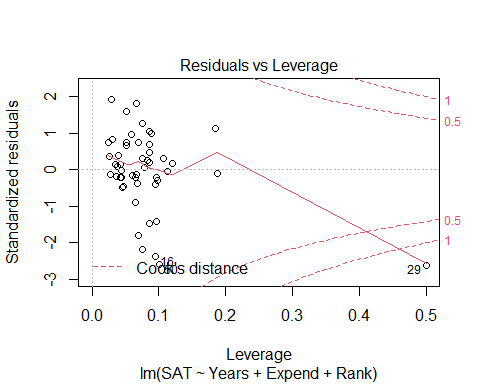
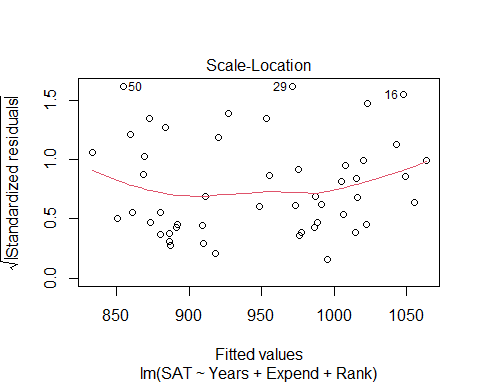
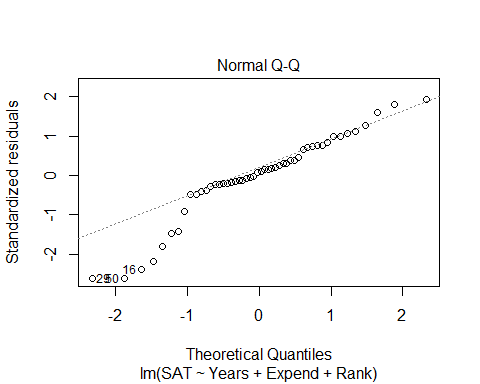
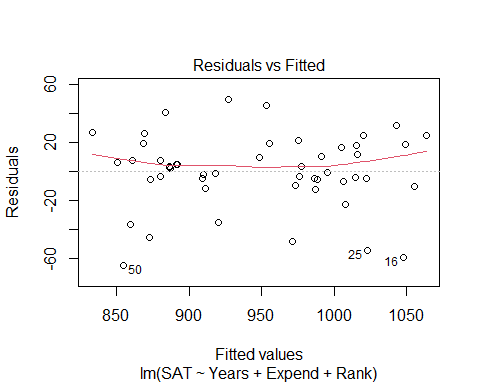
library(readr)  
library(leaps)  
  
StateSAT <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/StateSAT.csv")  
  
source("https://raw.githubusercontent.com/JA-McLean/STOR455/master/scripts/ShowSubsets.R")

**Example: State SAT** - Model #1: Y=SAT vs. X=Takers

mod = lm(SAT~Years+Expend+Rank, data=StateSAT)  
summary(mod)

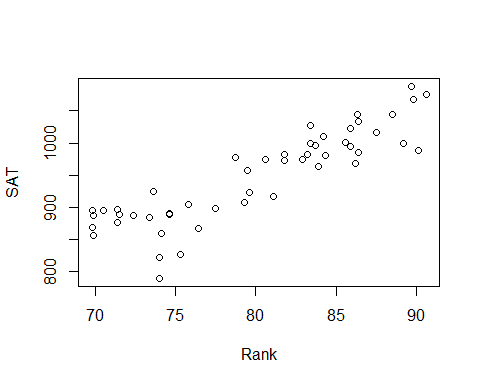
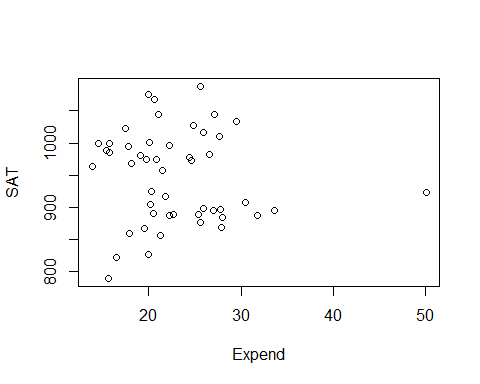
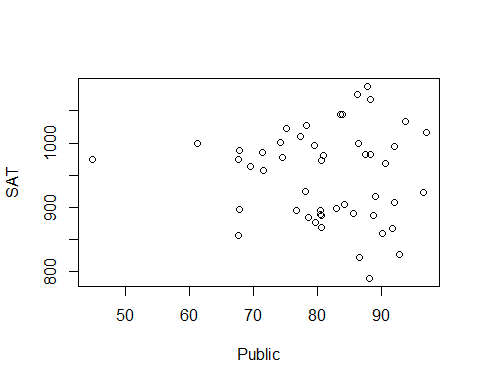
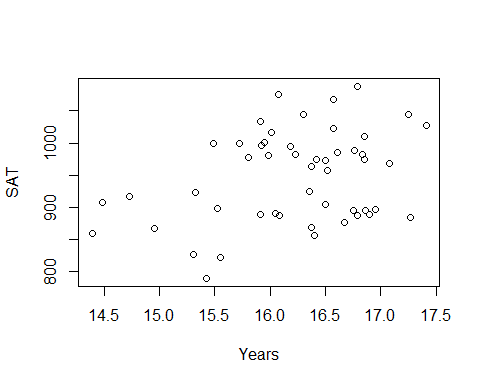
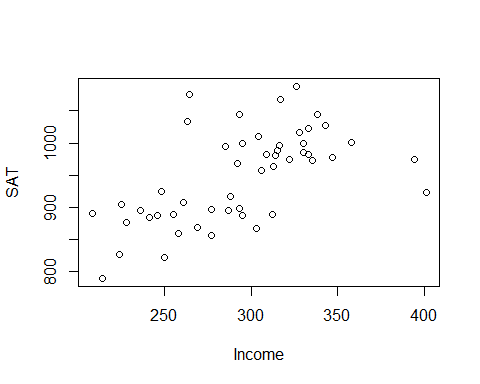
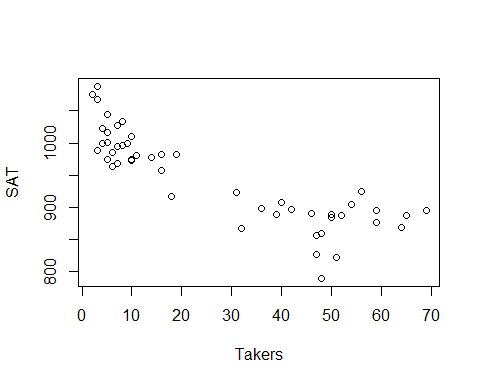
##   
## Call:  
## lm(formula = SAT ~ Years + Expend + Rank, data = StateSAT)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -64.802 -6.798 2.169 17.525 49.706   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -303.7243 97.8415 -3.104 0.00326 \*\*   
## Years 26.0952 5.3894 4.842 1.49e-05 \*\*\*  
## Expend 1.8609 0.6351 2.930 0.00526 \*\*   
## Rank 9.8258 0.5987 16.412 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 26.25 on 46 degrees of freedom  
## Multiple R-squared: 0.8711, Adjusted R-squared: 0.8627   
## F-statistic: 103.6 on 3 and 46 DF, p-value: < 2.2e-16

plot(mod)



# We have a curve in teh residual plot   
# The noramility is an issue   
# The one state has a lot of influence, WE think it's AK, that has a fewer precentage of the population in public schools

plot(SAT~., data=StateSAT[2:8])

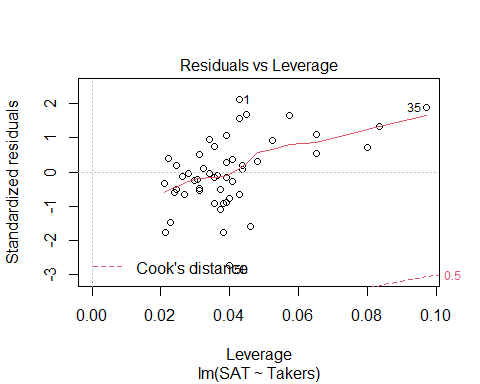
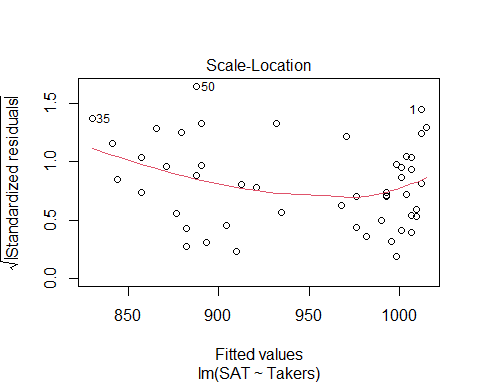
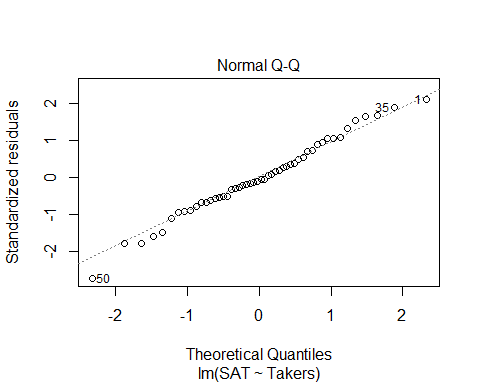
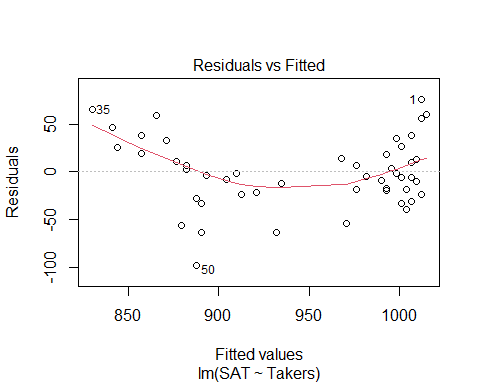


# Here, we see the correlation between each of the variables in the dataset (each of teh numerical variables in teh dataset)   
# Expend is iffy   
# Years is good  
# Takers looks like the best pattern, but we dont have it because it doesn't look lienar

# LEts look at just the taker's variable   
modSAT1 = lm(SAT~Takers, data=StateSAT)  
summary(modSAT1)

##   
## Call:  
## lm(formula = SAT ~ Takers, data = StateSAT)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -97.828 -21.387 -2.628 23.881 75.974   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1020.3062 8.1391 125.36 < 2e-16 \*\*\*  
## Takers -2.7600 0.2387 -11.56 1.77e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 36.8 on 48 degrees of freedom  
## Multiple R-squared: 0.7358, Adjusted R-squared: 0.7303   
## F-statistic: 133.7 on 1 and 48 DF, p-value: 1.768e-15

plot(modSAT1)



# Small pvlaue and high variability is described   
# Linearity condition is super messed up though, so what can we do? - Transofmrations!

**Polynomial Regression** For a single predictor X: 𝑌=𝛽\_𝑜+𝛽\_1 𝑋+𝛽\_2 𝑋^2+⋯+𝛽\_𝑝 𝑋^𝑝+𝜀

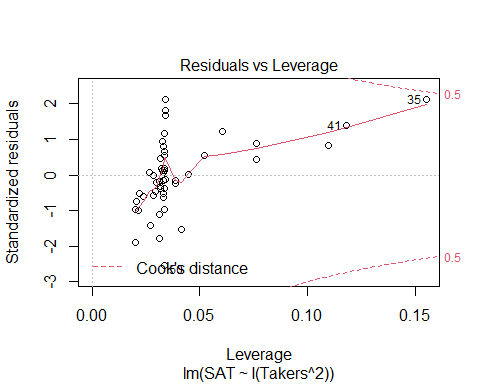
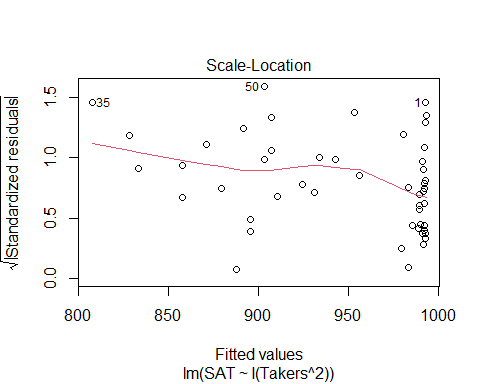
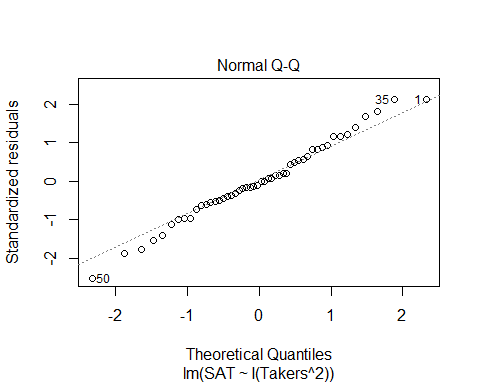
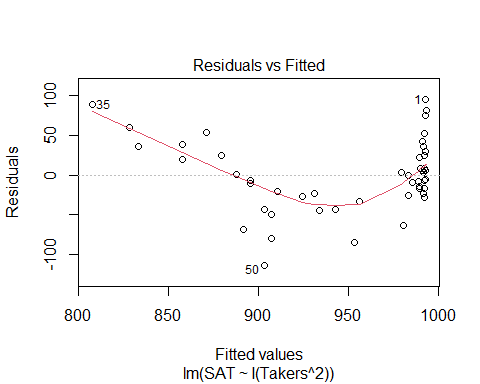
* **LINEAR** – 𝑌=𝛽\_𝑜+𝛽\_1 𝑋+𝜀
* **QUADRATIC** – 𝑌=𝛽\_𝑜+𝛽\_1 𝑋+𝛽\_2 𝑋^2+𝜀
* **CUBIC** –𝑌=𝛽\_𝑜+𝛽\_1 𝑋+𝛽\_2 𝑋^2+𝛽\_3 𝑋^3+𝜀

**Issues with Polynomial Regressionn** - We can move it up and down based on the intercept or make it widder or thinner based on the slope; - We can’t change where the vertex is

#What if we raise taker's the the 2nd power?   
# We have to insolute it so R will actually do it   
# THis made it worse, so to shift it to the right, we have to use a quadratic regression line   
modSAT2 = lm(SAT~I(Takers^2), data=StateSAT)  
summary(modSAT2)

##   
## Call:  
## lm(formula = SAT ~ I(Takers^2), data = StateSAT)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -113.361 -24.883 -2.685 28.102 94.990   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 993.361226 8.422344 117.944 < 2e-16 \*\*\*  
## I(Takers^2) -0.039063 0.004659 -8.385 5.81e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 45.6 on 48 degrees of freedom  
## Multiple R-squared: 0.5943, Adjusted R-squared: 0.5858   
## F-statistic: 70.3 on 1 and 48 DF, p-value: 5.811e-11

plot(modSAT2)



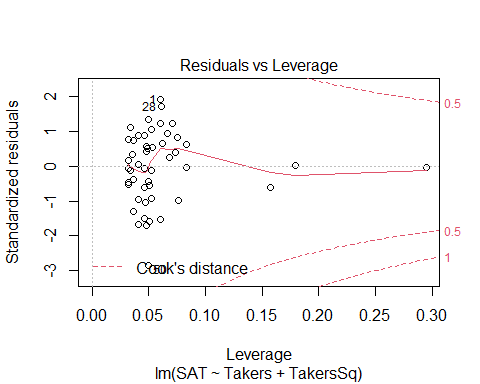
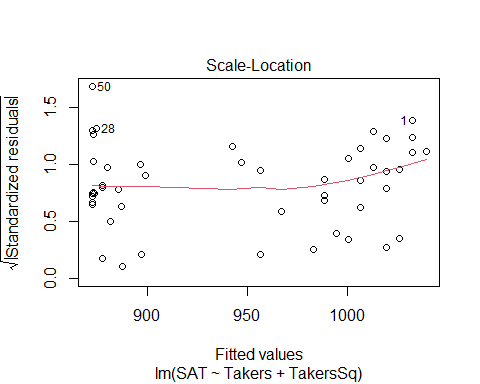
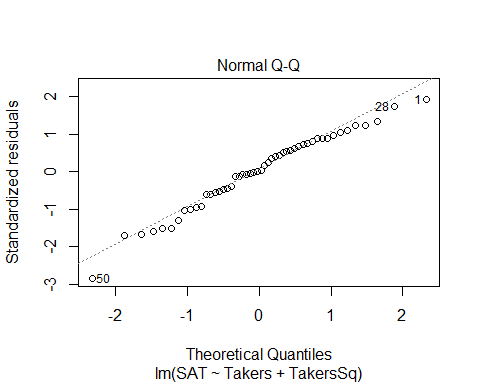
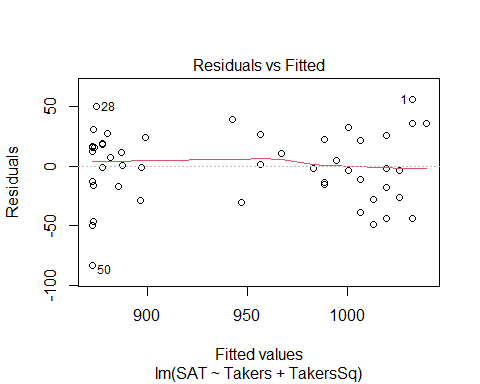
**Polynomial Regression in R** - We can add as many powers we want but then we might be over fitting, so that’s not always best

Method #1: Create new variables with predictor powers. - Create a new model of takers^2 - Use to shift to the right or left

StateSAT$TakersSq = StateSAT$Takers^2  
# ameks a new column of takers^2  
  
modSATquad1 = lm(SAT~Takers + TakersSq, data=StateSAT)  
summary(modSATquad1)

##   
## Call:  
## lm(formula = SAT ~ Takers + TakersSq, data = StateSAT)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -83.015 -16.636 0.783 22.167 55.714   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1053.13112 9.27372 113.561 < 2e-16 \*\*\*  
## Takers -7.16159 0.89220 -8.027 2.32e-10 \*\*\*  
## TakersSq 0.07102 0.01405 5.055 6.99e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 29.93 on 47 degrees of freedom  
## Multiple R-squared: 0.8289, Adjusted R-squared: 0.8216   
## F-statistic: 113.8 on 2 and 47 DF, p-value: < 2.2e-16

# Above, makes a model with takers^2  
# We see that it's a pretty sig model   
# It looks good, but does it help with the residuals?   
  
plot(modSATquad1)



# The linearity looks pretty good   
# Constance variance could be better because we dont have a lot of data   
# the normal, looks pretty good too

**Polynomial Regression in R** Method #2: Use I( )in the lm( ) - Does the same thing as a bove, but it does it with just the insulate function

# Quadratic model for SAT  
# (𝑆𝐴𝑇)̂=1053.1−7.1616𝑇𝑎𝑘𝑒𝑟𝑠+0.0710〖𝑇𝑎𝑘𝑒𝑟𝑠〗^2  
modSATquad2 = lm(SAT~ Takers+ I(Takers^2), data=StateSAT)  
summary(modSATquad2)

##   
## Call:  
## lm(formula = SAT ~ Takers + I(Takers^2), data = StateSAT)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -83.015 -16.636 0.783 22.167 55.714   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1053.13112 9.27372 113.561 < 2e-16 \*\*\*  
## Takers -7.16159 0.89220 -8.027 2.32e-10 \*\*\*  
## I(Takers^2) 0.07102 0.01405 5.055 6.99e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 29.93 on 47 degrees of freedom  
## Multiple R-squared: 0.8289, Adjusted R-squared: 0.8216   
## F-statistic: 113.8 on 2 and 47 DF, p-value: < 2.2e-16

**Polynomial Regression in R** Method #3: Use poly - Does the same thing as the other methods, but it just tells it to make a polynomial - This will be treated as one unit instead of separately

modSATquad3 = lm(SAT~poly(Takers, degree=2, raw=TRUE), data=StateSAT) # 2 = quadratic   
summary(modSATquad3)

##   
## Call:  
## lm(formula = SAT ~ poly(Takers, degree = 2, raw = TRUE), data = StateSAT)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -83.015 -16.636 0.783 22.167 55.714   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1053.13112 9.27372 113.561 < 2e-16  
## poly(Takers, degree = 2, raw = TRUE)1 -7.16159 0.89220 -8.027 2.32e-10  
## poly(Takers, degree = 2, raw = TRUE)2 0.07102 0.01405 5.055 6.99e-06  
##   
## (Intercept) \*\*\*  
## poly(Takers, degree = 2, raw = TRUE)1 \*\*\*  
## poly(Takers, degree = 2, raw = TRUE)2 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 29.93 on 47 degrees of freedom  
## Multiple R-squared: 0.8289, Adjusted R-squared: 0.8216   
## F-statistic: 113.8 on 2 and 47 DF, p-value: < 2.2e-16

# Same values

#ANOVA TREATS THE DIFFERENT METHODS DIFFERENTLY  
anova(modSATquad1)

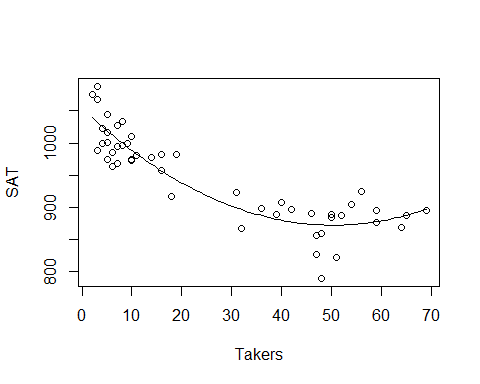
## Analysis of Variance Table  
##   
## Response: SAT  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Takers 1 181024 181024 202.089 < 2.2e-16 \*\*\*  
## TakersSq 1 22886 22886 25.549 6.992e-06 \*\*\*  
## Residuals 47 42101 896   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Looks at takers vs takers^2  
# tells you adding the squared term is useful for us   
anova(modSATquad3)

## Analysis of Variance Table  
##   
## Response: SAT  
## Df Sum Sq Mean Sq F value Pr(>F)   
## poly(Takers, degree = 2, raw = TRUE) 2 203910 101955 113.82 < 2.2e-16 \*\*\*  
## Residuals 47 42101 896   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Condesnes the terms into one   
# Jsut a test on the one model  
  
# Doing the same things, but the function treats it differently depending on the method you use

# Quadratic model for SAT  
plot(SAT~Takers, data=StateSAT) # Plot raw data  
  
# Pull out teh coeff for the terms for the quadratic model   
B0\_modSATquad2 = summary(modSATquad2)$coef[1,1]  
B1\_modSATquad2 = summary(modSATquad2)$coef[2,1]  
B2\_modSATquad2 = summary(modSATquad2)$coef[3,1]  
  
# curve(INtercept, coef\*x, coef\*x^2, add = TRUE)  
curve(B0\_modSATquad2 + B1\_modSATquad2\*x + B2\_modSATquad2\*x^2, add=TRUE)



# Looks like it fits really well

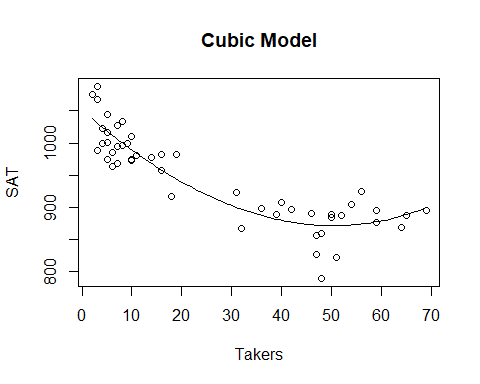
Would a Cubic work better?

#Cubic MOdel  
modSATcubic = lm(SAT~ Takers+ I(Takers^2) + I(Takers^3), data=StateSAT)  
summary(modSATcubic)

##   
## Call:  
## lm(formula = SAT ~ Takers + I(Takers^2) + I(Takers^3), data = StateSAT)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -82.267 -17.192 -0.321 21.610 56.676   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.051e+03 1.452e+01 72.366 < 2e-16 \*\*\*  
## Takers -6.753e+00 2.380e+00 -2.837 0.00676 \*\*   
## I(Takers^2) 5.631e-02 8.051e-02 0.699 0.48777   
## I(Takers^3) 1.408e-04 7.586e-04 0.186 0.85353   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 30.24 on 46 degrees of freedom  
## Multiple R-squared: 0.829, Adjusted R-squared: 0.8178   
## F-statistic: 74.33 on 3 and 46 DF, p-value: < 2.2e-16

# R and rsqaured are simular   
# Models of dif predictors, adj r squared is better measure   
# Its a little worse than a 2 model   
# High p value   
# tells us that not as sig

# Cubic MOdel  
plot(SAT~Takers, data=StateSAT, main="Cubic Model")  
  
B0\_modSATcubic = summary(modSATcubic)$coef[1,1]  
B1\_modSATcubic = summary(modSATcubic)$coef[2,1]  
B2\_modSATcubic = summary(modSATcubic)$coef[3,1]  
B3\_modSATcubic = summary(modSATcubic)$coef[4,1]  
  
curve(B0\_modSATcubic + B1\_modSATcubic\*x + B2\_modSATcubic\*x^2 + B3\_modSATcubic\*x^3, add=TRUE)



# Doesn't look super differnet   
# In the end, there's not a lot of change witht eh cube term   
# IT has a small coef compared to the others as well, so so not super big influence out the gate but we dont know if its a tually influencital we would haev to check other htings

anova(modSATcubic)

## Analysis of Variance Table  
##   
## Response: SAT  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Takers 1 181024 181024 197.9375 < 2.2e-16 \*\*\*  
## I(Takers^2) 1 22886 22886 25.0241 8.72e-06 \*\*\*  
## I(Takers^3) 1 32 32 0.0345 0.8535   
## Residuals 46 42069 915   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Tells you if all the other models are sig  
# Tells us that takers to the 3rd isn't useful

car::vif(modSATcubic)

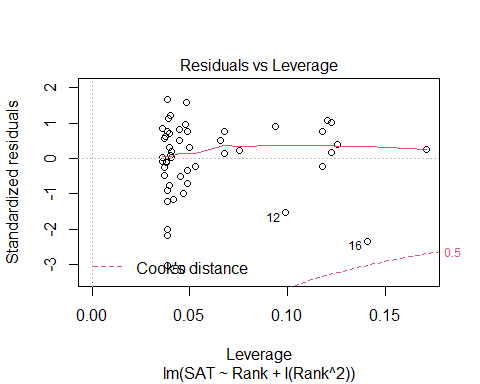
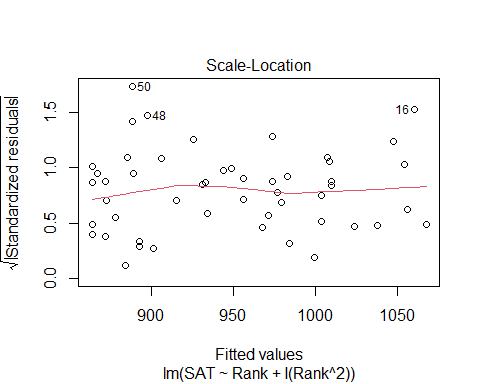
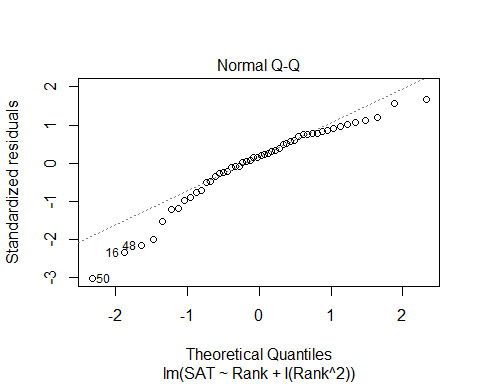
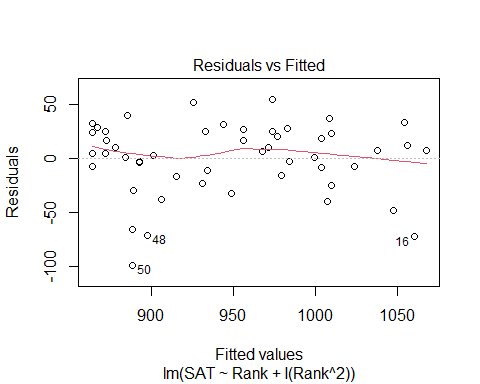
## Takers I(Takers^2) I(Takers^3)   
## 147.2369 678.9666 222.6922

# shows that there is a high multicollinearity with takers

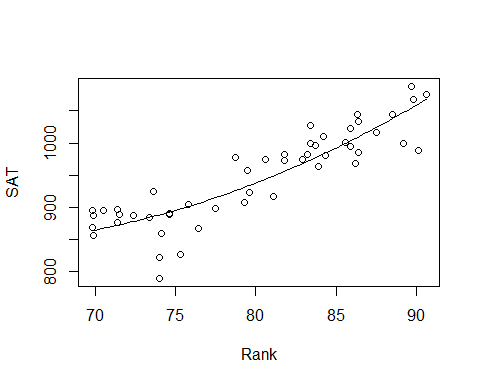
modSATquad4 = lm(SAT~ Rank+ I(Rank^2), data=StateSAT)  
summary(modSATquad4)

##   
## Call:  
## lm(formula = SAT ~ Rank + I(Rank^2), data = StateSAT)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -98.531 -14.457 5.853 24.304 54.192   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1692.9133 855.7237 1.978 0.0538 .  
## Rank -28.5644 21.5731 -1.324 0.1919   
## I(Rank^2) 0.2391 0.1352 1.768 0.0835 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 33.29 on 47 degrees of freedom  
## Multiple R-squared: 0.7883, Adjusted R-squared: 0.7793   
## F-statistic: 87.52 on 2 and 47 DF, p-value: < 2.2e-16

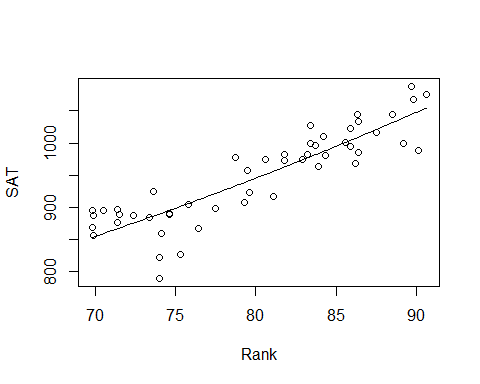
# Model looks pretty good, rank and rank^@ are not sig; btut ehre is a hug amount of multicollinearity   
# P value is close to 0, but the individual tests say things different because of the multicollinearity   
  
plot(modSATquad4)



# Residual anaysis isn't too bad  
# Normal is a little bit of an issue  
   
plot(SAT~Rank, data=StateSAT)  
# Slight curve when we raise to a power   
# Not a linear relationship   
# We could fit a line to it, but there may be some issues   
# THis is the more ideal situation than a line   
  
B0\_modSATquad4 = summary(modSATquad4)$coef[1,1]  
B1\_modSATquad4 = summary(modSATquad4)$coef[2,1]  
B2\_modSATquad4 = summary(modSATquad4)$coef[3,1]  
  
curve(B0\_modSATquad4 + B1\_modSATquad4\*x + B2\_modSATquad4\*x^2, add=TRUE)



# IF we jsut used a squared rank term, then we would   
# IT ssimilar, but not quite just right  
# WE have to flatten the parabola out more   
# Still centered at the zero, jsut stretched further and its less useful   
plot(SAT~Rank, data=StateSAT)  
mod2 = lm(SAT~I(Rank^2), data=StateSAT)  
  
B0\_mod2 = summary(mod2)$coef[1,1]  
B1\_mod2 = summary(mod2)$coef[2,1]  
  
curve(B0\_mod2 + B1\_mod2\*x^2, add=TRUE)



**Polynomial with one predictor** - We can use different order models that look at other models in a 3D space for predictors

**Second Order Models** Definition: A second order model for two quantitative predictors would be 𝑌=𝛽\_𝑜+𝛽\_1 𝑋\_1+𝛽\_2 𝑋\_2+𝛽\_3 𝑋\_1^2+𝛽\_4 𝑋\_2^2+𝛽\_5 𝑋\_1 𝑋\_2+𝜀 Y = INtercept + First Order + First Order + Quadratic + Quadratic + INteraction, where Wuadratic+ INTeraction = Second order

Example: Try a full second order model for Y=SAT using X1=Takers and X2=Expend

## STOR 455 - Class 21 - R Cross Validation

library(readr)  
StateSAT <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/StateSAT.csv")  
Pulse <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/Pulse.csv")

**Example: State SAT Scores** Source: Statistical Sleuth, Case 12.1 pg. 339  
Response Variable:  
SAT =Average combined SAT Score Potential Predictors:  
Takers = % taking the exam Income = median family income ($100’s) Years = avg. years of study (SS, NS, HU) Public = % public school Expend = spend per student ($100’s) Rank = median class rank of takers

**Second Order Models** - Definition: A second order model for two quantitative predictors would be Y = Bo + First Order + First Order + Quadratic + Quadratic + Interaction, where Quadratic + Quadratic \_ Interaction = Second ORder - Y = B0 + B1X1 + B2X2 + B3X1^2 + B4X2^2 + B5X1\*X2 + Error - Example: Try a full second order model for Y=SAT using X1=Takers and X2=Expend

*Previously on* - We did with taker’s vsariable - We looked at quadratic, transformation alone wasnt good enoughb/c we can’t change vertex - Cubic didn’t give us much there - we could keep going, but we’re probabbly going to overfit too much if we kept going *Polynomial regression in other dimensions* - It can extend to other dimensions - When we have 2 dimentions, it’s a plane, but as we add more, then it’s not technically a line, but we call it a line - A curved plane in 3D, we are building this polynomial with 2 extra predictors - We are looking at the terms themselves and also looking at the squared terms themsevles adnthe interaction between them (this is the second order model we are looking at)

# Full SEcond ORder Model   
modSAT2ndorder=lm(SAT~Takers+Expend+I(Takers^2)+I(Expend^2)+I(Takers\*Expend),data=StateSAT)  
summary(modSAT2ndorder)

##   
## Call:  
## lm(formula = SAT ~ Takers + Expend + I(Takers^2) + I(Expend^2) +   
## I(Takers \* Expend), data = StateSAT)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -50.472 -13.535 1.023 8.866 60.870   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 893.66283 36.14094 24.727 < 2e-16 \*\*\*  
## Takers -7.05561 0.83740 -8.426 9.96e-11 \*\*\*  
## Expend 10.33333 2.49600 4.140 0.000155 \*\*\*  
## I(Takers^2) 0.07725 0.01328 5.816 6.28e-07 \*\*\*  
## I(Expend^2) -0.11775 0.04426 -2.660 0.010851 \*   
## I(Takers \* Expend) -0.03344 0.03716 -0.900 0.373087   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 23.68 on 44 degrees of freedom  
## Multiple R-squared: 0.8997, Adjusted R-squared: 0.8883   
## F-statistic: 78.96 on 5 and 44 DF, p-value: < 2.2e-16

# Summary at bottom tells us the overall anova test   
# Ho: coef of all the coef are zero vs the Ha: at least one is non zero <- for the ANOVA test   
# Low pvalue - There is a relationship with at least one of these predictors   
# Looking at the individual predictors and there are many that are very low   
# The itneraction term is now a sig pvalue   
# FOr a model like this for polynomial regression or the compelte second order model, when we are looking at adding categorical variables, we don't want ot pick and choose different parts of this model and keep them   
# We want either a full polynomial model with all the degrees of the predictor in it or some other model   
# Even though the interaction term isn't sig, we don't want to get rid of it   
  
# Residual analysis   
# Linearity looks good   
# Normal looks pretty good, some devation but pretty good   
  
# overall: Seems like a good model   
  
# Adj R squared; almost 90% of the varability is predicted by this model   
# One warning error: There are some weird things with the plot, but don't worry about that; AK is just giving us a leverage warning   
  
# Do we need the interaction term?   
# YEs, if we have the other htings   
  
# Do we need the second order terms?  
# Are any useful? Takers and linear parts of it, and the squared value   
# Do the extra terms give us much improvement vs a model without it? We can't tell that from teh summary table, we have to do some drop in deviance test to tell that or a nested test   
  
# Do we need the terms with Expend?  
# A model with just tackers was a good model, and that is a subset of this one, so do we need the extra things?

**NEsted test to see if we don’t need the additional terms**

modSAT2ndorder\_Reduced=lm(SAT~Takers+Expend+I(Takers^2)+I(Expend^2),data=StateSAT)  
anova(modSAT2ndorder\_Reduced, modSAT2ndorder) # named mod2 in the lecutre, may refer to as mod2

## Analysis of Variance Table  
##   
## Model 1: SAT ~ Takers + Expend + I(Takers^2) + I(Expend^2)  
## Model 2: SAT ~ Takers + Expend + I(Takers^2) + I(Expend^2) + I(Takers \*   
## Expend)  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 45 25123   
## 2 44 24669 1 454 0.8098 0.3731

# If wanted to see the additional terms int eh second order, we can do an anova test with teh mod2 and the full model from the previoud chunck   
# THis is a hypo test askign if any of the additional second order model terms are useful for us   
# Null: Is the coef for takers squared, explanatory square and the itnearction all equal to zero   
# Alternative: Do we have evidence that they are not equal to zero? (Meaning that there is some sig)  
# Low pvalue - at least one is non zero   
# The full second order mdoel looks like an improvemnet instead of jsut using a linear model   
  
# What if we build a model with jsut the taker's terms in it?   
modSAT2ndorder\_Reduced2=lm(SAT~Takers+Expend, data=StateSAT)  
anova(modSAT2ndorder\_Reduced2,modSAT2ndorder)

## Analysis of Variance Table  
##   
## Model 1: SAT ~ Takers + Expend  
## Model 2: SAT ~ Takers + Expend + I(Takers^2) + I(Expend^2) + I(Takers \*   
## Expend)  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 47 56278   
## 2 44 24669 3 31609 18.793 5.404e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# The above anova test checks that   
# Ho: Is the coef for expend, expend^2, and interaction = 0   
# HA: Is coef for at least one != 0   
# Small pvalue, at least one is nonzero   
# This might be the more ideal model   
# WE shoudl also check the lienar model conditions  
  
# Linear model conditions are good   
modSAT2ndorder\_Reduced3 = lm(SAT~Takers+I(Takers^2),data=StateSAT)  
anova(modSAT2ndorder\_Reduced3, modSAT2ndorder)

## Analysis of Variance Table  
##   
## Model 1: SAT ~ Takers + I(Takers^2)  
## Model 2: SAT ~ Takers + Expend + I(Takers^2) + I(Expend^2) + I(Takers \*   
## Expend)  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 47 42101   
## 2 44 24669 3 17432 10.364 2.787e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# not going to look at very much, but this really is so you cna use it if you don't want to or transofmration just aren't doing it

**Would our model do a good job at predicting future if we got a different set of data?**

**Cross Validation** - Concern: A model may reflect the structure of a particular sample, but not generalize well to the population. - Cross validation checks for overfitting

**To see if this is a problem:** Split the original sample into two parts (a) A “training” sample to build a model (b) A “holdout” sample to test the model

1. Build model on a training set,

* subset it to build the model
* keep a holdout testing sampel to see how well the model does with teh new data
* sometimes you will have to break it up yourself

**Example: Pulse Rates** Response Variable:  
Active pulse Predictors:  
Resting pulse Hgt Wgt Sex (0=M, 1=F) Smoke (0=No, 1=Yes) Exercise (1=Slight, 2=Moderate, 3=Lots)

**Example: Active Pulse Rates** Training sample – first 300 cases (PulseTrain) Holdout sample – #301-375 (PulseHoldout)

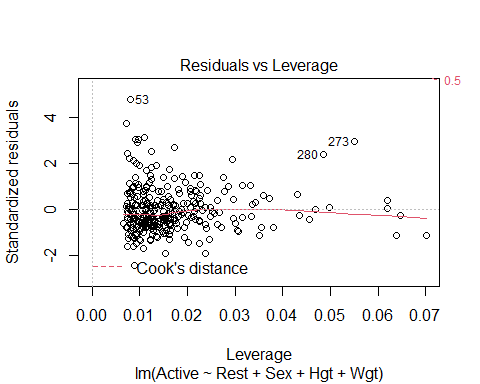
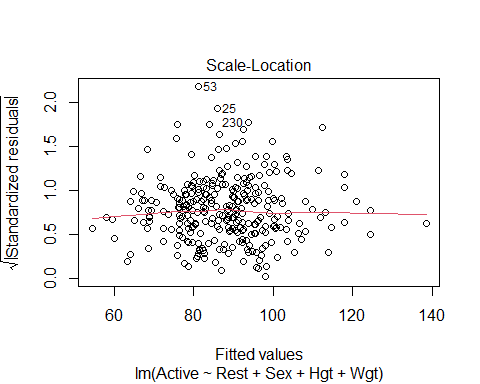
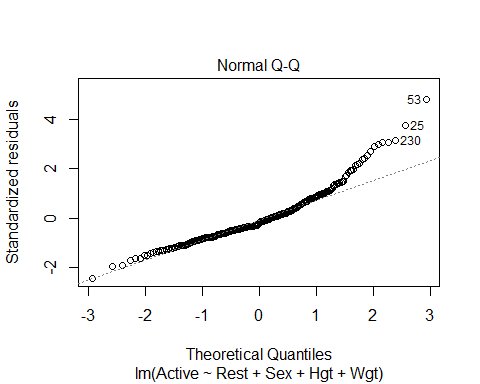
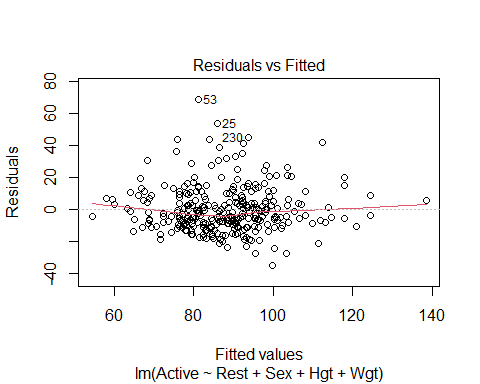
set.seed(12345) # Only set the seed to get the same results at the end, if you want actual random, then don't set seed  
# Want to do it randomly because we dont want to have a certain connection with the different rows  
rows <- sample(nrow(Pulse)) # Counts rows in pulse and takes 374 values without repleacement in a random order  
Pulse\_shuffled = Pulse[rows,] # reassign pulse in a different order here   
  
PulseTrain=Pulse\_shuffled[1:300,] # Make the training data, took about 75% of the data into the training   
PulseHoldout=Pulse\_shuffled[301:375,] # Holdout

What is the best model to predict Active pulse?

#"best" model  
# Ran a model selection sequence, we don't really care how we got this for the thing for this example  
PulseTrainMod=lm(Active~Rest+Sex+Hgt+Wgt,data=PulseTrain)  
  
summary(PulseTrainMod)

##   
## Call:  
## lm(formula = Active ~ Rest + Sex + Hgt + Wgt, data = PulseTrain)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -34.869 -8.916 -2.794 6.515 68.782   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 33.70967 23.49083 1.435 0.15234   
## Rest 1.19626 0.08650 13.830 < 2e-16 \*\*\*  
## Sex 4.30152 2.51881 1.708 0.08873 .   
## Hgt -0.69392 0.35542 -1.952 0.05184 .   
## Wgt 0.11892 0.04128 2.881 0.00426 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.44 on 295 degrees of freedom  
## Multiple R-squared: 0.4216, Adjusted R-squared: 0.4137   
## F-statistic: 53.75 on 4 and 295 DF, p-value: < 2.2e-16

plot(PulseTrainMod) # WE dont want to touch the holdout sample, not until the end



# Looking at the model   
# SUmmary   
# ANOVA test gives us evident that some of the things are nonzero   
# Height isn't good on its own - probably multicollinearity going on here   
# Things correlated with height and all that   
  
# Model conditions   
# Some curve that could be an ainssue, we could flatten it out with transformations, but we wont mess with that now   
# WE could have issues with constant variance, but we wont mess with either   
# Normal is an issue, skew on the right side

* We want to use the model made to predict the data in teh test data **Fit for Training Model** Another way to think of R2: 𝑅2 = square of correlation between 𝑌 and 𝑌̂
* If we use this model to predict all the values in the new data set, we can find the cross validation correlation
* how well correlated the actual values are predicted vs how re predict they are with the model constructed from the old data
* WE’re coming the R2 for the other models with the new model

**How does the training model work for the holdout sample?**  - Compute predicted values for the holdout sample using the fitted prediction equation from the training. – fitActive=predict(mod,newdata=PulseHoldout) - Compute the residuals for the holdout sample. – holdoutresid=PulseHoldout$Active - fitActive

We want to see how far off we are from teh testing data to the holdout sample

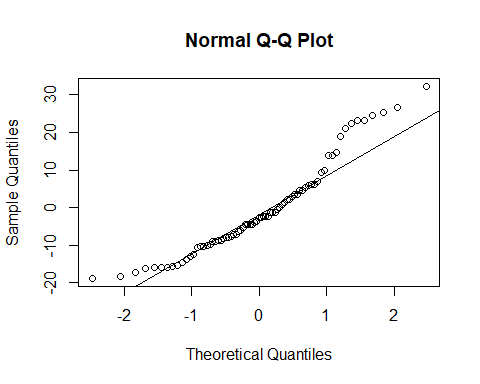
#predict active heart rates for data in the holdout sample with the model made from the training data  
fitActive=predict(PulseTrainMod,newdata=PulseHoldout)  
# Just all teh predctions   
  
#Actual active heart rates in holdout sample minus their predicted values  
# how far off are we?   
holdoutresid=PulseHoldout$Active - fitActive  
# Looks at the residuals; so the actual - the predicted   
  
# WE saw the linear was not prefect in teh past; does this model predict data in a simialr way   
# Is there as imilar center, and spread and the Residual SE the same?   
# IS it distributed nmuch the same way?   
# IS there the same type of skew in this or is it different?   
#So welook at the center, spread and shape below   
  
#Center, spread, and shape for the distribution of holdouts  
mean(holdoutresid)# Want value to be close to zero; because if its above or below it, there are bias in teh predicitons

## [1] -0.826801

# Big or little is subjective to teh data; this is saying we are off 1 beat per minute, if we are predicting GPA, then 1 point would be really bad and off, but this isn't too bad   
sd(holdoutresid)

## [1] 12.21473

# See how spread out htings are   
# THis says 12.21 doesn't say much on it's own   
# Is the similar spread as the orignal dat?   
# So look at teh orinigal training mod summary, and that is the REsd SE, and we want to see if the REsd SE is similar to teh spread of the SE of the holdout residusla   
# One is 12 and one is 14, so the dat appears more compact, we odnt knwo why, but its okay   
# IT doesn't llook drastically different   
# THis is saything that htere is some difference, but it looks okay   
  
# Distribution plots   
qqnorm(holdoutresid)  
qqline(holdoutresid)



# Above we see that the line looks okay,   
# It's not great, there are tail issues with skew, but its liek the orignial data, so there doesnt look to be mcuh difference in teh shape   
# This tells you should probably fix it in the oringial model   
# Check it with residuals to make sure no drastic differences between teh other predicitons in teh orinal model   
#Want th shape to be similar to the orinigal model

cor(PulseHoldout$Active,fitActive)

## [1] 0.6474953

# Correlation bt holdout and the predictions for the values   
# This tells you there is a 0.64 correlation bt the two values   
# THis means that   
# Compare teh value from a previous model   
# THis is the cross validation ocorreation   
#Look below

#Correlation between predicted and actual active heart rates  
crosscorr=cor(PulseHoldout$Active,fitActive)  
# In previous model the multi r squared did the same thing; it was jsut the square of teh predicted adn actual values   
crosscorr^2

## [1] 0.4192502

# Looking at the difference bt crosscorr^2 and what the oringial model says tells you the shrinkage from the orinigal and the test   
# Want to see how different the predicitons with teh model correlatioe with teh actual and predicted values   
# How well teh actual and predicted values correalte even though this model wasn't use to build the model   
# WE want these to be similar and very close values   
  
#Change in r^2 from the training to the holdout  
shrinkage = summary(PulseTrainMod)$r.squared-crosscorr^2  
shrinkage

## [1] 0.002325805

# Very clsoe to zero, less than 1 percentage  
# Gives a some measure how not as well the oringal model predcits teh new data   
# IF teh thing is big, then you're not predicitng very well for the new data vs the old data   
# If it is big you're probably overfitting the original model   
# You are really only get a value from 0 - 1, under 0.15 is pretty okay, but above 0.2 can be kinda bad   
# YOu could get a negative, which would tell you that you are predicting your stuff better than thte oringial model

## STOR 455 Class 22 Influential Points in Multiple Regression

library(readr)  
library(Stat2Data)  
library(car)

## Loading required package: carData

data("Houses")  
data("Perch")  
  
StateSAT <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/StateSAT.csv")

## Rows: 50 Columns: 8

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (1): State  
## dbl (7): SAT, Takers, Income, Years, Public, Expend, Rank

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

Pulse <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/Pulse.csv")

## Rows: 375 Columns: 7

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## dbl (7): Active, Rest, Smoke, Sex, Exercise, Hgt, Wgt

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

head(Perch)

## Obs Weight Length Width  
## 1 104 5.9 8.8 1.4  
## 2 105 32.0 14.7 2.0  
## 3 106 40.0 16.0 2.4  
## 4 107 51.5 17.2 2.6  
## 5 108 70.0 18.5 2.9  
## 6 109 100.0 19.2 3.3

**Types of “Unusual” Points in SLM** -Outlier: A data point that is far from the regression line. -Influential point: A data point that has a large effect on the regression fit.

How do we measure “far”? How do we measure “effect on the fit”?

* SOme things are the same and some are different when thinking about multiple linear regression
* the outliers are about the same think about
* the influential point are baout the same in the multipel as well,, but keep cook’s distance plot as the thing

**Detecting Unusual Cases - Overview** 1. Compute residuals “raw”, standardized, studentized 2. Plots of residuals (or std. residuals) Boxplot, scatterplot, normal plot 3. Leverage Unusual values for the predictors 4. Cook’s distance Cases with large influence

**Standardized Residuals** - DOESNT CHANGE WITH SIMPLE AND MULTIPLE LINEAR REGRESSION For residuals: mean=0 and std. dev. ≈𝜎^\_𝜀 Standardized Residual ~~ almost = (𝑦\_𝑖−𝑦)/𝜎\_𝜀 - Look for values beyond +/-2 (mild) or beyond +/-3

Definition: The standardized residuals are: 𝑠𝑡𝑑.𝑟𝑒𝑠\_𝑖=(𝑦\_𝑖−𝑦^\_𝑖)/(𝜎^\_𝜀 √(1−ℎ\_𝑖 ))

Hi = Leverage

**Studentized Residuals** - DOESN”T CHANGE WITH SIMPLE AND MUTIPLE LINEAR REGRESSION - Better IDea of influence Definition: The studentized residuals are: stud res = (y-yhat)/(stderror\*sqrt(1-hi)) stederror = using model fit without ith case

**Leverage in Simple Linear Regression** - HARDER TO THINK ABOUT IN MULTIPLE PREDICTIORS - if far of the L orR of the model, then you caould have the odel tilted towards you - Predictors on many axies tilting in different ways - a point could have a lot of leverage on one variable specifically - how does this change?

**For a simple linear model:** ℎ\_𝑖 = 1/𝑛+ 〖(𝑥\_𝑖− 𝑥)〗2/(∑▒〖(𝑥\_𝑖− 𝑥)〗2 )

∑▒ℎ\_𝑖 =∑▒1/𝑛+(∑▒〖(𝑥\_𝑖−𝑥)2 〗)/𝑆𝑆𝑋 = 1+ 1 = 2

* **For multiple linear regression** Look for: ℎ\_𝑖 > 2( 2/𝑛 ) OR ℎ\_𝑖 > 3( 2/𝑛 )

**Leverage in Multiple Regression** For a multiple regression with k predictors: ∑▒ℎ\_𝑖 =𝑘+1 “"Typical" leverage”=(𝑘+1)/𝑛 Look for: hi > (2(k+1))/n OR hi > (3(k+1))/n

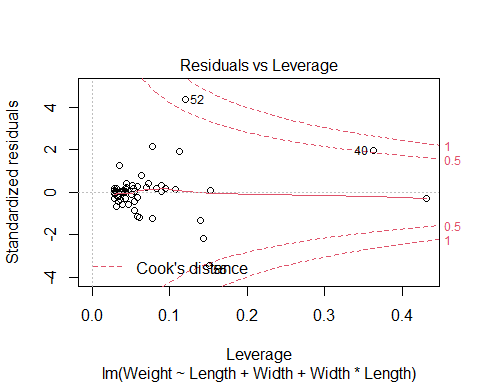
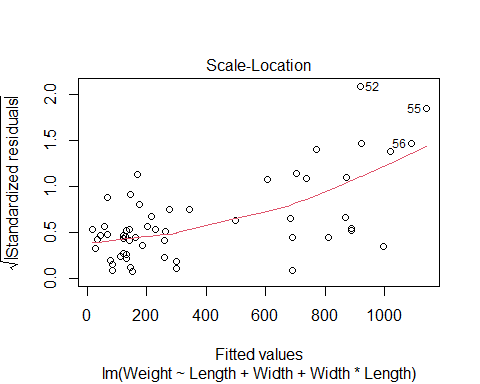
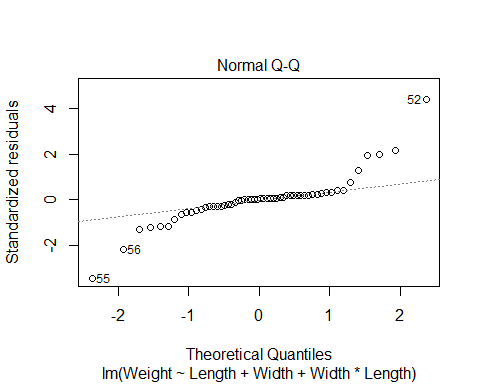
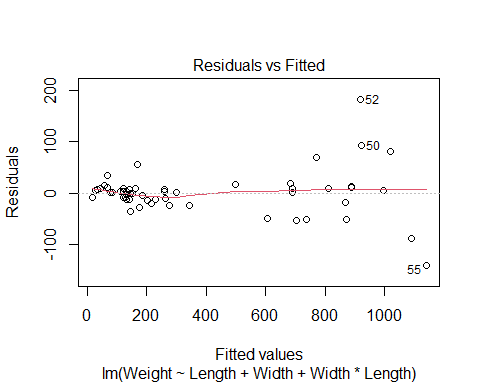
**Leverage in Multiple Regression: Perch** Perch\_lm = lm(Weight~Length+Width+Width\*Length, data=Perch)

* plot(Perch\_lm\*fitted.values) abline(0,0)

# model to predict weight by lenght, wdiget and the interaction between teh two variables  
Perch\_lm = lm(Weight~Length+Width+Width\*Length, data=Perch)  
summary(Perch\_lm)

##   
## Call:  
## lm(formula = Weight ~ Length + Width + Width \* Length, data = Perch)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -140.106 -12.226 1.230 8.489 181.408   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 113.9349 58.7844 1.938 0.058 .   
## Length -3.4827 3.1521 -1.105 0.274   
## Width -94.6309 22.2954 -4.244 9.06e-05 \*\*\*  
## Length:Width 5.2412 0.4131 12.687 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 44.24 on 52 degrees of freedom  
## Multiple R-squared: 0.9847, Adjusted R-squared: 0.9838   
## F-statistic: 1115 on 3 and 52 DF, p-value: < 2.2e-16

# Summary tells us   
# length doesn tappear to be a good preictor   
# The interaction appears sig though; but if we want to keep the interaction term, then we have to keep length   
# Multiple R squared says we are explain 98% in this model   
# WE also need to look at the model conditions so we can plot it   
  
# DO a regression analysis   
  
# Plots for residual analysis  
plot(Perch\_lm)



# Linear looks pretty good   
# Constance variance is bad; fish that are lower appears to be fanning pattern   
# TIghtly packed prediction compred to what is seen # Constance variance is not good   
# Normal; prettuy big issues   
# WE could try and ocrrect this iwth transofmraitons, log the weight would probably help this because it tend to calm right side skews   
# WE want to look at what points are having a lot fo infuence on this model

**Leverage in Multiple Regression: Perch**

# Double and triple the average leverage for 3 predictors  
# Tells you what kind of points have the potential to have leverage   
# These are the big boundaries for leverage   
2\*(3+1)/56

## [1] 0.1428571

3\*(3+1)/56

## [1] 0.2142857

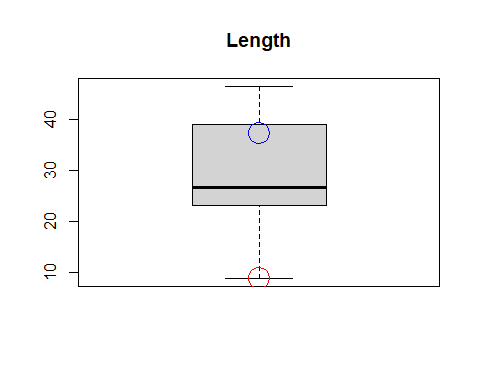
# Tells us which values in the dataset are over the highest boundary for leverage   
# Will give you waht is true   
Lev\_indices = which(hatvalues(Perch\_lm) >= 3\*(3+1)/56)  
  
#Two cases with high leverage  
# will tell you which of the perch values are potiential high leverage fish   
# What are diferent about these fish?   
Perch[Lev\_indices,]

## Obs Weight Length Width  
## 1 104 5.9 8.8 1.4  
## 40 143 840.0 37.3 7.8

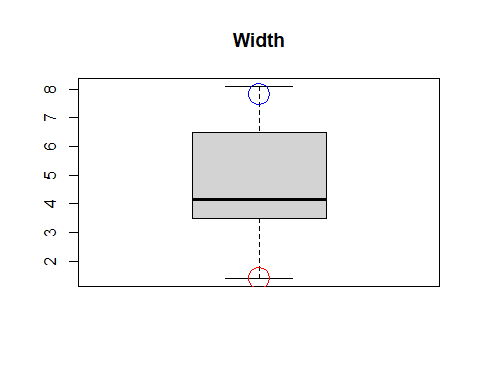
# Have fish 1 and 40   
# Fish 1 is kinda small   
# In teh slides we see how this are sorted differently

**Leverage in Multiple Regression: Perch**

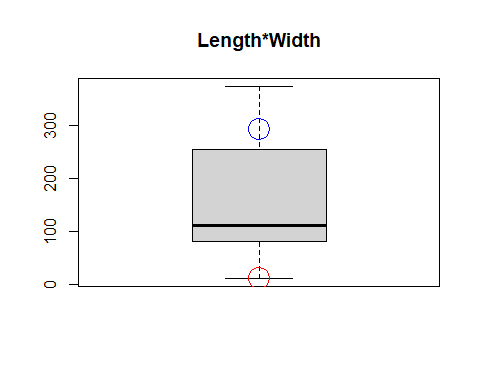
#boxplots for the three variables in the model  
#points() is used to show the values for cases 1 and 40 that have high leverage  
#cex=3 is the type of symbol to show in the plot  
  
# Tehse are box plots   
# first look at weight and length variable and made a box plot for each   
# Then drew some points for teh specific fish   
# The seond line points lenght 1 = the fish that is number 1   
# did the same thing for the 40th fish in blue   
  
# Fish 1 appears teh smallest fish in teh data   
# Fish 40 appears to be in the middle 50 in teh box   
# Fish 1 also appears to be the lowest width of a fish   
# FIsh 40 is a really high fish; it si pretty fat   
  
# DO the same thign for the intearction term; fish 1 appears to still be an aissue   
# Maybe fish 1 has influenc eon the model?   
# WE dont kno whtis yet, it might be on the prediciton line   
# If the regression line is righ tnext to it, it's not going to have influence, we dont know yet until we run more tests   
  
boxplot(Perch$Length, main="Length")  
points(Perch$Length[1], col="red", cex=3)  
points(Perch$Length[40], col="blue", cex=3)



boxplot(Perch$Width, main="Width")  
points(Perch$Width[1], col="red", cex=3)  
points(Perch$Width[40], col="blue", cex=3)



boxplot(Perch$Length\*Perch$Width, main="Length\*Width")  
points(Perch$Length[1]\*Perch$Width[1], col="red", cex=3)  
points(Perch$Length[40]\*Perch$Width[40], col="blue", cex=3)



# Gives you an idea of wehre things are   
# But you could have a big prediction difference between one variable and the other might not

**Cook’s Distance** How much would the fit change if one data value were omitted?

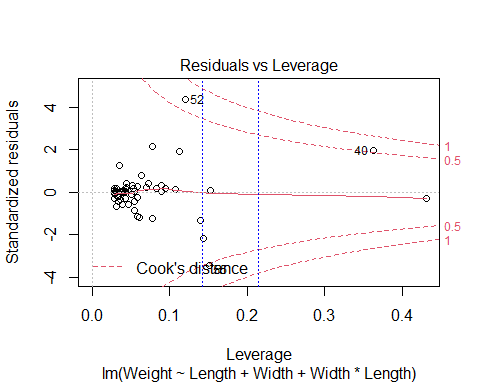
Cook’s Di = (((std.resi)^2)/(k+1))\*(hi/(1-hi)) Di increases with either poor fit (std.resi) and high leverage (hi).

1. Compare to other Di’s.
2. Study any case with Di > 0.5; worry if Di > 1.0.

# IF something has influence, but why   
# Is it due to outlier or other things?   
# Just use cook's distance   
  
# can ssee that fish 1 has high leveerage but no influence   
# Fish 52 and 55 appear to be right on the outside or righ ton the line   
# the two verical lines are the cut off for big leverages   
  
# Shows 3 cases with high Cook's Distance  
Cooks\_indices = which(cooks.distance(Perch\_lm) >= 0.5)  
# use the same logic as abouve to see which points are tur for being over the 0.5 cook's distance   
  
# Below tells you which points in perch has a cook's distance of over 0.5, which means that they have high influence   
Perch[Cooks\_indices,]

## Obs Weight Length Width  
## 40 143 840 37.3 7.8  
## 52 155 1100 44.6 6.9  
## 55 158 1000 46.0 8.1

# we see it's teh heavy fish with high weight that have influence   
  
# ',5' shows only the Cook's plot and not other residual diagnostics plots  
plot(Perch\_lm,5)  
  
# 'v' draws a vertical line  
# lty chooses the type of line to draw (dashes)  
abline(v = 2\*(3+1)/56, col="blue", lty=3)  
abline(v = 3\*(3+1)/56, col="blue", lty=3)



How to compre when we add a thing to the model, does any specific point have influence or is that teh trend of the data overall?

* Anwser this witht eh houses dataset

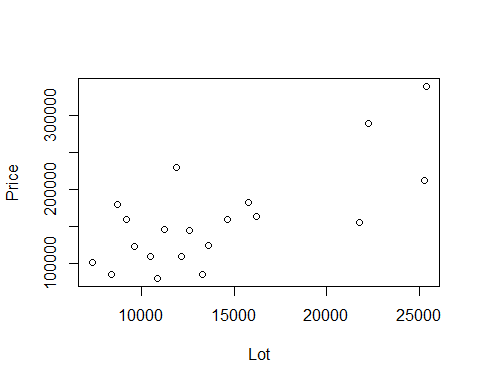
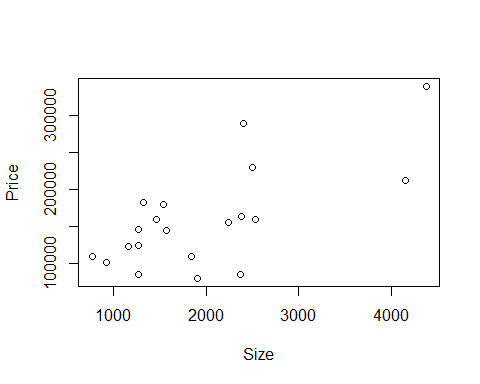
**Houses** Response variable: Y = House price Predictors: X1 = Size; X2 = Lot (size of the lot)

head(Houses)

## Price Size Lot  
## 1 212000 4148 25264  
## 2 230000 2501 11891  
## 3 339000 4374 25351  
## 4 289000 2398 22215  
## 5 160000 2536 9234  
## 6 85000 2368 13329

**Fitting the Multiple Regression Model**

plot(Price~Size+Lot, data=Houses)



Houses.lm=lm(Price~Size+Lot, data=Houses)  
summary(Houses.lm)

##   
## Call:  
## lm(formula = Price ~ Size + Lot, data = Houses)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -79532 -28464 3713 21450 73507   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 34121.649 29716.458 1.148 0.2668   
## Size 23.232 17.700 1.313 0.2068   
## Lot 5.657 3.075 1.839 0.0834 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 47400 on 17 degrees of freedom  
## Multiple R-squared: 0.5571, Adjusted R-squared: 0.505   
## F-statistic: 10.69 on 2 and 17 DF, p-value: 0.000985

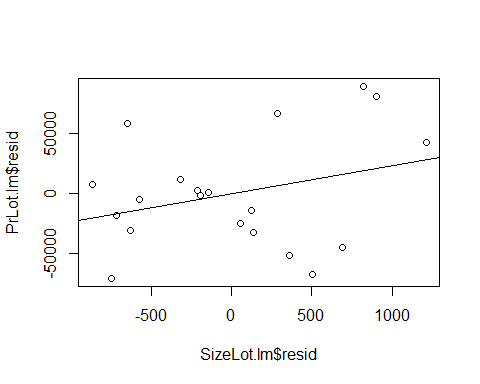
# based on teh slope it doesnt appear that size or ot is a good predictor forom price   
# looking at anova, we get that its fine   
# There is multicollinearity here

**Added Variable Plot** Say we want to add teh predcitor z, but want to see its impact You have to comapre teh residuals of the model with the z vs not with z SO find teh difference btween a mdoel with z vs without z

Basic idea: For any single predictor Z … 1. Fit a model for Y using all other predictors. Save residuals as error1 (what the other predictors don’t know about Y).

1. Fit a model for Z using all other predictors. Save residuals as error2 (what the other predictors don’t know about Z).
2. Plot error1 vs. error2 (what’s unique to Z that explains new variability in Y).

# 1. make a model with size not predited by lot   
PrLot.lm = lm(Price~Lot, data = Houses)  
#2. Size predicted by lot   
SizeLot.lm = lm(Size~Lot, data = Houses)  
  
  
# The residuals from this model—PrLot.lm —are saved as PrLot.lm$resid  
plot(PrLot.lm$resid~SizeLot.lm$resid)  
# The residuals from this model— SizeLot.lm —are saved as SizeLot.lm$resid  
  
# Plot PrLot.lm$resid vs. SizeLot.lm$resid  
model = lm(PrLot.lm$resid~SizeLot.lm$resid)  
abline(model)



summary(model)

##   
## Call:  
## lm(formula = PrLot.lm$resid ~ SizeLot.lm$resid)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -79532 -28464 3713 21450 73507   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1.772e-12 1.030e+04 0.000 1.000  
## SizeLot.lm$resid 2.323e+01 1.720e+01 1.351 0.194  
##   
## Residual standard error: 46070 on 18 degrees of freedom  
## Multiple R-squared: 0.09201, Adjusted R-squared: 0.04157   
## F-statistic: 1.824 on 1 and 18 DF, p-value: 0.1936

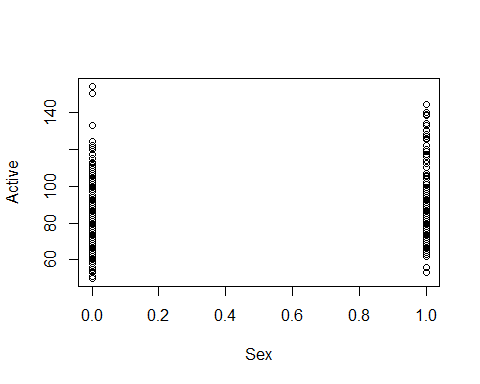
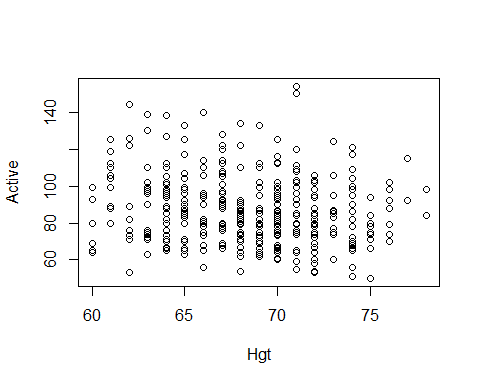
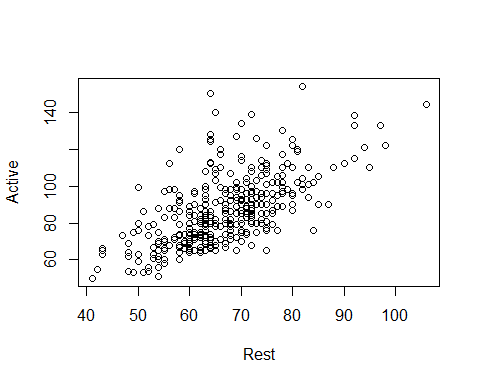
#Is there a relationship?   
# Equation of the line:   
# Looks familliar because the size estimate for the summary of the model above is the   
# New model fo teh residuals has an incet fo 0 and a slope of 23, and oges through gthe orignas, have teh same slope as size by lot   
# Do we see anythign wehre some values may be skewing out data or does it appear out thing follows the trend well   
# Then if there are no changes, then we can say tehre may not be a big influence   
# If we see something very different, then we can see that there may be useful to add the value to the newer model

**Pulse: More than Two Predictors** Response variable: Y = Active pulse Predictors: X1 = Resting pulse X2 = Hgt X3 = Sex (0 = M, 1 = F)

# Predicts the active heart rate by using rest, height and sex   
Pulse.lm = lm(Active~Rest+Hgt+Sex, data=Pulse)  
summary(Pulse.lm)

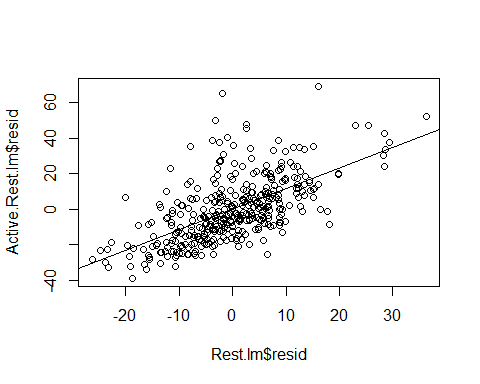
##   
## Call:  
## lm(formula = Active ~ Rest + Hgt + Sex, data = Pulse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.730 -9.381 -2.691 6.817 67.413   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.47044 20.40682 0.660 0.510   
## Rest 1.16249 0.07569 15.359 <2e-16 \*\*\*  
## Hgt -0.07440 0.26969 -0.276 0.783   
## Sex 1.90940 2.11311 0.904 0.367   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.17 on 371 degrees of freedom  
## Multiple R-squared: 0.4056, Adjusted R-squared: 0.4008   
## F-statistic: 84.38 on 3 and 371 DF, p-value: < 2.2e-16

# The summary tells us that we have an incept of 13.4   
# base don this we probably wont use these predictors   
# Height and sex doent appear to be useful int eh model   
# What if we make a varibale plot do some variables have extreme values that will effect the mdoel in some way?   
  
plot(Active~Rest+Hgt+Sex, data=Pulse)



**Pulse: More than Two Predictors** This idifferes from anova, becaus we are controlling which variables are dropped and stay WE are also not dropping them all, we are only cycling through what different variable combinations look like

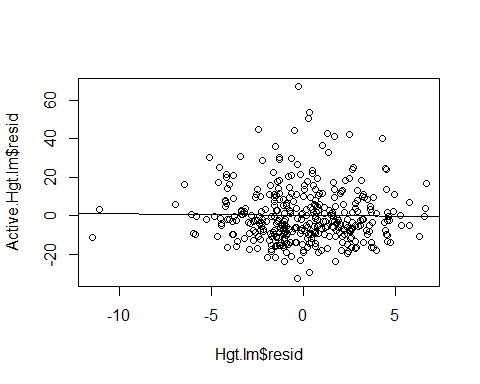
#Want to make a plot where actiec is predicted byt weverthing   
Active.Rest.lm = lm(Active~Hgt+Sex, data=Pulse)  
Rest.lm = lm(Rest~Hgt+Sex, data = Pulse)  
plot(Active.Rest.lm$resid~Rest.lm$resid)  
mod1 = lm(Active.Rest.lm$resid~Rest.lm$resid)  
abline(mod1)



summary(mod1)

##   
## Call:  
## lm(formula = Active.Rest.lm$resid ~ Rest.lm$resid)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.730 -9.381 -2.691 6.817 67.413   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.599e-15 7.300e-01 0.0 1   
## Rest.lm$resid 1.162e+00 7.548e-02 15.4 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.14 on 373 degrees of freedom  
## Multiple R-squared: 0.3887, Adjusted R-squared: 0.3871   
## F-statistic: 237.2 on 1 and 373 DF, p-value: < 2.2e-16

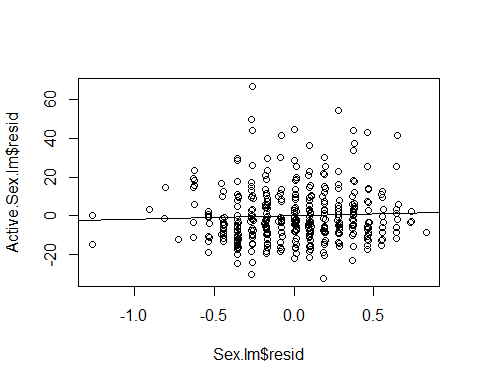
# Where active is predicted by rest and sex, not hgieht   
Active.Hgt.lm = lm(Active~Rest+Sex, data=Pulse)  
Hgt.lm = lm(Hgt~Rest+Sex, data = Pulse)  
plot(Active.Hgt.lm$resid~Hgt.lm$resid)  
mod2 = lm(Active.Hgt.lm$resid~Hgt.lm$resid)  
abline(mod2)



summary(mod2)

##   
## Call:  
## lm(formula = Active.Hgt.lm$resid ~ Hgt.lm$resid)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.730 -9.381 -2.691 6.817 67.413   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -2.064e-15 7.300e-01 0.000 1.000  
## Hgt.lm$resid -7.440e-02 2.690e-01 -0.277 0.782  
##   
## Residual standard error: 14.14 on 373 degrees of freedom  
## Multiple R-squared: 0.0002051, Adjusted R-squared: -0.002475   
## F-statistic: 0.07652 on 1 and 373 DF, p-value: 0.7822

# Then active is predicted by rest and hight   
Active.Sex.lm = lm(Active~Rest+Hgt, data=Pulse)  
Sex.lm = lm(Sex~Rest+Hgt, data = Pulse)  
plot(Active.Sex.lm$resid~Sex.lm$resid)  
mod3 = lm(Active.Sex.lm$resid~Sex.lm$resid)  
abline(mod3)

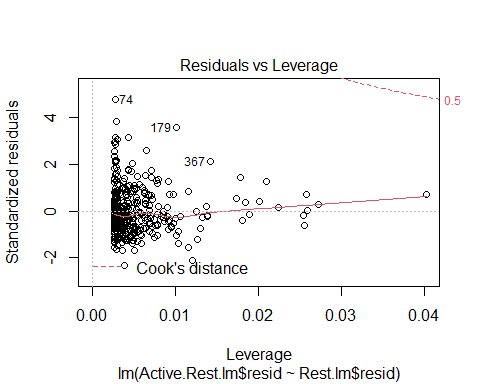


summary(mod3)

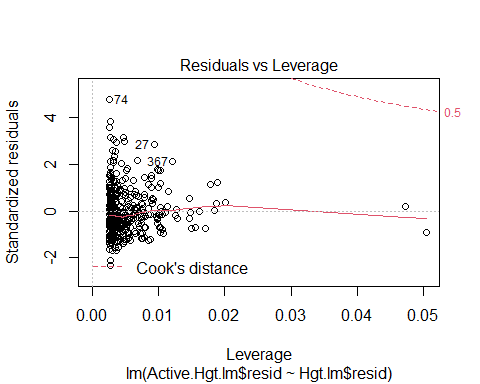
##   
## Call:  
## lm(formula = Active.Sex.lm$resid ~ Sex.lm$resid)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.730 -9.381 -2.691 6.817 67.413   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -9.785e-16 7.300e-01 0.000 1.000  
## Sex.lm$resid 1.909e+00 2.107e+00 0.906 0.366  
##   
## Residual standard error: 14.14 on 373 degrees of freedom  
## Multiple R-squared: 0.002196, Adjusted R-squared: -0.0004791   
## F-statistic: 0.8209 on 1 and 373 DF, p-value: 0.3655

# We see that there doesn't appear much difference when looking at teh sumamries of the the predictors   
# If we want to see if a point has influence because of all the model vs it just being one varibale, the added variable plot will tell you if its one variable that is being extra or if it is just the whole dataset

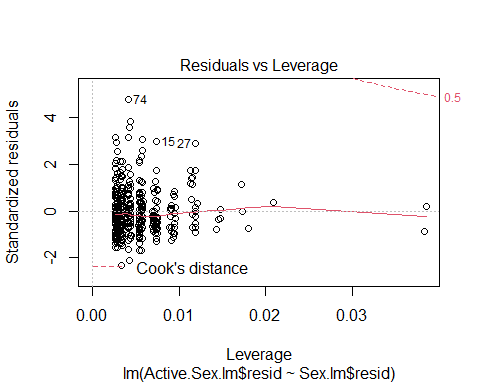
# Now we check each of these for leverage and influence   
plot(mod1, 5)



plot(mod2, 5)

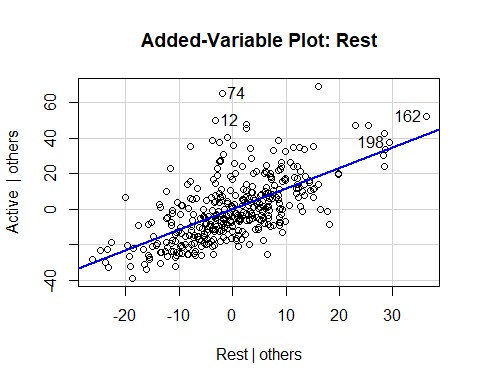


plot(mod3, 5)

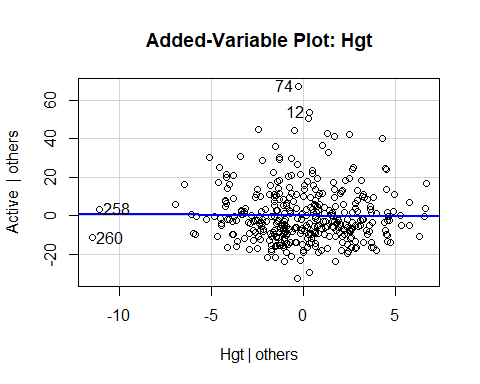


**Added Variable Plots** HOW TO DO THE ADDED VARIBALE PLOT EASILY AND QUICKLY IN R

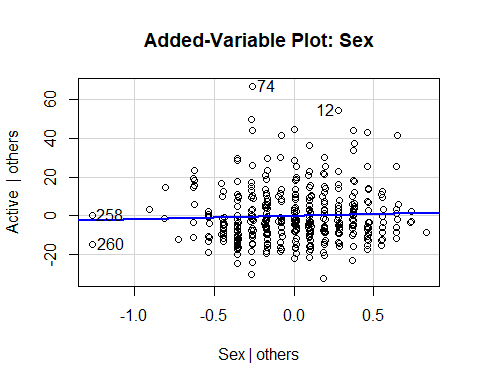
library(car)  
  
# Give the lienar model and what you want to look at as the added variabe   
# So it looks at the thing in pulse.lm, but takes out rest in the first one and then compares what the residual plots would look like between those two   
avPlot(Pulse.lm, "Rest") # need to give the string name of the variable in the lm model you're referencing from the first argument



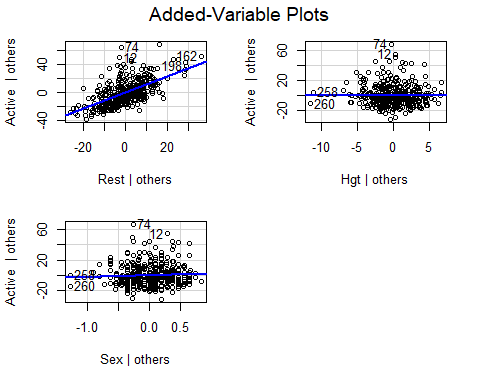
avPlot(Pulse.lm, "Hgt")



avPlot(Pulse.lm, "Sex")



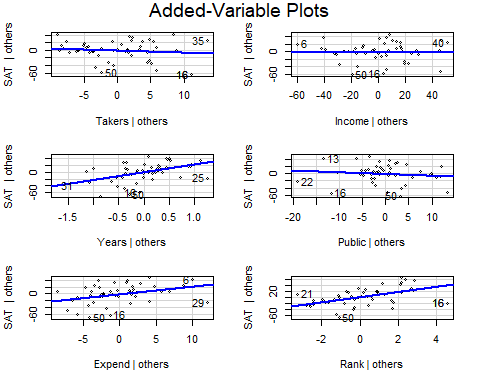
avPlots(Pulse.lm, ~.) # This is an avplot for all of the variables in the Pulse.lm linear regression model



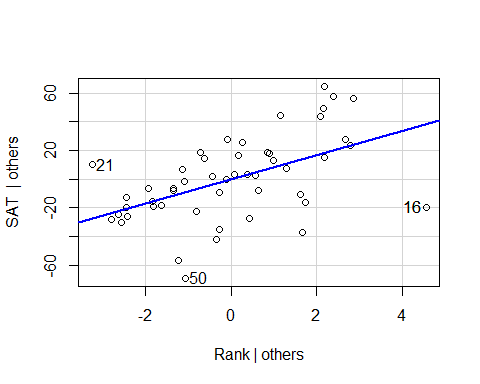
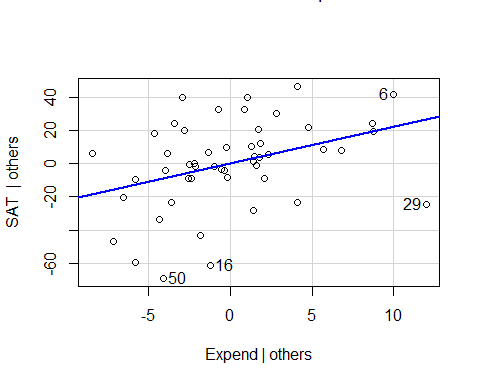
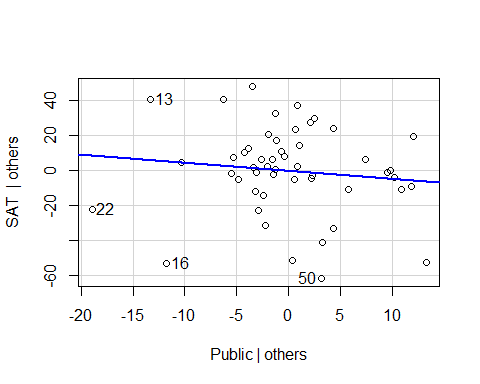
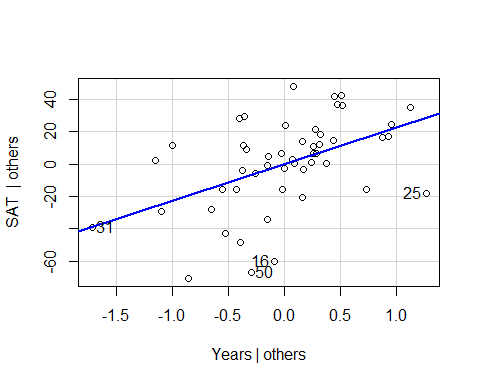
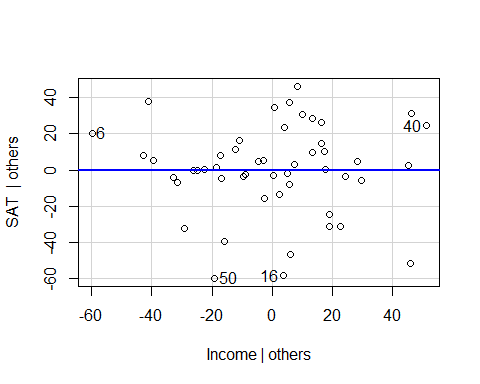
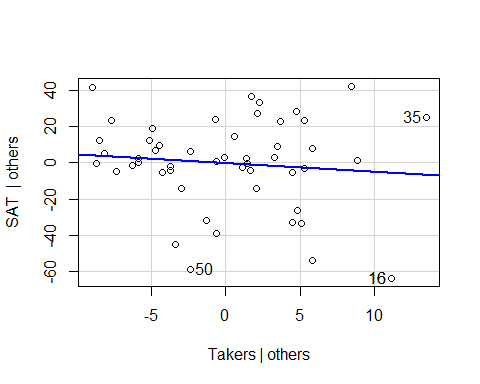
# This iwll make an avPlot for all the variables in teh pulse.lm linear regression model so you don't have to   
# THIS IS PLURARL   
# Its not super great in how it does it   
# YOu have to do it by all the predictors  
# issues: it's really hard to read it's really jumbled, you can read it   
# All the different plots are put in to one plot   
# This is okay for 3, but the more we get teh less nice it will be   
# So looka t teh SAT data for what this looks like less nice

**StateSAT: More than Two Predictors**

StateSAT.lm = lm(SAT~., data=StateSAT[,c(2:8)])  
avPlots(StateSAT.lm, ~.)



variables = colnames(StateSAT)  
  
# This does the avPlot, but it cycles through for each variable in the dataset for ytou (or at least teh ones between 3 - 8 )   
for(i in 3:8){  
 avPlots(StateSAT.lm, variables[i])  
}



# With the for statemtne we can see teh same plots as writing them one by one, you can see if there are any extreme values by these plots that may be guiding these values   
# WE dont see any in takers, but point 25 might draw the line down in some   
# Point 22 have an influence in the SAT others, can guess   
#State 29 - alaska

## STOR 455 Homework #4

**Situation:** Suppose that (again) you are interested in purchasing a used car. How much should you expect to pay? Obviously the price will depend on the type of car you get (the model) and how much it’s been used. For this assignment you will investigate how the price might depend on the age and mileage, as well as the state where the car is purchased.

**Data Source:** To get a sample of cars, begin with the UsedCars CSV file. The data was acquired by scraping TrueCar.com for used car listings on 9/24/2017 and contains more than 1.2 million used cars. For this assignment you should choose the same car *Model* and *State* that you initially chose for homework #2. You should again add a variable called *Age* which is 2017-year (since the data was scraped in 2017).

**Directions:** The code below can again be used to select data from a particular *Model* and *State* of your choice. The R chunk below begins with {r, eval=FALSE}. eval=FALSE makes these chunks not run when I knit the file. Before you run this chunk, you should revert it to {r}.

library(readr)

library(car)

ModelOfMyChoice = "Civic"

StateOfMyChoice = "NY"

UsedCars <- read\_csv("UsedCars.csv")

MyCars = subset(UsedCars, Model==ModelOfMyChoice & State==StateOfMyChoice)

MyCars$Age = 2017 - MyCars$Year

**MODEL #4: Use Age and Miles as predictors for Price**

1. Construct a model using two predictors (age and miles) with *Price* as the response variable and provide the summary output.

**0.5 points** model  
**0.5 points** summary output

modq1 = lm(Price~Age+Mileage, data=MyCars)

summary(modq1)

##

## Call:

## lm(formula = Price ~ Age + Mileage, data = MyCars)

##

## Residuals:

## Min 1Q Median 3Q Max

## -4694.7 -1443.3 -316.7 1099.8 7168.3

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 1.817e+04 1.660e+02 109.499 < 2e-16 \*\*\*

## Age -1.036e+03 6.319e+01 -16.396 < 2e-16 \*\*\*

## Mileage -2.543e-02 4.642e-03 -5.478 7.23e-08 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 1901 on 445 degrees of freedom

## Multiple R-squared: 0.7325, Adjusted R-squared: 0.7313

## F-statistic: 609.4 on 2 and 445 DF, p-value: < 2.2e-16

1. Assess the importance of each of the predictors in the regression model - be sure to indicate the specific value(s) from the summary output you are using to make the assessments. Include hypotheses and conclusions in context.

**2 point** (1 pt each) They should use the summary of the model to comment on the p-values for the individual slope tests. They do not need to specifically list the hypotheses being tested, just note if the p-values are small. If they instead do tests for correlation, I’ll allow that for full credit as well if they again comment on the p-values without the need to cite hypotheses. For my model both p-values for the *Age* and *Mileage* predictors are well below 0.05, hence useful in the model.

1. Assess the overall effectiveness of this model (with a formal test). Again, be sure to include hypotheses and the specific value(s) you are using from the summary output to reach a conclusion.

**2 points** I expect them to perform a hypothesis test with Null: βi = 0 for all i, Alternative βi ≠ 0 for some i. and draw a conclusion from the p-value of anova455() or the similar output in the summary table. They can write out these conclusion in words, such as the null hypothesis is that all coefficients are zero, the alternative is that at least one is nonzero. If p-values or significance are mentioned in the hypotheses, deduct 0.5 points. For my model the p-value is small (2.2e-16), so I have evidence to support the alternative, that at least one of the coefficients is nonzero.

**Note:** They should not draw this conclusion from the p-values in the anova() output. By default this uses a sequential sums method, which performs a series of nested F tests. You can take off 1 point for using anova() instead of anova455() or the summary() (Summary is also full credit).

**Note:** Throughout the assignment, anova455() may show different outputs in students’ notebooks vs the knitted html. In the notebook, the values will likely not go below 2.2e-16. When knit, this does not seem to be the case, If you see this come up, make sure **not** to take off credit when the students cite 2.2e-16 as the p-value, even the the output might show a lower number.

source("https://raw.githubusercontent.com/JA-McLean/STOR455/master/scripts/anova455.R")

anova455(modq1)

|  |
| --- |
|  |

|  | **Df**  **<dbl>** | **Sum Sq**  **<dbl>** | **Mean Sq**  **<dbl>** | **F value**  **<dbl>** | **P(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- |
| Model | 2 | 4402719393 | 2201359696 | 609.3773 | 0 |
| Error | 445 | 1607551043 | 3612474 | NA | NA |
| Total | 447 | 6010270436 | NA | NA | NA |

3 rows

summary(modq1)

##

## Call:

## lm(formula = Price ~ Age + Mileage, data = MyCars)

##

## Residuals:

## Min 1Q Median 3Q Max

## -4694.7 -1443.3 -316.7 1099.8 7168.3

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 1.817e+04 1.660e+02 109.499 < 2e-16 \*\*\*

## Age -1.036e+03 6.319e+01 -16.396 < 2e-16 \*\*\*

## Mileage -2.543e-02 4.642e-03 -5.478 7.23e-08 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 1901 on 445 degrees of freedom

## Multiple R-squared: 0.7325, Adjusted R-squared: 0.7313

## F-statistic: 609.4 on 2 and 445 DF, p-value: < 2.2e-16

1. Compute and interpret the variance inflation factor (VIF) for your predictors.

**1 point** compute VIF - most will use the vif() function from the car package. This package has issues intalling on some macs, so I also made available a script of this function, which they may use as well. It’s possible that they calculate it from the R2 of the model using Predictor1~Predictor2, then vif = 1/(1-R2). This is fine as well.

**1 point** discuss VIF - They should in some may say that there is little, or substantial multicollinearity based on the VIF value. We haven’t concrete cutoffs for this, but over 5 may be substantial, with lower as little multicollinearity. Here my VIF is fairly small, so there is little concern about multicollinearity.

vif(modq1)

## Age Mileage

## 2.756156 2.756156

1. Suppose that you are interested in purchasing a car of this model that is four years old (in 2017) with 31K miles. Determine each of the following: a 90% confidence interval for the mean price at this age and mileage, and a 90% prediction interval for the price of an individual car at this age and mileage. Write sentences that carefully interpret each of the intervals (in terms of car prices)

**1 points** new dataframe for 4 year old car with 31000 miles  
**0.5 points** confidence interval  
**0.5 points** prediction interval  
**1 point** (0.5 each) They should clearly distinguish that the confidence interval is predicting the mean price of four year old cars with 31K miles in their model, while the prediction interval is predicting the price of a specific car of their model that is four years old with 31K miles. The textbook also describes the prediction interval interval as predicting the interval where most cars of this age/model would be contained. This is fine as well.

oneCar2 = data.frame(Age = 4, Mileage=31000)

predict.lm(modq1, oneCar2, interval = "confidence", level=.9)

## fit lwr upr

## 1 13242.38 13064.6 13420.16

predict.lm(modq1, oneCar2, interval = "prediction", level=.9)

## fit lwr upr

## 1 13242.38 10104.52 16380.23

**MODEL #5: Now Include a Categorical predictor**

For this section you will combine both datasets used in Homework #2, as well as a third dataset. Each dataset from Homework #2 included cars from your specific *Model*, but from two different states. You should use the same code that you used in homework #2 to construct this second dataframe with cars from North Carolina, and a third dataframe with cars of your model from a third state of your choice. Then manipulate the code below to combine the three dataframes into one dataframe. Make sure to add the *Age* variable again to your dataframes for the additional states before binding them together. The R chunk below begins with {r, eval=FALSE}. eval=FALSE makes these chunks not run when I knit the file. Before you run this chunk, you should revert it to {r}.

MyCars2 = subset(UsedCars, Model==ModelOfMyChoice & State=="NC")

MyCars3 = subset(UsedCars, Model==ModelOfMyChoice & State=="CA")

MyCars2$Age = 2017 - MyCars2$Year

MyCars3$Age = 2017 - MyCars3$Year

State1 = MyCars

State2 = MyCars2

State3 = MyCars3

# rbind combines the rows in one dataframe, assuming that the columns are the same.

CombinedStates = rbind(State1, State2, State3)

1. Fit a multiple regression model using *Age*, *Mileage*, and *State* to predict the *Price* of the car.

**1 pt** - code for model. They may factor() *State*, but this is redundant.

modq6 = lm(Price~Age+Mileage+State, data=CombinedStates)

1. Perform a hypothesis test to determine the importance of *State* terms in the model constructed in question 6. List your hypotheses, p-value, and conclusion.

**2 pt** - Construct a reduced model and use anova().  
**1 pt** - (0.5 points each) - hypotheses - could be in symbolic form as below, or in words citing the coefficients for all (two) State terms. Take off 0.5 points if they state that there is only one coefficient being tested.  
**0.5** pts - conclusion

H0: β3 = β4A: β3 ≠ 0 or β4 ≠ 0;

Reject the null. There is statistically significant evidence (2.127e-12) to suggest that at least one coefficient of a State variable is nonzero.

modq7 = lm(Price~Age+Mileage, data=CombinedStates)

anova(modq7, modq6)

|  |
| --- |
|  |

|  | **Res.Df**  **<dbl>** | **RSS**  **<dbl>** | **Df**  **<dbl>** | **Sum of Sq**  **<dbl>** | **F**  **<dbl>** | **Pr(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 2489 | 11752614042 | NA | NA | NA | NA |
| 2 | 2487 | 11501323368 | 2 | 251290674 | 27.16904 | 2.126612e-12 |

2 rows

1. Fit a multiple regression model using *Age*, *Mileage*, *State*, and the interactions between *Age* and *State*, and *Mileage* and *State* to predict the *Price* of the car.

**2 pt** - code for model. They may factor() *State*, but this is redundant . Note that if they only include the interaction terms, the lm function will ‘fill in the blanks’ for them and create a model using the individual terms as well. So Price ~ both interaction terms would produce the correct model as well.

modq8 = lm(Price~Age+Mileage+State + Age\*State + Mileage \* State, data=CombinedStates)

# Also correct

modq8.1 = lm(Price~Age\*State + Mileage \* State, data=CombinedStates)

1. Perform a hypothesis test to determine the importance of *State* terms in the model constructed in question 8. List your hypotheses, p-value, and conclusion.

**1 pt** - hypotheses. Take off 0.5 points if they state that there is only two coefficients being tested.  
**1 pt** - (0.5 points each) - hypotheses - could be in symbolic form as below, or in words citing the coefficients for all (six) Model terms.  
**0.5 pts** - conclusion

H0: βi = 0; for all i, i=(3,4,5,6,7)

HA: βi ≠ 0; for at least one i, i=(3,4,5,7)

The 3rd through 7th terms of the model contain a State term.

Reject the null. There is statistically significant evidence (9.002e-15) to suggest that at least one of the coefficients for a term with State in the linear model is nonzero.

modq9 = lm(Price~Age+Mileage, data=CombinedStates)

anova(modq9, modq8)

|  |
| --- |
|  |

|  | **Res.Df**  **<dbl>** | **RSS**  **<dbl>** | **Df**  **<dbl>** | **Sum of Sq**  **<dbl>** | **F**  **<dbl>** | **Pr(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 2489 | 11752614042 | NA | NA | NA | NA |
| 2 | 2483 | 11389191105 | 6 | 363422937 | 13.2052 | 9.001764e-15 |

2 rows

**MODEL #6: Polynomial models**

One of the drawbacks of the linear model in homework #2 was the “free car” phenomenon where the predicted price is eventually negative as the line decreases for older cars. Let’s see if adding one or more polynomial terms might help with this. For this section you should use the dataset with cars from three states that you used for model 5.

1. Fit a quadratic model using *Age* to predict *Price* and examine the residuals. Construct a scatterplot of the data with the quadratic fit included. You do not need to specifically cite all conditions for the linear model, but should discuss any issues that you see in the conditions.

**1.5 pt** - code for quadratic model (may use poly function or create new squared vairable in the dataframe)  
**2 pt** - plot with quadratic curve  
**1 pt** - conditions for model. They do not need to comment on all of the conditions for the linear model. You can give the full point for some discussion of the residuals in terms of the 3 conditions. This could be as simple as saying that for my data most of the conditions look good, with the possible the qqnorm plot showing a slight deviation from the qqline at the right tail, which would impact the normality of the residuals.

modq10 = lm(Price~Age+I(Age^2), data=CombinedStates)

# alternative method using the poly() function

# Must have Raw=TRUE or the two methods will not be the same

modq10poly = lm(Price~poly(Age, degree=2, raw=TRUE), data=CombinedStates)

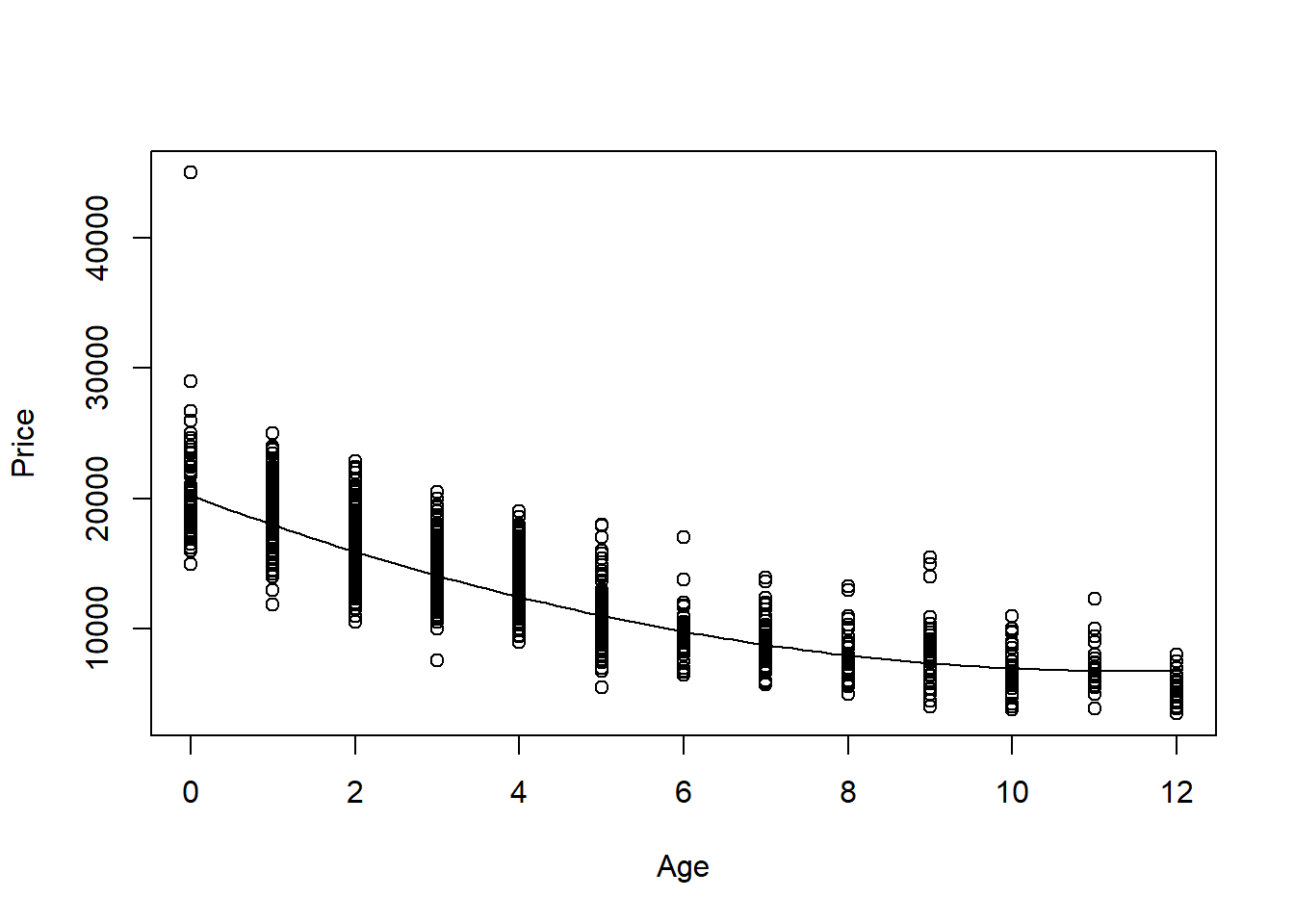
plot(Price~Age, data=CombinedStates)

a = summary(modq10)$coef[3]

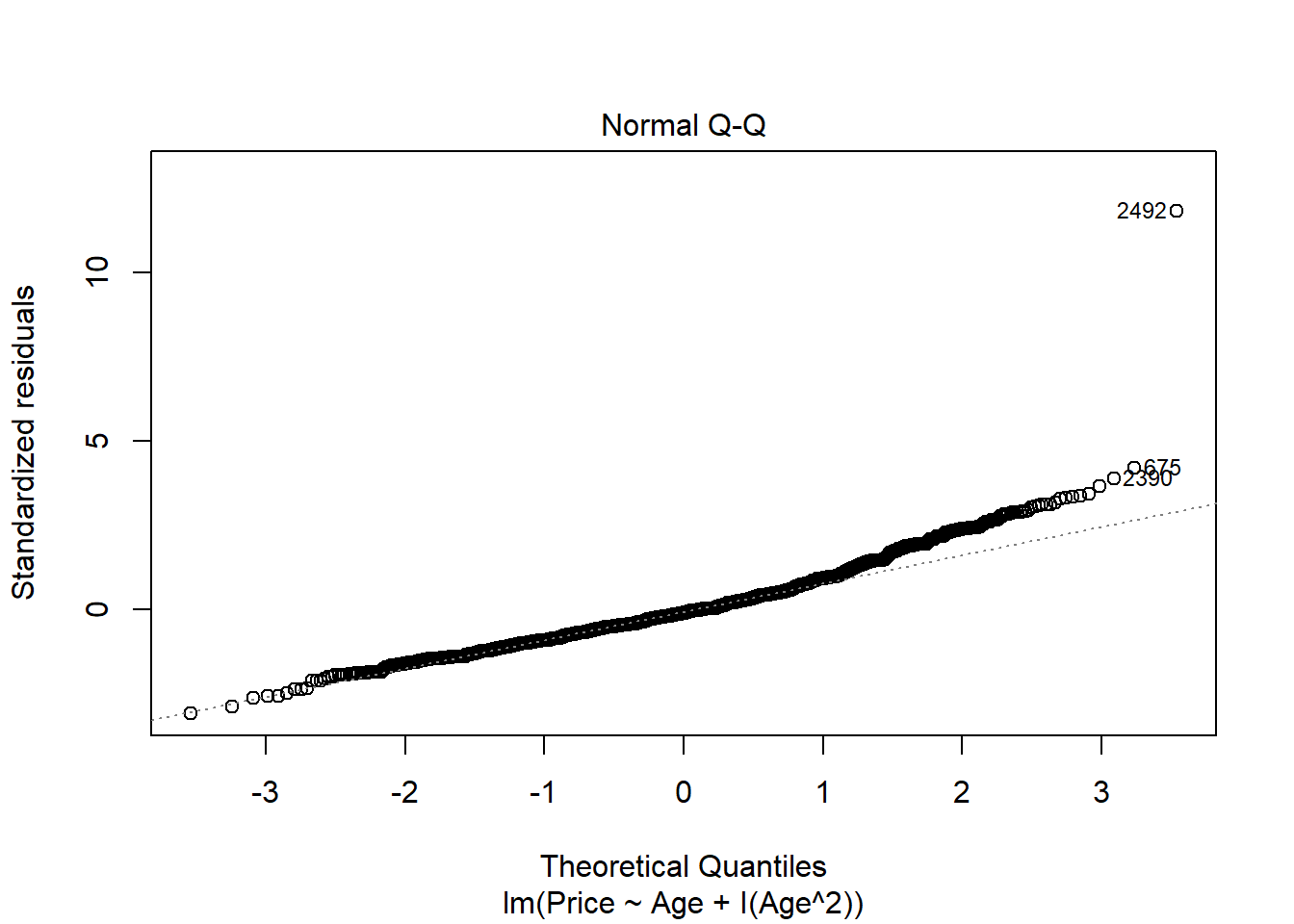
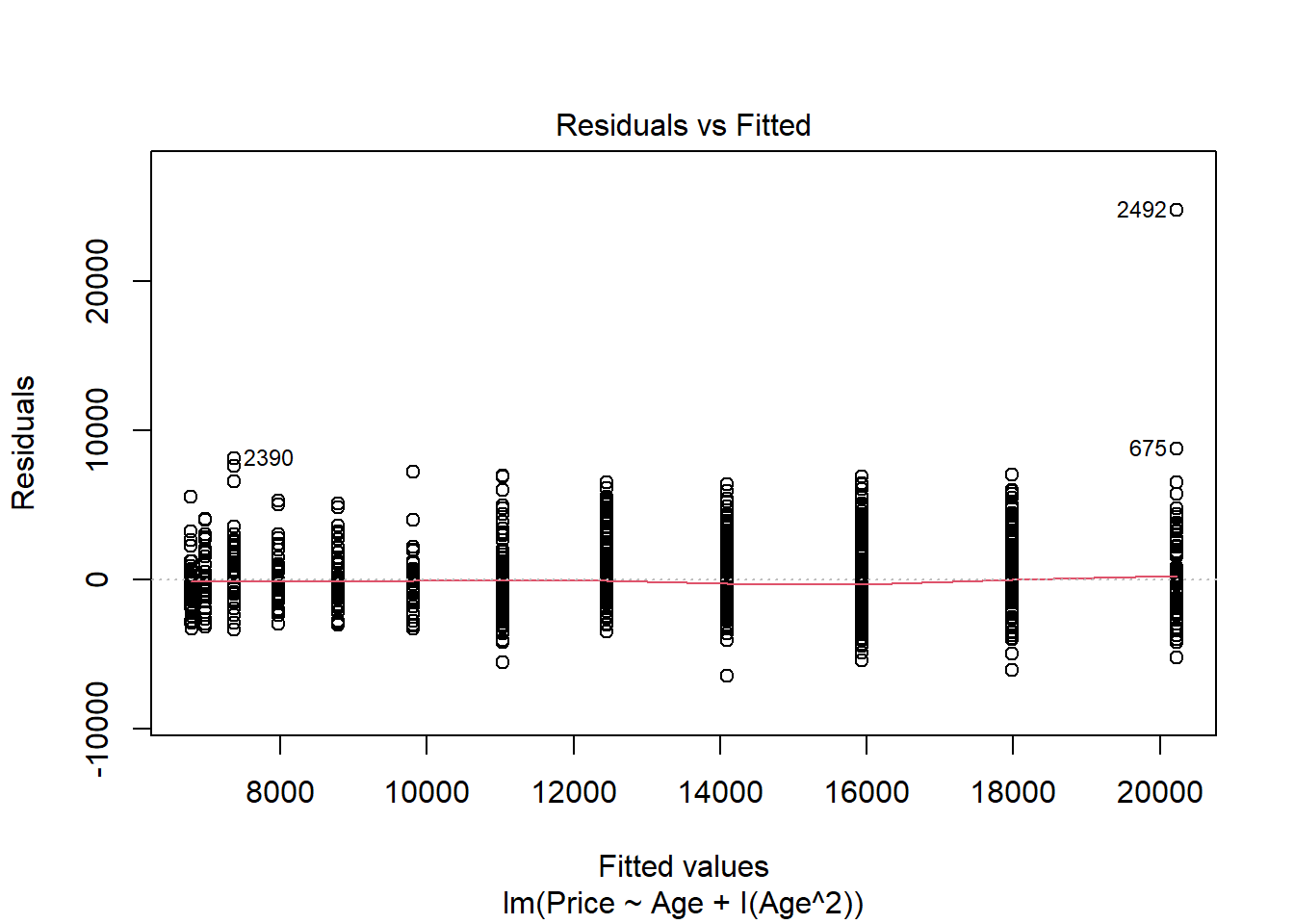
b = summary(modq10)$coef[2]

c = summary(modq10)$coef[1]

curve(a\*x^2 + b\*x + c, add=TRUE)



plot(modq10, c(1,2))



1. Perform a hypothesis test to determine if this model is significant. List your hypotheses, p-value, and conclusion.

**1 pt** - hypotheses  
**0.5 pts** - anova455 or anova test from summary to get the p-value if they built their model without using poly(). If they use anova() on a non poly() model, they are not doing the correct test. If they use anova() on a poly() model, the result is correct.  
**0.5 pts** - conclusion

H0: βi = 0; for all i

HA: βi ≠ 0; for at least one i

Reject the null. There is statistically significant evidence (p-value=2.2e-16) to suggest that at least one coefficient in the model is nonzero.

anova455(modq10)

|  |
| --- |
|  |

|  | **Df**  **<dbl>** | **Sum Sq**  **<dbl>** | **Mean Sq**  **<dbl>** | **F value**  **<dbl>** | **P(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- |
| Model | 2 | 29206450923 | 14603225462 | 3330.633 | 0 |
| Error | 2489 | 10913068062 | 4384519 | NA | NA |
| Total | 2491 | 40119518985 | NA | NA | NA |

3 rows

# or

summary(modq10)

##

## Call:

## lm(formula = Price ~ Age + I(Age^2), data = CombinedStates)

##

## Residuals:

## Min 1Q Median 3Q Max

## -6491.9 -1347.2 -238.2 1022.5 24760.8

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 20227.163 114.134 177.22 <2e-16 \*\*\*

## Age -2353.440 54.602 -43.10 <2e-16 \*\*\*

## I(Age^2) 102.785 4.843 21.23 <2e-16 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 2094 on 2489 degrees of freedom

## Multiple R-squared: 0.728, Adjusted R-squared: 0.7278

## F-statistic: 3331 on 2 and 2489 DF, p-value: < 2.2e-16

# or

anova(modq10poly)

|  |
| --- |
|  |

|  | **Df**  **<int>** | **Sum Sq**  **<dbl>** | **Mean Sq**  **<dbl>** | **F value**  **<dbl>** | **Pr(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- |
| poly(Age, degree = 2, raw = TRUE) | 2 | 29206450923 | 14603225462 | 3330.633 | 0 |
| Residuals | 2489 | 10913068062 | 4384519 | NA | NA |

2 rows

# incorrect

anova(modq10)

|  |
| --- |
|  |

|  | **Df**  **<int>** | **Sum Sq**  **<dbl>** | **Mean Sq**  **<dbl>** | **F value**  **<dbl>** | **Pr(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- |
| Age | 1 | 27231216805 | 27231216805 | 6210.7648 | 0.000000e+00 |
| I(Age^2) | 1 | 1975234118 | 1975234118 | 450.5019 | 4.971935e-92 |
| Residuals | 2489 | 10913068062 | 4384519 | NA | NA |

3 rows

1. You are looking at a 4-year-old car of your model and want to find an interval that is likely to contain its *Price* using your quadratic model. Construct an interval to predict the value of this car, and include an interpretive sentence in context.

**1 pts** - new dataframe wth age=4 car  
**0.5 pts** - prediction interval at any confidence level (not confidence interval!)  
**0.5 pts** - conclusion specific to the prediction of this one particular car’s price

ThisCar = data.frame(Age=4)

predict.lm(modq10, ThisCar, interval="prediction")

## fit lwr upr

## 1 12457.97 8350.305 16565.63

1. Does the quadratic model allow for some *Age* where a car has a zero or negative predicted price? Justify your answer using a calculation or graph.

**2 pt** - yes/no with some justification. Some students had concave down parabalas, which they could note would clearly go below zero. Other have concave up, and may need to in some way find the roots of the equation, or plot the curve in some way that it is clear if it crosses below the horizontal axis.

Note: We did not use the polyroot function in class, so they likely found the roots some other way. The roots are imaginary, so the Price never goes below zero. This is also seen in the plot below. My roots are imaginary, showing that the parabol never crosses zero.

# shows only imaginary roots, hence Price never equals 0

polyroot(c(c, b, a))

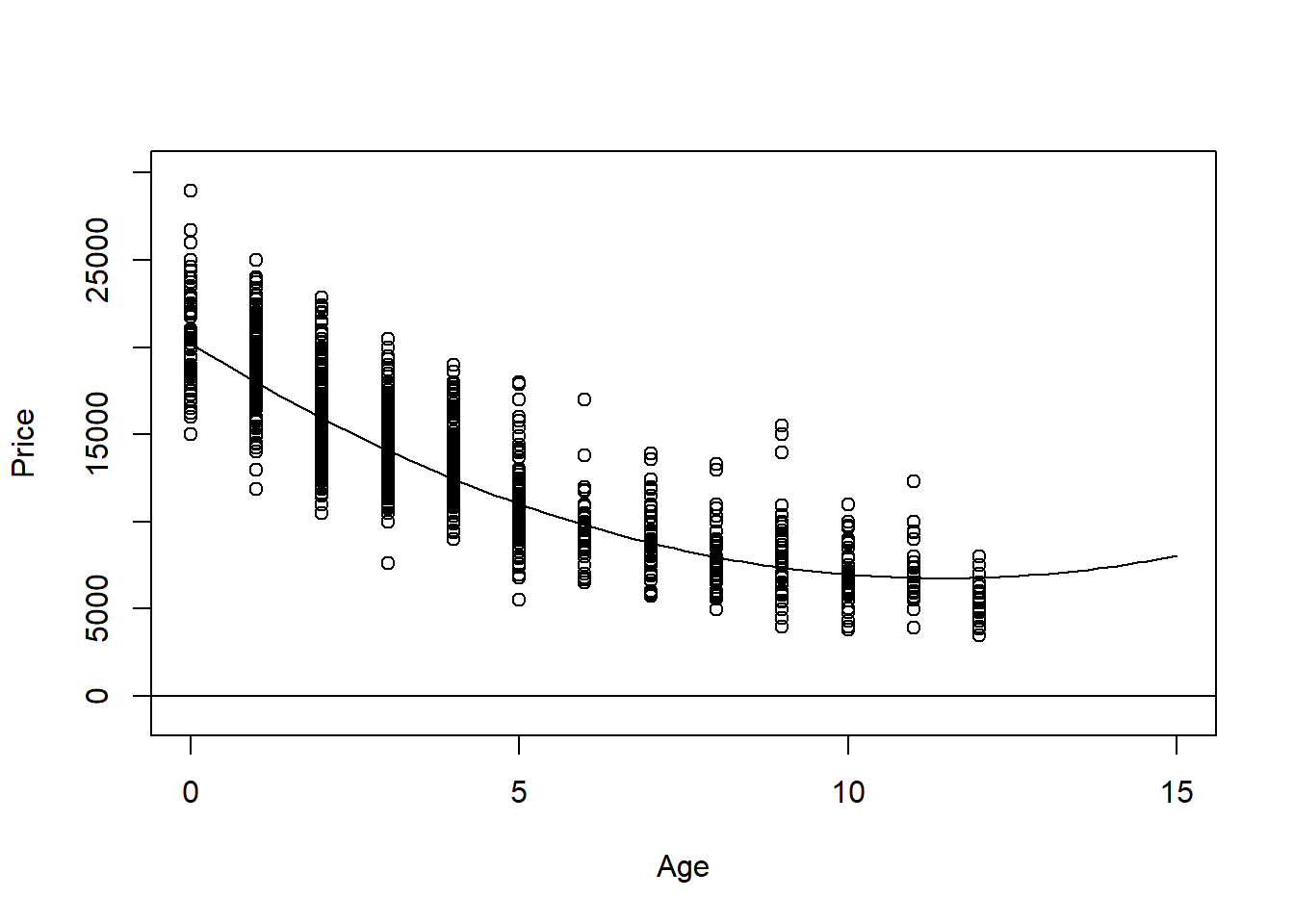
## [1] 11.44833+8.10717i 11.44833-8.10717i

# Or plot

plot(Price~Age, data=CombinedStates, xlim=c(0,15), ylim=c(-1000, 30000))

curve(a\*x^2 + b\*x + c, add=TRUE)

abline(0,0)



1. Would the fit improve significantly if you also included a cubic term? Does expanding your polynomial model to use a quartic term make significant improvements? Justify your answer.

**2 pts** - There are many ways that students could reasonably answer this. They do not need to note specific hypotheses for hypothesis tests, but they should draw their conclusion from performing a hypothesis test. This could be with a nested F test for quadratic and cubic models, quadratic and quartic models, or cubic and quartic with the anova() function. using anova(quartic model) would also perform these tests for students to interpret. If they only add one term at a time, this same p-value can be found in the summary and coefficients table as well. For my data, the p-value is small for the nested tests showing the addition of the cubic term (but not quartic). THe cubic term significantly improves this model.

modq14c = lm(Price~Age+I(Age^2)+I(Age^3), data=CombinedStates)

modq14q = lm(Price~Age+I(Age^2)+I(Age^3)+I(Age^4), data=CombinedStates)

anova(modq10, modq14c)

|  |
| --- |
|  |

|  | **Res.Df**  **<dbl>** | **RSS**  **<dbl>** | **Df**  **<dbl>** | **Sum of Sq**  **<dbl>** | **F**  **<dbl>** | **Pr(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 2489 | 10913068062 | NA | NA | NA | NA |
| 2 | 2488 | 10801715538 | 1 | 111352525 | 25.64825 | 4.396797e-07 |

2 rows

anova(modq10, modq14q)

|  |
| --- |
|  |

|  | **Res.Df**  **<dbl>** | **RSS**  **<dbl>** | **Df**  **<dbl>** | **Sum of Sq**  **<dbl>** | **F**  **<dbl>** | **Pr(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 2489 | 10913068062 | NA | NA | NA | NA |
| 2 | 2487 | 10800874898 | 2 | 112193164 | 12.91675 | 2.6258e-06 |

2 rows

anova(modq14c, modq14q)

|  |
| --- |
|  |

|  | **Res.Df**  **<dbl>** | **RSS**  **<dbl>** | **Df**  **<dbl>** | **Sum of Sq**  **<dbl>** | **F**  **<dbl>** | **Pr(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 2488 | 10801715538 | NA | NA | NA | NA |
| 2 | 2487 | 10800874898 | 1 | 840639.8 | 0.193565 | 0.6600042 |

2 rows

# OR

# Check if addition of quartic term to cubic model is significant, etc...

summary(modq14q)

##

## Call:

## lm(formula = Price ~ Age + I(Age^2) + I(Age^3) + I(Age^4), data = CombinedStates)

##

## Residuals:

## Min 1Q Median 3Q Max

## -6362.6 -1338.7 -191.7 1036.4 24322.2

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 20665.7920 175.4985 117.755 <2e-16 \*\*\*

## Age -2777.4334 203.2146 -13.667 <2e-16 \*\*\*

## I(Age^2) 191.4913 76.1030 2.516 0.0119 \*

## I(Age^3) -2.8970 10.3100 -0.281 0.7787

## I(Age^4) -0.1982 0.4505 -0.440 0.6600

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 2084 on 2487 degrees of freedom

## Multiple R-squared: 0.7308, Adjusted R-squared: 0.7303

## F-statistic: 1688 on 4 and 2487 DF, p-value: < 2.2e-16

**MODEL #7: Complete second order model**

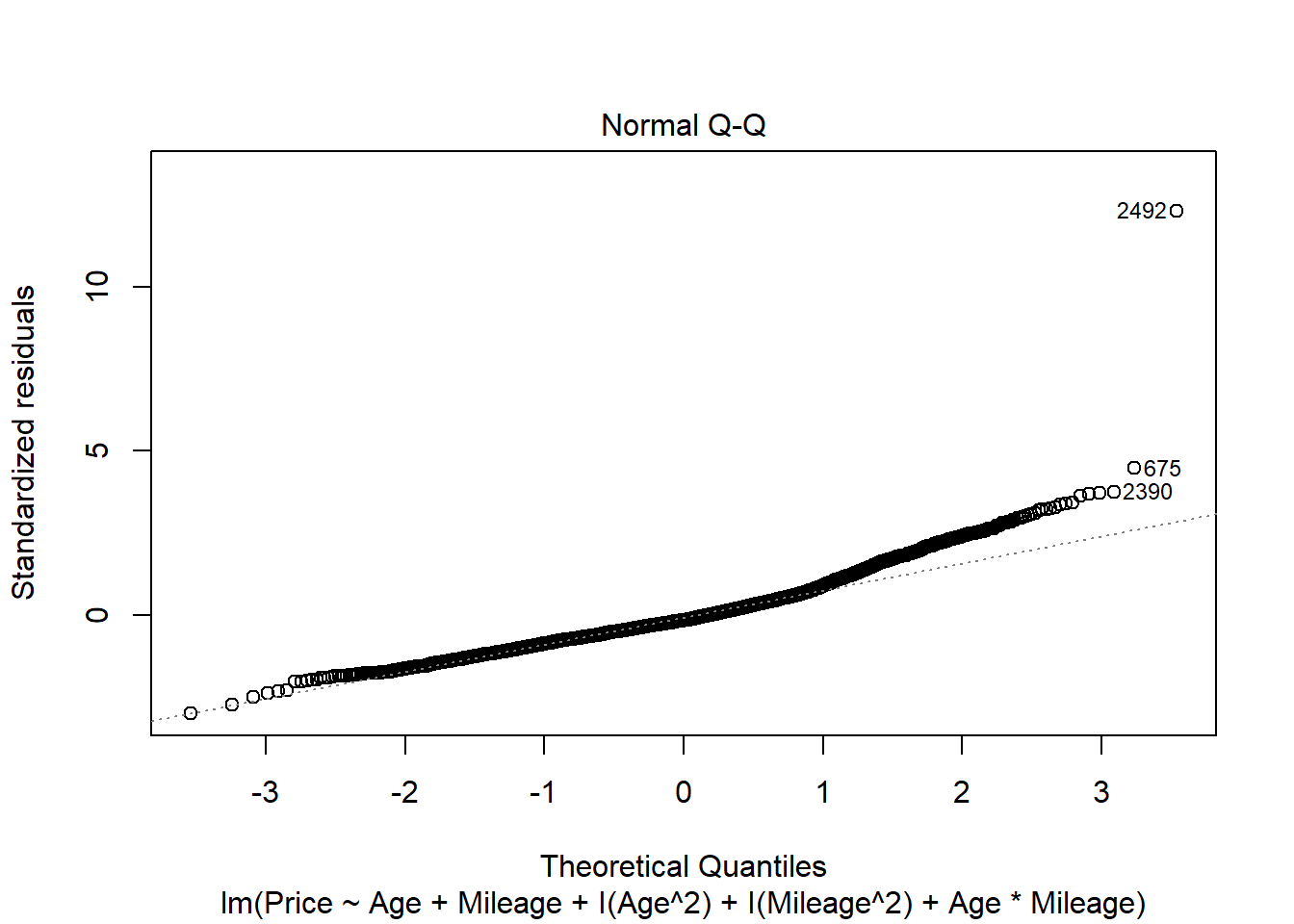
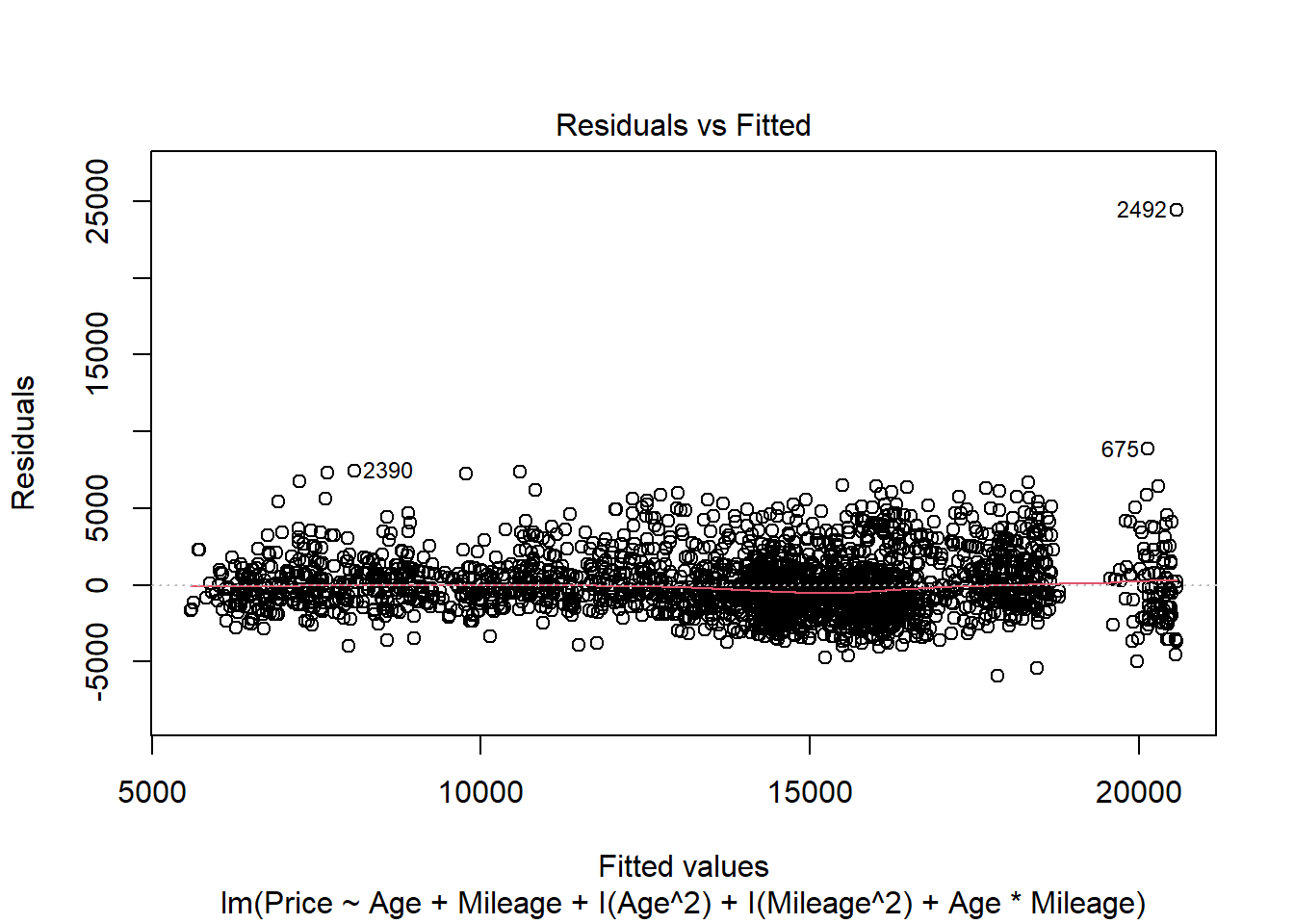
For this section you should again use the dataset with cars from three states that you used for models 5 and 6.

1. Fit a complete second order model for predicting a used car *Price* based on *Age* and *Mileage* and examine the residuals. You do not need to specifically cite all conditions for the linear model, but should discuss any issues that you see in the conditions.

**1.5 pt** - code for model. may also use polym() function although I did not do this in class.  
**1 pt** - conditions for model. They do not need to comment on all of the conditions for the linear model. You can give the full point for some discussion of the residuals in terms of the 3 conditions. As for my earlier model, the conditions seem decently met with the normality of the residuals still being a bit problematic due to the right tail deviating from the qqline.

modq15 = lm(Price~Age+Mileage+I(Age^2)+I(Mileage^2)+Age\*Mileage, data=CombinedStates)

plot(modq15, c(1,2))



# or

modq15poly = lm(Price~polym(Age, Mileage, degree=2, raw=TRUE), data=CombinedStates)

1. Perform a hypothesis test to determine if the model constructed in question 15 is significant. List your hypotheses, p-value, and conclusion.

**1 pt** - (0.5 points each) - hypotheses - could be in symbolic form as below, or in words citing the coefficients for all terms.  
**0.5 pts** - anova455 or anova test from summary to get the p-value if function was made without polym(). If the model is made from polym(), then the anova() function can be used.  
**0.5 pts** - conclusion

H0: βi = 0; for all i

HA: βi ≠ 0; for at least one i

Reject the null. There is statistically significant evidence (p-value=2.2e-16) to suggest that at least one coefficient in the model is nonzero.

anova455(modq15)

|  |
| --- |
|  |

|  | **Df**  **<dbl>** | **Sum Sq**  **<dbl>** | **Mean Sq**  **<dbl>** | **F value**  **<dbl>** | **P(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- |
| Model | 5 | 30335970693 | 6067194139 | 1541.674 | 0 |
| Error | 2486 | 9783548292 | 3935458 | NA | NA |
| Total | 2491 | 40119518985 | NA | NA | NA |

3 rows

#or

summary(modq15)

##

## Call:

## lm(formula = Price ~ Age + Mileage + I(Age^2) + I(Mileage^2) +

## Age \* Mileage, data = CombinedStates)

##

## Residuals:

## Min 1Q Median 3Q Max

## -5955.4 -1272.1 -315.6 938.5 24426.2

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 2.057e+04 1.150e+02 178.839 < 2e-16 \*\*\*

## Age -1.862e+03 6.284e+01 -29.629 < 2e-16 \*\*\*

## Mileage -4.556e-02 3.983e-03 -11.440 < 2e-16 \*\*\*

## I(Age^2) 8.830e+01 7.222e+00 12.225 < 2e-16 \*\*\*

## I(Mileage^2) 8.971e-08 1.719e-08 5.220 1.94e-07 \*\*\*

## Age:Mileage 1.869e-04 6.011e-04 0.311 0.756

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 1984 on 2486 degrees of freedom

## Multiple R-squared: 0.7561, Adjusted R-squared: 0.7556

## F-statistic: 1542 on 5 and 2486 DF, p-value: < 2.2e-16

#or

anova(modq15poly)

|  |
| --- |
|  |

|  | **Df**  **<int>** | **Sum Sq**  **<dbl>** | **Mean Sq**  **<dbl>** | **F value**  **<dbl>** | **Pr(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- |
| polym(Age, Mileage, degree = 2, raw = TRUE) | 5 | 30335970693 | 6067194139 | 1541.674 | 0 |
| Residuals | 2486 | 9783548292 | 3935458 | NA | NA |

2 rows

#incorrect

anova(modq15)

|  |
| --- |
|  |

|  | **Df**  **<int>** | **Sum Sq**  **<dbl>** | **Mean Sq**  **<dbl>** | **F value**  **<dbl>** | **Pr(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- |
| Age | 1 | 27231216805 | 27231216805 | 6.919453e+03 | 0.000000e+00 |
| Mileage | 1 | 1135688138 | 1135688138 | 2.885784e+02 | 2.559451e-61 |
| I(Age^2) | 1 | 1838648320 | 1838648320 | 4.672006e+02 | 4.338885e-95 |
| I(Mileage^2) | 1 | 130037013 | 130037013 | 3.304241e+01 | 1.011954e-08 |
| Age:Mileage | 1 | 380417 | 380417 | 9.666399e-02 | 7.558964e-01 |
| Residuals | 2486 | 9783548292 | 3935458 | NA | NA |

6 rows

1. Perform a hypothesis test to determine the importance of just the second order terms (quadratic and interaction) in the model constructed in question 15. List your hypotheses, p-value, and conclusion.

**1 pt** - (0.5 points each) - hypotheses - could be in symbolic form as below, or in words citing the coefficients for all (three) second order terms.  
**0.5 pts** - anova() nested test code  
**0.5 pts** - conclusion

Note: If students construct an incorrect model, but perform the test correctly on that model, they could receive points for all parts except for constructing the model.

H0: βi = 0; for all i, i=(3,4,5)

HA: βi ≠ 0; for at least one i, i=(3,4,5)

The 3rd, 4th, and 5th terms of the model are the second order terms.

Reject the null. There is statistically significant evidence (p-value=2.2e-16) to suggest that at least one of the second order coefficients in the model is nonzero.

modq17 = lm(Price~Age+Mileage, data=CombinedStates)

anova(modq17, modq15)

|  |
| --- |
|  |

|  | **Res.Df**  **<dbl>** | **RSS**  **<dbl>** | **Df**  **<dbl>** | **Sum of Sq**  **<dbl>** | **F**  **<dbl>** | **Pr(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 2489 | 11752614042 | NA | NA | NA | NA |
| 2 | 2486 | 9783548292 | 3 | 1969065750 | 166.7799 | 1.669608e-98 |

2 rows

1. Perform a hypothesis test to determine the importance of just the terms that involve *Mileage* in the model constructed in question 15. List your hypotheses, p-value, and conclusion.

**1 pt** - (0.5 points each) - hypotheses - could be in symbolic form as below, or in words citing the coefficients for all (three) Mileage terms.  
**0.5 pts** - anova() nested test code  
**0.5 pts** - conclusion

H0: βi = 0; for all i, i=(2,4,5)

HA: βi ≠ 0; for at least one i, i=(2,4,5)

The 2nd, 4th, and 5th terms of the model contain a Mileage term.

Reject the null. There is statistically significant evidence (p-value=2.2e-16) to suggest that at least one of the coefficients for a term with Mileage in the model is nonzero.

anova(modq10, modq15)

|  |
| --- |
|  |

|  | **Res.Df**  **<dbl>** | **RSS**  **<dbl>** | **Df**  **<dbl>** | **Sum of Sq**  **<dbl>** | **F**  **<dbl>** | **Pr(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 2489 | 10913068062 | NA | NA | NA | NA |
| 2 | 2486 | 9783548292 | 3 | 1129519770 | 95.67034 | 1.342345e-58 |

2 rows

## Homework 5

STOR 455 Homework #5   
40 points - Due 10/20 at 5:00pm   
Directions: For parts 6 and 9 you may work together, but they should be submitted   
individually by each group member. For parts 7 and 8, you should have only one   
submission per group. There will be separate places on Gradescope to submit the   
individual vs group work.   
Situation: Can we predict the selling price of a house in Ames, Iowa based on recorded   
features of the house? That is your task for this assignment. Each team will get a dataset   
with information on forty potential predictors and the selling price (in $1,000’s) for a   
sample of homes. The data sets for your group are AmesTrain??.csv and AmesTest??.csv   
(where ?? corresponds to your group number) A separate file identifies the variables in the   
Ames Housing data and explains some of the coding.   
library(readr)   
library(corrplot)   
## corrplot 0.90 loaded   
library(leaps)   
library(car)   
## Loading required package: carData   
Part 6. Cross-validation:   
In some situations, a model might fit the peculiarities of a specific sample of data well, but   
not reflect structure that is really present in the population. A good test for how your model   
might work on “real” house prices can be simulated by seeing how well your fitted model   
does at predicting prices that were NOT in your original sample. This is why we reserved   
an additional 200 cases as a holdout sample in AmesTest??.csv. Import your holdout test   
data and   
setwd("C:/Users/adeve/Desktop")   
amestrain24 <- read.csv("AmesTrain24.csv")   
amestest24 <- read.csv("AmesTest24.csv")   
• Compute the predicted Price for each of the cases in the holdout test sample, using   
your model resulting from the initial fit and residual analysis in parts 1 and 2 of   
Homework #3. This should be done with the same AmesTrain??.csv dataset that you   
used for homework #3, with your assignment #3 group numbe, and AmesTrain?? also   
using your assignment #3 group number.   
allsubmod = lm(Price~Fireplaces+GarageSF+GroundSF, amestrain24)   
  
ames.test.predict <- predict(allsubmod, newdata=amestest24)

• Compute the residuals for the 200 holdout cases.   
ames.test.residual = amestest24$Price - ames.test.predict   
• Compute the mean and standard deviation of these residuals. Are they close to what   
you expect from the training model?   
From the summary of the allsubmod, we would expect a residual standard error of 46.14.   
Since the ames.test.residual is 37.61 and we are talking about thousands of dollars when   
referring to houses, the residual is roughly close enough to what we would expect from the   
training model.   
mean(ames.test.residual)   
## [1] 4.265713   
sd(ames.test.residual)   
## [1] 37.61052   
summary(allsubmod)   
##   
## Call:   
## lm(formula = Price ~ Fireplaces + GarageSF + GroundSF, data = amestrain24)   
##   
## Residuals:   
## Min 1Q Median 3Q Max   
## -189.853 -26.864 -1.401 21.907 233.684   
##   
## Coefficients:   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.891286 6.291353 1.095 0.274   
## Fireplaces 19.901124 3.513306 5.665 2.3e-08 \*\*\*   
## GarageSF 0.142456 0.009959 14.305 < 2e-16 \*\*\*   
## GroundSF 0.062734 0.004650 13.490 < 2e-16 \*\*\*   
## ---   
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1   
##   
## Residual standard error: 46.14 on 596 degrees of freedom   
## Multiple R-squared: 0.6301, Adjusted R-squared: 0.6283   
## F-statistic: 338.5 on 3 and 596 DF, p-value: < 2.2e-16   
• Construct a plot of the residuals to determine if they are normally distributed. Is this   
plot what you expect to see considering the training model? The residuals in the testing   
data are more spread out than in the training data. Furthermore, the right tail of the   
QQNorm plot on the testing data is much more prominent than the training data. This   
suggests that there may be a skew that a model that is fitted to the testing data that may   
not be accounted for in the other data.   
plot(allsubmod)

## Quiz 2A

**Directions:** This quiz is open books, notes, internet, and all things other than direct communication with others. The Turtle\_Quiz2.csv dataset is needed to complete the quiz. This dataset can be imported from the chunk below or from the csv file, also attached in this Sakai assignment. You should complete the quiz in this R Notebook, including all code, plots, and explanations. For your submission, you should knit the notebook and submit it as a pdf to Gradescope.

library(readr)

turtles = read\_csv('https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/Turtle\_Quiz2.csv')

Back to Eastern Box Turtles

The Box Turtle Connection is a long-term study anticipating at least 100 years of data collection on box turtles. Their purpose is to learn more about the status and trends in box turtle populations, identify threats, and develop strategies for long-term conservation of the species. Eastern Box Turtle populations are in decline in North Carolina and while they are recognized as a threatened species by the International Union for Conservation of Nature, the turtles have no protection in North Carolina. There are currently more than 30 active research study sites across the state of North Carolina. Turtles are weighed, measured, photographed, and permanently marked. These data, along with voucher photos (photos that document sightings), are then entered into centralized database managed by the NC Wildlife Resources Commission. The Turtles\_Quiz2.csv dataset contains data collected at The Piedmont Wildlife Center in Durham.

In Homework #1, you may have have constructed a linear model with log(Annuli) as the response and log(Mass) as the predictor. This was a means to roughly estimate the age of the turtle from its mass. Is it possible that other variables have a significant interaction with the relationship of these two variables?

1. Construct a linear model with log(Annuli) as the response and log(Mass), CaptureMethod, and the interaction between log(Mass) and CaptureMethod as the predictors. Include the output for the summary of the model. Note that CaptureMethod uses the following numerical codes for how the turtle was captured: 1=road capture; 2=while mowing; 3=active search; 4=incidental; 5=radio signal; and 6=dog. 3pts

mod1 = lm(log(Annuli)~log(Mass)\*factor(CaptureMethod), data=turtles)

summary(mod1)

##

## Call:

## lm(formula = log(Annuli) ~ log(Mass) \* factor(CaptureMethod),

## data = turtles)

##

## Residuals:

## Min 1Q Median 3Q Max

## -0.93813 -0.20087 0.00084 0.16491 1.21063

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 2.76209 3.81809 0.723 0.470

## log(Mass) 0.04149 0.64992 0.064 0.949

## factor(CaptureMethod)2 -15.76746 18.25595 -0.864 0.388

## factor(CaptureMethod)3 -2.85796 4.00950 -0.713 0.477

## factor(CaptureMethod)4 -3.86608 3.82433 -1.011 0.313

## factor(CaptureMethod)5 -0.80322 4.77116 -0.168 0.866

## factor(CaptureMethod)6 -1.25611 4.69701 -0.267 0.789

## log(Mass):factor(CaptureMethod)2 2.60284 3.07732 0.846 0.398

## log(Mass):factor(CaptureMethod)3 0.46047 0.68259 0.675 0.500

## log(Mass):factor(CaptureMethod)4 0.64108 0.65107 0.985 0.326

## log(Mass):factor(CaptureMethod)5 0.12084 0.80719 0.150 0.881

## log(Mass):factor(CaptureMethod)6 0.23448 0.80032 0.293 0.770

##

## Residual standard error: 0.3261 on 294 degrees of freedom

## (21 observations deleted due to missingness)

## Multiple R-squared: 0.5349, Adjusted R-squared: 0.5175

## F-statistic: 30.74 on 11 and 294 DF, p-value: < 2.2e-16

1. How many predictor variables are in the model that you constructed in question #1? 0.5pts

11 - as shown by the variables in the summary of the model.

1. Specifically for turtles who were captured while mowing, for each 1 unit increase in log(Mass), what does your model from question #1 predict will be turtles’ increase in log(Annuli) (Your answer should be numerical)? 1.5pts

# coefficient for log(mass) for left out indicator (road capture) +

# coefficient for change in log(mass) predictor for mowing

summary(mod1)$coef[2] + summary(mod1)$coef[8]

## [1] 2.644326

1. Perform a hypothesis test to determine the importance of just the terms that involve CaptureMethod in the model constructed in question #1. List your hypotheses (be specific!), p-value, and conclusion. 3pts

H0: βi = 0; for all i in (2:11) The coefficients involving CaptureMethod  
HA: βi ≠ 0; for at least one i in (2:11) The coefficients involving CaptureMethod

Since the p-value (0.4678) is greater than 0.05, there is not evidence to suggest that at least one of the coefficients of a CaptureMethod term is nonzero.

mod4 = lm(log(Annuli)~log(Mass), data=turtles)

anova(mod4, mod1)

|  |
| --- |
|  |

|  | **Res.Df**  **<dbl>** | **RSS**  **<dbl>** | **Df**  **<dbl>** | **Sum of Sq**  **<dbl>** | **F**  **<dbl>** | **Pr(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 304 | 32.30365 | NA | NA | NA | NA |
| 2 | 294 | 31.26970 | 10 | 1.033948 | 0.9721249 | 0.4677988 |

2 rows

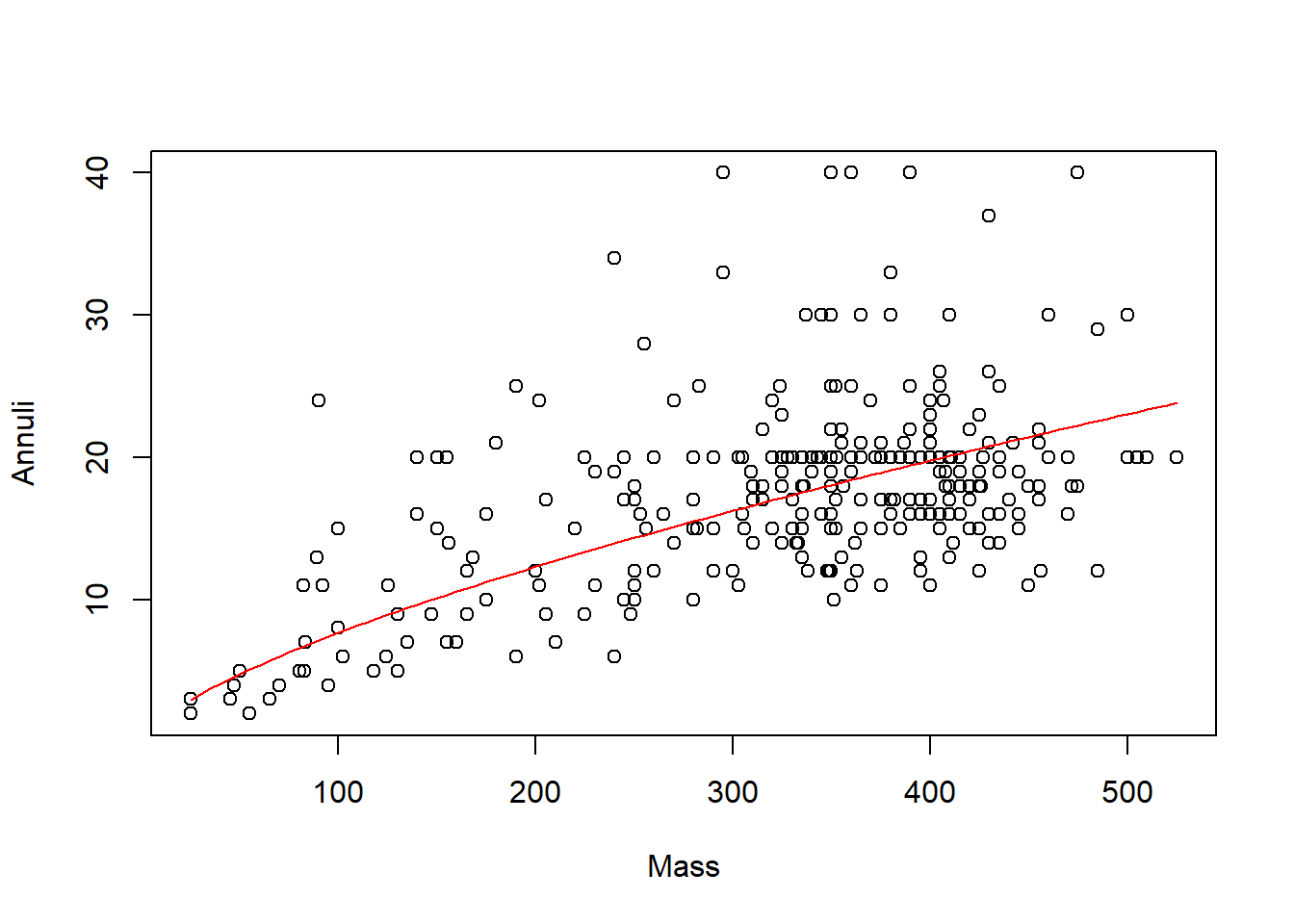
1. Plot Annuli on the vertical axis and Mass on the horizontal axis (Not the log of these variables!). Add the model that you are asked to construct in question #1 to this plot for specifically when the CaptureMethod was incidental. 2 pts

plot(Annuli~Mass, data=turtles)

B0 = summary(mod1)$coef[1,1] + summary(mod1)$coef[5,1]

B1 = summary(mod1)$coef[2,1] + summary(mod1)$coef[10,1]

curve(exp(B0)\*x^(B1), add=TRUE, col='red')



# Unit 5: Logistic Regression

## Class 26 R Logistic Regression

library(titanic)  
data("titanic\_train")  
head(titanic\_train)

## PassengerId Survived Pclass  
## 1 1 0 3  
## 2 2 1 1  
## 3 3 1 3  
## 4 4 1 1  
## 5 5 0 3  
## 6 6 0 3  
## Name Sex Age SibSp Parch  
## 1 Braund, Mr. Owen Harris male 22 1 0  
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female 38 1 0  
## 3 Heikkinen, Miss. Laina female 26 0 0  
## 4 Futrelle, Mrs. Jacques Heath (Lily May Peel) female 35 1 0  
## 5 Allen, Mr. William Henry male 35 0 0  
## 6 Moran, Mr. James male NA 0 0  
## Ticket Fare Cabin Embarked  
## 1 A/5 21171 7.2500 S  
## 2 PC 17599 71.2833 C85 C  
## 3 STON/O2. 3101282 7.9250 S  
## 4 113803 53.1000 C123 S  
## 5 373450 8.0500 S  
## 6 330877 8.4583 Q

**Logistic Regression** In all of our regression models (so far) the response variable, Y, has been quantitative. What if we want to model a categorical response?

**Categorical Response Variables** - Ways you can think about categorical response variables - WE will only focus on binary responses

* Binary Response: Whether or not a person smokes and Success of a medical treatment, where Y is divided into NOn-smoker vs smoker and X is divided into Durvies vs Dies
* Ordinal Response: Opinion Poll responses: Where Y = Agree, Netural, and Disagree
* Nominal Response: Political preference; where y = Democrat, Republican, independent

**Binary Logistic Regression** - Response variable (Y) is categorical with just two categories (yes/no or success/failure or 0/1 …). - One approach: Code the response Y as a (0,1) dummy (indicator) variable. - Assume we have a single quantitative predictor X.

**Titanic Survival** Y = Survived (0 = no; 1 = yes) X = Fare (ticket cost in dollars) - Want to predict if the people on the titance survived based on how much they paid

# amkes a table that tells you how many people survived out of the titanic overall   
# Just how many people survived total vs died  
table(titanic\_train$Survived)

##   
## 0 1   
## 549 342

# Survival related to ticket   
# SO a table on if survived based on what type of ticket class they bought   
# Shows a rough relatioship between the pclass and the others   
table(titanic\_train$Survived, titanic\_train$Pclass)

##   
## 1 2 3  
## 0 80 97 372  
## 1 136 87 119

# The below caompres the survided to teh class,   
  
# the first calss,   
# third class a lot didn't survived   
# Wouldnt an underlying variable be that there are more people buying lower class tickets than higher class tickets? I think people should look at the proportion   
# WEll, based on teh propprtions, it still looks like the upper class surived more, hmm wonder why

Below: low pvale; we ave strong ev to say that it si nonzero and there is some linear relationship

* if we look at teh model we can plot survied by fare;

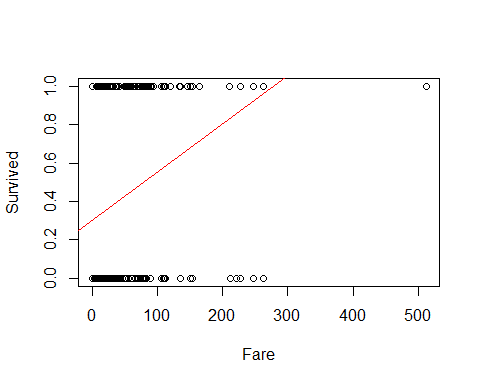
Titanic\_mod=lm(Survived ~ Fare, data=titanic\_train)  
summary(Titanic\_mod)

##   
## Call:  
## lm(formula = Survived ~ Fare, data = titanic\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.9653 -0.3391 -0.3222 0.6044 0.6973   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.3026994 0.0187849 16.114 < 2e-16 \*\*\*  
## Fare 0.0025195 0.0003174 7.939 6.12e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4705 on 889 degrees of freedom  
## Multiple R-squared: 0.06621, Adjusted R-squared: 0.06516   
## F-statistic: 63.03 on 1 and 889 DF, p-value: 6.12e-15

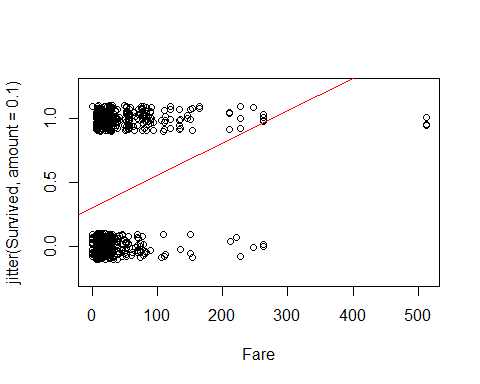
plotting surivied by fare; hard to see how dense it is; there are a lot of calues on top of the m on the bottom;

if we jitter teh data, it moves it up and down a random amount so we can see the visual difference; it doens’t chnage the data, it just changes the visual of it.

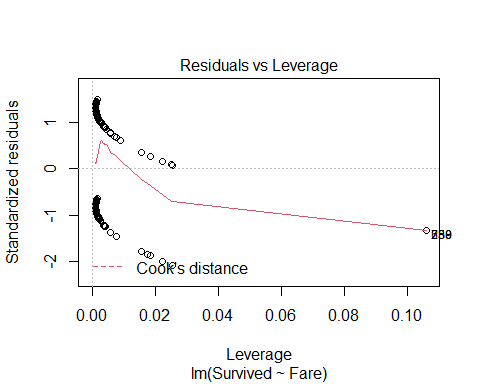
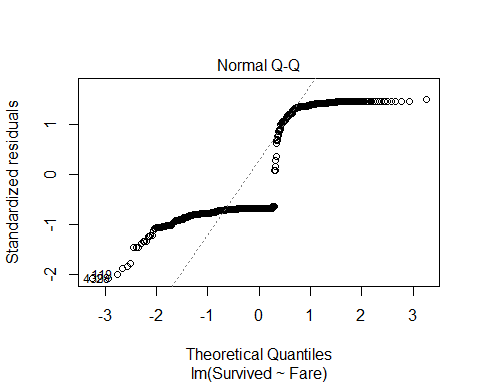
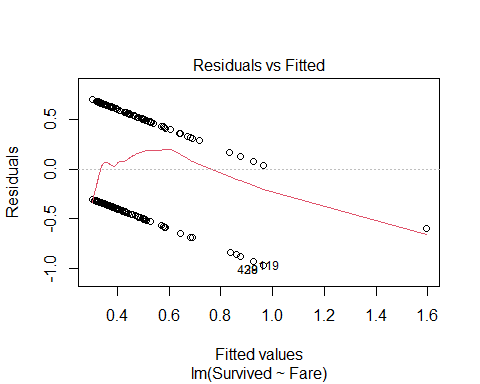
plot(Survived ~ Fare, data=titanic\_train)  
abline(Titanic\_mod, col="red")

 Jitter shows that there is a trend; there appeares to be if you paied more for your ticket, you appear to survied more - residual analysis helps see the difference in plottin gof the things

plot(  
 jitter(Survived, amount=0.1) ~ Fare,   
 ylim = c(-0.25,1.25),   
 data=titanic\_train  
 )  
  
 abline(Titanic\_mod, col="red")



plot(Titanic\_mod, c(1, 2, 5))



The aove shows that teh resultial by fitted has a path residuals are NOT normally distributed and the cook’s distance doesn’t hae any poitns of influence;

but bottom line the model doesn’t work very well

**Binary Logistic Regression Model** Y = Binary response X = Quantitative predictor π = proportion of 1’s (yes, success,…) at any x Probability form 𝜋=𝑒^(𝛽\_𝑜+𝛽\_1 𝑥)/(1+𝑒^(𝛽\_𝑜+𝛽\_1 𝑥) ) - curve(exp(B0+B1*x)/(1+exp(B0+B1*x)),add=TRUE, col=“red”)

below is a model of it by fare ; yo umaek teh family binomial What does it mean when you make the family binomial? Does that make it binary? - No it’s not binary, it means that it is a squared plot

Titanic\_logitmod = glm(Survived ~ Fare, family = binomial, data=titanic\_train)  
# Darws a curve that has a curve in teh middle with a similar likelihood of surviving or not suriving   
# We are looking at the model and predicting teh pi outcome   
# The probabiltiy of that outcome   
# Predict prob of 0 - 1someone who fits this fare would survive based ont eh model we have created   
# WE have to use teh glm function to tdo that   
  
summary(Titanic\_logitmod)

##   
## Call:  
## glm(formula = Survived ~ Fare, family = binomial, data = titanic\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.4906 -0.8878 -0.8531 1.3429 1.5942   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.941330 0.095129 -9.895 < 2e-16 \*\*\*  
## Fare 0.015197 0.002232 6.810 9.79e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1186.7 on 890 degrees of freedom  
## Residual deviance: 1117.6 on 889 degrees of freedom  
## AIC: 1121.6  
##   
## Number of Fisher Scoring iterations: 4

# this can be aline if we want it to be   
# But we have to look at it differently

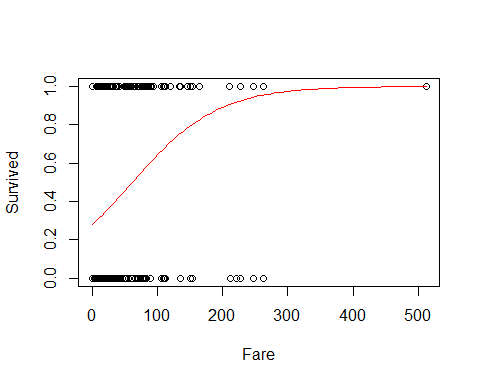
ABove, we are going to claim that the above is a linear model; we;re going to replot the data with teh jitter data; then see if the curve function equation looks nicer

we will learn later where teh curve function is coming from

we want to plot the curve = (exp(B0+B1*x))/(1+exp(B0+B1*x)) (Can see this formula filled in below)

**Predicting Proportion of “Success”** In regression the model predicts the mean Y for any combination of predictors. - What’s the “mean” of a 0/1 indicator variable? 𝑦̄=(∑𝑦\_𝑖 )/𝑛=(#” of ” 1′𝑠)/(#” of trials” )=“Proportion of "success"” - Goal for this model: Predict the “true” proportion of success, π, at any value of the predictor.

plot(Survived ~ Fare, data=titanic\_train)  
  
B0 = summary(Titanic\_logitmod)$coef[1]  
B1 = summary(Titanic\_logitmod)$coef[2]  
  
curve(exp(B0+B1\*x)/(1+exp(B0+B1\*x)),add=TRUE, col="red")



# predicitng the changes of dying based ont eh ticket you bought   
# WE say that there is about a 20% chance of dying if you payed a ceratin amount. THat's what the red line says; at what price of your ticket would oyu have X precentage of curiviing or dying

set.seed(10012020)  
passenger = titanic\_train[sample(nrow(titanic\_train),1),]  
passenger

## PassengerId Survived Pclass Name Sex Age SibSp  
## 622 622 1 1 Kimball, Mr. Edwin Nelson Jr male 42 1  
## Parch Ticket Fare Cabin Embarked  
## 622 0 11753 52.5542 D19 S

# We are randomly selecting one person so that we can check the residuals for a random value

predict(Titanic\_logitmod, passenger, type="response")

## 622   
## 0.4643927

# This is telling us, what do we predict a person who bought a certain amount's chance of surviving?   
# This looks at how much they paid for their ticket and tells us where on teh red curve we would expect this dude to fall   
# So thsi tells us that the dude has a 46% chance of surviving if he paid X amount for his ticket

**Binary Logistic Regression Model** Y = Binary response X = Quantitative predictor π = proportion of 1’s (yes, success,…) at any x Probability form: 𝜋=𝑒^(𝛽\_𝑜+𝛽\_1 𝑥)/(1+𝑒^(𝛽\_𝑜+𝛽\_1 𝑥) ) Logit form: log⁡(𝜋/(1−𝜋))=𝛽\_0+𝛽\_1 𝑥 **NOTE** The logit form can be solved to be in linear form, which is why we can use linear regression rules with it.

**Binary Logistic Regression Model** **Probability Form:** P(X) = ((e(Bo+B1X))/(e(B0+B1X)+1)) **Logit Form:** ln(p/(1-p)) = B0 + B1X

**Example: Golf Putts** Build a model to predict the proportion of putts made (success) based on length (in feet). Data are in Putts1 of Stat2Data.

library(Stat2Data)  
  
data("Putts1")  
head(Putts1)

## Length Made  
## 1 3 1  
## 2 3 1  
## 3 3 1  
## 4 3 1  
## 5 3 1  
## 6 3 1

**Logistic Regression for Putting**

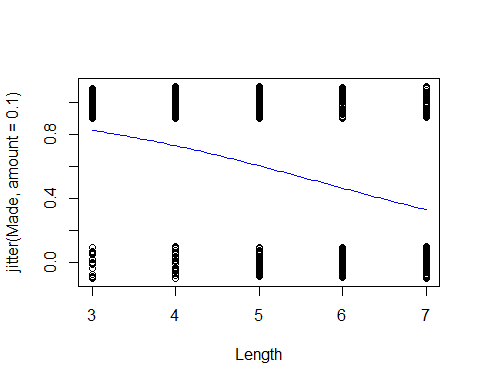
modPutt=glm(Made~Length,family=binomial,data=Putts1)  
summary(modPutt)

##   
## Call:  
## glm(formula = Made ~ Length, family = binomial, data = Putts1)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.8705 -1.1186 0.6181 1.0026 1.4882   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.25684 0.36893 8.828 <2e-16 \*\*\*  
## Length -0.56614 0.06747 -8.391 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 800.21 on 586 degrees of freedom  
## Residual deviance: 719.89 on 585 degrees of freedom  
## AIC: 723.89  
##   
## Number of Fisher Scoring iterations: 4

# pvalues are small   
# So we like these realtionships, but we should plot it to see what it actually loks like and that is done below

logit = function(B0, B1, x)  
 {  
 exp(B0+B1\*x)/(1+exp(B0+B1\*x))  
}  
# THis function will make the curve that we need on the curve above   
#SO , this is the same as above

B0 = summary(modPutt)$coef[1]  
B1 = summary(modPutt)$coef[2]  
  
plot(jitter(Made,amount=0.1)~Length,data=Putts1)  
# These lines overall plot what the data looks like   
curve(exp(B0+B1\*x)/(1+exp(B0+B1\*x)),add=TRUE, col="red")  
# This line is the line that we would use to predict someone will make a put based on the distance from the hole.   
  
# Can also use the logit function   
curve(logit(B0, B1, x), add = TRUE, col = "blue")



# THis will make the same line as above

logit = function(B0, B1, x)  
 {  
 exp(B0+B1\*x)/(1+exp(B0+B1\*x))  
}  
# THis function will make the curve that we need on the curve above   
#SO , this is the same as above

**Golf Putts Probabilities** 𝜋̂=𝑒(3.257−0.5661𝐿𝑒𝑛𝑔𝑡ℎ)/(1+𝑒(3.257−0.5661𝐿𝑒𝑛𝑔𝑡ℎ) ) Where phat = 𝑝̂=(# 𝑚𝑎𝑑𝑒)/(# 𝑡𝑟𝑖𝑎𝑙𝑠)

**Golf Putts Probabilities** Length: 3,4,5,6,7 phat: 0.835, 0.739, 0.565, 0.488, 0.328 pihat: 0.826, 0.730, 0.605, 0.465, 0.330

Making a table of those that are made vs failed at different lengths

Putts.table = table(Putts1$Made, Putts1$Length)  
Putts.table

##   
## 3 4 5 6 7  
## 0 17 31 47 64 90  
## 1 84 88 61 61 44

p.hat = as.vector(Putts.table[2,]/colSums(Putts.table))  
# Make it a vector because we want to be able to use it with dataframes   
p.hat

## [1] 0.8316832 0.7394958 0.5648148 0.4880000 0.3283582

pi.hat=0  
  
# Compare the predictions, so 3 - 7 feet; so from 3 ft to 4 ft to 5 ft to 6 ft etc.   
# Will make the pihat for each of these   
# Pihat = the probability of sucess at a certain feet distance   
# Pi = success/trials   
for(i in 3:7)  
 {  
 pi.hat[i-2] = logit(B0, B1, i)  
 }  
  
pi.hat

## [1] 0.8261256 0.7295364 0.6049492 0.4650541 0.3304493

# Makea a dataframe that tells you the pihat values and the p hat values   
# We dont know the difference btween pi and p hat   
Putts = data.frame(  
 "Length" = c(3:7),   
 "p.hat" = p.hat,   
 "pi.hat" = pi.hat)  
  
head(Putts)

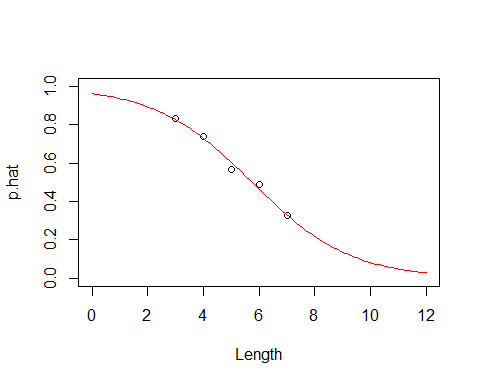
## Length p.hat pi.hat  
## 1 3 0.8316832 0.8261256  
## 2 4 0.7394958 0.7295364  
## 3 5 0.5648148 0.6049492  
## 4 6 0.4880000 0.4650541  
## 5 7 0.3283582 0.3304493

* Probability form of puttin gmodel\*
* etended from 0 - 12; the points on teh graph are the actual proprtions that were made; the p hat values;

the line shows the pi hat values; teh line on the 4 = what we predict; teh dot = the actual value - these are close adnthis is how we test teh linear model - we are going to put it back to the logit form so that we cna put this on a line and we want to see if teh data fits teh lie - if it doens’t then we will have to do transformations - it gets a lot math-y-er we agoig to spend next class talking more about the math how to lok at hypothesis testing and anova doesn’t make sence here anymore because teh residual doesn’t work each point represntes differen combination of data points.

**Probability Form of Putting Model**

plot(p.hat~Length,ylim=c(0,1), xlim=c(0,12), data=Putts)  
curve(logit(B0, B1, x),add=TRUE, col="red")



## Class 27 R Logistic Regression: Odds Ratio And Inferences

library(Stat2Data)  
  
data("Putts1")  
head(Putts1)

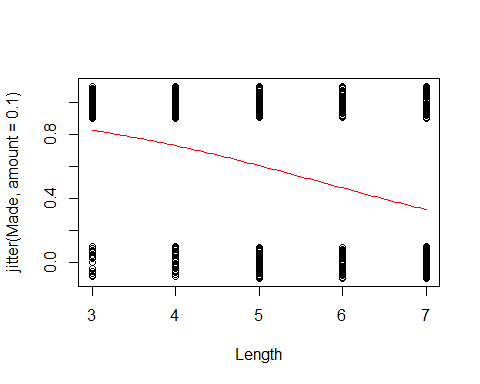
## Length Made  
## 1 3 1  
## 2 3 1  
## 3 3 1  
## 4 3 1  
## 5 3 1  
## 6 3 1

**Logistic Regression for Putting**

# Use glm for different types of graphs  
modPutt=glm(Made~Length,family=binomial,data=Putts1)  
summary(modPutt)

##   
## Call:  
## glm(formula = Made ~ Length, family = binomial, data = Putts1)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.8705 -1.1186 0.6181 1.0026 1.4882   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.25684 0.36893 8.828 <2e-16 \*\*\*  
## Length -0.56614 0.06747 -8.391 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 800.21 on 586 degrees of freedom  
## Residual deviance: 719.89 on 585 degrees of freedom  
## AIC: 723.89  
##   
## Number of Fisher Scoring iterations: 4

B0 = summary(modPutt)$coef[1]  
B1 = summary(modPutt)$coef[2]  
  
plot(jitter(Made,amount=0.1)~Length,data=Putts1)  
curve(exp(B0+B1\*x)/(1+exp(B0+B1\*x)),add=TRUE, col="red")



**Golf Putts Probabilities** 𝜋̂=𝑒(3.257−0.5661𝐿𝑒𝑛𝑔𝑡ℎ)/(1+𝑒(3.257−0.5661𝐿𝑒𝑛𝑔𝑡ℎ) ) 𝑝̂=(# 𝑚𝑎𝑑𝑒)/(# 𝑡𝑟𝑖𝑎𝑙𝑠) - THis is also a part of Class 26

# This makes a table so we can then make the proportion of success for the golf putts probabilities   
  
Putts.table = table(Putts1$Made, Putts1$Length)  
Putts.table

##   
## 3 4 5 6 7  
## 0 17 31 47 64 90  
## 1 84 88 61 61 44

# Proportion made for each distance   
# takes the probabilities from the putts table and see the proportion for each distance so it's the total for distance 3  
# P(success)/Ntrials  
p.hat = as.vector(Putts.table[2,]/colSums(Putts.table))  
p.hat

## [1] 0.8316832 0.7394958 0.5648148 0.4880000 0.3283582

logit = function(B0, B1, x)  
 {  
 exp(B0+B1\*x)/(1+exp(B0+B1\*x))  
 }

pi.hat=0  
  
for(i in 3:7)  
 {  
 pi.hat[i-2] = logit(B0, B1, i)  
 }  
  
pi.hat

## [1] 0.8261256 0.7295364 0.6049492 0.4650541 0.3304493

Putts = data.frame(  
 "Length" = c(3:7),   
 "p.hat" = p.hat,   
 "pi.hat" = pi.hat)  
  
head(Putts)

## Length p.hat pi.hat  
## 1 3 0.8316832 0.8261256  
## 2 4 0.7394958 0.7295364  
## 3 5 0.5648148 0.6049492  
## 4 6 0.4880000 0.4650541  
## 5 7 0.3283582 0.3304493

he above is all review from last class that we didn’t get to p-hat = the 3:7, length is 3:7; this sis from teh putts data,

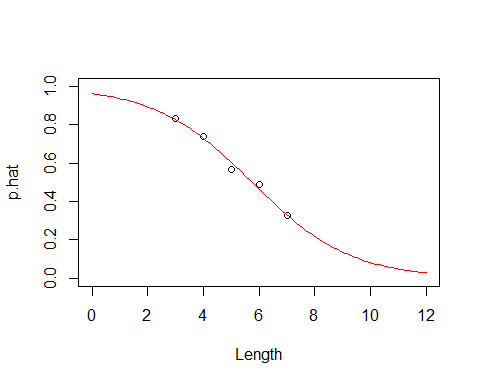
the below plots the curve ontop of it

logit, logs the thing, i think; see th formula above code

if we change x limits, it shows the smaller vs bigger graph

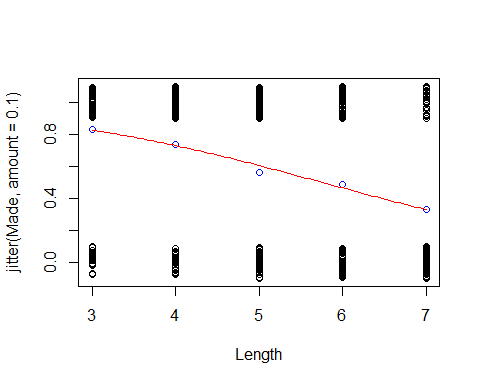
**Probability Form of Putting Model**

plot(p.hat~Length,ylim=c(0,1), xlim=c(0,12), data=Putts)  
# Shows the proportion we are predicting to the prediction plot that we have   
curve(logit(B0, B1, x),add=TRUE, col="red")

 Putts1 = a different dataset; we want to putthe same points on this differently

the blue points are the mean values for each value of x

plot(jitter(Made,amount=0.1)~Length,data=Putts1)  
  
points(p.hat~Length, data=Putts, col='blue')  
  
curve(logit(B0, B1, x),add=TRUE, col="red")

 THink about the Odds of something happening; - the odds vs probability - the odds against are 4:1; they expect those one horse to lose 4/5 vs 1/5 of the time

if pi = proportion of yes (success 1, etc)

the odds of yes = P(pi)/(P(1-pi)) odds of yes = P(yes)/P(no)

logit form = log(odds of yes) = B0 + B1X Below adds 2 new columns to teh dataframe, this messes with teh data

one is probability of it happened over teh probaility that it doesnt haveppen

pi = predicted

**Odds** The odds against a certain horse winning a race are 4 to 1.  
- What does that mean? – 4 losses for every 1 win – P(Win) = 1/5 – P(Loss) = 4/5

𝑂𝑑𝑑𝑠= (𝑃(𝑊𝑖𝑛))/(𝑃(𝐿𝑜𝑠𝑠))=(1/5)/(4/5)=1/4

**Odds** If pi = proportion of “yes” (success, 1, ….) the odds of yes are(is)

(𝑃(𝑦𝑒𝑠))/(𝑃(𝑛𝑜))=𝜋/(1−𝜋)

With a little bit of algebra… 𝑜𝑑𝑑𝑠=𝜋/(1−𝜋)⇔𝜋=𝑜𝑑𝑑𝑠/(1+𝑜𝑑𝑑𝑠)

**Odds and Logistic Regression** Logit form: log⁡(𝜋/(1−𝜋))=𝛽\_𝑜+𝛽\_1 𝑋 -The logistic model assumes a linear relationship between the predictor and log(odds). - log⁡( 𝑜𝑑𝑑𝑠)=𝛽\_𝑜+𝛽\_1 𝑋

**Logit Form of Putting Model**

**Back to Putting Data** Since we have lots of putts, we can estimate 𝑝̂ (proportion of putts made) at each length 𝑝̂=(# 𝑚𝑎𝑑𝑒)/(# 𝑡𝑟𝑖𝑎𝑙𝑠) and the odds (𝑜𝑑𝑑𝑠)̂=(# 𝑚𝑎𝑑𝑒)/(# 𝑚𝑖𝑠𝑠𝑒𝑑)=𝑝̂/(1−𝑝̂ ) and find log⁡((𝑜𝑑𝑑𝑠̂) at each length.

**Golf Putts Odds** (𝑜𝑑𝑑𝑠)̂=(# 𝑚𝑎𝑑𝑒)/(# 𝑚𝑖𝑠𝑠𝑒𝑑)=𝑝̂/(1−𝑝̂ ) (from sample) (𝑜𝑑𝑑𝑠)̂=𝜋̂/(1−𝜋̂ ) (from logistic regression) Length: 3,4,5,6,7 oddshat(from sample): 4.94,2.84,1.30,0.95,0.49 oddshat(from regression): 4.75,2.70,1.53,0.87,0.49

Putts$p.Odds = Putts$p.hat/(1-Putts$p.hat)  
Putts$pi.Odds = Putts$pi.hat/(1-Putts$pi.hat)  
  
head(Putts)

## Length p.hat pi.hat p.Odds pi.Odds  
## 1 3 0.8316832 0.8261256 4.9411765 4.751277  
## 2 4 0.7394958 0.7295364 2.8387097 2.697355  
## 3 5 0.5648148 0.6049492 1.2978723 1.531320  
## 4 6 0.4880000 0.4650541 0.9531250 0.869348  
## 5 7 0.3283582 0.3304493 0.4888889 0.493539

**Plot for Putts Data** Plot log⁡((𝑜𝑑𝑑𝑠̂) versus Length (3, 4, 5, 6, 7) Add a line with intercept and slope from the logistic model.

The below code does something

the line tells you how the probaility chanes as teh rate of other things change.

so we need to think fo teh odds ratio, a common way to compare two groiups is to look at a ratio of their odds

odds ratio (OR) = Odd.R = Odd1/Odd2

odds using data from 4 ft = 2.84 odds using data from 3 feet = 4.94

odds ratio ( 4 to 3) = 2.84/4.94 = 0.57

the odds of making a putt from 4 feet are 57% of the odds of making from 3 feet

*Interpreting Slope Using Odds Ratio*

log(Odds) = B0+B1X -> odds = e^(B0+B1\*X)

*CI for Slope and ODds Ratio* - Using teh SE for the slope, find a CI for B1 with:

B-hat1 +/- z-star \* SE

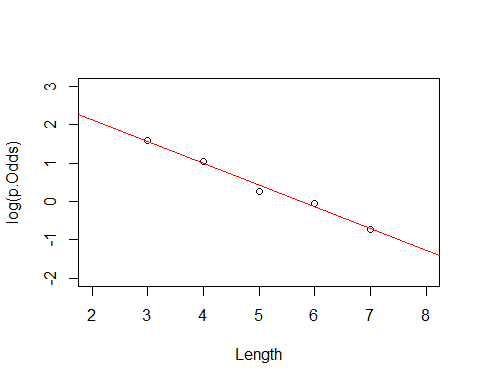
the above will get you theCI for teh odds ratio (E^B1) exponentiate the CI for B1

ex:

CI for slope = (0.5, 0.6) CI for OR = e^0.5, e^0.6 = (0.497, 0.648)

**Logit Form of Putting Model**

plot(log(p.Odds)~Length, data=Putts, xlim=c(2,8), ylim=c(-2,3))  
abline(B0, B1, col="red")



**Odds Ratio** A common way to compare two groups is to look at the ratio of their odds “Odds Ratio”=” OR “=(”Odd” “s” \_1)/(“Odd” “s” \_2 )

**Putting Data** Odds using data from 4 feet = 2.84 Odds using data from 3 feet = 4.94 Odds ratio (4 ft to 3 ft) = 2.84/4.94=0.57 The odds of making a putt from 4 feet are 57% of the odds of making from 3 feet.

**Odds Ratios for Putts** From fitted logistic: Length: 3,4,5,6,7, pihat: 0.826,0.730,0.605,0.465,0.331 odds hat: 4.75,2.70,1.53,0.84,0.49

To Odds Ratio Length: 4-3 ft, 5-4 ft, 6-5 ft, 7-6 ft Odds Ratio: 0.57, 0.57, 0.57, 0.57

In a logistic model, the odds ratio when changing the predictor by one is constant.

**Odds Ratios for Putts** From samples at each distance: Length: 3,4,5,6,7 phat: 0.832, 0.739, 0.565, 0.488, 0.328 oddshat: 4.94, 2.84, 1.30, 0.95, 0.49

To Odds Ratio: Length: 4-3ft, 5-4 ft, 6-5ft, 7-6ft Odds Ratio: 0.57, 0.46, 0.73, 0.51

**Interpreting “Slope” using Odds Ratio** log⁡(𝑜𝑑𝑑𝑠)=𝛽\_0+𝛽\_1 𝑥goes to 𝑜𝑑𝑑𝑠=𝑒^(𝛽\_0+𝛽\_1 𝑥) What happens when we increase x by one? 𝑒^(𝛽\_0+𝛽\_1 (𝑥+1))=𝑒^(𝛽\_0+𝛽\_1 𝑥)∙𝑒^(𝛽\_1 ) When we increase x by one, the odds increase/decrease by a factor of 𝑒^(𝛽\_1 ) (odds ratio). For putts: The odds of making a putt decrease by a factor of 0.57 (𝑒^(−0.566)) for every extra foot of length.

**CI for Slope and Odds Ratio** Using the SE for the slope, find a CI for β1 with 𝛽̂\_1±𝑧^∗∙𝑆𝐸 <- this is just the formula for confidence intervals To get CI for the odds ratio (𝑒^(𝛽\_1 )) exponentiate the CI for β1

CI for slope: −0.566±1.96(0.06747) =(−0.698,−0.434)

CI for OR: 〖(𝑒〗(−0.698),𝑒(−0.438)) =(0.497, 0.648)

SE\_B1 = summary(modPutt)$coef[2,2]  
exp(B1 - SE\_B1\*qnorm(0.975))

## [1] 0.4973894

exp(B1 + SE\_B1\*qnorm(0.975))

## [1] 0.6479761

in practice we are not going to use teh confint.default function; because the default forces the thing to use z scores; and teh not by default uses some lo glikelihoods to get this thing

**Similar tests/measures for logistic regression?** Recall: “Ordinary” Regression lm(formula = Active ~ Rest)

Coefficients: Estimate Std. Error t value Pr(>|t|)  
(Intercept) 8.75340 5.60773 1.561 0.12  
Rest 1.18387 0.08214 14.413 <2e-16 \*\*Rest, above, tests for individual coefficients\*

*For the first three italics, there are for comparing the models* Residual standard error: *14.39* on 310 degrees of freedom Multiple R-squared: *0.4012*, Adjusted R-squared: *0.3993* F-statistic: *207.7* on 1 and 310 DF, p-value: *< 2.2e-16* <- test overall fit

**Test for Individual Coefficients** Ho: Bi = 0 Ha: Bi != 0

t.s. = Bhat/SEofBhat (R will give you all of these variables) Interpret as with individual t-tests in ordinary regression P-value = 2P( Z > |t.s.| )

Estimate Std. Error z value Pr(>|z|)

(Intercept) 3.25684 0.36893 8.828 <2e-16  ***Length -0.56614 0.06747 -8.391 <2e-16***

**Estimating Parameters in Ordinary Regression** Coefficients are chosen to minimize the sum of the squared errors in the observed sample. (Least Squares Estimation) 𝑆𝑆𝐸=〖Σ(𝑦−𝑦̂)〗^2 **WE WANT A SMALL SSE**

**Test for Overall Fit** Ho: B1 = 0 -> log(odds) = Bo Ha: B1 != 0 -> log(Odds) = Bo + B1X

How much “better” does the linear model do than one with a constant? Is it “significantly” better?

**Maximizing the Likelihood of the Sample** - Suppose that there are three decks of cards: 1. Standard 52 card deck 2. Euchre deck (9, 10, J, Q, K, A) 3. Deck with all red cards If two cards were drawn from a deck (without replacement), a Jack of Hearts, then a King of Hearts, from which deck do you think that there were chosen?

* Suppose that there are three decks of cards:

1. Standard 52 card deck; (1/52)(1/51)≈“0.000377”
2. Euchre deck (9, 10, J, Q, K, A); (1/24)(1/23)≈“0.001812”
3. Deck with all red cards; (1/26)(1/25)≈“0.001538”

**Estimating Parameters in Logistic Regression** Parameters are chosen to maximize the likelihood of the observed sample. (Maximum Likelihood Estimation) If the ith data point is YES (yi=1), calculate 𝜋̂\_𝑖 If the ith data point is NO (yi=0), calculate 1−𝜋̂\_𝑖

Likelihood:𝐿=∏〖𝜋̂\_𝑖〗^(𝑦\_𝑖 ) (1−𝜋̂\_𝑖 )^(1−𝑦\_𝑖 ) **WE WANT A HIGH LIKELIHOOD**

**Test for Overall Fit** Length: 3,4,5,6,7, Made: 84,88,61,61,44 Missed: 17,31,47,64,90 Ratio: 0.826, 0.730, 0.605, 0.465, 0.330

𝐿=∏〖𝜋̂\_𝑖〗^(𝑦\_𝑖 ) (1−𝜋̂\_𝑖 )^(1−𝑦\_𝑖 ) Ho: B1 = 0 -> log(odds) = Bo L = .576^338\*(1-.576)^249

Ha: B1 != 0 -> log(Odds) = Bo + B1X L = (0.826^84 \* 0.174^17) \* (0.730^88 \* 0.270^31) \* (0.605^61 \* 0.395^47) \* (0.465^61 \* 0.535^64) \* (0.330^44 \* 0.670^90)

exp(confint.default(modPutt))

## 2.5 % 97.5 %  
## (Intercept) 12.6006177 53.5133410  
## Length 0.4973894 0.6479761

exp(confint(modPutt))

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) 12.7974573 54.4505172  
## Length 0.4960611 0.6464444

just keep in mind the units for the CI

Similar tests/measures for logistic regression

recall: “Ordinary” regression

*Test for idividual coeff*

Ho: Bi = 0  
Ha: Bi -/= 0

t.s = B-hati/SE(B-hati)

R WIll give you all of these numbers

interpret as with individual t tests in ordinary regression

p-value = 2P(Z>abs(t.s))

*Estimating Parameters in ORd Regression*

Coeff are chosen to min the sum of the squared errors in teh observed sample (LEast Squares Estimation

SEE = sum(y-y-hat)^2

We want a small SSE

*Test for Overall Fit* H0: B1 = 0 Ha: B1 =/= 0 log(odds) = B0 log(odds) = B0 + B1X; these are competing models

how much better does the lienar mdoel do than one with a constatst? IS sit sig better?

*Estimating Parameters in Logistic Regression* Parameters are chosen to max the likelihood of the observed sample (MAx likelihood estimation)

If teh ith data poin is YES (yi = 1), calc pi-hati

If teh ith data point is No (yi = 0), calc 1-pi-hati

We want L to be big

THis is where the table(Putts1$MAde) starts

summary(modPutt)

##   
## Call:  
## glm(formula = Made ~ Length, family = binomial, data = Putts1)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.8705 -1.1186 0.6181 1.0026 1.4882   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.25684 0.36893 8.828 <2e-16 \*\*\*  
## Length -0.56614 0.06747 -8.391 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 800.21 on 586 degrees of freedom  
## Residual deviance: 719.89 on 585 degrees of freedom  
## AIC: 723.89  
##   
## Number of Fisher Scoring iterations: 4

**−2 ln⁡(𝐿) for Constant (H0) Model** For a constant model:  
𝐿\_0=𝜋̂^(#𝑦𝑒𝑠) 〖(1−𝜋̂)〗^(𝑛−#𝑦𝑒𝑠) log⁡(𝐿\_0 )=#𝑦𝑒𝑠∙log⁡𝜋̂ )+#𝑛𝑜∙log-pihat

Combining all putts: 338 made out of 587 𝜋hat =338/587=0.5758 𝐿\_0=〖0.5758〗^338 〖0.4242〗^249

log⁡(𝐿\_0 )=338 log⁡(0.576)+249 log⁡(0.424)=−400.1 〖−2log〗⁡(𝐿\_0 )=800.2

**Putts1: Made~Length** lmodPutt=glm(Made~Length,family=binomial,data=Putts1) summary(lmodPutt)

**Example: Golf Putts** 𝐿=∏〖𝜋̂\_𝑖〗^(𝑦\_𝑖 ) (1−𝜋̂\_𝑖 )^(1−𝑦\_𝑖 ) 𝐿=〖0.826〗^84 〖0.174〗^17 〖0.730〗^88 〖0.270〗31⋯〖0.330〗44 〖0.670〗^90 log⁡(𝐿)=84 log⁡(0.826)+17 log⁡(0.174)+⋯ +44 log⁡(0.330)+90 log⁡(0.670)=−359.9 Coefficients are chosen to get 𝑙𝑜𝑔(𝐿) as big as possible 〖−2log〗⁡(𝐿)=718. 8 <- Minimize residual deviance

* How much “improvement” with the predictor?
* Compare the null deviance with the residual deviance; subtract the two to get your Gstatistic

lmodPutt=glm(Made~Length,data=Putts1,family=binomial,data=Putts1) summary(lmodPutt)

Coefficients: Estimate Std. Error z value Pr(>|z|) (Intercept) 3.25684 0.36893 8.828 <2e-16 Length -0.56614 0.06747 -8.391 <2e-16

Null deviance: 800.21 on 586 degrees of freedom

Residual deviance: 719.89 on 585 degrees of freedom

−2 l𝑜𝑔⁡(𝐿\_0 )−(−2 log⁡(𝐿) )=800.2−719.99=80.3 This difference is called the G statistic.

**Evaluating Overall Fit** Test for overall fit (Similar to regression ANOVA) t.s. = G = improvement in –2log(L) over a model with just a constant term Compare to y2 with k d.f. (chi-square) - k = number of predictiors

The null sys tat it doens’t matter how far we are from teh hole, while teh laternative says that it does matter

Bo = 0 Bo =/= 0

table(Putts1$Made)

##   
## 0 1   
## 249 338

338/(338+249)

## [1] 0.5758092

L.null = (.576)^338\*(1-.576)^249  
L.null

## [1] 1.725431e-174

-2\*log(L.null)

## [1] 800.2087

if the distance matters, then teh difference lengts =will ahev different values

we first calc how you got the sample from 3 ft putts

based on data, we made 0.73 putts at 3ft, then the probabiltiy of making it

the log(L) below is a little bigger than thte above L.null, which means taht we like the second L better

L = 0.826^84\*0.174^17\*0.730^88\*0.270^31\*.605^61\*.395^47\*.465^61\*.535^64\*0.330^44\*0.670^90  
L

## [1] 4.765502e-157

-2\*log(L)

## [1] 719.8889

do things with chi-squared; it lieks chi squared, so we like log?

we cna look at this like a chi squared

how likeily is it to get this on a chi squared distribution

summary(modPutt)

##   
## Call:  
## glm(formula = Made ~ Length, family = binomial, data = Putts1)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.8705 -1.1186 0.6181 1.0026 1.4882   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.25684 0.36893 8.828 <2e-16 \*\*\*  
## Length -0.56614 0.06747 -8.391 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 800.21 on 586 degrees of freedom  
## Residual deviance: 719.89 on 585 degrees of freedom  
## AIC: 723.89  
##   
## Number of Fisher Scoring iterations: 4

1-pchisq(80.3,1)

## [1] 0

summary(modPutt)

##   
## Call:  
## glm(formula = Made ~ Length, family = binomial, data = Putts1)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.8705 -1.1186 0.6181 1.0026 1.4882   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.25684 0.36893 8.828 <2e-16 \*\*\*  
## Length -0.56614 0.06747 -8.391 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 800.21 on 586 degrees of freedom  
## Residual deviance: 719.89 on 585 degrees of freedom  
## AIC: 723.89  
##   
## Number of Fisher Scoring iterations: 4

G = summary(modPutt)$null.deviance - summary(modPutt)$deviance  
  
1 - pchisq(G,1)

## [1] 0

the below gives you how likly we would see this by chacne; we if small p value; then we can reject Ho

**Evaluating Overall Fit** Ho: Bi = 0 Ha: Bi != 0

log⁡(𝜋/(1−𝜋))=𝛽\_𝑜+𝛽\_1 𝑋

anova(modPutt, test="Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: Made  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 586 800.21   
## Length 1 80.317 585 719.89 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Class 28 R Logistic Regression Assessing the model

library(Stat2Data)  
library(readr)  
  
GoldenBalls <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/GoldenBalls.csv")  
  
logit = function(B0, B1, x)  
 {  
 exp(B0+B1\*x)/(1+exp(B0+B1\*x))  
 }

**Checking Linearity** Three methods depending on the type of dataset: - Datasets with a binary predictor – nothing to check! - Datasets with a quantitative predictor with many response values for each predictor - Datasets with a quantitative predictor with many values for the predictor but few response values for each predictor value.

*Example Golf Putts* We looked at the log odds formula there wre no big deviations

what about the Datsets with binary predicotrs? We are going to look at golden balls; split or steal the money how likeyly are they to split or steal based on teh age of teh contestant?

head(GoldenBalls)

## # A tibble: 6 x 2  
## Over40 Split  
## <dbl> <dbl>  
## 1 1 1  
## 2 1 1  
## 3 1 1  
## 4 1 1  
## 5 1 1  
## 6 1 1

table(GoldenBalls$Split, GoldenBalls$Over40)

##   
## 0 1  
## 0 195 76  
## 1 187 116

Claim: There is a difference in the proportion of people who would split or steal based on their age.

He table above shows the table for if the preson is over40 and has split or not; so row, column; where row = if they split and coloumn = age over or below 40. over forty = 1; under = 0

we see that there is a low p value below that says that we have evidence to say there is a realtionship between age and their saying yes or no to steal

then we pull out teh coeffs. we plot the raw data with some jitter, but we dont think tis super needed right now.

if we want to make a logit plot,then we want to find out where the table is, why do we table 2 and coloumn; p-hat will give you the samplel porp[tion, he is just pulling from the table

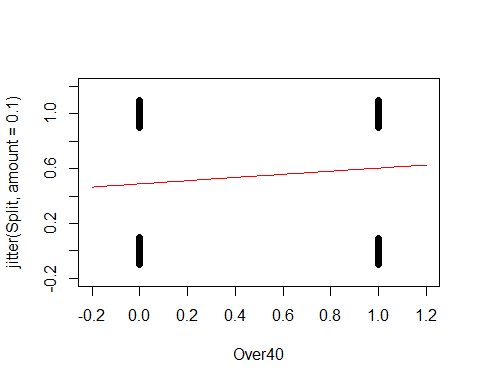
then goldenball odds is just loged

**Golden Balls: Logistic Regression**

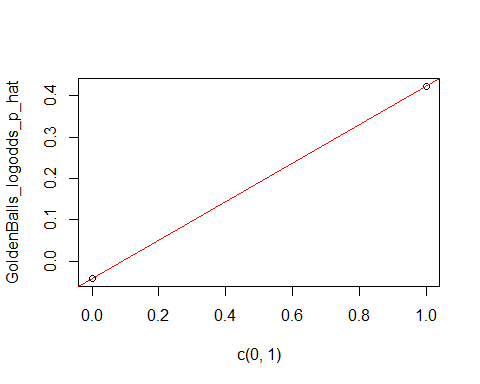
GBmod = glm(Split~Over40, data=GoldenBalls, family=binomial)  
# Make logistic model  
summary(GBmod)

##   
## Call:  
## glm(formula = Split ~ Over40, family = binomial, data = GoldenBalls)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.361 -1.160 1.004 1.195 1.195   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.04189 0.10235 -0.409 0.68233   
## Over40 0.46475 0.17960 2.588 0.00966 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 793.95 on 573 degrees of freedom  
## Residual deviance: 787.17 on 572 degrees of freedom  
## AIC: 791.17  
##   
## Number of Fisher Scoring iterations: 4

# See summary of logistic mdoel   
  
B0 = summary(GBmod)$coef[1] # Pull out intercept  
B1 = summary(GBmod)$coef[2] # Pull out slope  
  
#Plot the GBMod data  
plot(jitter(Split,amount=0.1)~Over40,data=GoldenBalls, xlim=c(-.2, 1.2), ylim=c(-.2, 1.2) )  
# Plot the GBMod  
curve(logit(B0, B1, x),add=TRUE, col="red")



GoldenBalls\_table=table(GoldenBalls$Split, GoldenBalls$Over40)  
GoldenBalls\_p\_hat=as.vector(GoldenBalls\_table[2,]/colSums(GoldenBalls\_table))  
GoldenBalls\_logodds\_p\_hat = log(GoldenBalls\_p\_hat/(1-GoldenBalls\_p\_hat))  
  
plot(GoldenBalls\_logodds\_p\_hat~c(0,1))  
abline(B0, B1, col="red")



**Quantitative predictor: Few response values for each predictor** - This process breaks down if there are not many values of the response, but the previous process can be mimicked by manipulating the data first.

* We can manipulate the data by:

1. Slicing the x-axis into intervals.
2. Compute the average x-value and empirical logit foreach slice
3. Plot the values as before

Medical School Acceptance Dataset: Is GPA a useful predictor for acceptance to medical school?

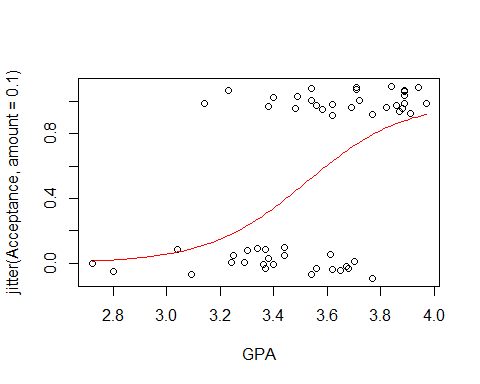
data("MedGPA")  
head(MedGPA)

## Accept Acceptance Sex BCPM GPA VR PS WS BS MCAT Apps  
## 1 D 0 F 3.59 3.62 11 9 9 9 38 5  
## 2 A 1 M 3.75 3.84 12 13 8 12 45 3  
## 3 A 1 F 3.24 3.23 9 10 5 9 33 19  
## 4 A 1 F 3.74 3.69 12 11 7 10 40 5  
## 5 A 1 F 3.53 3.38 9 11 4 11 35 11  
## 6 A 1 M 3.59 3.72 10 9 7 10 36 5

WE’re hoping that therea re different cases for the low, medium and high; if they didn’t follow the logit curve, ten they would be weird and you cant use logit

below we are predicting the mdoel with acceptabnce and gpa

MedGPA.glm = glm(Acceptance~GPA, data=MedGPA, family = binomial)  
  
B0 = summary(MedGPA.glm)$coef[1]  
B1 = summary(MedGPA.glm)$coef[2]  
  
plot(jitter(Acceptance,amount=0.1)~GPA,data=MedGPA)  
curve(logit(B0, B1, x),add=TRUE, col="red")

 It makes sense that teh logit model fits here, the ost people who get in have high gpa it looks like its a jittered acceptabnce rate we dont know other things about these people, there are probablyother reasons they arent getting or they gota ccepted into med school

below we are shorting based on how good their gpa is, this is just ordering it by gpa; so its going to look at the levels **Create a new dataframe with the predictor sorted smallest to largest**

sorted.MedGPA = MedGPA[order(MedGPA$GPA),]  
GPA = sorted.MedGPA$GPA  
Acceptance = sorted.MedGPA$Acceptance  
  
#we want to pull out GPA so we can just work with that; so we do gpa = sorted.medgpa$gpa   
#then we also want to know if thery got acepted, so see teh Acceptance object in R   
  
# Select a number of “slices” or groups for the data and find the mean value of the predictor for each slice  
# WE slect slices so we can look at the mean of the groups   
# We want to see if our model follows the means of the groups well   
  
groups = 5  
group.size = 11  
  
GPA.means = 0  
Acceptance.sums = 0  
  
for(i in 1:groups){  
 GPA.means[i] = mean(  
 GPA[((group.size\*i)-(group.size-1)):(group.size\*i)])  
 }  
  
GPA.means

## [1] 3.130909 3.410000 3.585455 3.734545 3.905455

above does a loop where it takes teh mean of teh first 11 elements, tehn it’s teh second 11 elements; so its not easy to read this, there is a nicer way to do this:

library(TTR)  
runMean(GPA, 11)

## [1] NA NA NA NA NA NA NA NA  
## [9] NA NA 3.130909 3.189091 3.240909 3.270909 3.297273 3.319091  
## [17] 3.334545 3.349091 3.366364 3.380000 3.396364 3.410000 3.426364 3.441818  
## [25] 3.457273 3.473636 3.490000 3.506364 3.525455 3.541818 3.558182 3.570909  
## [33] 3.585455 3.597273 3.610000 3.623636 3.636364 3.650000 3.661818 3.671818  
## [41] 3.685455 3.699091 3.717273 3.734545 3.751818 3.769091 3.786364 3.803636  
## [49] 3.820000 3.836364 3.851818 3.864545 3.880000 3.893636 3.905455

the above will do the same thing as teh loop, but it’s from the TTR package this gives us all of the 11, we jsut want the 11th, 22th, 33rd, and 44th values; how do we get that? look below

runMean(GPA, 11)[seq(11,length(GPA),11)]

## [1] 3.130909 3.410000 3.585455 3.734545 3.905455

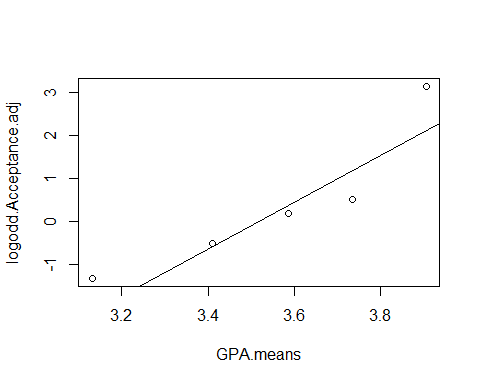
THe above gives you what we want

Accepted sums **Find the number of acceptances for each slice**

for(i in 1:groups){  
 Acceptance.sums[i] =sum(   
 Acceptance[((group.size\*i)-(group.size-1)):(group.size\*i)])  
 }  
  
Acceptance.sums

## [1] 2 4 6 7 11

# “Fudge” the proportions slightly and find the log of the predicted odds  
# Why “fudge”?  
# Proportions of 0 and 1 cause issues.  
  
Acceptance.prop.adj = (Acceptance.sums +0.5)/(group.size+1)  
  
logodd.Acceptance.adj = log(Acceptance.prop.adj/(1-Acceptance.prop.adj))  
  
# Plot the logodds of the adjusted proportions by the means of the predictor variables and a linear model  
# ASk self if the data ppears linear and if the group numbers matter (YES! THEY DO)  
plot(logodd.Acceptance.adj~GPA.means)  
abline(B0,B1)

 above does it in loop format, but you can do it better with teh ttr package

acceptance.sums = runSum(Acceptance, 11)  
acceptance.sums

## [1] NA NA NA NA NA NA NA NA NA NA 2 2 2 2 3 2 2 2 2 2 3 4 5 6 6  
## [26] 6 6 6 6 6 7 7 6 5 4 5 4 5 5 6 7 6 6 7 8 9 9 10 10 10  
## [51] 10 10 11 11 11

THis is saysing tha ttherea re not enought o give me an 11 people sum; the above gives you a running cumumlaitve sum for 11, then 12, then 13 then etc

so to get what we want do below:

acceptance.sums = runSum(Acceptance, 11)[seq(11,length(GPA), 11)]  
acceptance.sums

## [1] 2 4 6 7 11

we need to fudge so we tell R to nto give us exactly 0 and not eactly 1 we get log of 0 andlog of 1 aer issues;

we are just goint o add 0.5 to the sums and divivde by the group size sot aht we never get exactly 1 or a little less than that.

look at teh group size and numbers way above.

WE’re a bit lost because he deviated adn I’m not paying as much attention as I could

acceptance.propo.ad = (Acceptance.sums + 0.5)/(group.size + 1)  
acceptance.propo.ad

## [1] 0.2083333 0.3750000 0.5416667 0.6250000 0.9583333

logodd.accept.ad = log(acceptance.propo.ad/(1-acceptance.propo.ad))  
logodd.accept.ad

## [1] -1.3350011 -0.5108256 0.1670541 0.5108256 3.1354942

below, we want to make sure we load stat2data

we give it teh raw data, so that it does the thing taht we want itt to; we want to group it by 5, but i dont know why

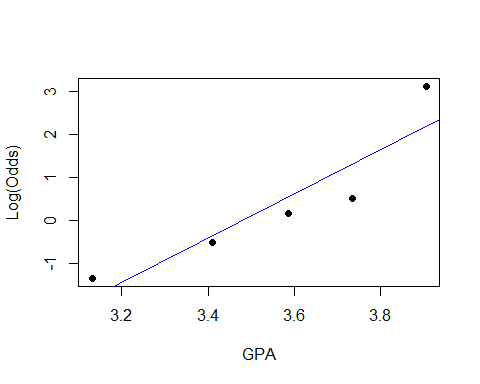
we choose 5 because that’s how we split teh data earlier in the code;

this will give you the same plot, but we the othere xtra work in teh past; its a fsater way to do the thing

there may be issues with this on teh data; if you try and slice the data ti migth overlap and cause errors; certain groupsing might work differently for different data; it defpending on teh datatset

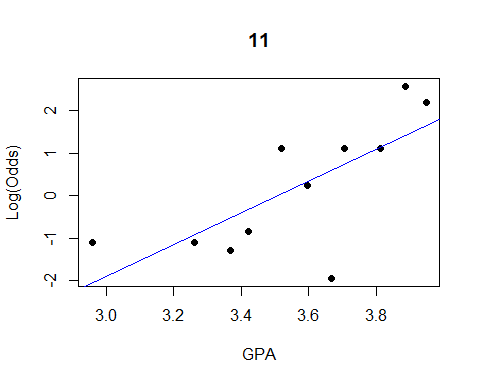
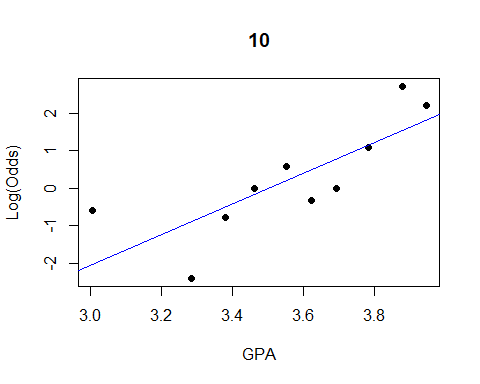
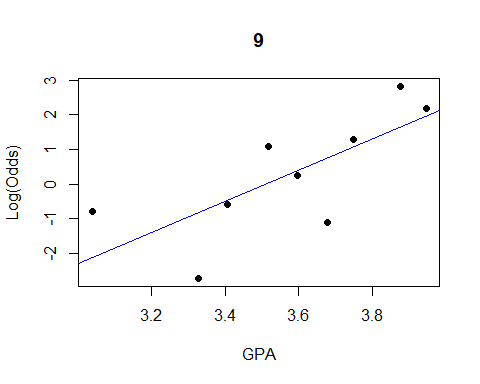
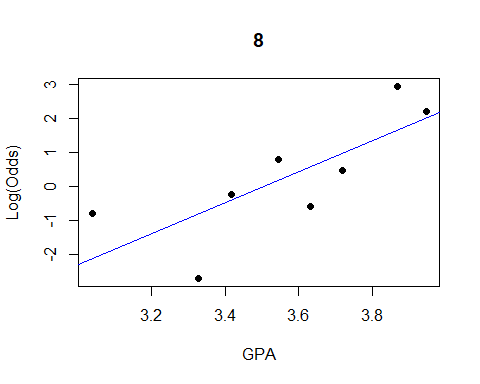
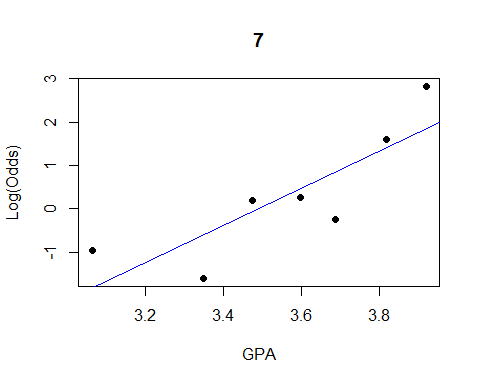
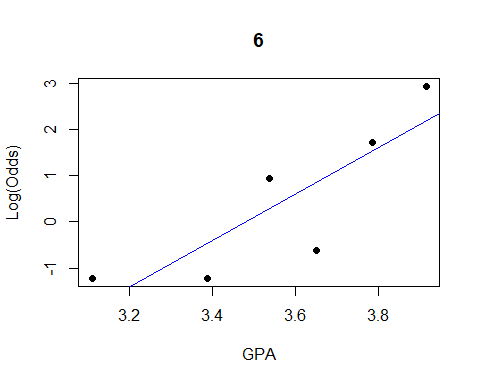
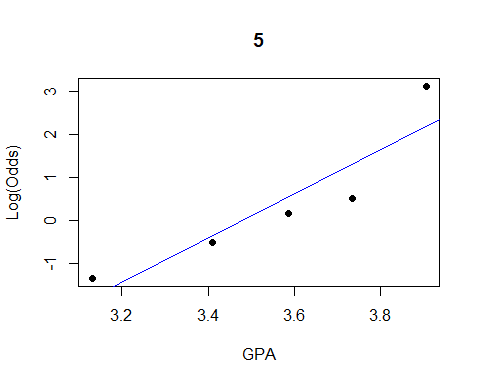
you just haev to ttiral and error it **Check linear conditions**

emplogitplot1(Acceptance~GPA, data=MedGPA, ngroups=5)

 this will check the groups of the other names; this is how you test other nimber grouping, which is useful

these arent residuals, but we think about them that way; if tehre are different patterns, we could try transformations; the logit plot works teh same as teh full things we did in the other classes **This checks different grouping types**

for(j in 5:11){emplogitplot1(Acceptance~GPA, data=MedGPA, ngroups=j, main=j)}



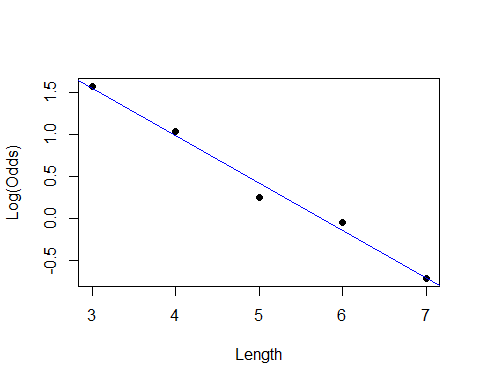
below shows you a shortcut on how to do the long thing int eh short; with teh logitplot function

ngroups = all shows that all possible predictor wvalue and will make teh own group

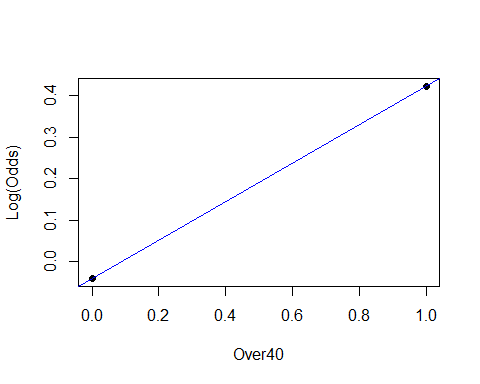
that doesnt work for medgpa data beause the grouping are different; there are vary few outcomes for each logodd for med gpa and it doesnt tell us much aout ht eoutcomes

the goldenballgraph is less exciting because it’s a graph with a thing between two lines

data("Putts1")  
emplogitplot1(Made~Length, data=Putts1, ngroups="all")



emplogitplot1(Split~Over40, data=GoldenBalls, ngroups="all")



## STOR 455 Class 29 R Multiple Logistic Regression

library(Stat2Data)  
library(leaps)  
  
source("https://raw.githubusercontent.com/JA-McLean/STOR455/master/scripts/ShowSubsets.R")  
  
logit = function(B0, B1, x)  
{  
 exp(B0+B1\*x)/(1+exp(B0+B1\*x))  
}

**Categorical Predictors with Multiple Categories in Logistic Regression** Example: Predicting survival in an intensive care unit (ICU) Response: Survive = 0 for dead and 1 for lived Predictor: AgeGroup = 1 for YOung, 2 for middle, 3 for old

data("ICU")  
head(ICU)

## ID Survive Age AgeGroup Sex Infection SysBP Pulse Emergency  
## 1 4 0 87 3 1 1 80 96 1  
## 2 8 1 27 1 1 1 142 88 1  
## 3 12 1 59 2 0 0 112 80 1  
## 4 14 1 77 3 0 0 100 70 0  
## 5 27 0 76 3 1 1 128 90 1  
## 6 28 1 54 2 0 1 142 103 1

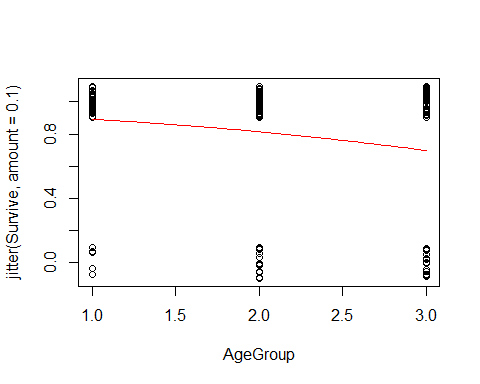
**Categorical Predictors with Multiple Categories in Logistic Regression** - Two approaches: 1. **Method #1:** Logistic regression for Survive with AgeGroup as a quantitative predictor. 2. **Method #2:** Use dummy (indicator) variables for the age categories as predictors in a logistic regression model for Survive.

**Method #1: AgeGroup as Quantitative Pred**

ICUmod = glm(Survive~AgeGroup, data=ICU, family=binomial)  
  
summary(ICUmod)

##   
## Call:  
## glm(formula = Survive ~ AgeGroup, family = binomial, data = ICU)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.1120 0.4769 0.6414 0.6414 0.8484   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.7566 0.5732 4.809 1.52e-06 \*\*\*  
## AgeGroup -0.6399 0.2414 -2.651 0.00802 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 200.16 on 199 degrees of freedom  
## Residual deviance: 192.66 on 198 degrees of freedom  
## AIC: 196.66  
##   
## Number of Fisher Scoring iterations: 4

B0 = summary(ICUmod)$coef[1]  
B1 = summary(ICUmod)$coef[2]  
  
plot(jitter(Survive,amount=0.1)~AgeGroup,data=ICU)  
curve(logit(B0, B1, x),add=TRUE, col="red")

 The above is a log mod that predicets survive y age group with fam = bi; if we don’t tell it fam = bi, then it will only give us a line and we wont get teh curve we want

if it’s non zero, tehn there is a change in teh log odds based on surviving based on teh age group when we plot this we can look at it and see the coeffs.

**Method #1: AgeGroup as Quantitative Pred**

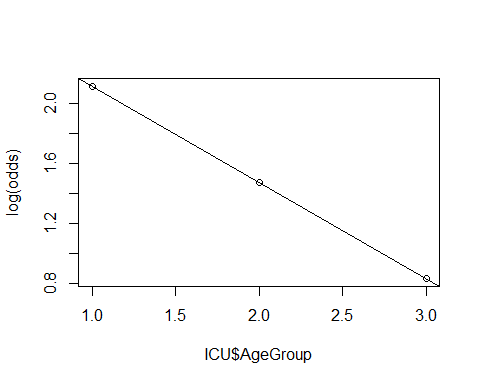
pi = logit(B0, B1, ICU$AgeGroup)  
head(pi)

## [1] 0.6977833 0.8925107 0.8140745 0.6977833 0.6977833 0.8140745

odds = pi/(1-pi)  
head(odds)

## [1] 2.308884 8.303252 4.378498 2.308884 2.308884 4.378498

plot(log(odds)~ICU$AgeGroup)  
abline(B0,B1)

 The above shows what we are predicting ; the odds are teh predicitng/ 1-odds predicted.

ploting the logodds with teh mdel on top of it

plotting teh log odds against teh other things here.

the predicts are right on this line; as we goes from young to middle to old we follow this ration

we miht not have this be true we might be forcing a relationship

its like when we were looking at active vs resting heartrate

its assuming a consistent rate of chaneg between age groups

if we lok at how the data actuallyuly looks with teh table; we can see that the actual counts are

we want to see the proportions are they different from teh predicted values and how much?

so we are going to make a table that are the proportions

so 54/59; etc etc. the prop.table will make this prop table for us

we want to lok at teh column proportion for those who surived adn that’s why we have a 2 in the code below

**Two-way Table: Survive by AgeGroup**

# Two way table of Counts  
ICU.table = table(ICU$Survive, ICU$AgeGroup)  
ICU.table

##   
## 1 2 3  
## 0 5 17 18  
## 1 54 60 46

# Two way table of Column Proportions  
ICU.prop.table = prop.table(ICU.table,2)  
ICU.prop.table

##   
## 1 2 3  
## 0 0.08474576 0.22077922 0.28125000  
## 1 0.91525424 0.77922078 0.71875000

# Two way table of Column logodds  
logodds.ICU.table = log(ICU.prop.table/(1-ICU.prop.table))  
logodds.ICU.table

##   
## 1 2 3  
## 0 -2.3795461 -1.2611312 -0.9382696  
## 1 2.3795461 1.2611312 0.9382696

above we can see in teh actual data ,the ic propo will tell us teh proportions; we have

if we plot all these together then we get a log odds table; that lets us plot it all together logodds proportion/ 1-proprotion

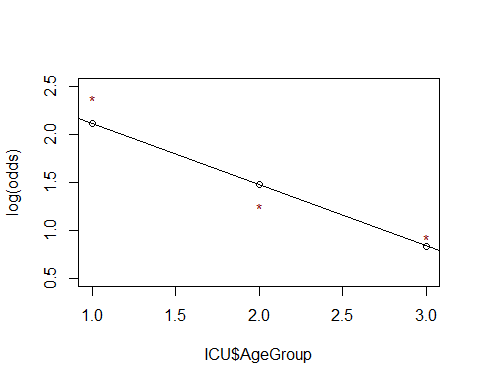
we want to be able to plot this, but it wont work well in a table format, so we need to make thi a dataframe.

we want teh columsn transposed; t = transponse **Two-way Table: Survive by AgeGroup**

logodds.ICU.df = t(as.data.frame.matrix(logodds.ICU.table))  
head(logodds.ICU.df)

## 0 1  
## 1 -2.3795461 2.3795461  
## 2 -1.2611312 1.2611312  
## 3 -0.9382696 0.9382696

plot(log(odds)~ICU$AgeGroup, ylim=c(.5, 2.5))  
abline(B0,B1)  
points(logodds.ICU.df[,2], col="dark red",pch="\*")



the above pulls out all the log odds rows and makes them red so they stand out this is so that you can do somethign else

what if we wanted two age groups; we could make 1 age group for young, and one for old; and if its’ not either then it has to be middle; but we have used this to be 1 = young and the other is middle, then old would be both = 0 **Method #2: Survive ~ Middle + Old**

ICUmod.2 = glm(Survive~factor(AgeGroup), data=ICU, family=binomial)  
summary(ICUmod.2)

##   
## Call:  
## glm(formula = Survive ~ factor(AgeGroup), family = binomial,   
## data = ICU)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.2218 0.4208 0.7063 0.7063 0.8127   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.3795 0.4675 5.090 3.57e-07 \*\*\*  
## factor(AgeGroup)2 -1.1184 0.5422 -2.063 0.03915 \*   
## factor(AgeGroup)3 -1.4413 0.5439 -2.650 0.00805 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 200.16 on 199 degrees of freedom  
## Residual deviance: 191.59 on 197 degrees of freedom  
## AIC: 197.59  
##   
## Number of Fisher Scoring iterations: 5

**Dummy Indicators for Multiple Categories** For a categorical predictor with k levels, we should use k − 1 dummy indicators. X = 1 if group 1, 0 if otherwise Xi-1 = 1 if in group k-1, 0 if otherwise

What happens to Group #k? That is teh reference group

Constant term is an estimate for Group #k and other coefficients are the differences from it.

* The coef for age 2 and 3 are the log odds for each in relation to the survive
* we dont want to lok at certain age groups we want ot know if age group as a whole is a good predicotr the ines dont give us that

**Binary Logistic Regression Model** Y = Binary response X1,X2,…,Xk = Multiple predictors π = proportion of 1’s at any x1, x2, …, xk Equivalent forms of the logistic regression model: Logit form: log⁡(𝜋/(1−𝜋))=𝛽\_0+𝛽\_1 𝑥\_1+𝛽\_2 𝑥\_2+⋯+𝛽\_𝑘 𝑥\_𝑘

Probability form: 𝜋=𝑒^(𝛽\_𝑜+𝛽\_1 𝑥\_1+𝛽\_2 𝑥\_2+⋯+𝛽\_𝑘 𝑥\_𝑘 )/(1+𝑒^(𝛽\_𝑜+𝛽\_1 𝑥\_1+𝛽\_2 𝑥\_2+⋯+𝛽\_𝑘 𝑥\_𝑘 ) )

y = binary response; X1, X2, Xk = mult predictor

pi = propotion of 1 at any xi

this is equal to the log reg mod

log form = log(pi/1-pi) = B0\_B1X1+B2X2 +…BkXk

prob form = pi = (e(B0+B1X1+…+BkXk)/1-e(same as num))

we can also use anova below to do the hypothesisi test; there aren’t teh samekind of residuals

the chisq thing will tell it;

recall nested f-test basic idea: Is teh improvement (reduction in SEE) Sig for teh number of extra preditores?

compare full model to reduced model = use t.s. = F - ratio (interpret similar to ANOVA)

**Interpreting Individual Tests** Similar issues to ordinary regression: - Is the predictor helpful, given the other predictors are already in the model? - Beware of problems due to multicollinearity. - Try to keep the model simple.

**G-Test for Overall Fit** H0:β1=β2=…=βk=0 vs. Ha: Some βi ≠ 0 t.s. = G = improvement in –2log(L) over a model with just a constant term Compare to 2 with k d.f.

Null deviance: 200.16 on 199 degrees of freedom

Residual deviance: 191.59 on 197 degrees of freedom 𝐺=200.16−191.59=8.57

1-pchisq(8.57,2) [1] 0.01377362 <- Reject H0

**Method #2: Survive ~ Middle + Old** Coefficients: Estimate Std. Error z value Pr(>|z|)  
(Intercept) **2.3795** 0.4675 5.090 3.57e-07 \*\*\* <- Log(oods) young factor(AgeGroup)2 **-1.1184** 0.5422 -2.063 0.03915 \*  
factor(AgeGroup)3 **-1.4413** 0.5439 -2.650 0.00805 \*\*

The factor age group bolded = the change in log(odds) for middle and old compared to young

anova(ICUmod.2, test="Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: Survive  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 199 200.16   
## factor(AgeGroup) 2 8.5721 197 191.59 0.01376 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Recall: Nested F-test** Purpose: Test a subset of predictors Ex: 𝑌=𝛽1𝑋1+𝛽2𝑋2+𝛽3𝑋3+𝛽4𝑋4+𝛽5𝑋5 + 𝜀  
𝐻0:𝛽3=𝛽4=𝛽5=0 vs. 𝐻𝑎: 𝑆𝑜𝑚𝑒 𝛽𝑖 ≠ 0 for i>2

Basic idea: Is the improvement (reduction in SSE) “significant” for the number of extra predictors? i.e. Compare “full” model to “reduced” model

t.s.= F-ratio (interpret similar to ANOVA)

**Nested LRT for Logistic Regression(Likelihood Ratio Test)** Purpose: Test a subset of predictors Ex: log⁡(𝑜𝑑𝑑𝑠)=𝛽1𝑋1+𝛽2𝑋2+𝛽3𝑋3+𝛽4𝑋4+𝛽5𝑋5  
𝐻0:𝛽3=𝛽4=𝛽5=0 vs. 𝐻𝑎: 𝑆𝑜𝑚𝑒 𝛽𝑖 ≠ 0 for i>2

Basic idea: Is the improvement, change in –2log⁡(𝐿), “significant” for the number of extra predictors? i.e. Compare “reduced” model to “full” model

𝜒^2=–2log⁡(𝐿𝑅𝑒𝑑𝑢𝑐𝑒𝑑) – (–2log⁡(𝐿𝐹𝑢𝑙𝑙))

Chi-square d.f.=#extra predictors tested

**Comparing Full to Reduced Models** ICUMod 3 = full and ICUMod2 = reduced

𝐻0:𝛽3=0 vs. 𝐻𝑎: 𝛽3 ≠ 0

ICUmod.3 = glm(Survive~factor(AgeGroup)+Emergency, data=ICU, family=binomial)  
summary(ICUmod.3)

##   
## Call:  
## glm(formula = Survive ~ factor(AgeGroup) + Emergency, family = binomial,   
## data = ICU)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.4388 0.2632 0.4469 0.8536 1.0137   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 4.7771 0.8801 5.428 5.7e-08 \*\*\*  
## factor(AgeGroup)2 -1.4317 0.5527 -2.590 0.009585 \*\*   
## factor(AgeGroup)3 -1.8557 0.5606 -3.310 0.000931 \*\*\*  
## Emergency -2.5234 0.7538 -3.347 0.000816 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 200.16 on 199 degrees of freedom  
## Residual deviance: 171.16 on 196 degrees of freedom  
## AIC: 179.16  
##   
## Number of Fisher Scoring iterations: 6

use anova for a drop in dev test;

this tells us

ICU mod 2 = reduced and 3 = full with emergency

we are going to see that just the two models you get the two residuals deviationces, it tells you df difference; its teh 1 bc its jstthe emerg var; the

doesnt give a p value ebcause we didnt give it a test

if we tell it the test is chisq, then we will get teh pvaleu

there are small values and they are different; tehre are different assumptions being made; it prob wont change the decision, but ti could be difference value thatn teh summaru **Drop in Deviance Test**

1 - pchisq(summary(ICUmod.2)$deviance - summary(ICUmod.3)$deviance, 1)

## [1] 6.187652e-06

#Reject H0 (p-value= 6.187652e-06). The Emergency term significantly improves the model.  
# This is also often called a “Drop-in-Deviance” test.

anova(ICUmod.2, ICUmod.3, test="Chisq")

## Analysis of Deviance Table  
##   
## Model 1: Survive ~ factor(AgeGroup)  
## Model 2: Survive ~ factor(AgeGroup) + Emergency  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 197 191.59   
## 2 196 171.16 1 20.429 6.188e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Example: Predicting Medical School Acceptance** Data: MedGPA  
Accept Status: A=accepted to medical school or D=denied admission Acceptance Indicator for Accept: 1=accepted or 0=denied Sex F=female or M=male BCPM Bio/Chem/Physics/Math grade point average GPA College grade point average VR Verbal reasoning (subscore) PS Physical sciences (subscore) WS Writing sample (subcore) BS Biological sciences (subscore) MCAT Score on the MCAT exam (sum of CR+PS+WS+BS) Apps Number of medical schools applied to

Goal: Find the “best” model for Acceptance using some or all of these predictors.

NOw, what if instead i did the anova of mod3; with a test = chisq; that is going to give su s a table tha tdeos teh test but compares with teh factor with teh null and tehn comp emergenc withw factor age grouo it everytime i add a thing then it des a nested test

useful only if you want to test things in order

if we want to test different order tehnw e have to do something difference. ’

data(MedGPA)  
head(MedGPA)

## Accept Acceptance Sex BCPM GPA VR PS WS BS MCAT Apps  
## 1 D 0 F 3.59 3.62 11 9 9 9 38 5  
## 2 A 1 M 3.75 3.84 12 13 8 12 45 3  
## 3 A 1 F 3.24 3.23 9 10 5 9 33 19  
## 4 A 1 F 3.74 3.69 12 11 7 10 40 5  
## 5 A 1 F 3.53 3.38 9 11 4 11 35 11  
## 6 A 1 M 3.59 3.72 10 9 7 10 36 5

**Criteria to Compare Models for Ordinary Multiple Regression** - Look for large R2 – But R2 is always best for the model with all predictors - Look for large adjusted R2 – Helps factor in the number of predictors in the model - Look at individual t-tests – Might be susceptible to multicollinearity problems

*-How to Choose Models to Compare for Ordinary Multiple Regression?* Method #1: All Subsets! Consider all possible combinations of predictors. How many are there? Pool of k predictors -> 2𝑘−1 subsets

Advantage: Find the best model for your criteria Disadvantage: LOTS of computation

* Note: requires leaps package

all = regsubsets(Acceptance~., data=MedGPA[,2:11])

## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =  
## force.in, : 1 linear dependencies found

## Reordering variables and trying again:

ShowSubsets(all)

## SexM BCPM GPA VR PS WS BS MCAT Apps Rsq adjRsq Cp  
## 1 ( 1 ) \* 30.57 29.24 11.08  
## 2 ( 1 ) \* \* 39.37 36.99 5.34  
## 3 ( 1 ) \* \* \* 43.75 40.37 3.49  
## 4 ( 1 ) \* \* \* \* 46.40 42.02 3.16  
## 5 ( 1 ) \* \* \* \* \* 48.87 43.55 2.98  
## 6 ( 1 ) \* \* \* \* \* \* 49.59 43.16 4.35  
## 7 ( 1 ) \* \* \* \* \* \* \* 49.99 42.38 6.00  
## 8 ( 1 ) \* \* \* \* \* \* \* \* 49.99 41.10 8.00

# This “works” in the sense that it runs, but creates a linear not a logistic model…

Will learn later how to automate the chosing the best model for other types of models

## STOR 455 Class 30 Multiple Logistic Regression (Again)

library(readr)  
library(bestglm)  
library(Stat2Data)  
  
insurance <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/insurance.csv")

* Looking at something like mallowCp and see which one is most likely based on teh precitors and a few other things -

1. Only the resposne and possible predictors variables should be withing the datagrame
2. te response variable must be the last column in teh dataframe.

We need to tell R. We need to think what we don’t want in the model ; if we have accept anad acceptance in teh model,then we are going to get some errors abecause they are the same things; there are issues because the logistical model wouldn’t work as well because it would be a straight vertial line.

WE could chosoe teh specific columns we want or choose teh ones we dont’ want with negative index ;

**Example: Predicting Medical School Acceptance** Data: MedGPA  
Accept Status: A=accepted to medical school or D=denied admission Acceptance Indicator for Accept: 1=accepted or 0=denied Sex F=female or M=male BCPM Bio/Chem/Physics/Math grade point average GPA College grade point average VR Verbal reasoning (subscore) PS Physical sciences (subscore) WS Writing sample (subcore) BS Biological sciences (subscore) MCAT Score on the MCAT exam (sum of CR+PS+WS+BS) Apps Number of medical schools applied to

Find the “best” model for Acceptance using some or all of these predictors.

data(MedGPA)  
head(MedGPA)

## Accept Acceptance Sex BCPM GPA VR PS WS BS MCAT Apps  
## 1 D 0 F 3.59 3.62 11 9 9 9 38 5  
## 2 A 1 M 3.75 3.84 12 13 8 12 45 3  
## 3 A 1 F 3.24 3.23 9 10 5 9 33 19  
## 4 A 1 F 3.74 3.69 12 11 7 10 40 5  
## 5 A 1 F 3.53 3.38 9 11 4 11 35 11  
## 6 A 1 M 3.59 3.72 10 9 7 10 36 5

below shows how to set accetance to null, that deltes teh accept vars.

best glm wants the response in the specific part of the dataframe, wants it in teh last section ; if your thing is named soemthign sepciifc, it sometimes acts differently, but mostly this is different.

THe second part of the code below reorders teh columns with teh response value last so that the glm is better. THere are other ways that you can do this, but this is for consistency.

**bestglm for Model Selection** Requirements to use bestglm() 1. Only the response and possible predictor variables should be within the dataframe

MedGPA.1 = within(MedGPA, {Accept = NULL}) #delete Accept variable  
head(MedGPA.1)

## Acceptance Sex BCPM GPA VR PS WS BS MCAT Apps  
## 1 0 F 3.59 3.62 11 9 9 9 38 5  
## 2 1 M 3.75 3.84 12 13 8 12 45 3  
## 3 1 F 3.24 3.23 9 10 5 9 33 19  
## 4 1 F 3.74 3.69 12 11 7 10 40 5  
## 5 1 F 3.53 3.38 9 11 4 11 35 11  
## 6 1 M 3.59 3.72 10 9 7 10 36 5

Above we could have just overwritten the thing; but this is easy to make the running the cell a lot of times and then it will be fine.

using the best glm fucntion; just like when makign teh lienar model, we need to tell it which family of functions to draw from; it’s going to look at a LSRL if we dno’t tell it otherwise

family = binomial tells you to make it logistics.

1. The response variable must be the last column in the dataframe.

MedGPA.2 = MedGPA.1[,c(2:10,1)] #reorder columns with response last  
#bestglm for Model Selection  
head(MedGPA.2)

## Sex BCPM GPA VR PS WS BS MCAT Apps Acceptance  
## 1 F 3.59 3.62 11 9 9 9 38 5 0  
## 2 M 3.75 3.84 12 13 8 12 45 3 1  
## 3 F 3.24 3.23 9 10 5 9 33 19 1  
## 4 F 3.74 3.69 12 11 7 10 40 5 1  
## 5 F 3.53 3.38 9 11 4 11 35 11 1  
## 6 M 3.59 3.72 10 9 7 10 36 5 1

Tell em about teh BIC and BICQ DO the same thing, but same it as an object

The best nmodels will tell you how many best models there arel

the top rowis the best model that you would like

the next four best models are the other best models

BIC = the baysian information criteria ; we are going to use it like mallowCp

calculated like: klog(n) - 2log(L(alpha)); n = sample size k = number of predictors alpha = set of all paramets L(alpha) = probability of obtraining the data which you have, supposing the modelbeing tested was given

*SMaller values indicate preferred models*

tells you we got teh data ttha we give given the model

there si going to be a best, but there migh tnot be a stat difference between teh things;

it’s saying on ce we take teh neg 2log, that its nmmore likely that it gen teh data thta we got the value is based on teh samp size and num predictors it could still bea g odoo number if we have different predictors

most are within 0-2 BIC, so there are not much difference between them .

there sin’t much difference between teh models its easier to get teh data, but it’s not very stats; E don’t really know that, but these look pretty similar

**bestglm for Model Selection** BIC = Bayesian Information Criteria

bestglm(MedGPA.2, family=binomial)

## Morgan-Tatar search since family is non-gaussian.

## BIC  
## BICq equivalent for q in (0.407407122288894, 0.830512766582046)  
## Best Model:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -39.4708940 12.2144951 -3.231480 0.001231510  
## SexM -2.8403423 1.1580871 -2.452616 0.014182182  
## GPA 5.3344003 2.4807386 2.150327 0.031529326  
## PS 1.0247592 0.4722984 2.169728 0.030027451  
## WS -0.7177605 0.3496614 -2.052730 0.040098780  
## BS 1.7914617 0.6434984 2.783941 0.005370279

**Bayesian Information Criteria** k log(n)- 2log(L(θ̂))

n : sample size k : number of predictors θ : set of all parameters. L(θ̂) :probability of obtaining the data which you have, supposing the model being tested was a given.

Selection criteria, similar to Mallow’s Cp Smaller values indicate preferred models

**Comparing Models by BIC** Change in BIC; Evidence against hiher BIC 0-2; Little 2-6; POsitive 6-10; Strong greater than 10; Very strong

MedGPA.2.bestglm = bestglm(MedGPA.2, family=binomial)

## Morgan-Tatar search since family is non-gaussian.

MedGPA.2.bestglm$BestModels

## Sex BCPM GPA VR PS WS BS MCAT Apps Criterion  
## 1 TRUE FALSE TRUE FALSE TRUE TRUE TRUE FALSE FALSE 51.35809  
## 2 TRUE FALSE TRUE FALSE TRUE FALSE TRUE FALSE FALSE 52.67338  
## 3 TRUE FALSE TRUE TRUE FALSE TRUE FALSE TRUE FALSE 52.81895  
## 4 TRUE FALSE TRUE FALSE FALSE FALSE TRUE FALSE FALSE 52.85687  
## 5 TRUE TRUE FALSE FALSE TRUE TRUE TRUE FALSE FALSE 53.46655

**Example: Predicting Survival** Data: ICU  
ID Patient ID code Survive 1=patient survived to discharge or 0=patient died Age Age (in years) AgeGroup 1= young (under 50), 2= middle (50-69), 3 = old (70+) Sex 1=female or 0=male Infection 1=infection suspected or 0=no infection SysBP Systolic blood pressure (in mm of Hg) Pulse Heart rate (beats per minute) Emergency 1=emergency admission or 0=elective admission

Find the “best” model for Survival using some or all of these predictors.

data("ICU")  
head(ICU)

## ID Survive Age AgeGroup Sex Infection SysBP Pulse Emergency  
## 1 4 0 87 3 1 1 80 96 1  
## 2 8 1 27 1 1 1 142 88 1  
## 3 12 1 59 2 0 0 112 80 1  
## 4 14 1 77 3 0 0 100 70 0  
## 5 27 0 76 3 1 1 128 90 1  
## 6 28 1 54 2 0 1 142 103 1

#Requirements to use bestglm()  
#1. Only the response and possible predictor variables should be within the dataframe  
ICU.1 <- within(ICU, {ID = NULL}) #delete ID variable  
  
#delete ID variable  
# WHy do we delete teh ID Variable? We probably don't need it because each row = the incident number  
  
#2. The response variable must be the last column in the dataframe.  
#reorder columns with response last; column 1 is now the survived column because the ID column was deleted.  
ICU.2 = ICU.1[,c(2:8,1)] #reorder columns with response last  
  
# AgeGroup is Treated as Quantitative   
head(ICU.2)

## Age AgeGroup Sex Infection SysBP Pulse Emergency Survive  
## 1 87 3 1 1 80 96 1 0  
## 2 27 1 1 1 142 88 1 1  
## 3 59 2 0 0 112 80 1 1  
## 4 77 3 0 0 100 70 0 1  
## 5 76 3 1 1 128 90 1 0  
## 6 54 2 0 1 142 103 1 1

bestglm(ICU.2, family=binomial)

## Morgan-Tatar search since family is non-gaussian.

## BIC  
## BICq equivalent for q in (0.0293273190867612, 0.516811637275042)  
## Best Model:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 5.3032038 0.9752351 5.437872 5.392065e-08  
## AgeGroup -0.8430258 0.2515652 -3.351123 8.048461e-04  
## Emergency -2.5144865 0.7576616 -3.318746 9.042257e-04

# THis tells me that teh best variable to predict survived is Emergency  
  
bestglm(ICU.2, family=binomial)$BestModels

## Morgan-Tatar search since family is non-gaussian.

## Age AgeGroup Sex Infection SysBP Pulse Emergency Criterion  
## 1 FALSE TRUE FALSE FALSE FALSE FALSE TRUE 183.3483  
## 2 FALSE TRUE FALSE FALSE TRUE FALSE TRUE 183.4829  
## 3 TRUE FALSE FALSE FALSE FALSE FALSE TRUE 183.6723  
## 4 TRUE FALSE FALSE FALSE TRUE FALSE TRUE 183.7191  
## 5 FALSE TRUE FALSE TRUE FALSE FALSE TRUE 186.7208

# The criteria doesn't change very much between teh first three models   
# Criteria is teh BIC; we want this to be low   
  
#THe data is teaching Age group as a numerical verabiel, we need to cahnge it to a cateorical variable if we want to look at each age group

*BElow is how to make agegroup a categorical variable* We are reassingin tee variable age group as teh factor of age group, so this breaks it into whatever age groups that are under agegroup category.

ICU\_factor\_AgeGroup = ICU.2   
ICU\_factor\_AgeGroup$AgeGroup = factor(ICU\_factor\_AgeGroup$AgeGroup)  
  
head(ICU\_factor\_AgeGroup)

## Age AgeGroup Sex Infection SysBP Pulse Emergency Survive  
## 1 87 3 1 1 80 96 1 0  
## 2 27 1 1 1 142 88 1 1  
## 3 59 2 0 0 112 80 1 1  
## 4 77 3 0 0 100 70 0 1  
## 5 76 3 1 1 128 90 1 0  
## 6 54 2 0 1 142 103 1 1

below is running the log model on the log model, but wiht age group sections differentiated.

bestglm(ICU\_factor\_AgeGroup, family=binomial)

## Morgan-Tatar search since family is non-gaussian.

## Note: factors present with more than 2 levels.

## BIC  
## Best Model:  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Age 1 1.149 1.1486 8.004 0.00515 \*\*   
## Emergency 1 2.581 2.5811 17.987 3.42e-05 \*\*\*  
## Residuals 197 28.270 0.1435   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

now it’s not using age group; but it’s giving more datapoints there is a change in the amount of predictirs we dont expect the below to be the same as the above ones, because we added more varaibles by levling teh age group

this goes with us a warning: “Factors rpesent with more than 2 level” it’s saying one thing is more than 2 levels; we its telling us taht there are more to teh columns that they give us

ICU\_factor\_AgeGroup\_bestglm = bestglm(ICU\_factor\_AgeGroup, family=binomial)

## Morgan-Tatar search since family is non-gaussian.

## Note: factors present with more than 2 levels.

ICU\_factor\_AgeGroup\_bestglm$BestModels

## Age AgeGroup Sex Infection SysBP Pulse Emergency Criterion  
## 1 TRUE FALSE FALSE FALSE FALSE FALSE TRUE 183.6723  
## 2 TRUE FALSE FALSE FALSE TRUE FALSE TRUE 183.7191  
## 3 FALSE TRUE FALSE FALSE FALSE FALSE TRUE 187.0545  
## 4 FALSE TRUE FALSE FALSE TRUE FALSE TRUE 187.3861  
## 5 TRUE FALSE FALSE TRUE FALSE FALSE TRUE 187.4172

Below is making the age groups, assigning numbers; so if tha agegroup was 2, then put a 1, if it was 3, then put a 1, then the last code removes the agegroup column because we don’t need age group anymore since we included teh dummy predictors in the first two lines of code below.

below is what bestglm is doing. This looked at the data tiself. IT didn’t look at atransformation if we ignore tranformation, then we have the ebst model here.

But should we ignore tranofmromatio?

NOt always.

#Requirements to use bestglm()  
# 3. Create dummy variables for non binary categorical variables.  
  
ICU.2$AgeGroup2 = ifelse(ICU.2$AgeGroup==2,1,0)  
ICU.2$AgeGroup3 = ifelse(ICU.2$AgeGroup==3,1,0)  
ICU.3 <- within(ICU.2, {AgeGroup = NULL}) #delete AgeGroup variable  
ICU.4 = ICU.3[,c(1:6,8,9,7)] #reorder columns with response last  
  
head(ICU.4)

## Age Sex Infection SysBP Pulse Emergency AgeGroup2 AgeGroup3 Survive  
## 1 87 1 1 80 96 1 0 1 0  
## 2 27 1 1 142 88 1 0 0 1  
## 3 59 0 0 112 80 1 1 0 1  
## 4 77 0 0 100 70 0 0 1 1  
## 5 76 1 1 128 90 1 0 1 0  
## 6 54 0 1 142 103 1 1 0 1

bestglm(ICU.4, family=binomial)

## Morgan-Tatar search since family is non-gaussian.

## BIC  
## BICq equivalent for q in (0.0343073257857045, 0.505855168373752)  
## Best Model:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 5.50876248 1.03351106 5.330144 9.813508e-08  
## Age -0.03401617 0.01069436 -3.180759 1.468899e-03  
## Emergency -2.45353515 0.75256981 -3.260209 1.113300e-03

# Comparing Models by BIC  
ICU.4.bestglm = bestglm(ICU.4, family=binomial)

## Morgan-Tatar search since family is non-gaussian.

ICU.4.bestglm$BestModels

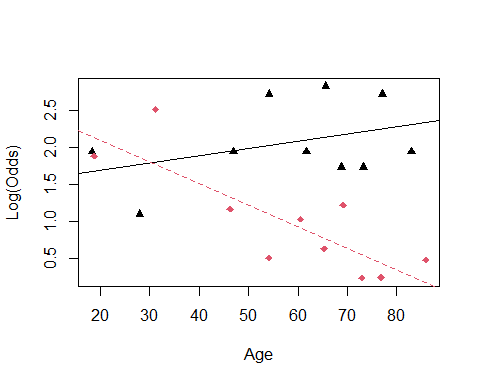
## Age Sex Infection SysBP Pulse Emergency AgeGroup2 AgeGroup3 Criterion  
## 1 TRUE FALSE FALSE FALSE FALSE TRUE FALSE FALSE 183.6723  
## 2 TRUE FALSE FALSE TRUE FALSE TRUE FALSE FALSE 183.7191  
## 3 FALSE FALSE FALSE FALSE FALSE TRUE TRUE TRUE 187.0545  
## 4 FALSE FALSE FALSE TRUE FALSE TRUE TRUE TRUE 187.3861  
## 5 TRUE FALSE TRUE FALSE FALSE TRUE FALSE FALSE 187.4172

We are assuming that age has teh same impact on surivial as old people; so age in general causes teh same surivial results.

WE can guess tho; if older peopple come in that is going to be different than if younger people are going in for an enermcy. WE can do that with an emperical logit plot.

THis logitplot will help us split by a factor for those brough tin with emergency and not emergency. if you run into errors with emplogitplot, then you can just factor the variables and sometimes that helps. Factor the last variable only, if that doesnt work, then factor others **bestglm for Model Selection**

emplogitplot2(Survive~Age+factor(Emergency), data=ICU.4, ngroups=10)

 We are assuming that age has teh same impact on surivial as old people; so age in general causes teh same surivial results.

WE can guess tho; if older peopple come in that is going to be different than if younger people are going in for an enermcy. WE can do that with an emperical logit plot.

THis logitplot will help us split by a factor for those brough tin with emergency and not emergency. if you run into errors with emplogitplot, then you can just factor the variables and sometimes that helps. Factor the last variable only, if that doesnt work, then factor others

ICU.4$EMAGE = ICU.4$Age\*ICU.4$Emergency  
head(ICU.4)

## Age Sex Infection SysBP Pulse Emergency AgeGroup2 AgeGroup3 Survive EMAGE  
## 1 87 1 1 80 96 1 0 1 0 87  
## 2 27 1 1 142 88 1 0 0 1 27  
## 3 59 0 0 112 80 1 1 0 1 59  
## 4 77 0 0 100 70 0 0 1 1 0  
## 5 76 1 1 128 90 1 0 1 0 76  
## 6 54 0 1 142 103 1 1 0 1 54

ICU.5 = ICU.4[,c(1:8,10,9)] # THis moves surive to teh end of the columns, so that wecan keep doing the code with bestgml.   
head(ICU.5)

## Age Sex Infection SysBP Pulse Emergency AgeGroup2 AgeGroup3 EMAGE Survive  
## 1 87 1 1 80 96 1 0 1 87 0  
## 2 27 1 1 142 88 1 0 0 27 1  
## 3 59 0 0 112 80 1 1 0 59 1  
## 4 77 0 0 100 70 0 0 1 0 1  
## 5 76 1 1 128 90 1 0 1 76 0  
## 6 54 0 1 142 103 1 1 0 54 1

x = bestglm(ICU.5, family=binomial)

## Morgan-Tatar search since family is non-gaussian.

x$BestModels

## Age Sex Infection SysBP Pulse Emergency AgeGroup2 AgeGroup3 EMAGE  
## 1 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE  
## 2 FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE TRUE  
## 3 FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE TRUE  
## 4 TRUE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE  
## 5 FALSE FALSE FALSE TRUE TRUE FALSE FALSE FALSE TRUE  
## Criterion  
## 1 179.1958  
## 2 179.2383  
## 3 183.0363  
## 4 183.6723  
## 5 183.6742

Someitmes best subsets isn’t as useful as we think so; for example: when you have categorical variables, soemtimes they are not immediately reflected through the best mdoels

*Use bestglm when you have binary categorical variables and when you have quantitative variables* IF you want to add interections adn transformations, then it will cause issues

THe ICU dataset was really nice, ti was really clean and easy to work with, but the below is less clean; insurance. How much you pay is based on a huge amount of htings; WHo elese do we haev data on that is like you and how much do we think that you and them are going toet in an accident and cost us moeny

index is just a number, there are 8k people; target flag = accident or no ltager amount= insurance costs - the first 6 rows, the first frow only 1 there are a lot of other types of things; red care = more insuance; previous thing; own a home, etc. so much we could deal with when making this.

THere are some probelms: 1. a lot of the these money variables, are character vectors and not numerical 2. some variabels are not binary, which is okay, but we also see thery’re saved as characters - characters and factors are different, and bestglm doesn’t like characters, they like factors.

WIll look back at this next class

*How to find the variables for the logistical regression models* - Bestglm - backgwards - formard - stepwise

head(insurance)

## # A tibble: 6 x 26  
## INDEX TARGET\_FLAG TARGET\_AMT KIDSDRIV AGE HOMEKIDS YOJ INCOME PARENT1  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr> <chr>   
## 1 1 0 0 0 60 0 11 $67,349 No   
## 2 2 0 0 0 43 0 11 $91,449 No   
## 3 4 0 0 0 35 1 10 $16,039 No   
## 4 5 0 0 0 51 0 14 <NA> No   
## 5 6 0 0 0 50 0 NA $114,986 No   
## 6 7 1 2946 0 34 1 12 $125,301 Yes   
## # ... with 17 more variables: HOME\_VAL <chr>, MSTATUS <chr>, SEX <chr>,  
## # EDUCATION <chr>, JOB <chr>, TRAVTIME <dbl>, CAR\_USE <chr>, BLUEBOOK <chr>,  
## # TIF <dbl>, CAR\_TYPE <chr>, RED\_CAR <chr>, OLDCLAIM <chr>, CLM\_FREQ <dbl>,  
## # REVOKED <chr>, MVR\_PTS <dbl>, CAR\_AGE <dbl>, URBANICITY <chr>

**Issues with Insurance Data for bestglm**

* Stepwise Regression (Linear Regression) Basic idea: Alternate forward selection and backward elimination

1. Use forward selection to choose a new predictor and check its significance.
2. Use backward elimination to see if predictors already in the model can be dropped.

**Is there a package in R to automate this process?** Yes! The stepAIC function in the MASS package can be used. - But we dont learn how to use it yet

Your task is to investigate the stepAIC function to determine how it can be used to determine the best logistic regression model using the insurance data

Currency\_Convert <- function(Field){  
 Field <- as.numeric(gsub("\\$|,","", Field))  
}  
  
#Change factors to numbers  
insurance$HOME\_VAL\_num = Currency\_Convert(insurance$HOME\_VAL)  
insurance$INCOME\_num = Currency\_Convert(insurance$INCOME)  
insurance$BLUEBOOK\_num = Currency\_Convert(insurance$BLUEBOOK)  
insurance$OLDCLAIM\_num = Currency\_Convert(insurance$OLDCLAIM)  
  
#remove unneeded variables  
insurance.1 = within(insurance,   
 {INDEX = NULL  
 TARGET\_AMT = NULL  
 HOME\_VAL = NULL  
 INCOME = NULL   
 BLUEBOOK = NULL  
 OLDCLAIM = NULL})  
  
  
head(insurance.1)

## # A tibble: 6 x 24  
## TARGET\_FLAG KIDSDRIV AGE HOMEKIDS YOJ PARENT1 MSTATUS SEX EDUCATION   
## <dbl> <dbl> <dbl> <dbl> <dbl> <chr> <chr> <chr> <chr>   
## 1 0 0 60 0 11 No z\_No M PhD   
## 2 0 0 43 0 11 No z\_No M z\_High School  
## 3 0 0 35 1 10 No Yes z\_F z\_High School  
## 4 0 0 51 0 14 No Yes M <High School   
## 5 0 0 50 0 NA No Yes z\_F PhD   
## 6 1 0 34 1 12 Yes z\_No z\_F Bachelors   
## # ... with 15 more variables: JOB <chr>, TRAVTIME <dbl>, CAR\_USE <chr>,  
## # TIF <dbl>, CAR\_TYPE <chr>, RED\_CAR <chr>, CLM\_FREQ <dbl>, REVOKED <chr>,  
## # MVR\_PTS <dbl>, CAR\_AGE <dbl>, URBANICITY <chr>, HOME\_VAL\_num <dbl>,  
## # INCOME\_num <dbl>, BLUEBOOK\_num <dbl>, OLDCLAIM\_num <dbl>

insurance.2 = insurance.1[,c(2:24,1)]   
head(insurance.2)

## # A tibble: 6 x 24  
## KIDSDRIV AGE HOMEKIDS YOJ PARENT1 MSTATUS SEX EDUCATION JOB TRAVTIME  
## <dbl> <dbl> <dbl> <dbl> <chr> <chr> <chr> <chr> <chr> <dbl>  
## 1 0 60 0 11 No z\_No M PhD Profe~ 14  
## 2 0 43 0 11 No z\_No M z\_High Sc~ z\_Blu~ 22  
## 3 0 35 1 10 No Yes z\_F z\_High Sc~ Cleri~ 5  
## 4 0 51 0 14 No Yes M <High Sch~ z\_Blu~ 32  
## 5 0 50 0 NA No Yes z\_F PhD Doctor 36  
## 6 0 34 1 12 Yes z\_No z\_F Bachelors z\_Blu~ 46  
## # ... with 14 more variables: CAR\_USE <chr>, TIF <dbl>, CAR\_TYPE <chr>,  
## # RED\_CAR <chr>, CLM\_FREQ <dbl>, REVOKED <chr>, MVR\_PTS <dbl>, CAR\_AGE <dbl>,  
## # URBANICITY <chr>, HOME\_VAL\_num <dbl>, INCOME\_num <dbl>, BLUEBOOK\_num <dbl>,  
## # OLDCLAIM\_num <dbl>, TARGET\_FLAG <dbl>

#Sad trombone, because best gml wont run here  
insurance.2 = as.data.frame(insurance.2)  
#bestglm(insurance.2, family=binomial)

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

insurance.2.1 = insurance.2 %>% mutate\_if(is.character, factor)  
head(insurance.2.1)

## KIDSDRIV AGE HOMEKIDS YOJ PARENT1 MSTATUS SEX EDUCATION JOB  
## 1 0 60 0 11 No z\_No M PhD Professional  
## 2 0 43 0 11 No z\_No M z\_High School z\_Blue Collar  
## 3 0 35 1 10 No Yes z\_F z\_High School Clerical  
## 4 0 51 0 14 No Yes M <High School z\_Blue Collar  
## 5 0 50 0 NA No Yes z\_F PhD Doctor  
## 6 0 34 1 12 Yes z\_No z\_F Bachelors z\_Blue Collar  
## TRAVTIME CAR\_USE TIF CAR\_TYPE RED\_CAR CLM\_FREQ REVOKED MVR\_PTS CAR\_AGE  
## 1 14 Private 11 Minivan yes 2 No 3 18  
## 2 22 Commercial 1 Minivan yes 0 No 0 1  
## 3 5 Private 4 z\_SUV no 2 No 3 10  
## 4 32 Private 7 Minivan yes 0 No 0 6  
## 5 36 Private 1 z\_SUV no 2 Yes 3 17  
## 6 46 Commercial 1 Sports Car no 0 No 0 7  
## URBANICITY HOME\_VAL\_num INCOME\_num BLUEBOOK\_num OLDCLAIM\_num  
## 1 Highly Urban/ Urban 0 67349 14230 4461  
## 2 Highly Urban/ Urban 257252 91449 14940 0  
## 3 Highly Urban/ Urban 124191 16039 4010 38690  
## 4 Highly Urban/ Urban 306251 NA 15440 0  
## 5 Highly Urban/ Urban 243925 114986 18000 19217  
## 6 Highly Urban/ Urban 0 125301 17430 0  
## TARGET\_FLAG  
## 1 0  
## 2 0  
## 3 0  
## 4 0  
## 5 0  
## 6 1

#Sadder trombone; because bestglm won't run   
insurance.2.1 = as.data.frame(insurance.2.1)  
#bestglm(insurance.2.1, family=binomial)

## STOR 455 Class 31 R Multiple Logistic Regression Again Agian

library(readr)  
library(bestglm)  
library(MASS)  
  
insurance <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/insurance.csv")

**Issues with Insurance Data for bestglm**

head(insurance)

## # A tibble: 6 x 26  
## INDEX TARGET\_FLAG TARGET\_AMT KIDSDRIV AGE HOMEKIDS YOJ INCOME PARENT1  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr> <chr>   
## 1 1 0 0 0 60 0 11 $67,349 No   
## 2 2 0 0 0 43 0 11 $91,449 No   
## 3 4 0 0 0 35 1 10 $16,039 No   
## 4 5 0 0 0 51 0 14 <NA> No   
## 5 6 0 0 0 50 0 NA $114,986 No   
## 6 7 1 2946 0 34 1 12 $125,301 Yes   
## # ... with 17 more variables: HOME\_VAL <chr>, MSTATUS <chr>, SEX <chr>,  
## # EDUCATION <chr>, JOB <chr>, TRAVTIME <dbl>, CAR\_USE <chr>, BLUEBOOK <chr>,  
## # TIF <dbl>, CAR\_TYPE <chr>, RED\_CAR <chr>, OLDCLAIM <chr>, CLM\_FREQ <dbl>,  
## # REVOKED <chr>, MVR\_PTS <dbl>, CAR\_AGE <dbl>, URBANICITY <chr>

Predict if people are going to get into a car accident How likely that these people get in a car accident based on teh data we are working with

# If it sees a dollar sign or a comma, it changes things to a number like as.numeric   
# So this makes us just have numbers   
# Want 4 new variables in teh dataframe   
Currency\_Convert <- function(Field){  
 Field <- as.numeric(gsub("\\$|,","", Field))  
}  
  
#Change factors to numbers  
insurance$HOME\_VAL\_num = Currency\_Convert(insurance$HOME\_VAL)  
insurance$INCOME\_num = Currency\_Convert(insurance$INCOME)  
insurance$BLUEBOOK\_num = Currency\_Convert(insurance$BLUEBOOK)  
insurance$OLDCLAIM\_num = Currency\_Convert(insurance$OLDCLAIM)  
  
#remove unneeded variables  
# Got rid of them because it would be not known at teh time we wanted to make prediction   
# But we now technically have this data   
# So we are jus getting rid of it   
insurance.1 = within(insurance,   
 {INDEX = NULL  
 TARGET\_AMT = NULL  
 HOME\_VAL = NULL  
 INCOME = NULL   
 BLUEBOOK = NULL  
 OLDCLAIM = NULL})  
  
  
head(insurance.1)

## # A tibble: 6 x 24  
## TARGET\_FLAG KIDSDRIV AGE HOMEKIDS YOJ PARENT1 MSTATUS SEX EDUCATION   
## <dbl> <dbl> <dbl> <dbl> <dbl> <chr> <chr> <chr> <chr>   
## 1 0 0 60 0 11 No z\_No M PhD   
## 2 0 0 43 0 11 No z\_No M z\_High School  
## 3 0 0 35 1 10 No Yes z\_F z\_High School  
## 4 0 0 51 0 14 No Yes M <High School   
## 5 0 0 50 0 NA No Yes z\_F PhD   
## 6 1 0 34 1 12 Yes z\_No z\_F Bachelors   
## # ... with 15 more variables: JOB <chr>, TRAVTIME <dbl>, CAR\_USE <chr>,  
## # TIF <dbl>, CAR\_TYPE <chr>, RED\_CAR <chr>, CLM\_FREQ <dbl>, REVOKED <chr>,  
## # MVR\_PTS <dbl>, CAR\_AGE <dbl>, URBANICITY <chr>, HOME\_VAL\_num <dbl>,  
## # INCOME\_num <dbl>, BLUEBOOK\_num <dbl>, OLDCLAIM\_num <dbl>

# Now we have numeric data where we need it

insurance.2 = insurance.1[,c(2:24,1)]   
# For best glm, we want the last column of the data as teh predicted, so we need that to be at t eh end of the dataframe   
# We want target flag at the end   
head(insurance.2)

## # A tibble: 6 x 24  
## KIDSDRIV AGE HOMEKIDS YOJ PARENT1 MSTATUS SEX EDUCATION JOB TRAVTIME  
## <dbl> <dbl> <dbl> <dbl> <chr> <chr> <chr> <chr> <chr> <dbl>  
## 1 0 60 0 11 No z\_No M PhD Profe~ 14  
## 2 0 43 0 11 No z\_No M z\_High Sc~ z\_Blu~ 22  
## 3 0 35 1 10 No Yes z\_F z\_High Sc~ Cleri~ 5  
## 4 0 51 0 14 No Yes M <High Sch~ z\_Blu~ 32  
## 5 0 50 0 NA No Yes z\_F PhD Doctor 36  
## 6 0 34 1 12 Yes z\_No z\_F Bachelors z\_Blu~ 46  
## # ... with 14 more variables: CAR\_USE <chr>, TIF <dbl>, CAR\_TYPE <chr>,  
## # RED\_CAR <chr>, CLM\_FREQ <dbl>, REVOKED <chr>, MVR\_PTS <dbl>, CAR\_AGE <dbl>,  
## # URBANICITY <chr>, HOME\_VAL\_num <dbl>, INCOME\_num <dbl>, BLUEBOOK\_num <dbl>,  
## # OLDCLAIM\_num <dbl>, TARGET\_FLAG <dbl>

#Sad trombone  
insurance.2 = as.data.frame(insurance.2)  
# bestglm(insurance.2, family=binomial)  
# This doesn't work because the structure of teh dataframe is maniuplated a bit weird; have to make it a dataframe; it still wont make it work all the time, but it might fix it   
# THis time it wont fix this issue

# So to fix that, we will use dplyr to make all the character vectors as factors   
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':  
##   
## select

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

insurance.2.1 = insurance.2 %>% mutate\_if(is.character, factor)  
head(insurance.2.1)

## KIDSDRIV AGE HOMEKIDS YOJ PARENT1 MSTATUS SEX EDUCATION JOB  
## 1 0 60 0 11 No z\_No M PhD Professional  
## 2 0 43 0 11 No z\_No M z\_High School z\_Blue Collar  
## 3 0 35 1 10 No Yes z\_F z\_High School Clerical  
## 4 0 51 0 14 No Yes M <High School z\_Blue Collar  
## 5 0 50 0 NA No Yes z\_F PhD Doctor  
## 6 0 34 1 12 Yes z\_No z\_F Bachelors z\_Blue Collar  
## TRAVTIME CAR\_USE TIF CAR\_TYPE RED\_CAR CLM\_FREQ REVOKED MVR\_PTS CAR\_AGE  
## 1 14 Private 11 Minivan yes 2 No 3 18  
## 2 22 Commercial 1 Minivan yes 0 No 0 1  
## 3 5 Private 4 z\_SUV no 2 No 3 10  
## 4 32 Private 7 Minivan yes 0 No 0 6  
## 5 36 Private 1 z\_SUV no 2 Yes 3 17  
## 6 46 Commercial 1 Sports Car no 0 No 0 7  
## URBANICITY HOME\_VAL\_num INCOME\_num BLUEBOOK\_num OLDCLAIM\_num  
## 1 Highly Urban/ Urban 0 67349 14230 4461  
## 2 Highly Urban/ Urban 257252 91449 14940 0  
## 3 Highly Urban/ Urban 124191 16039 4010 38690  
## 4 Highly Urban/ Urban 306251 NA 15440 0  
## 5 Highly Urban/ Urban 243925 114986 18000 19217  
## 6 Highly Urban/ Urban 0 125301 17430 0  
## TARGET\_FLAG  
## 1 0  
## 2 0  
## 3 0  
## 4 0  
## 5 0  
## 6 1

#Sadder trombone  
insurance.2.1 = as.data.frame(insurance.2.1)  
#bestglm(insurance.2.1, family=binomial)  
# When trying again, it doesn't appear to work still; so we have to try something different   
# NEw error, Error: p = 23, much be <= 15   
# tells us that there are too many variables, so we can't run it

When best glm doesn’t work, let’s use different stepwise. We will do AIC Stepwise

full\_insurance = glm(TARGET\_FLAG~., data=insurance.1, family="binomial")  
summary(full\_insurance)

##   
## Call:  
## glm(formula = TARGET\_FLAG ~ ., family = "binomial", data = insurance.1)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.5538 -0.7033 -0.3906 0.6213 3.1736   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.474e-01 3.122e-01 -0.472 0.636775   
## KIDSDRIV 3.194e-01 7.059e-02 4.524 6.06e-06 \*\*\*  
## AGE -3.903e-03 4.681e-03 -0.834 0.404452   
## HOMEKIDS 2.313e-02 4.275e-02 0.541 0.588473   
## YOJ -9.356e-03 9.775e-03 -0.957 0.338476   
## PARENT1Yes 4.109e-01 1.266e-01 3.246 0.001170 \*\*   
## MSTATUSz\_No 4.266e-01 1.006e-01 4.240 2.24e-05 \*\*\*  
## SEXz\_F -2.076e-01 1.287e-01 -1.613 0.106647   
## EDUCATIONBachelors -3.762e-01 1.325e-01 -2.840 0.004508 \*\*   
## EDUCATIONMasters -4.433e-01 2.140e-01 -2.072 0.038312 \*   
## EDUCATIONPhD 9.120e-02 2.635e-01 0.346 0.729208   
## EDUCATIONz\_High School -1.126e-03 1.068e-01 -0.011 0.991590   
## JOBDoctor -8.849e-01 3.279e-01 -2.699 0.006958 \*\*   
## JOBHome Maker -3.239e-01 1.669e-01 -1.941 0.052309 .   
## JOBLawyer -1.491e-01 2.120e-01 -0.703 0.481916   
## JOBManager -1.075e+00 1.650e-01 -6.519 7.07e-11 \*\*\*  
## JOBProfessional -2.843e-01 1.418e-01 -2.006 0.044906 \*   
## JOBStudent -3.615e-01 1.528e-01 -2.365 0.018010 \*   
## JOBz\_Blue Collar -1.931e-01 1.204e-01 -1.603 0.108890   
## TRAVTIME 1.569e-02 2.192e-03 7.159 8.15e-13 \*\*\*  
## CAR\_USEPrivate -8.292e-01 1.060e-01 -7.822 5.18e-15 \*\*\*  
## TIF -5.225e-02 8.544e-03 -6.115 9.64e-10 \*\*\*  
## CAR\_TYPEPanel Truck 6.955e-01 1.948e-01 3.570 0.000357 \*\*\*  
## CAR\_TYPEPickup 5.556e-01 1.154e-01 4.816 1.47e-06 \*\*\*  
## CAR\_TYPESports Car 1.107e+00 1.466e-01 7.553 4.27e-14 \*\*\*  
## CAR\_TYPEVan 5.678e-01 1.497e-01 3.793 0.000149 \*\*\*  
## CAR\_TYPEz\_SUV 8.265e-01 1.257e-01 6.575 4.88e-11 \*\*\*  
## RED\_CARyes -2.278e-01 1.032e-01 -2.208 0.027231 \*   
## CLM\_FREQ 2.004e-01 3.320e-02 6.036 1.58e-09 \*\*\*  
## REVOKEDYes 8.521e-01 1.075e-01 7.930 2.19e-15 \*\*\*  
## MVR\_PTS 1.161e-01 1.587e-02 7.312 2.63e-13 \*\*\*  
## CAR\_AGE -3.920e-03 8.897e-03 -0.441 0.659514   
## URBANICITYz\_Highly Rural/ Rural -2.306e+00 1.244e-01 -18.537 < 2e-16 \*\*\*  
## HOME\_VAL\_num -1.420e-06 4.287e-07 -3.312 0.000927 \*\*\*  
## INCOME\_num -3.429e-06 1.432e-06 -2.394 0.016680 \*   
## BLUEBOOK\_num -2.257e-05 6.102e-06 -3.699 0.000216 \*\*\*  
## OLDCLAIM\_num -1.310e-05 4.574e-06 -2.864 0.004186 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 6990.9 on 6044 degrees of freedom  
## Residual deviance: 5362.3 on 6008 degrees of freedom  
## (2116 observations deleted due to missingness)  
## AIC: 5436.3  
##   
## Number of Fisher Scoring iterations: 5

#More sad trombone  
#stepAIC(full\_insurance)  
# Says that some blank values are messing us up, so lets get rid of them

insurance.forstepAIC = na.omit(insurance.1) # Gets rid of the na values   
  
full\_insurance.2 = glm(TARGET\_FLAG~., data=insurance.forstepAIC, family="binomial")  
  
# Given a full model - Backwards selection is the default  
stepAIC(full\_insurance.2)

## Start: AIC=5436.27  
## TARGET\_FLAG ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + PARENT1 + MSTATUS +   
## SEX + EDUCATION + JOB + TRAVTIME + CAR\_USE + TIF + CAR\_TYPE +   
## RED\_CAR + CLM\_FREQ + REVOKED + MVR\_PTS + CAR\_AGE + URBANICITY +   
## HOME\_VAL\_num + INCOME\_num + BLUEBOOK\_num + OLDCLAIM\_num  
##   
## Df Deviance AIC  
## - CAR\_AGE 1 5362.5 5434.5  
## - HOMEKIDS 1 5362.6 5434.6  
## - AGE 1 5363.0 5435.0  
## - YOJ 1 5363.2 5435.2  
## <none> 5362.3 5436.3  
## - SEX 1 5364.9 5436.9  
## - RED\_CAR 1 5367.1 5439.1  
## - INCOME\_num 1 5368.1 5440.1  
## - OLDCLAIM\_num 1 5370.6 5442.6  
## - PARENT1 1 5372.8 5444.8  
## - HOME\_VAL\_num 1 5373.3 5445.3  
## - EDUCATION 4 5382.1 5448.1  
## - BLUEBOOK\_num 1 5376.2 5448.2  
## - MSTATUS 1 5380.1 5452.1  
## - KIDSDRIV 1 5382.6 5454.6  
## - CLM\_FREQ 1 5398.3 5470.3  
## - TIF 1 5401.0 5473.0  
## - JOB 7 5425.0 5485.0  
## - TRAVTIME 1 5413.7 5485.7  
## - MVR\_PTS 1 5416.3 5488.3  
## - REVOKED 1 5424.1 5496.1  
## - CAR\_USE 1 5424.5 5496.5  
## - CAR\_TYPE 5 5444.9 5508.9  
## - URBANICITY 1 5842.8 5914.8  
##   
## Step: AIC=5434.47  
## TARGET\_FLAG ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + PARENT1 + MSTATUS +   
## SEX + EDUCATION + JOB + TRAVTIME + CAR\_USE + TIF + CAR\_TYPE +   
## RED\_CAR + CLM\_FREQ + REVOKED + MVR\_PTS + URBANICITY + HOME\_VAL\_num +   
## INCOME\_num + BLUEBOOK\_num + OLDCLAIM\_num  
##   
## Df Deviance AIC  
## - HOMEKIDS 1 5362.8 5432.8  
## - AGE 1 5363.2 5433.2  
## - YOJ 1 5363.4 5433.4  
## <none> 5362.5 5434.5  
## - SEX 1 5365.1 5435.1  
## - RED\_CAR 1 5367.3 5437.3  
## - INCOME\_num 1 5368.4 5438.4  
## - OLDCLAIM\_num 1 5370.8 5440.8  
## - PARENT1 1 5373.0 5443.0  
## - HOME\_VAL\_num 1 5373.3 5443.3  
## - BLUEBOOK\_num 1 5376.3 5446.3  
## - MSTATUS 1 5380.3 5450.3  
## - EDUCATION 4 5387.3 5451.3  
## - KIDSDRIV 1 5382.8 5452.8  
## - CLM\_FREQ 1 5398.4 5468.4  
## - TIF 1 5401.3 5471.3  
## - JOB 7 5425.2 5483.2  
## - TRAVTIME 1 5413.8 5483.8  
## - MVR\_PTS 1 5416.5 5486.5  
## - REVOKED 1 5424.2 5494.2  
## - CAR\_USE 1 5424.6 5494.6  
## - CAR\_TYPE 5 5445.2 5507.2  
## - URBANICITY 1 5843.0 5913.0  
##   
## Step: AIC=5432.76  
## TARGET\_FLAG ~ KIDSDRIV + AGE + YOJ + PARENT1 + MSTATUS + SEX +   
## EDUCATION + JOB + TRAVTIME + CAR\_USE + TIF + CAR\_TYPE + RED\_CAR +   
## CLM\_FREQ + REVOKED + MVR\_PTS + URBANICITY + HOME\_VAL\_num +   
## INCOME\_num + BLUEBOOK\_num + OLDCLAIM\_num  
##   
## Df Deviance AIC  
## - YOJ 1 5363.5 5431.5  
## - AGE 1 5364.0 5432.0  
## <none> 5362.8 5432.8  
## - SEX 1 5365.4 5433.4  
## - RED\_CAR 1 5367.6 5435.6  
## - INCOME\_num 1 5368.6 5436.6  
## - OLDCLAIM\_num 1 5371.1 5439.1  
## - HOME\_VAL\_num 1 5373.8 5441.8  
## - BLUEBOOK\_num 1 5376.6 5444.6  
## - PARENT1 1 5377.6 5445.6  
## - MSTATUS 1 5380.7 5448.7  
## - EDUCATION 4 5387.8 5449.8  
## - KIDSDRIV 1 5390.0 5458.0  
## - CLM\_FREQ 1 5398.8 5466.8  
## - TIF 1 5401.6 5469.6  
## - JOB 7 5425.4 5481.4  
## - TRAVTIME 1 5414.0 5482.0  
## - MVR\_PTS 1 5416.9 5484.9  
## - REVOKED 1 5424.8 5492.8  
## - CAR\_USE 1 5425.1 5493.1  
## - CAR\_TYPE 5 5445.7 5505.7  
## - URBANICITY 1 5843.1 5911.1  
##   
## Step: AIC=5431.49  
## TARGET\_FLAG ~ KIDSDRIV + AGE + PARENT1 + MSTATUS + SEX + EDUCATION +   
## JOB + TRAVTIME + CAR\_USE + TIF + CAR\_TYPE + RED\_CAR + CLM\_FREQ +   
## REVOKED + MVR\_PTS + URBANICITY + HOME\_VAL\_num + INCOME\_num +   
## BLUEBOOK\_num + OLDCLAIM\_num  
##   
## Df Deviance AIC  
## - AGE 1 5365.0 5431.0  
## <none> 5363.5 5431.5  
## - SEX 1 5366.1 5432.1  
## - RED\_CAR 1 5368.3 5434.3  
## - INCOME\_num 1 5369.8 5435.8  
## - OLDCLAIM\_num 1 5372.0 5438.0  
## - HOME\_VAL\_num 1 5374.5 5440.5  
## - BLUEBOOK\_num 1 5377.4 5443.4  
## - PARENT1 1 5377.9 5443.9  
## - EDUCATION 4 5388.3 5448.3  
## - MSTATUS 1 5382.8 5448.8  
## - KIDSDRIV 1 5390.4 5456.4  
## - CLM\_FREQ 1 5399.6 5465.6  
## - TIF 1 5402.6 5468.6  
## - JOB 7 5425.5 5479.5  
## - TRAVTIME 1 5414.6 5480.6  
## - MVR\_PTS 1 5417.9 5483.9  
## - REVOKED 1 5425.6 5491.6  
## - CAR\_USE 1 5426.2 5492.2  
## - CAR\_TYPE 5 5446.8 5504.8  
## - URBANICITY 1 5843.5 5909.5  
##   
## Step: AIC=5431  
## TARGET\_FLAG ~ KIDSDRIV + PARENT1 + MSTATUS + SEX + EDUCATION +   
## JOB + TRAVTIME + CAR\_USE + TIF + CAR\_TYPE + RED\_CAR + CLM\_FREQ +   
## REVOKED + MVR\_PTS + URBANICITY + HOME\_VAL\_num + INCOME\_num +   
## BLUEBOOK\_num + OLDCLAIM\_num  
##   
## Df Deviance AIC  
## <none> 5365.0 5431.0  
## - SEX 1 5367.1 5431.1  
## - RED\_CAR 1 5369.7 5433.7  
## - INCOME\_num 1 5371.0 5435.0  
## - OLDCLAIM\_num 1 5373.6 5437.6  
## - HOME\_VAL\_num 1 5376.9 5440.9  
## - BLUEBOOK\_num 1 5380.7 5444.7  
## - MSTATUS 1 5383.5 5447.5  
## - PARENT1 1 5383.9 5447.9  
## - EDUCATION 4 5390.0 5448.0  
## - KIDSDRIV 1 5391.4 5455.4  
## - CLM\_FREQ 1 5401.0 5465.0  
## - TIF 1 5404.0 5468.0  
## - TRAVTIME 1 5415.8 5479.8  
## - JOB 7 5428.4 5480.4  
## - MVR\_PTS 1 5420.1 5484.1  
## - REVOKED 1 5427.4 5491.4  
## - CAR\_USE 1 5427.8 5491.8  
## - CAR\_TYPE 5 5447.1 5503.1  
## - URBANICITY 1 5846.8 5910.8

##   
## Call: glm(formula = TARGET\_FLAG ~ KIDSDRIV + PARENT1 + MSTATUS + SEX +   
## EDUCATION + JOB + TRAVTIME + CAR\_USE + TIF + CAR\_TYPE + RED\_CAR +   
## CLM\_FREQ + REVOKED + MVR\_PTS + URBANICITY + HOME\_VAL\_num +   
## INCOME\_num + BLUEBOOK\_num + OLDCLAIM\_num, family = "binomial",   
## data = insurance.forstepAIC)  
##   
## Coefficients:  
## (Intercept) KIDSDRIV   
## -3.905e-01 3.307e-01   
## PARENT1Yes MSTATUSz\_No   
## 4.730e-01 4.146e-01   
## SEXz\_F EDUCATIONBachelors   
## -1.861e-01 -3.995e-01   
## EDUCATIONMasters EDUCATIONPhD   
## -4.969e-01 4.064e-02   
## EDUCATIONz\_High School JOBDoctor   
## -7.396e-03 -9.037e-01   
## JOBHome Maker JOBLawyer   
## -2.872e-01 -1.650e-01   
## JOBManager JOBProfessional   
## -1.091e+00 -2.958e-01   
## JOBStudent JOBz\_Blue Collar   
## -3.133e-01 -2.020e-01   
## TRAVTIME CAR\_USEPrivate   
## 1.557e-02 -8.324e-01   
## TIF CAR\_TYPEPanel Truck   
## -5.237e-02 7.121e-01   
## CAR\_TYPEPickup CAR\_TYPESports Car   
## 5.525e-01 1.090e+00   
## CAR\_TYPEVan CAR\_TYPEz\_SUV   
## 5.778e-01 8.112e-01   
## RED\_CARyes CLM\_FREQ   
## -2.236e-01 2.001e-01   
## REVOKEDYes MVR\_PTS   
## 8.551e-01 1.170e-01   
## URBANICITYz\_Highly Rural/ Rural HOME\_VAL\_num   
## -2.308e+00 -1.469e-06   
## INCOME\_num BLUEBOOK\_num   
## -3.471e-06 -2.370e-05   
## OLDCLAIM\_num   
## -1.326e-05   
##   
## Degrees of Freedom: 6044 Total (i.e. Null); 6012 Residual  
## Null Deviance: 6991   
## Residual Deviance: 5365 AIC: 5431

# trace=FALSE will show only the final model, not each step.  
# Looks at new insurance data   
# Big model at teh end   
stepAIC(full\_insurance.2, trace=FALSE)

##   
## Call: glm(formula = TARGET\_FLAG ~ KIDSDRIV + PARENT1 + MSTATUS + SEX +   
## EDUCATION + JOB + TRAVTIME + CAR\_USE + TIF + CAR\_TYPE + RED\_CAR +   
## CLM\_FREQ + REVOKED + MVR\_PTS + URBANICITY + HOME\_VAL\_num +   
## INCOME\_num + BLUEBOOK\_num + OLDCLAIM\_num, family = "binomial",   
## data = insurance.forstepAIC)  
##   
## Coefficients:  
## (Intercept) KIDSDRIV   
## -3.905e-01 3.307e-01   
## PARENT1Yes MSTATUSz\_No   
## 4.730e-01 4.146e-01   
## SEXz\_F EDUCATIONBachelors   
## -1.861e-01 -3.995e-01   
## EDUCATIONMasters EDUCATIONPhD   
## -4.969e-01 4.064e-02   
## EDUCATIONz\_High School JOBDoctor   
## -7.396e-03 -9.037e-01   
## JOBHome Maker JOBLawyer   
## -2.872e-01 -1.650e-01   
## JOBManager JOBProfessional   
## -1.091e+00 -2.958e-01   
## JOBStudent JOBz\_Blue Collar   
## -3.133e-01 -2.020e-01   
## TRAVTIME CAR\_USEPrivate   
## 1.557e-02 -8.324e-01   
## TIF CAR\_TYPEPanel Truck   
## -5.237e-02 7.121e-01   
## CAR\_TYPEPickup CAR\_TYPESports Car   
## 5.525e-01 1.090e+00   
## CAR\_TYPEVan CAR\_TYPEz\_SUV   
## 5.778e-01 8.112e-01   
## RED\_CARyes CLM\_FREQ   
## -2.236e-01 2.001e-01   
## REVOKEDYes MVR\_PTS   
## 8.551e-01 1.170e-01   
## URBANICITYz\_Highly Rural/ Rural HOME\_VAL\_num   
## -2.308e+00 -1.469e-06   
## INCOME\_num BLUEBOOK\_num   
## -3.471e-06 -2.370e-05   
## OLDCLAIM\_num   
## -1.326e-05   
##   
## Degrees of Freedom: 6044 Total (i.e. Null); 6012 Residual  
## Null Deviance: 6991   
## Residual Deviance: 5365 AIC: 5431

final\_model\_backwards=stepAIC(full\_insurance.2, trace=FALSE)   
  
summary(final\_model\_backwards)

##   
## Call:  
## glm(formula = TARGET\_FLAG ~ KIDSDRIV + PARENT1 + MSTATUS + SEX +   
## EDUCATION + JOB + TRAVTIME + CAR\_USE + TIF + CAR\_TYPE + RED\_CAR +   
## CLM\_FREQ + REVOKED + MVR\_PTS + URBANICITY + HOME\_VAL\_num +   
## INCOME\_num + BLUEBOOK\_num + OLDCLAIM\_num, family = "binomial",   
## data = insurance.forstepAIC)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.5691 -0.7024 -0.3901 0.6201 3.1495   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.905e-01 2.287e-01 -1.707 0.087770 .   
## KIDSDRIV 3.307e-01 6.390e-02 5.176 2.27e-07 \*\*\*  
## PARENT1Yes 4.730e-01 1.089e-01 4.342 1.41e-05 \*\*\*  
## MSTATUSz\_No 4.146e-01 9.568e-02 4.334 1.47e-05 \*\*\*  
## SEXz\_F -1.861e-01 1.272e-01 -1.464 0.143274   
## EDUCATIONBachelors -3.995e-01 1.231e-01 -3.245 0.001174 \*\*   
## EDUCATIONMasters -4.969e-01 1.905e-01 -2.609 0.009094 \*\*   
## EDUCATIONPhD 4.064e-02 2.461e-01 0.165 0.868809   
## EDUCATIONz\_High School -7.396e-03 1.063e-01 -0.070 0.944554   
## JOBDoctor -9.037e-01 3.273e-01 -2.761 0.005755 \*\*   
## JOBHome Maker -2.872e-01 1.569e-01 -1.831 0.067098 .   
## JOBLawyer -1.650e-01 2.112e-01 -0.781 0.434854   
## JOBManager -1.091e+00 1.641e-01 -6.650 2.93e-11 \*\*\*  
## JOBProfessional -2.958e-01 1.410e-01 -2.098 0.035938 \*   
## JOBStudent -3.133e-01 1.446e-01 -2.166 0.030305 \*   
## JOBz\_Blue Collar -2.020e-01 1.201e-01 -1.682 0.092562 .   
## TRAVTIME 1.557e-02 2.189e-03 7.114 1.13e-12 \*\*\*  
## CAR\_USEPrivate -8.324e-01 1.059e-01 -7.861 3.82e-15 \*\*\*  
## TIF -5.237e-02 8.537e-03 -6.134 8.56e-10 \*\*\*  
## CAR\_TYPEPanel Truck 7.121e-01 1.944e-01 3.664 0.000249 \*\*\*  
## CAR\_TYPEPickup 5.525e-01 1.153e-01 4.793 1.65e-06 \*\*\*  
## CAR\_TYPESports Car 1.090e+00 1.451e-01 7.512 5.84e-14 \*\*\*  
## CAR\_TYPEVan 5.778e-01 1.494e-01 3.866 0.000110 \*\*\*  
## CAR\_TYPEz\_SUV 8.112e-01 1.247e-01 6.508 7.63e-11 \*\*\*  
## RED\_CARyes -2.236e-01 1.031e-01 -2.168 0.030141 \*   
## CLM\_FREQ 2.001e-01 3.318e-02 6.032 1.62e-09 \*\*\*  
## REVOKEDYes 8.551e-01 1.074e-01 7.965 1.65e-15 \*\*\*  
## MVR\_PTS 1.170e-01 1.585e-02 7.383 1.54e-13 \*\*\*  
## URBANICITYz\_Highly Rural/ Rural -2.308e+00 1.244e-01 -18.552 < 2e-16 \*\*\*  
## HOME\_VAL\_num -1.469e-06 4.267e-07 -3.443 0.000575 \*\*\*  
## INCOME\_num -3.471e-06 1.424e-06 -2.439 0.014744 \*   
## BLUEBOOK\_num -2.370e-05 6.031e-06 -3.930 8.51e-05 \*\*\*  
## OLDCLAIM\_num -1.326e-05 4.569e-06 -2.903 0.003700 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 6990.9 on 6044 degrees of freedom  
## Residual deviance: 5365.0 on 6012 degrees of freedom  
## AIC: 5431  
##   
## Number of Fisher Scoring iterations: 5

none = glm(TARGET\_FLAG~1, data=insurance.forstepAIC, family="binomial")  
  
# Tell it to do forward and stepwise below for an AIC model   
final\_model\_forwards = stepAIC(none, scope=list(upper=full\_insurance.2), direction="forward", trace=FALSE)  
final\_model\_both = stepAIC(none, scope=list(upper=full\_insurance.2), direction = "both", trace=FALSE)

summary(final\_model\_forwards)

##   
## Call:  
## glm(formula = TARGET\_FLAG ~ URBANICITY + JOB + MVR\_PTS + CAR\_TYPE +   
## MSTATUS + REVOKED + CAR\_USE + TRAVTIME + TIF + KIDSDRIV +   
## INCOME\_num + CLM\_FREQ + EDUCATION + BLUEBOOK\_num + PARENT1 +   
## HOME\_VAL\_num + OLDCLAIM\_num + RED\_CAR + SEX, family = "binomial",   
## data = insurance.forstepAIC)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.5691 -0.7024 -0.3901 0.6201 3.1495   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.905e-01 2.287e-01 -1.707 0.087770 .   
## URBANICITYz\_Highly Rural/ Rural -2.308e+00 1.244e-01 -18.552 < 2e-16 \*\*\*  
## JOBDoctor -9.037e-01 3.273e-01 -2.761 0.005755 \*\*   
## JOBHome Maker -2.872e-01 1.569e-01 -1.831 0.067098 .   
## JOBLawyer -1.650e-01 2.112e-01 -0.781 0.434854   
## JOBManager -1.091e+00 1.641e-01 -6.650 2.93e-11 \*\*\*  
## JOBProfessional -2.958e-01 1.410e-01 -2.098 0.035938 \*   
## JOBStudent -3.133e-01 1.446e-01 -2.166 0.030305 \*   
## JOBz\_Blue Collar -2.020e-01 1.201e-01 -1.682 0.092562 .   
## MVR\_PTS 1.170e-01 1.585e-02 7.383 1.54e-13 \*\*\*  
## CAR\_TYPEPanel Truck 7.121e-01 1.944e-01 3.664 0.000249 \*\*\*  
## CAR\_TYPEPickup 5.525e-01 1.153e-01 4.793 1.65e-06 \*\*\*  
## CAR\_TYPESports Car 1.090e+00 1.451e-01 7.512 5.84e-14 \*\*\*  
## CAR\_TYPEVan 5.778e-01 1.494e-01 3.866 0.000110 \*\*\*  
## CAR\_TYPEz\_SUV 8.112e-01 1.247e-01 6.508 7.63e-11 \*\*\*  
## MSTATUSz\_No 4.146e-01 9.568e-02 4.334 1.47e-05 \*\*\*  
## REVOKEDYes 8.551e-01 1.074e-01 7.965 1.65e-15 \*\*\*  
## CAR\_USEPrivate -8.324e-01 1.059e-01 -7.861 3.82e-15 \*\*\*  
## TRAVTIME 1.557e-02 2.189e-03 7.114 1.13e-12 \*\*\*  
## TIF -5.237e-02 8.537e-03 -6.134 8.56e-10 \*\*\*  
## KIDSDRIV 3.307e-01 6.390e-02 5.176 2.27e-07 \*\*\*  
## INCOME\_num -3.471e-06 1.424e-06 -2.439 0.014744 \*   
## CLM\_FREQ 2.001e-01 3.318e-02 6.032 1.62e-09 \*\*\*  
## EDUCATIONBachelors -3.995e-01 1.231e-01 -3.245 0.001174 \*\*   
## EDUCATIONMasters -4.969e-01 1.905e-01 -2.609 0.009094 \*\*   
## EDUCATIONPhD 4.064e-02 2.461e-01 0.165 0.868809   
## EDUCATIONz\_High School -7.396e-03 1.063e-01 -0.070 0.944554   
## BLUEBOOK\_num -2.370e-05 6.031e-06 -3.930 8.51e-05 \*\*\*  
## PARENT1Yes 4.730e-01 1.089e-01 4.342 1.41e-05 \*\*\*  
## HOME\_VAL\_num -1.469e-06 4.267e-07 -3.443 0.000575 \*\*\*  
## OLDCLAIM\_num -1.326e-05 4.569e-06 -2.903 0.003700 \*\*   
## RED\_CARyes -2.236e-01 1.031e-01 -2.168 0.030141 \*   
## SEXz\_F -1.861e-01 1.272e-01 -1.464 0.143274   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 6990.9 on 6044 degrees of freedom  
## Residual deviance: 5365.0 on 6012 degrees of freedom  
## AIC: 5431  
##   
## Number of Fisher Scoring iterations: 5

summary(final\_model\_both)

##   
## Call:  
## glm(formula = TARGET\_FLAG ~ URBANICITY + JOB + MVR\_PTS + CAR\_TYPE +   
## MSTATUS + REVOKED + CAR\_USE + TRAVTIME + TIF + KIDSDRIV +   
## INCOME\_num + CLM\_FREQ + EDUCATION + BLUEBOOK\_num + PARENT1 +   
## HOME\_VAL\_num + OLDCLAIM\_num + RED\_CAR + SEX, family = "binomial",   
## data = insurance.forstepAIC)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.5691 -0.7024 -0.3901 0.6201 3.1495   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.905e-01 2.287e-01 -1.707 0.087770 .   
## URBANICITYz\_Highly Rural/ Rural -2.308e+00 1.244e-01 -18.552 < 2e-16 \*\*\*  
## JOBDoctor -9.037e-01 3.273e-01 -2.761 0.005755 \*\*   
## JOBHome Maker -2.872e-01 1.569e-01 -1.831 0.067098 .   
## JOBLawyer -1.650e-01 2.112e-01 -0.781 0.434854   
## JOBManager -1.091e+00 1.641e-01 -6.650 2.93e-11 \*\*\*  
## JOBProfessional -2.958e-01 1.410e-01 -2.098 0.035938 \*   
## JOBStudent -3.133e-01 1.446e-01 -2.166 0.030305 \*   
## JOBz\_Blue Collar -2.020e-01 1.201e-01 -1.682 0.092562 .   
## MVR\_PTS 1.170e-01 1.585e-02 7.383 1.54e-13 \*\*\*  
## CAR\_TYPEPanel Truck 7.121e-01 1.944e-01 3.664 0.000249 \*\*\*  
## CAR\_TYPEPickup 5.525e-01 1.153e-01 4.793 1.65e-06 \*\*\*  
## CAR\_TYPESports Car 1.090e+00 1.451e-01 7.512 5.84e-14 \*\*\*  
## CAR\_TYPEVan 5.778e-01 1.494e-01 3.866 0.000110 \*\*\*  
## CAR\_TYPEz\_SUV 8.112e-01 1.247e-01 6.508 7.63e-11 \*\*\*  
## MSTATUSz\_No 4.146e-01 9.568e-02 4.334 1.47e-05 \*\*\*  
## REVOKEDYes 8.551e-01 1.074e-01 7.965 1.65e-15 \*\*\*  
## CAR\_USEPrivate -8.324e-01 1.059e-01 -7.861 3.82e-15 \*\*\*  
## TRAVTIME 1.557e-02 2.189e-03 7.114 1.13e-12 \*\*\*  
## TIF -5.237e-02 8.537e-03 -6.134 8.56e-10 \*\*\*  
## KIDSDRIV 3.307e-01 6.390e-02 5.176 2.27e-07 \*\*\*  
## INCOME\_num -3.471e-06 1.424e-06 -2.439 0.014744 \*   
## CLM\_FREQ 2.001e-01 3.318e-02 6.032 1.62e-09 \*\*\*  
## EDUCATIONBachelors -3.995e-01 1.231e-01 -3.245 0.001174 \*\*   
## EDUCATIONMasters -4.969e-01 1.905e-01 -2.609 0.009094 \*\*   
## EDUCATIONPhD 4.064e-02 2.461e-01 0.165 0.868809   
## EDUCATIONz\_High School -7.396e-03 1.063e-01 -0.070 0.944554   
## BLUEBOOK\_num -2.370e-05 6.031e-06 -3.930 8.51e-05 \*\*\*  
## PARENT1Yes 4.730e-01 1.089e-01 4.342 1.41e-05 \*\*\*  
## HOME\_VAL\_num -1.469e-06 4.267e-07 -3.443 0.000575 \*\*\*  
## OLDCLAIM\_num -1.326e-05 4.569e-06 -2.903 0.003700 \*\*   
## RED\_CARyes -2.236e-01 1.031e-01 -2.168 0.030141 \*   
## SEXz\_F -1.861e-01 1.272e-01 -1.464 0.143274   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 6990.9 on 6044 degrees of freedom  
## Residual deviance: 5365.0 on 6012 degrees of freedom  
## AIC: 5431  
##   
## Number of Fisher Scoring iterations: 5

anova(final\_model\_backwards, test="Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: TARGET\_FLAG  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 6044 6990.9   
## KIDSDRIV 1 42.25 6043 6948.6 8.032e-11 \*\*\*  
## PARENT1 1 121.66 6042 6826.9 < 2.2e-16 \*\*\*  
## MSTATUS 1 32.03 6041 6794.9 1.517e-08 \*\*\*  
## SEX 1 1.26 6040 6793.7 0.261667   
## EDUCATION 4 142.63 6036 6651.0 < 2.2e-16 \*\*\*  
## JOB 7 90.04 6029 6561.0 < 2.2e-16 \*\*\*  
## TRAVTIME 1 10.04 6028 6550.9 0.001530 \*\*   
## CAR\_USE 1 86.88 6027 6464.1 < 2.2e-16 \*\*\*  
## TIF 1 39.38 6026 6424.7 3.488e-10 \*\*\*  
## CAR\_TYPE 5 121.74 6021 6302.9 < 2.2e-16 \*\*\*  
## RED\_CAR 1 0.97 6020 6302.0 0.324811   
## CLM\_FREQ 1 230.56 6019 6071.4 < 2.2e-16 \*\*\*  
## REVOKED 1 80.01 6018 5991.4 < 2.2e-16 \*\*\*  
## MVR\_PTS 1 89.60 6017 5901.8 < 2.2e-16 \*\*\*  
## URBANICITY 1 476.35 6016 5425.5 < 2.2e-16 \*\*\*  
## HOME\_VAL\_num 1 26.75 6015 5398.7 2.317e-07 \*\*\*  
## INCOME\_num 1 9.38 6014 5389.3 0.002193 \*\*   
## BLUEBOOK\_num 1 15.73 6013 5373.6 7.304e-05 \*\*\*  
## OLDCLAIM\_num 1 8.60 6012 5365.0 0.003368 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

G = final\_model\_backwards$null.deviance - final\_model\_backwards$deviance  
Gdf = final\_model\_backwards$df.null - final\_model\_backwards$df.residual  
  
1-pchisq(G, Gdf)

## [1] 0

# how likely this random person that we made up would be to get into an accident   
  
some\_person = data.frame(   
   
 KIDSDRIV = 0,   
 KIDSDRIV = "No",  
 PARENT1 = "No",  
 MSTATUS = "z\_No",   
 SEX="z\_F",   
 EDUCATION = "Masters",   
 JOB= "Professional",   
 TRAVTIME = 15,   
 CAR\_USE = "Private",   
 TIF = 5,   
 CAR\_TYPE = "z\_SUV",   
 RED\_CAR = "no",   
 CLM\_FREQ = 0,   
 REVOKED = "No",   
 MVR\_PTS = 0,   
 URBANICITY = "Highly Urban/ Urban",   
 HOME\_VAL\_num = 258000,  
 INCOME\_num = 82000,   
 BLUEBOOK\_num = 16400,   
 OLDCLAIM\_num = 0  
)

# Predicts logodds  
# Using the final backwards model, this predicts how likely that the person we made up will get into an accident Put type = REsponse if you want probabilitiy   
  
predict(final\_model\_backwards, some\_person)

## 1   
## -2.05658

odds = exp(predict(final\_model\_backwards, some\_person))  
  
odds/(1+odds)

## 1   
## 0.1133892

# Gets teh odds of getting into an accident   
# Same as the predict probability function below

# Predicts probability  
# this predicts the probability that the person we made up will get into an accident  
predict(final\_model\_backwards, some\_person, type = "response")

## 1   
## 0.1133892

predict(final\_model\_forwards, some\_person, type = "response")

## 1   
## 0.1133892

predict(final\_model\_both, some\_person, type = "response")

## 1   
## 0.1133892

## **STOR 455 Homework 6**

Are Emily and Greg More Employable Than Lakisha and Jamal?

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review, 94*(4), pp. 991-1013.

*Abstract*

We perform a field experiment to measure racial discrimination in the labor market. We respond with fictitious resumes to help-wanted ads in Boston and Chicago newspapers. To manipulate perception of race, each resume is randomly assigned either a very African American sounding name or a very White sounding name. The results show significant discrimination against African-American names: White names receive 50 percent more callbacks for interviews. We also find that race affects the benefits of a better resume. For White names, a higher quality resume elicits 30 percent more callbacks whereas for African Americans, it elicits a far smaller increase. Applicants living in better neighborhoods receive more callbacks but, interestingly, this effect does not differ by race. The amount of discrimination is uniform across occupations and industries. Federal contractors and employers who list “Equal Opportunity Employer” in their ad discriminate as much as other employers. We find little evidence that our results are driven by employers inferring something other than race, such as social class, from the names. These results suggest that racial discrimination is still a prominent feature of the labor market.

| **Variables** | **Descriptions** |
| --- | --- |
| *call* | Was the applicant called back? (1 = yes; 0 = no) |
| *ethnicity* | indicating ethnicity (i.e., “Caucasian-sounding” vs. “African-American sounding” first name) |
| *sex* | indicating sex |
| *quality* | Indicating quality of resume. |
| *experience* | Number of years of work experience on the resume |
| *equal* | Is the employer EOE (equal opportunity employment)? |

Use the *ResumeNames455* found at the address below:

<https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/ResumeNames455.csv>

library(readr)

ResumeNames455 = read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/ResumeNames455.csv")

1. Construct a logistic model to predict if the job applicant was called back using *experience* as the predictor variable.

mod1 = glm(call~experience, data=ResumeNames455, family = binomial)

summary(mod1)

##

## Call:

## glm(formula = call ~ experience, family = binomial, data = ResumeNames455)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -0.7780 -0.4075 -0.3924 -0.3779 2.3598

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -2.75960 0.09620 -28.687 < 2e-16 \*\*\*

## experience 0.03908 0.00918 4.257 2.07e-05 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 2726.9 on 4869 degrees of freedom

## Residual deviance: 2710.2 on 4868 degrees of freedom

## AIC: 2714.2

##

## Number of Fisher Scoring iterations: 5

1. Plot the raw data and the logistic curve on the same axes.

plot(jitter(call, amount=0.1)~experience, data=ResumeNames455)

logit = function(B0, B1, x){

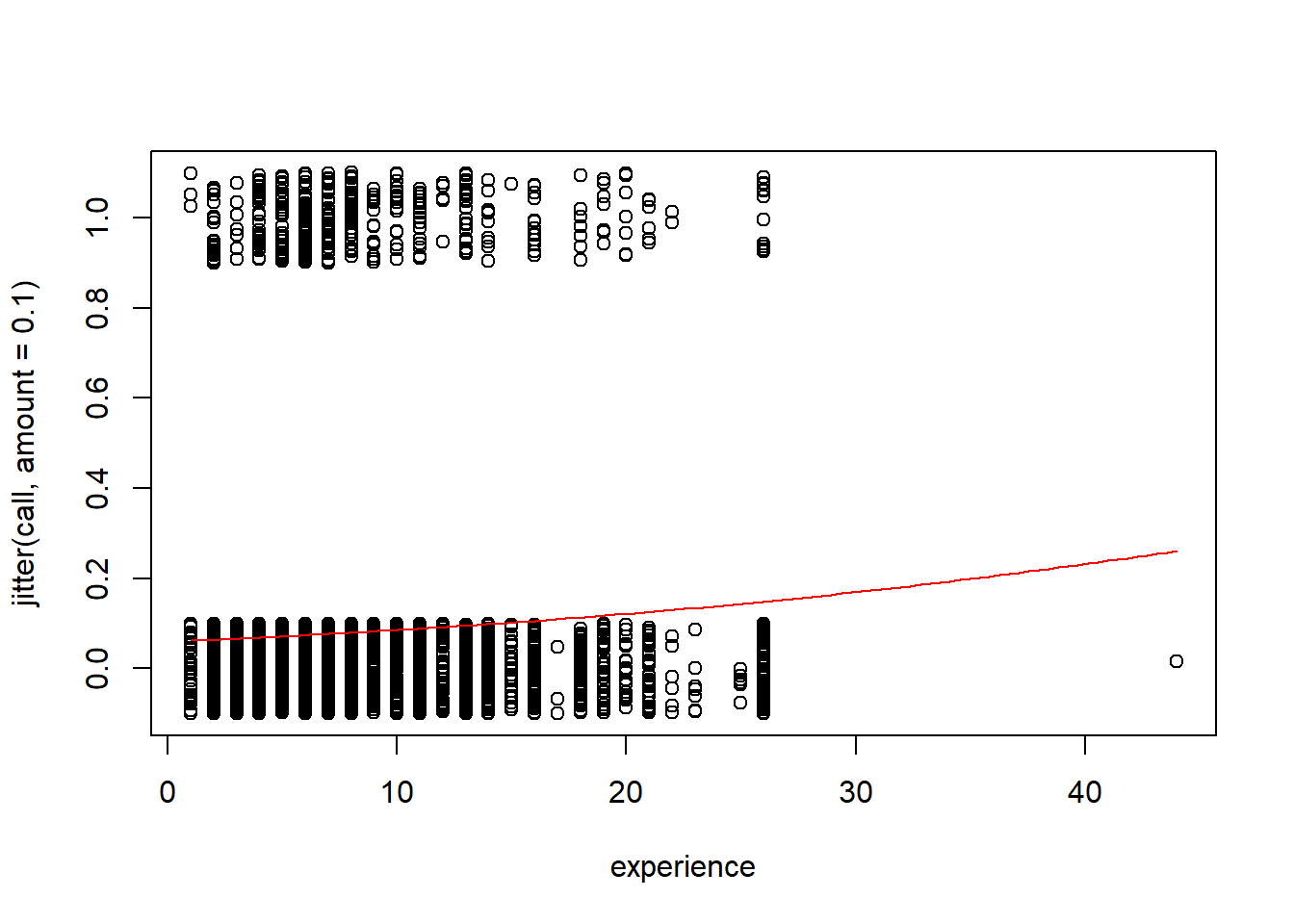
exp(B0 + B1\*x)/(1 + exp(B0 + B1\*x))

}

B0 = summary(mod1)$coef[1]

B1 = summary(mod1)$coef[2]

curve(logit(B0, B1, x), add=TRUE, col="red")



1. For an applicant with 3 years of experience, what does your model predict is the probability of this applicant getting called back?

# without type='response' you are predicting the log(odds)

applicant = data.frame(experience = 3)

predict(mod1, applicant, type="response")

## 1

## 0.06646115

1. Construct an empirical logit plot and comment on the linearity of the data.

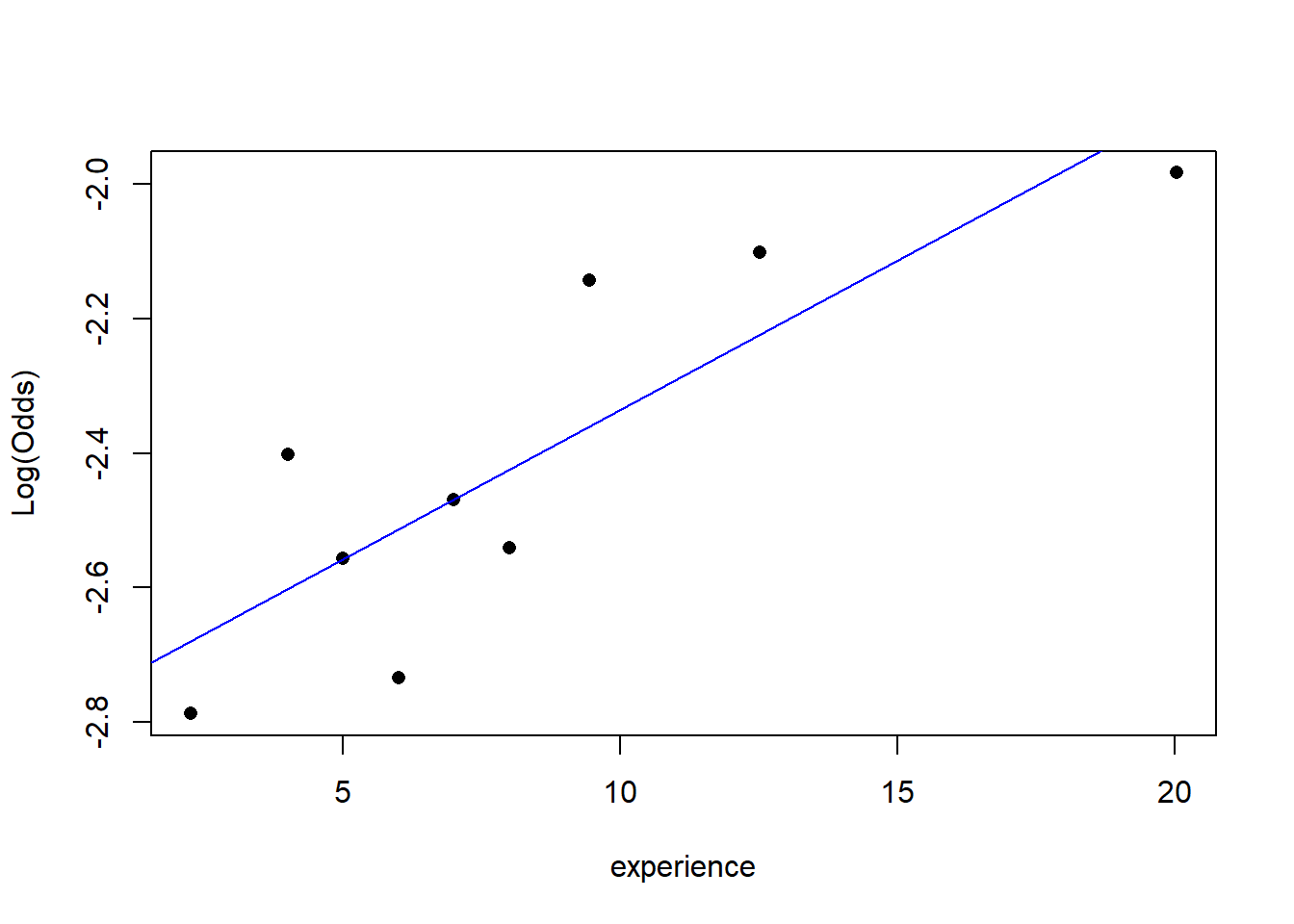
There are no clear nonlinear patterns in the data, so the logistic model seems appropriate

library(Stat2Data)

# ngroups=9 is arbitrarily chosen.

# This data was causes error in the function for many choices of breaks

emplogitplot1(call~experience, data=ResumeNames455, ngroups = 9)



1. Use the model from question #1 to perform a hypothesis test to determine if there is significant evidence of a relationship between *call* and *experience*. Cite your hypotheses, p-value, and conclusion in context.

H0: β1 = 0  
HA: β1 ≠ 0

Since the p-value (2.07e-05 using the summary or 4.298e-05 using log likelihood from anova) is less than 0.05, there is evidence to suggest that the coefficient of the experience term in the binary logistic model is nonzero.

# Either method could be used

summary(mod1)

##

## Call:

## glm(formula = call ~ experience, family = binomial, data = ResumeNames455)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -0.7780 -0.4075 -0.3924 -0.3779 2.3598

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -2.75960 0.09620 -28.687 < 2e-16 \*\*\*

## experience 0.03908 0.00918 4.257 2.07e-05 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 2726.9 on 4869 degrees of freedom

## Residual deviance: 2710.2 on 4868 degrees of freedom

## AIC: 2714.2

##

## Number of Fisher Scoring iterations: 5

anova(mod1, test="Chisq")

|  |
| --- |
|  |

|  | **Df**  **<int>** | **Deviance**  **<dbl>** | **Resid. Df**  **<int>** | **Resid. Dev**  **<dbl>** | **Pr(>Chi)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- |
| NULL | NA | NA | 4869 | 2726.921 | NA |
| experience | 1 | 16.73516 | 4868 | 2710.186 | 4.2977e-05 |

2 rows

1. Construct a confidence interval for the odds ratio for your model and include a sentence interpreting the interval in the context.

We are 95% confident that for each 1 year increase in experience, the odds of getting called back will increase by a factor between approximately 1.02 and 1.06.

exp(confint(mod1))

## Waiting for profiling to be done...

## 2.5 % 97.5 %

## (Intercept) 0.05235403 0.07634629

## experience 1.02097262 1.05841947

1. For each 5-year increase in *experience*, how does your model predict the odds will change for the applicant getting called back?

For each 5 year increase in experience, the odds of getting called back will increase by a factor of approximately 1.22.

exp(summary(mod1)$coef[2,1]) ^ 5

## [1] 1.215796

# or by multiplying before using exp()

exp(5 \* summary(mod1)$coef[2,1])

## [1] 1.215796

In homework #7 we will continue with this data to investigate how the other variables impact an applicant’s chances of being called back.

## STOR 455 Practice Exam Solutions

library(readr)  
library(leaps)  
library(bestglm)  
library(MASS)  
  
BirthWeight <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/BirthWeight.csv")  
abalone\_train <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/abalone\_train.csv")  
abalone\_test <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/abalone\_test.csv")

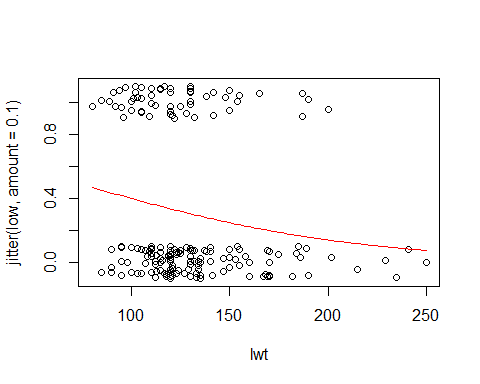
#### Question 1

Low birth weight is an outcome that has been of concern to physicians for years. This is due to the fact that infant mortality rates and birth defect rates are very high for low birth weight babies. Behavior during pregnancy (including diet, smoking habits, and receiving prenatal care) can greatly alter the chances of carrying the baby to term and, consequently, of delivering a baby of normal birth weight. Data were collected at Baystate Medical Center, Springfield, Massachusetts, in 1986 for variables (shown in the table below) that have been shown to be associated with low birth weight in the obstetrical literature.

| Variable | Description |
| --- | --- |
| low | indicator of child’s birth weight less than 2.5 kg. |
| age | mother’s age in years. |
| lwt | mother’s weight in pounds at last menstrual period. |
| race | mother’s race (1 = white, 2 = black, 3 = other). |
| smoke | smoking status during pregnancy. |
| ptl | number of previous premature labours. |
| ht | history of hypertension |
| ui | presence of uterine irritability. |
| ftv | number of physician visits during the first trimester. |
| bwt | child’s birth weight in grams. |

1. Construct and plot a model using the indicator for a child’s low birth weight, *low*, as the response variable, and the mother’s weight in pounds at last menstrual period, *lwt*, and the predictor.

mod1A=glm(low~lwt, data=BirthWeight, family="binomial")  
plot(jitter(low, amount=.1)~lwt, data=BirthWeight)  
  
logit = function(B0, B1, x)  
{  
 exp(B0+B1\*x)/(1+exp(B0+B1\*x))  
}  
  
B0 = summary(mod1A)$coef[1]  
B1 = summary(mod1A)$coef[2]  
  
curve(logit(B0, B1, x),add=TRUE, col="red")



1. Construct a model using the indicator for a child’s low birth weight, *low*, as the response variable, and the mother’s weight in pounds at last menstrual period, age, smoking status during pregnancy, and race as the predictor variables.

mod1B=glm(low~lwt+age+smoke+factor(race), data=BirthWeight, family="binomial")

1. Is there evidence to suggest that the model constructed in part (B) is significantly better than the model constructed in part (A)? Conduct the appropriate hypothesis test. State hypotheses, and provide a conclusion in the context of the data. *6 pts*

Null: The coefficients for the variables age, smoke, and (both dummy) races are 0;  
Alternative: The coefficients for at least one of the variables age, smoke, and race are not 0.  
Statistically significant evidence suggests that at least one of the additional terms has a nonzero coefficient, thus making for a better model than the one with a single predictor.

anova(mod1A, mod1B, test="Chisq")

## Analysis of Deviance Table  
##   
## Model 1: low ~ lwt  
## Model 2: low ~ lwt + age + smoke + factor(race)  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 187 228.69   
## 2 183 214.58 4 14.113 0.006942 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#You may instead examine the summaries of these models to find the difference between the residual deviances, and compare this G statistic to the chi-squared distribution with 4 df.  
  
summary(mod1A)

##   
## Call:  
## glm(formula = low ~ lwt, family = "binomial", data = BirthWeight)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.0951 -0.9022 -0.8018 1.3609 1.9821   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.99831 0.78529 1.271 0.2036   
## lwt -0.01406 0.00617 -2.279 0.0227 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 234.67 on 188 degrees of freedom  
## Residual deviance: 228.69 on 187 degrees of freedom  
## AIC: 232.69  
##   
## Number of Fisher Scoring iterations: 4

summary(mod1B)

##   
## Call:  
## glm(formula = low ~ lwt + age + smoke + factor(race), family = "binomial",   
## data = BirthWeight)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.5173 -0.9065 -0.5865 1.3035 2.0401   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.332452 1.107672 0.300 0.76407   
## lwt -0.012526 0.006386 -1.961 0.04982 \*   
## age -0.022478 0.034170 -0.658 0.51065   
## smoke 1.054439 0.380000 2.775 0.00552 \*\*  
## factor(race)2 1.231671 0.517152 2.382 0.01724 \*   
## factor(race)3 0.943263 0.416232 2.266 0.02344 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 234.67 on 188 degrees of freedom  
## Residual deviance: 214.58 on 183 degrees of freedom  
## AIC: 226.58  
##   
## Number of Fisher Scoring iterations: 4

G = summary(mod1A)$deviance - summary(mod1B)$deviance  
1 - pchisq(G, 4)

## [1] 0.006941683

1. Use one of the model selection procedures covered in class to determine the best model to predict the indicator for a child’s low birth weight, *low*.

The “best” models produced each way look a bit different, since AIC and BIC values are not directly comparable. Each method has positives and negatives. The models are quite different! BIC penalizes model complexity more heavily, hence the “best” models have fewer terms.

#With bestglm  
  
# Must factor race so it is considered categorical  
  
BirthWeight$race = as.factor(BirthWeight$race)  
  
# Must move low to last column and remove bwt.   
  
# bwt is the baby's birth weight, which will directly correspond  
# to low and cause an error. When predicting if a baby has low   
# birth weight, you won;t know their birth weight first  
  
BirthWeight\_forbestglm = BirthWeight[,c(2:9, 1)]  
head(BirthWeight\_forbestglm)

## # A tibble: 6 x 9  
## age lwt race smoke ptl ht ui ftv low  
## <dbl> <dbl> <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 19 182 2 0 0 0 1 0 0  
## 2 33 155 3 0 0 0 0 3 0  
## 3 20 105 1 1 0 0 0 1 0  
## 4 21 108 1 1 0 0 1 2 0  
## 5 18 107 1 1 0 0 1 0 0  
## 6 21 124 3 0 0 0 0 0 0

#This line is sometimes need to restore structure to the dataframe  
BirthWeight\_forbestglm = as.data.frame(BirthWeight\_forbestglm)  
  
bestglm1D = bestglm(BirthWeight\_forbestglm, family=binomial)

## Morgan-Tatar search since family is non-gaussian.

## Note: factors present with more than 2 levels.

bestglm1D$BestModels

## age lwt race smoke ptl ht ui ftv Criterion  
## 1 FALSE TRUE FALSE FALSE FALSE TRUE FALSE FALSE 231.6256  
## 2 FALSE TRUE FALSE FALSE TRUE TRUE FALSE FALSE 231.6890  
## 3 FALSE TRUE FALSE FALSE FALSE TRUE TRUE FALSE 232.3381  
## 4 FALSE TRUE FALSE TRUE FALSE TRUE FALSE FALSE 232.5830  
## 5 FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE 233.1344

#With stepAIC  
  
BirthWeight\_foretepAIC = BirthWeight[1:9]  
  
# I factored the variable race first above.   
# You could have factored it inside of the glm() function.  
  
mod1D = glm(low~., data=BirthWeight\_foretepAIC, family="binomial")  
none = glm(low~1, data=BirthWeight\_foretepAIC, family="binomial")  
  
# You could have dine this with backwards or forwards as well.  
  
stepAIC(none, scope = list(upper = mod1D), trace=0)

##   
## Call: glm(formula = low ~ ptl + lwt + ht + race + smoke + ui, family = "binomial",   
## data = BirthWeight\_foretepAIC)  
##   
## Coefficients:  
## (Intercept) ptl lwt ht race2 race3   
## -0.08655 0.50321 -0.01591 1.85504 1.32572 0.89708   
## smoke ui   
## 0.93873 0.78570   
##   
## Degrees of Freedom: 188 Total (i.e. Null); 181 Residual  
## Null Deviance: 234.7   
## Residual Deviance: 202 AIC: 218

#### Question 2

Abalone are marine gastropod molluscs, which means they are marine snails.The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope – a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem.

| Variable | Description |
| --- | --- |
| Sex | M, F |
| Length | longest shell measurement in mm |
| Diameter | perpendicular to length in mm |
| Height | with meat in shell in mm |
| Whole weight | whole abalone in g |
| Shucked weight | weight of meat in g |
| Viscera weight | gut weight (after bleeding) in g |
| Shell weight | after being dried in g |
| Rings | number of rings |

1. Construct a model to predict abalones’ age (using *rings* as the response) with the lowest Mallow’s Cp using any/all of the variables in the *abalone train* dataset. Do not use transformations, or second or greater order terms, or perform an analysis of the residuals.

# model with regsubsets  
  
regsubsets2A=regsubsets(rings~., data=abalone\_train)  
  
source("https://raw.githubusercontent.com/JA-McLean/STOR455/master/scripts/ShowSubsets.R")  
ShowSubsets(regsubsets2A)

## sexM length diameter height weight\_whole weight\_shucked weight\_viscera  
## 1 ( 1 )   
## 2 ( 1 ) \* \*   
## 3 ( 1 ) \* \* \*  
## 4 ( 1 ) \* \* \* \*  
## 5 ( 1 ) \* \* \* \* \*  
## 6 ( 1 ) \* \* \* \* \*  
## 7 ( 1 ) \* \* \* \* \* \*  
## 8 ( 1 ) \* \* \* \* \* \* \*  
## weight\_shell Rsq adjRsq Cp  
## 1 ( 1 ) \* 19.40 19.29 227.63  
## 2 ( 1 ) 35.60 35.41 44.03  
## 3 ( 1 ) 37.49 37.22 24.31  
## 4 ( 1 ) 39.05 38.70 8.45  
## 5 ( 1 ) 39.53 39.10 4.96  
## 6 ( 1 ) \* 39.69 39.17 5.11  
## 7 ( 1 ) \* 39.70 39.09 7.03  
## 8 ( 1 ) \* 39.70 39.00 9.00

mod2A = lm(rings~length+height+weight\_whole+weight\_shucked+weight\_viscera, data=abalone\_train)  
  
# model with step - produces the same best model  
  
none2 = lm(rings~1, data=abalone\_train)  
full = lm(rings~., data=abalone\_train)  
MSE = (summary(full)$sigma)^2  
  
step(none2,scope=list(upper=full),scale=MSE, trace=0)

##   
## Call:  
## lm(formula = rings ~ weight\_shucked + weight\_whole + height +   
## weight\_viscera + length, data = abalone\_train)  
##   
## Coefficients:  
## (Intercept) weight\_shucked weight\_whole height weight\_viscera   
## 4.331 -24.898 14.057 20.404 -14.205   
## length   
## 5.742

#model with forward - produces slightly different model  
  
step(none2,scope=list(upper=full),scale=MSE, direction="forward", trace=0)

##   
## Call:  
## lm(formula = rings ~ weight\_shell + weight\_shucked + weight\_whole +   
## height + weight\_viscera + length, data = abalone\_train)  
##   
## Coefficients:  
## (Intercept) weight\_shell weight\_shucked weight\_whole height   
## 4.534 3.942 -22.957 11.906 18.939   
## weight\_viscera length   
## -12.338 5.383

#model with backward - produces the same best model  
  
step(full,scale=MSE, trace=0)

##   
## Call:  
## lm(formula = rings ~ length + height + weight\_whole + weight\_shucked +   
## weight\_viscera, data = abalone\_train)  
##   
## Coefficients:  
## (Intercept) length height weight\_whole weight\_shucked   
## 4.331 5.742 20.404 14.057 -24.898   
## weight\_viscera   
## -14.205

1. A second dataset, abalone\_test, contains additional data for 500 more abalone. Use this dataset, and your model constructed in part (A), to perform a cross validation analysis of your model. Calcuate and comment on the cross-validation correlation, shrinkage, and analysis of holdout residuals. Does the model constructed in part (A) appear to be similarly effective for predicting the number of rings for abalone?

Holdout residual mean relatively close to zero (close is relative)  
Holdout standard deviation is very similar to the standard error of the regression line for the original model constructed from the training data.  
The shape of the holdout residulas is approximately normally distributed, but might indicate a slight bias and the center seems to be shifted left.  
Shrinkage is near 0.10, which isn’t as small as it could be, but suggests that the model predicts the new data similarly as well as the old.

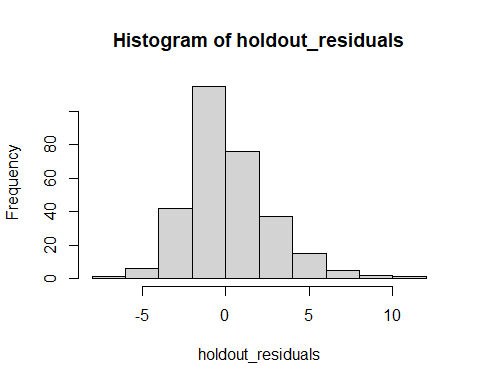
fit = predict(mod2A, abalone\_test)  
  
holdout\_residuals = abalone\_test$rings - fit  
  
mean(holdout\_residuals)

## [1] 0.1286006

sd(holdout\_residuals)

## [1] 2.596266

hist(holdout\_residuals)



summary(mod2A)

##   
## Call:  
## lm(formula = rings ~ length + height + weight\_whole + weight\_shucked +   
## weight\_viscera, data = abalone\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.8210 -1.5879 -0.3408 1.0211 11.9233   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.3314 0.9636 4.495 8.16e-06 \*\*\*  
## length 5.7423 2.4497 2.344 0.019357 \*   
## height 20.4039 5.7006 3.579 0.000369 \*\*\*  
## weight\_whole 14.0574 1.1463 12.263 < 2e-16 \*\*\*  
## weight\_shucked -24.8985 1.5680 -15.879 < 2e-16 \*\*\*  
## weight\_viscera -14.2047 2.8975 -4.902 1.18e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.439 on 694 degrees of freedom  
## Multiple R-squared: 0.3953, Adjusted R-squared: 0.391   
## F-statistic: 90.74 on 5 and 694 DF, p-value: < 2.2e-16

cv\_corr = cor(fit, abalone\_test$rings)  
  
summary(mod2A)$r.squared - cv\_corr^2

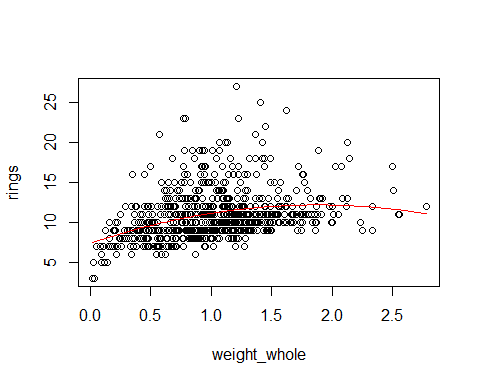
## [1] 0.09549159

1. Linearity is an issue in any abalone model that uses the various measures of weight to predict the number of rings. Would a polynomial model be more appropriate? Contruct a quadratic model using *rings* as the response and the *weight whole* as the predictor. Plot the data and the curve on the same axes. Use the *abalone train* dataset.

mod2C = lm(rings~weight\_whole+I(weight\_whole^2), data=abalone\_train)  
  
summary(mod2C)

##   
## Call:  
## lm(formula = rings ~ weight\_whole + I(weight\_whole^2), data = abalone\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.8056 -1.9640 -0.8028 1.0129 15.4238   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.3778 0.4598 16.045 < 2e-16 \*\*\*  
## weight\_whole 5.1079 0.8353 6.115 1.61e-09 \*\*\*  
## I(weight\_whole^2) -1.3569 0.3569 -3.802 0.000156 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.952 on 697 degrees of freedom  
## Multiple R-squared: 0.1104, Adjusted R-squared: 0.1079   
## F-statistic: 43.26 on 2 and 697 DF, p-value: < 2.2e-16

a = summary(mod2C)$coef[3,1]  
b = summary(mod2C)$coef[2,1]  
c = summary(mod2C)$coef[1,1]  
  
plot(rings~weight\_whole, data=abalone\_train)  
curve(a\*x^2 + b\*x + c, add=TRUE, col="red")



1. Consider a model that uses the log( *rings* ) as the response variable. The predictor variables for the model are *diameter*, *length*, *sex*, and the interactions between *sex* and each other predictor variable. Perform a hypothesis test to determine if the model including the interaction terms is significantly better than a model including the same variables but without the interactions. Include the hypotheses and conclusion.

null: coefficients for the interaction terms are 0;  
alternative: the coefficients for at least one interaction term is nonzero.  
Since the p-value is small, there is significant evidence to suggest that at least one of the interaction terms have a nonzero coefficient.

mod2D1 = lm(log(rings)~length+diameter+sex+length\*sex+diameter\*sex, data=abalone\_train)  
  
mod2D2 = lm(log(rings)~length+diameter+sex, data=abalone\_train)  
  
anova(mod2D2, mod2D1)

## Analysis of Variance Table  
##   
## Model 1: log(rings) ~ length + diameter + sex  
## Model 2: log(rings) ~ length + diameter + sex + length \* sex + diameter \*   
## sex  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 696 40.884   
## 2 694 40.153 2 0.73099 6.3172 0.001911 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## **STOR 455 Exam #2**

**Directions:** This exam is open books, notes, internet, and all things other than direct communication with others. The *LAdata.csv* dataset is needed to complete the exam. This dataset can be imported from the web address below or from the csv file, also attached in this Sakai assignment. You should complete the exam in this R Notebook, including all code, plots, and explanations. For your submission, you should knit the notebook and submit it as a pdf to Gradescope. If you are unable to knit your exam, you should submit the RMD file to Sakai under the ‘Unable to Knit’ tab. The dataset can be found at GitHub at the address below:

<https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/LAdata.csv>

**Should STOR have Undergraduate Learning Assitants? (YES! Starting Fall 2022…)**

Large introductory STEM courses historically have high failure rates, and failing such courses often leads students to change majors or even drop out of college. Instructional innovations such as the Learning Assistant model can influence this trend by changing institutional norms. In collaboration with faculty who teach large-enrollment introductory STEM courses, undergraduate learning assistants (LAs) use research-based instructional strategies designed to encourage active student engagement and elicit student thinking. These instructional innovations help students master the types of skills necessary for college success such as critical thinking and defending ideas. A study was conducted to investigate the relationship between exposure to LA support in large introductory STEM courses and general failure rates in introductory courses at University of Colorado Boulder.

Alzen, J.L., Langdon, L.S. & Otero, V.K. (2018) A logistic regression investigation of the relationship between the Learning Assistant model and failure rates in introductory STEM courses. *International Journal of STEM Education*, *56*(5). <https://doi.org/10.1186/s40594-018-0152-1>

The *LAdata.csv* dataset represent a subset of the variables examined in this study and a random sample of the data for students in one course, MATH 1300.

| **Variables** | **Descriptions** |
| --- | --- |
| *la\_stud* | Did the student’s course have a learning assistant? (1=yes; 0=no) |
| *sex* | Identified sex of the student (1=male; 0=female) |
| *nonwhite* | Does the student identify as not white? (1=yes; 0=no) |
| *first.gen* | Is the student a first generation college student? (1=yes; 0=no) |
| *finaid\_ever* | Has the student ever receive financial aid? (1=yes; 0=no) |
| *act\_new* | ACT score for the student |
| *hs\_gpa* | High school GPA for the student |
| *credits\_entry* | College credits earned before entering the university |
| *grade* | Grade in the course on a 0.0 to 4.0 scale |
| *fail* | Did the student fail the course? (1=yes; 0=no) |

library(readr)

LAdata <- read\_csv('https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/LAdata.csv')

1. Use the *LAdata.csv* dataset to construct a model to predict if students will *fail* a class using *la\_stud*, *sex*, *nonwhite*, *first.gen*, as well as the interactions between *la\_stud* with each of *sex*, *nonwhite* and *first.gen* as predictors (ie 3 interactions of two variables, with *la\_stud* and each one of the other predictors). Include a summary of this model. *6 pts*

mod1 = glm(fail ~

la\_stud +

nonwhite +

first.gen +

sex +

la\_stud\*nonwhite +

la\_stud\*first.gen +

la\_stud\*sex,

data=LAdata,

family=binomial)

summary(mod1)

##

## Call:

## glm(formula = fail ~ la\_stud + nonwhite + first.gen + sex + la\_stud \*

## nonwhite + la\_stud \* first.gen + la\_stud \* sex, family = binomial,

## data = LAdata)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -0.9828 -0.6412 -0.5567 -0.3909 2.3218

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -1.22376 0.20397 -6.000 1.98e-09 \*\*\*

## la\_stud -0.44374 0.25337 -1.751 0.0799 .

## nonwhite 0.74711 0.34294 2.179 0.0294 \*

## first.gen -0.09224 0.40773 -0.226 0.8210

## sex -1.30957 0.32798 -3.993 6.53e-05 \*\*\*

## la\_stud:nonwhite -0.55691 0.40915 -1.361 0.1735

## la\_stud:first.gen 0.70932 0.47646 1.489 0.1366

## la\_stud:sex 0.91348 0.38946 2.346 0.0190 \*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 959.54 on 1068 degrees of freedom

## Residual deviance: 926.12 on 1061 degrees of freedom

## AIC: 942.12

##

## Number of Fisher Scoring iterations: 5

1. Conduct a hypothesis test at the 0.05 significance level to determine the effectiveness of the *la\_stud* terms in the model constructed in question 1. Cite your hypotheses, p-value, and conclusion in context. *8pts*

*H*0:*β*1=*β*5=*β*6=*β*7=0

*HA*:*At* *least* *one* *of* *β*1,*β*5,*β*6,*β*7≠0

Since the p-value is small (0.04286 < 0.05), there is statistically significant evidence to claim that at least one of the la\_stud terms have a nonzero coefficient.

mod2 = glm(fail ~

nonwhite +

first.gen +

sex,

data=LAdata,

family=binomial)

anova(mod2, mod1, test='Chisq')

|  |
| --- |
|  |

|  | **Resid. Df**  **<dbl>** | **Resid. Dev**  **<dbl>** | **Df**  **<dbl>** | **Deviance**  **<dbl>** | **Pr(>Chi)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- |
| 1 | 1065 | 935.9784 | NA | NA | NA |
| 2 | 1061 | 926.1187 | 4 | 9.859697 | 0.04285856 |

2 rows

# Can also be done this way

G = summary(mod2)$deviance - summary(mod1)$deviance

1 - pchisq(G, 4)

## [1] 0.04285856

1. For non first generation nonwhite female students, what does the model from question 1 predict will be the probability that these students will **pass** the course for courses **with** a learning assistant? What does the model from question 1 predict will be the probability that these students will **pass** the course for courses **without** a learning assistant? *8pts*

s1 = data.frame(sex=0, nonwhite=1, first.gen=0, la\_stud=1)

s2 = data.frame(sex=0, nonwhite=1, first.gen=0, la\_stud=0)

#Probability non first generation nonwhite female student passes with LA

1 - predict(mod1, s1, type='response')

## 1

## 0.8141637

#Probability non first generation nonwhite female student passes without LA

1 - predict(mod1, s2, type='response')

## 1

## 0.6169563

1. Construct a model to predict students’ *grade* in the course using all of the variables (except for *fail*, *hs\_gpa*, and *act\_new*) and including the interactions between *la\_stud* and all of the other variables (except for *fail*, *hs\_gpa*, and *act\_new* and as with question 1, the interactions between *la\_stud* and each one of the other predictors). You should not include transformations, nor a residual analysis. Include a summary of this model. *5 pts*

mod4 = lm(grade~. + la\_stud\*., data=LAdata[-c(6, 7, 10)])

summary(mod4)

##

## Call:

## lm(formula = grade ~ . + la\_stud \* ., data = LAdata[-c(6, 7,

## 10)])

##

## Residuals:

## Min 1Q Median 3Q Max

## -3.2440 -0.6333 0.1551 0.8389 2.2176

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 2.3955419 0.1124618 21.301 < 2e-16 \*\*\*

## la\_stud 0.0026514 0.1343917 0.020 0.984264

## sex 0.4466620 0.1224744 3.647 0.000278 \*\*\*

## nonwhite -0.3225878 0.1495213 -2.157 0.031194 \*

## first.gen 0.1253382 0.1706271 0.735 0.462762

## finaid\_ever -0.4158908 0.1284067 -3.239 0.001237 \*\*

## credits\_entry 0.0272103 0.0084656 3.214 0.001348 \*\*

## la\_stud:sex -0.2341253 0.1478221 -1.584 0.113531

## la\_stud:nonwhite 0.1872411 0.1767231 1.060 0.289606

## la\_stud:first.gen -0.3123343 0.2060254 -1.516 0.129819

## la\_stud:finaid\_ever 0.3500524 0.1542170 2.270 0.023416 \*

## la\_stud:credits\_entry 0.0006063 0.0097250 0.062 0.950300

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 1.106 on 1057 degrees of freedom

## Multiple R-squared: 0.08483, Adjusted R-squared: 0.0753

## F-statistic: 8.907 on 11 and 1057 DF, p-value: 2.621e-15

1. Using **only** the summary output of the model constructed in question 4, What does the coefficient of the *la\_stud* variable (not the interactions, nor the test statistic or p-value) tell you about the relationship between the *la\_stud* and *grade* variables in this model, as well as what specific students this coefficient applies to? *5pts*

The la\_stud coefficient (0.0026514) in the increase in GPA for students in a course with a learning assistant, specifically for students that are female (sex=0), not nonwhite (nonwhite=0), not first generation college students (first.gen=0), have not had financial aid ever (finaid\_ever=0), and have no college credits before entering the university (credits\_entry=0).

1. Perform a **backwards** model selection method (for the lowest Mallow’s Cp) using all of the terms in the model that you constructed in question 4 as possible predictors. Construct this best model and include a summary of this model. *6pts*

MSE = (summary(mod4)$sigma)^2

mod5 = step(mod4, scale=MSE, trace=FALSE)

summary(mod5)

##

## Call:

## lm(formula = grade ~ la\_stud + sex + nonwhite + finaid\_ever +

## credits\_entry + la\_stud:sex + la\_stud:finaid\_ever, data = LAdata[-c(6,

## 7, 10)])

##

## Residuals:

## Min 1Q Median 3Q Max

## -3.1515 -0.6342 0.1717 0.8420 2.2277

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 2.382222 0.108682 21.919 < 2e-16 \*\*\*

## la\_stud 0.019275 0.127257 0.151 0.879639

## sex 0.446036 0.121615 3.668 0.000257 \*\*\*

## nonwhite -0.200140 0.078110 -2.562 0.010536 \*

## finaid\_ever -0.409754 0.122709 -3.339 0.000869 \*\*\*

## credits\_entry 0.028039 0.004157 6.744 2.52e-11 \*\*\*

## la\_stud:sex -0.241861 0.146793 -1.648 0.099725 .

## la\_stud:finaid\_ever 0.314556 0.146807 2.143 0.032369 \*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 1.105 on 1061 degrees of freedom

## Multiple R-squared: 0.08155, Adjusted R-squared: 0.07549

## F-statistic: 13.46 on 7 and 1061 DF, p-value: < 2.2e-16

1. Conduct a hypothesis test at the 0.05 significance level to determine the effectiveness of the *la\_stud* terms in the model constructed in question 6. Cite your hypotheses, p-value, and conclusion in context. *8pts*

*H*0:*β*1=*β*6=*β*7=0

*HA*:*At* *least* *one* *of* *β*1,*β*6,*β*7≠0

Since the p-value is small (0.0447 < 0.05), there is statistically significant evidence to claim that at least one of the la\_stud terms have a nonzero coefficient.

mod7 = lm(grade~sex+nonwhite+finaid\_ever+credits\_entry, data=LAdata)

anova(mod7, mod5)

|  |
| --- |
|  |

|  | **Res.Df**  **<dbl>** | **RSS**  **<dbl>** | **Df**  **<dbl>** | **Sum of Sq**  **<dbl>** | **F**  **<dbl>** | **Pr(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 1064 | 1306.490 | NA | NA | NA | NA |
| 2 | 1061 | 1296.602 | 3 | 9.887836 | 2.697048 | 0.0447025 |

2 rows

1. In question 6 I specifically asked you to perform a backwards model selection method rather than a forward or stepwise method. Using the backwards method, *la\_stud* terms were included in your best model. If you had used a forwards or stepwise method, you would find that *la\_stud* terms would **not** be included in your best models. Using the summary of the model from question 6, and knowledge of how the backwards, forwards, and stepwise procedures determine the best model, why would the *la\_stud* terms be included in the backwards model selection output, but not in the forward and stepwise model selection output? *4pts*

backwards will first look to remove the interaction terms and not the individual LA\_stud terms. Since the interaction terms are not all removed from the model in backwards selection, it is never an option to remove the LA\_stud term. For forwards and stepwise, you begin with not predictors. THe interaction terms are not considered unless the individual terms are first added to the model. LA\_stud alone has a high p-value and was never added to the model, so the interaction terms were never considered.

step(mod4, scale=MSE)

## Start: AIC=12

## grade ~ la\_stud + sex + nonwhite + first.gen + finaid\_ever +

## credits\_entry + la\_stud \* (la\_stud + sex + nonwhite + first.gen +

## finaid\_ever + credits\_entry)

##

## Df Sum of Sq RSS Cp

## - la\_stud:credits\_entry 1 0.0048 1292.0 10.004

## - la\_stud:nonwhite 1 1.3721 1293.3 11.123

## <none> 1292.0 12.000

## - la\_stud:first.gen 1 2.8092 1294.8 12.298

## - la\_stud:sex 1 3.0662 1295.0 12.508

## - la\_stud:finaid\_ever 1 6.2977 1298.3 15.152

##

## Step: AIC=10

## grade ~ la\_stud + sex + nonwhite + first.gen + finaid\_ever +

## credits\_entry + la\_stud:sex + la\_stud:nonwhite + la\_stud:first.gen +

## la\_stud:finaid\_ever

##

## Df Sum of Sq RSS Cp

## - la\_stud:nonwhite 1 1.369 1293.3 9.1243

## <none> 1292.0 10.0039

## - la\_stud:first.gen 1 2.830 1294.8 10.3189

## - la\_stud:sex 1 3.075 1295.1 10.5199

## - la\_stud:finaid\_ever 1 6.305 1298.3 13.1617

## - credits\_entry 1 53.907 1345.9 52.1067

##

## Step: AIC=9.12

## grade ~ la\_stud + sex + nonwhite + first.gen + finaid\_ever +

## credits\_entry + la\_stud:sex + la\_stud:first.gen + la\_stud:finaid\_ever

##

## Df Sum of Sq RSS Cp

## - la\_stud:first.gen 1 2.113 1295.5 8.8532

## <none> 1293.3 9.1243

## - la\_stud:sex 1 3.061 1296.4 9.6289

## - nonwhite 1 6.838 1300.2 12.7189

## - la\_stud:finaid\_ever 1 7.168 1300.5 12.9887

## - credits\_entry 1 54.251 1347.6 51.5084

##

## Step: AIC=8.85

## grade ~ la\_stud + sex + nonwhite + first.gen + finaid\_ever +

## credits\_entry + la\_stud:sex + la\_stud:finaid\_ever

##

## Df Sum of Sq RSS Cp

## - first.gen 1 1.134 1296.6 7.7812

## <none> 1295.5 8.8532

## - la\_stud:sex 1 3.322 1298.8 9.5709

## - la\_stud:finaid\_ever 1 5.588 1301.1 11.4249

## - nonwhite 1 6.607 1302.1 12.2582

## - credits\_entry 1 54.412 1349.9 51.3691

##

## Step: AIC=7.78

## grade ~ la\_stud + sex + nonwhite + finaid\_ever + credits\_entry +

## la\_stud:sex + la\_stud:finaid\_ever

##

## Df Sum of Sq RSS Cp

## <none> 1296.6 7.7812

## - la\_stud:sex 1 3.317 1299.9 8.4953

## - la\_stud:finaid\_ever 1 5.610 1302.2 10.3711

## - nonwhite 1 8.023 1304.6 12.3450

## - credits\_entry 1 55.584 1352.2 51.2557

##

## Call:

## lm(formula = grade ~ la\_stud + sex + nonwhite + finaid\_ever +

## credits\_entry + la\_stud:sex + la\_stud:finaid\_ever, data = LAdata[-c(6,

## 7, 10)])

##

## Coefficients:

## (Intercept) la\_stud sex

## 2.38222 0.01927 0.44604

## nonwhite finaid\_ever credits\_entry

## -0.20014 -0.40975 0.02804

## la\_stud:sex la\_stud:finaid\_ever

## -0.24186 0.31456

none = lm(grade~1, data=LAdata[-c(6, 7, 10)])

mod8 = step(none, scope = list(upper=mod4), direction='both')

## Start: AIC=299.28

## grade ~ 1

##

## Df Sum of Sq RSS AIC

## + credits\_entry 1 64.112 1347.6 251.60

## + sex 1 26.302 1385.4 281.18

## + nonwhite 1 12.199 1399.5 292.01

## + finaid\_ever 1 11.344 1400.4 292.66

## + first.gen 1 7.351 1404.4 295.70

## <none> 1411.7 299.28

## + la\_stud 1 2.234 1409.5 299.59

##

## Step: AIC=251.6

## grade ~ credits\_entry

##

## Df Sum of Sq RSS AIC

## + sex 1 20.158 1327.5 237.49

## + finaid\_ever 1 13.418 1334.2 242.90

## + nonwhite 1 11.980 1335.6 244.06

## + first.gen 1 5.499 1342.1 249.23

## <none> 1347.6 251.60

## + la\_stud 1 0.628 1347.0 253.10

## - credits\_entry 1 64.112 1411.7 299.28

##

## Step: AIC=237.49

## grade ~ credits\_entry + sex

##

## Df Sum of Sq RSS AIC

## + finaid\_ever 1 13.289 1314.2 228.73

## + nonwhite 1 11.216 1316.2 230.42

## + first.gen 1 6.512 1321.0 234.23

## <none> 1327.5 237.49

## + la\_stud 1 1.259 1326.2 238.47

## - sex 1 20.158 1347.6 251.60

## - credits\_entry 1 57.968 1385.4 281.18

##

## Step: AIC=228.73

## grade ~ credits\_entry + sex + finaid\_ever

##

## Df Sum of Sq RSS AIC

## + nonwhite 1 7.686 1306.5 224.46

## + first.gen 1 2.550 1311.6 228.66

## <none> 1314.2 228.73

## + la\_stud 1 0.876 1313.3 230.02

## - finaid\_ever 1 13.289 1327.5 237.49

## - sex 1 20.029 1334.2 242.90

## - credits\_entry 1 59.944 1374.1 274.42

##

## Step: AIC=224.46

## grade ~ credits\_entry + sex + finaid\_ever + nonwhite

##

## Df Sum of Sq RSS AIC

## <none> 1306.5 224.46

## + first.gen 1 1.157 1305.3 225.52

## + la\_stud 1 1.079 1305.4 225.58

## - nonwhite 1 7.686 1314.2 228.73

## - finaid\_ever 1 9.759 1316.2 230.42

## - sex 1 19.404 1325.9 238.22

## - credits\_entry 1 59.589 1366.1 270.14

# Unit 6: ANOVA

## STOR 455 Class 34 R One Way ANOVA

library(readr)  
  
Exams4 <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/Exams4.csv")  
  
Exams4

## # A tibble: 20 x 3  
## Student Exam Grade  
## <chr> <dbl> <dbl>  
## 1 Barb 1 62  
## 2 Barb 2 87  
## 3 Barb 3 74  
## 4 Barb 4 77  
## 5 Betsy 1 94  
## 6 Betsy 2 95  
## 7 Betsy 3 86  
## 8 Betsy 4 89  
## 9 Bill 1 68  
## 10 Bill 2 93  
## 11 Bill 3 82  
## 12 Bill 4 73  
## 13 Bob 1 86  
## 14 Bob 2 97  
## 15 Bob 3 70  
## 16 Bob 4 79  
## 17 Bud 1 50  
## 18 Bud 2 63  
## 19 Bud 3 28  
## 20 Bud 4 47

*Investigating relationships* 1. Y = Quantitative, X = Categorical - “dummy” variable and regression 2. Y = Binary Categorical, X = Quantitative - logistic Regression 3. Difference in two means - 2 sample t test 4. Difference in more than 2 means - ANOVA for Means

Goal TOday: See if there is evidence Many difference groups from same population and see if they have evidence that the means are difference btween the groups

*Samples from K Different Groups* Test: Ho: mu1 = mu2 = mu k Ha: some mui =/= muj

For each row is a different student

Is there a sig difference in average grade among the four exams?

Pull in the data

means = tapply(Exams4$Grade, Exams4$Exam, mean)  
  
tapply(Exams4$Grade, Exams4$Exam, mean)

## 1 2 3 4   
## 72 87 68 73

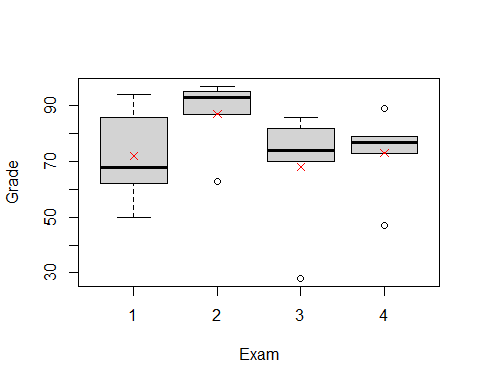
tapply(Exams4$Grade, Exams4$Exam, sd)

## 1 2 3 4   
## 17.88854 13.92839 23.23790 15.68439

tapply(Exams4$Grade, Exams4$Exam, length)

## 1 2 3 4   
## 5 5 5 5

boxplot(Grade ~ Exam, data = Exams4)  
points(means, col="red", pch=4)

 We want think about two completing model s to describe teh situtaion

maybe if i just looked at teh mean overall score, just added them all up = the scores in teh population

completing Ha: Instead, I should see each mu as based on teh exam score for the people who took that exam; just use the mean from each group individually

Note: The groups in teh mu in the group are going to be better overall, but we want to see if it’s a stat sig difference to see if we want to use either one.

*Predicting in ANOVA Model* - If the group means are the same (Ho); if H0 is true - yhat = ybar; for all groups > residual = y - ybar

* If the group means can be different (Ha); if Ha is true
  + yhat = ybar for the ith group = residual = y-ybar
* Do we do sig better with separate means?
  + Compare the sums of squared residuals
  + SSTotal = sum(y-ybar)^2 vs
  + SSE = sum(y-ybari)^2 (how different they are from the new model)

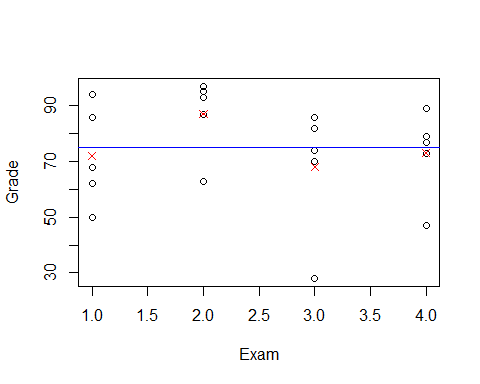
The SSE should be small = when we look at inviduval groups = support Ha

*Partitioning Variability* Data = Model + Error Y = mu i + E Total varation in response, Y = variation explained by MODEL, mui + Unexplained variation in RESIDUALS

Y=Mui +E (y-ybar) = (ybari - ybar) + (y-ybari) Sum(line 77)^2 each thing in parenthesis

SSTOTAL = SSGROUPS + SSE

plot(Grade ~ Exam, data = Exams4)  
points(means, col="red", pch=4)  
abline(h = mean(Exams4$Grade), col = "blue")



The above plots points for grades by exam; we are trying to see fi we can pair each group to the red xs and get a better line than comparing it to the horizontial line

If SSE = big, it tells us that this model may not be some much better than the null model; so we might not want to use the Ha.

We have to factor exam below so that it looks at thte groups, otherwise it will assume some kind of relationship in the exams category

Beklow we see that the sum of squares have a lot

Below the bottom row is so much bigger, so we are not havcing a lot of variability explained by the model

amodG=aov(Grade~factor(Exam),data=Exams4)  
summary(amodG)

## Df Sum Sq Mean Sq F value Pr(>F)  
## factor(Exam) 3 1030 343.3 1.056 0.395  
## Residuals 16 5200 325.0

*Alternate Form: ANOVA MOdel For Means*

Y = mu + aplhai+ E Y = grand mean + effect for ith group + Random Error

mu-hat = y-bar alpha-hat = y-bari - y-bar

Hypothesis Testing: Ho: mu1 = mu2 = muk Ha: some Mui =/= muj

Above Hypo test is interchangable with Ho: alpha1 = alha2 - alphak =0 Ha: Some alphai =/= 0

We think about the effect of the difference. This is how we have phrased things for regression in teh past; so it’s like how we did regression; this is just a different way to do the same thing

We will do more pairwise comparisions in the future classes

*This about in a simualation persepctive* Going to StatKey website

If all of the exam scores are equal, then it wont matter where they are assigned. When you do a random; ANOVA for a difference in means; it randomly scatteres the values to see if the relation is stat sig. We are taking the mean of grades and splitting it up by exam

The ANOVA table in StatKey; looking at teh SSE - how far away each of those values are from their group means squared adn summed

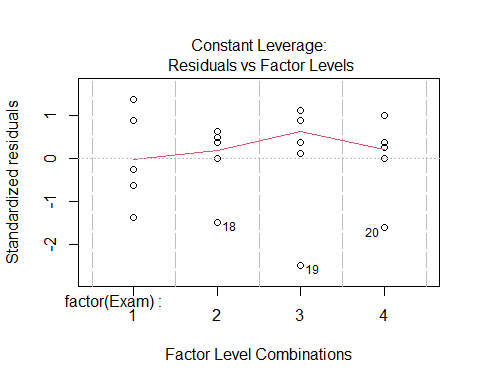
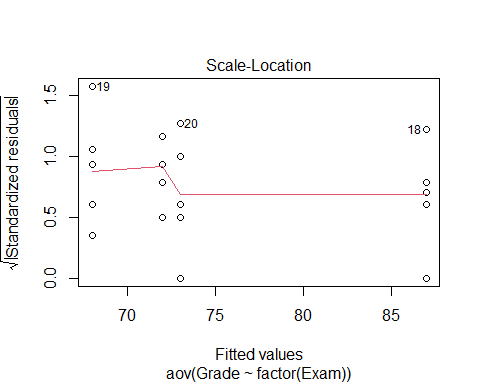
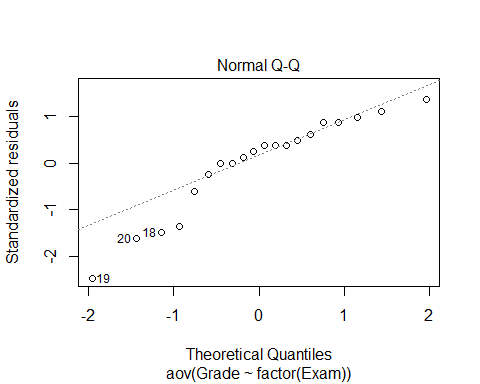
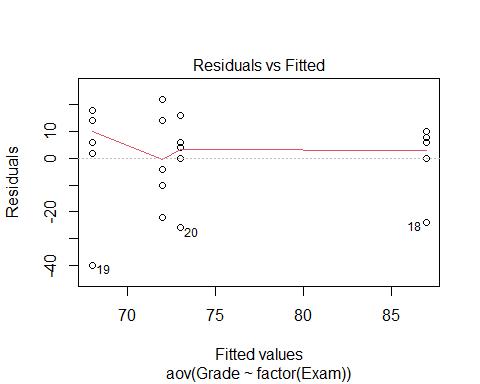
Total - how far are away from total mean, squared adn sum We cna see there are a lot of numbers in teh SSE term, which means that there isn’t a very sig relationship here.

MSE = Sum of squares/dof

F test = MSGroup/MSE

*Checking Conditions for ANOVA* E~N(0,stdE) <- check with residuals 1. Zero Mean: Always holds for sample residuals 2. Constant Variance: Plot residuals vs fits and/or compare std. dev’s og groups (Check if some other groups si is more than two another) 3. Normaility: histogram/normal plot of residuals 4. Indepdnence: Pay attention to data colelction

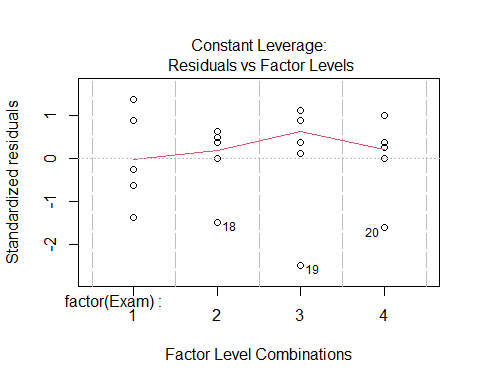
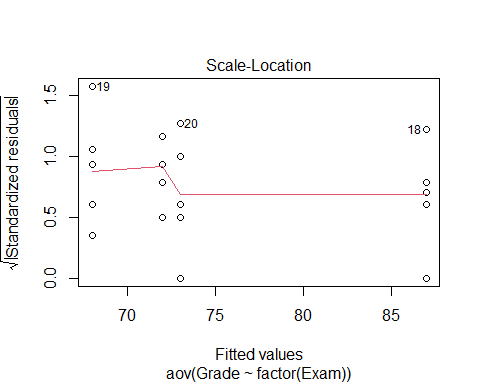
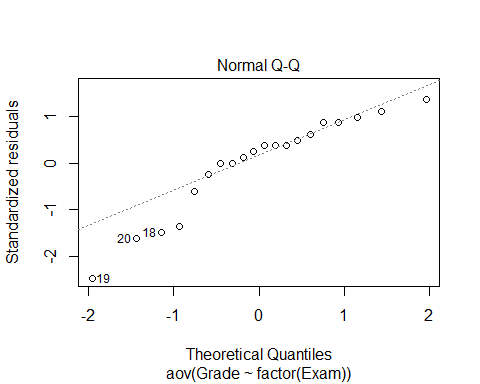
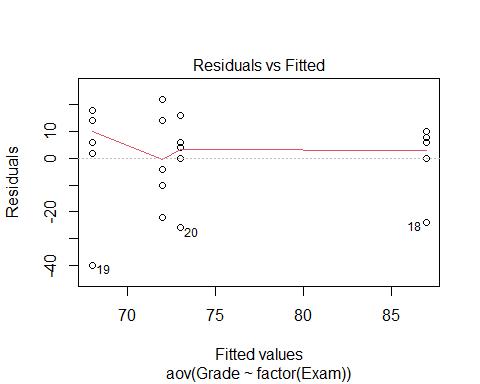
plot(amodG)



round(tapply(Exams4$Grade,Exams4$Exam,sd),2)

## 1 2 3 4   
## 17.89 13.93 23.24 15.68

plot(amodG)



round(tapply(Exams4$Grade, Exams4$Exam, sd),2)

## 1 2 3 4   
## 17.89 13.93 23.24 15.68

What about the other direction?

Is there a sig dif in average grade among the 5 students? construct a model by student

There is a stat sig relationship we ahve evidence to say at least 1 is stat sig

but where are those differences?

Bud is probably lower than the rest, but do the other 5 students have smaller things?

There are a lot of different ways to look at this

amodS=aov(Grade~Student,data=Exams4)  
summary(amodS)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Student 4 4480 1120.0 9.6 0.000468 \*\*\*  
## Residuals 15 1750 116.7   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

NOw that we know there is a difference, hwere is the difference?

N groups \* (n-1)/2 = how many comparisions we have to do

In this example we have 5 students, so 5\*4/2 = 10 combinations

If we do lots of test, we might run into an error; if we are saying then

If an outcome would happen only 5% if the null is true, we’re going to reject it and say teh alternative is suported instead. Well, that could happen, but we might do a type I error; we don’t know that we a re doing a type I error though,

Each time there is a change of making at ype 1 error when you test a combination; each time there is a 5% chance of making a type 1 error when you are working with an alpha level of 0.05.

There is a 95% chance no type 1 Error; so 0.95^20 = chance of making no type 1 errros in a test that has 20 combinations; We have a 0.358 chance of making no errors; we have a 0.6415 chance of making AT LEAST 1 TYPE 1 ERROR.

This is an issue; we need to look at how can we decrease the chance of mkaing type I errors.

*Ways to possibly fix this* a) do a few preplanned comparisions - we cna’t always just test teh difference because we are cheating and we cant really find wehre other differences are b) Adjust teh sig level used for each test - make the sig level higher; so you have a decreased chance of mkaing a type I error; with 1% sig level, so we have a 0.99 chance of NOT makign a TYPE 1 Erric

So we would still have a chance of getting a type 1 Error, but ti woudl be lower.

*Above* Shows one way to show no sig dif vt exames and sig dif bt students but: *Question*: Can we use both factors to help explain teh variability int eh exam scores?

*Looking forwards* - will look at teh ANOVA model for dif in means with 1 predi - how to dod comparisions if there are differenecnes in data - all have difference ways of makign errors and how to avoid it

## STOR 455 Class 35 One Way ANOVA (Again)

Homework 8 comments - you have to pick 6 dif models of car and use a random sample selection of 50 cars - choose these cars that are from jap or us companies; they are not part of the sample, but you have the car names - you’ll have to deduce which are jap vs us cars; do based on what they car type name is; think about the company - choose 3 jap cars that are a compact, midsize and suv - you have to find this yourself - Goal: take samples of 50 and find the compact, midsize, and suv for - use rbind like HW 4 to put all in the same dataframe (left join) - make a new variable - a bit of a setup - reference hw 2 and 4

library(readr)  
  
Exams4 <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/Exams4.csv")  
  
Diet <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/Diet.csv")

Last time talked about anova for a difference in k mean - if we split it up into groups and classify the mean for the group that it is in, does that explain a significant amount more of the variability? Aka: Should we classify things by their group mean or the overall mean?

*Hypo test* Ho: mu1 = mu2=…muk Ha: some mui != muj

We looked to see if there is sig dif in the average grade among the four exams? Or are teh results by chance?

*Partiioning Variability* Data = Model + Error Y = mui + E total varaition in the response, Y = variation explained by MODEL + Unexplained variation in Residuals

Question: Does the model explain a sig amount of the total variability?

We want to know if we should

the error term is what variabiltiy there is grom each group mean

*Partitioning Variability ANOVA for Group Means* Y = muk + E SSTotal = SSGroups + SSE

The SSTotal = the SSdifferences from the overal gran mean SSE = sum of the difference from each indiivdual group mean SSGroup = the variation explained by the groups that we have imposed on this model

If we have sig evi of the diff bt the means, then the model will have SSE = small compared to the SSGroup; b/c the SSGroup tells us if a lot of the variability has been explained by using the group means

*Below is the model we made last class*

Ho: all exam means are equal in population Ha: at least one exam mean is different

The pvalue is big, so there is no evidece for sig dif

**Below:** IF we sum these, then we get the total SS; SSTotal; if we have a good model, then we would want a lot of variability explained by our groups factor(Exam) and we see there is a lot in the residuals, which is below factor(Exam) - this is why we have a large p value

amodG=aov(Grade~factor(Exam),data=Exams4)  
summary(amodG)

## Df Sum Sq Mean Sq F value Pr(>F)  
## factor(Exam) 3 1030 343.3 1.056 0.395  
## Residuals 16 5200 325.0

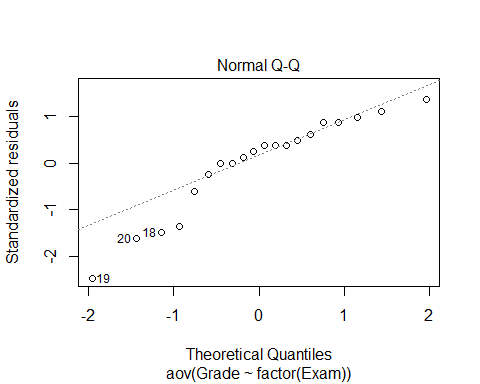
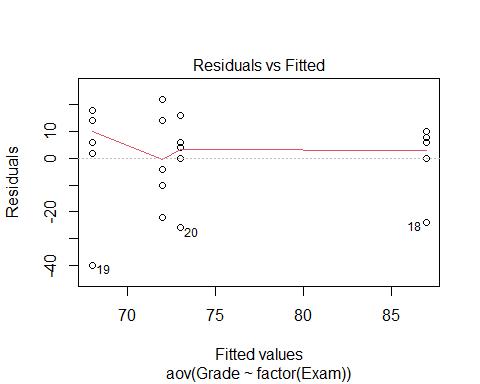
*Checking conditions for ANOVA* *Mostly have to focus on constant variance and normality* -Zero mean: Always holds for sample residuals - Constant Variance = plot residuals vs fits and/or compare std dev.’s of group (check if some groups is is more than twice another) *can be hard to check since we ar looking at groups; and sometimes its just teh mean values; we want to see a similar spread between the groups; can also think that as long as the largest standard deviation is not double the smallest standard deviation, then we ar egood enough for constant variance.* (Think more about the data, because if you have a really small dataset, you might have to throw that idea out the window) - normality: histogram/normal QQplot of residuals - Independence: Pay attention to how the data was collected

**Below: Checking Residuals** This amount of devation is not a problem

**Below: We check to see wht the std are of the functions**

This will split it by categories.

plot(amodG, 1:2)



round(tapply(Exams4$Grade,Exams4$Exam,sd),2)

## 1 2 3 4   
## 17.89 13.93 23.24 15.68

*Question* Is there sig dif in the average grade among the five students? Is it a sig dif?

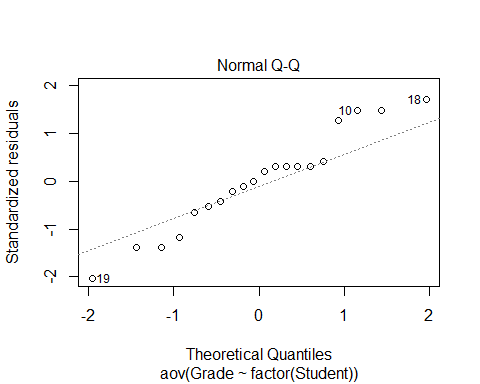
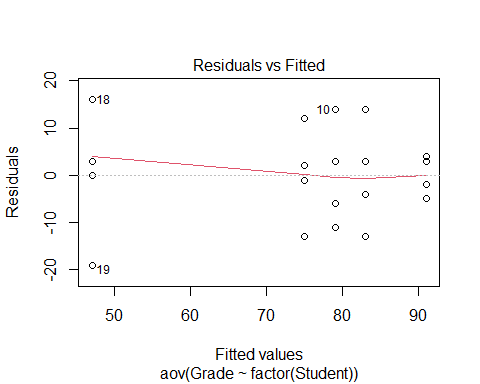
We can build the same model, but just change the group by student level instead of grade. we have evidence to say there is a sig dif.

This plot tells us where each persono’s scores are plotted in relation to each other

modS=aov(Grade~factor(Student),data=Exams4)  
summary(modS)

## Df Sum Sq Mean Sq F value Pr(>F)   
## factor(Student) 4 4480 1120.0 9.6 0.000468 \*\*\*  
## Residuals 15 1750 116.7   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

plot(modS, 1:2)



round(tapply(Exams4$Grade,Exams4$Student,sd),2)

## Barb Betsy Bill Bob Bud   
## 10.30 4.24 10.98 11.40 14.45

*How many comparisions?* Which one of the indivates caused the sig difference?

There could be a lot of comparisions to do.

*Problem of Multiplicity* When we are doing many pairwaise comparisions we are more likely to make a Tpye I Effor (finding a false difference)

**Possible fixes:** - Do only a few pre planned comparisons – If we know what we want to focus on, then we can do that - Adjust the sig level used for each test

* THe more testes we do, the more like we are th make a type 1 error = null is true, but we find evidence to reject it

*Example:* If we did tests at a .05 sig level, then there is a 5% chance that we will make a type 1 error; so there is a 95% chance that we won’t make a type 1 error.

We have to find out how we can minimize the chance of making a tpye 1 error.

# The probability of not making a type 1 error if you run 10 tests  
(1 - 0.05)^10

## [1] 0.5987369

# Chance that we make at least 1 type 1 error when we do 10 tests  
1 - (1 - 0.05)^10

## [1] 0.4012631

# The probability of not making a type 1 error if you run 100 tests   
(1 - 0.05)^100

## [1] 0.005920529

# Chance that we make at least 1 type 1 error when we do 100 tests  
1 - (1 - 0.05)^100

## [1] 0.9940795

*Pairwise comparision AFTER ANOVA* This is where you do every single test on the data

Compute a C.I for mui - muj

Pairwaise t-test for difference in means Ho: mui = muj Ha: mui != muj

*Modifications* - estimate any std with sqrt(MSE) = Se - Use teh error d.f for any t-values

We assumed the null came from a normal population.

Rather than using these ind groups to see the std, you can use the pooled value of all of the groups; we can estim the population std using the std of the residuals in teh model

*Pairwise Inference After ANOVA* C.I for mui - muj = (ybari - ybarj) +/- t-star*sqrt(MSE)*sqrt((1/ni)+(1/nj)) #Note: use the error d.f for the t-star in both the C.I and Test # If 0 is in the C.I then that would give me evidence for a stat sig dif. # The hypothesis testing tells you how unlikely it would be to get that result

Test: Ho: mui = muj Ho: mui != muj

t = (ybari-ybarj)/(sqrt(MSE)\*sqrt((1/ni)+(1/nj)))

**T-Star tells you how many standard errors i need to go in each direction depending on the sampel size; gives you an approximation for what you think the population difference is**

* Bigger sample, fewer in each direction = trend towards a normal distribution
* MSE = standard devation of the residuals and it is weighted based on the size of the groups
* If zero is in the C.I, then we have evidence

**T-Test Stat** **Is what the actual difference is divided by that measure of the standard deviation scales in this way. Tells me how many standard deviations away from the null this data is**

*Fisher’s LSD* The least sig difference: adding or subing to those differences to see how much variability there is

LSD = t-star*sqrt(MSE)*sqrt((1/ni)+(1/nj)) # concludes mui differs from uj if abs(ybari-ybarj)>LSD

*If looking at the t-distribution* The thing is centered at zero and we want to know how far in each direction we need to go to get the middle some percent of the data (whatever level you’re doing the sig test at); 95% = 0.05 sig level

We want to figure out what t-star is, because it will tell me how many i need to go in each direction. We want the value of t-star is 97.5% of the data under that curve to the left of it. **Look below, this is where we do the code for this**

# qt, give it an area under the curve, under the t-dist; and it will tell me the t value that is the bound for the left of it.   
  
#We also need df, and it's coming back from the anova model we made. We want the residuals df. We have 5 groups, because 5 students. This just gives us an idea of how much we trust the variability of the sample and how well it can predict the population   
  
# Big df = big sample, so the sample = more like the population   
# Small sample sizes = variability might not predict population as well   
  
qt(1-0.05/2, 15) #Tells you if you go OUTPUT number of standard deviations in each direction of this distribution, then I will capture the middle 95% of the data.

## [1] 2.13145

t\_LSD = qt(1 - 0.05/2, modS$df.residual)  
  
#WE want to make a prediction for the difference in population of the two groups, so we need to go 2.xxx standard deviations in each direction; we will calc the std by the sqrt(MSE)\*sqrt((1/ni)+(1/nj)) ((This is the last half of the LSD model))  
  
MSE = summary(modS)[[1]][,3][2] # This looks at the anova table output from the summary, that's what teh 1 is; the 3 gives me the third column and all rows; the 2 gives the second element of the 3rd column is what we are pulling out  
  
LSD = t\_LSD \* sqrt(MSE)\*sqrt(1/4 + 1/4) # it is 4 because each student took 4 exams   
LSD

## [1] 16.27921

*Interpreting the results above:* If we do each test basically at the 5% sig level, if any of your group means are more than this amount different, we have sig evidence to say there is a difference there;

we are predicting the mean s bt any two groups are this much or less by chance

*ANOVA for Grades vs Students* - Every compariasion we make; everyone’s score ocmpared with bud’s are more than 16.33 apart; so we have evideicent to say barb and bettey are different and others as well.

* THe multiplicity is an issue here; We are doing these tests at the 0.05 level, there are chances of making a type 1 error. This is a small example, but as it gets bigger it could be an error

*The Problem of Multiplicity* - When doing many pairwise comparisions there is an increased change to make a type 1 error (find a false difference) - Fisher’s LSD may be too lenient

*Possible fizes* - Do only a few preplaed comparsions - Use a smaller alpha for each test

OR

**Bonferroni Adjustment** When doing m tests with a *overall* error rate of alpha-star, use alpha = alpha-star/m for each test

* *This adjustment, basically lowering the levels of each one of the individual tests. So the chance of making a tpye 1 error for all the tests together can be defined and then from there you can figure out how twhat a sig level you need for each individual test*
* Bonferroni is going to let us do everything at the 0.05 level, and not a bigger aggregate of this **That’s really useful**
* IT does this by reducing the smaller alpha for each test. Reduce it by dividing the sig level by whatever number of comparisions you want to do.
* If I want the overall sig to be 0.05 I need to divide that by 10 (because there are 10 comparisons), then each test will be done at the 0.005 level.

#Bonferroni Method  
# No type 1 errors  
(1 - 0.05/10)^10

## [1] 0.9511101

# At least 1 type 1 error  
1 - (1 - 0.05/10)^10

## [1] 0.04888987

All that really changes, are the number of std that you have to go to trap that middle value gets bigger

We have to go much wider to track Bonferroni effect; and by going wider, we ar emaking it less likely that we make a type 1 error, but we are also demanding that we have stronger evidence of a difference to see that sig difference

*ANOVA for Grades vs Students*

*How to calculate the Bonferroni t-value* See below:

We get a value of 25, so using this method, we have to be 25 difference in term of average score to show some sig dif bt groups

The results are the same because of our data, but for other examples there may be a different conclusion

**Things to not do** - DOnt p-value hack; don’t choose what fits your narrative

t\_bf = qt(1 - 0.05/10/2, modS$df.residual) # What changed is the division by 10 ebcause of the total number of comparisions we wnat to do   
  
BSD = t\_bf \* sqrt(MSE)\*sqrt(1/4 + 1/4)  
BSD

## [1] 25.09753

*Tukey’s HSD: Honestly Sig Dif* *Replace t-star with value of q-star from teh studentized range distribution with R*

HSD = (q-star/sqrt(2))*sqrt(MSE)*sqrt((1/ni)+(1/nj))

q-star depends on slpha, # groups = K, and error df

tukey’s test is a bit more in the middle when it comes up with getting the dif values

* trying to get some middle group where we reduce the chance of a type 1 error, but not reducing it as much as the last case where we are inflating the case of making a type 2 error.
* the only change you see here is how many standard deivations you need to go in each difrection; it’s somewhere no quite in the middle; but its not as big as the last value
* Tukey’s is more the middle ground for what we would expect to see by chance

HSD = qtukey(1-0.05, 5, modS$df.residual) \* sqrt(MSE)\*sqrt(1/4 + 1/4)

*How can this be automated better by R?* *Automating the LSD and Bonferroni in R*

* Som eof the pvalues have changed

**LSD Method:**

pairwise.t.test(Exams4$Grade, Exams4$Student, p.adj = 'none') # What this is doing is doing each individual pairs of hypothesis test. It's not calcuing the conf int. It's teh same idea; where if the pvalue is below 0.05, then the dif bt these must be bigger than that 16.3 or something in the previous calculations.

##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: Exams4$Grade and Exams4$Student   
##   
## Barb Betsy Bill Bob   
## Betsy 0.05357 - - -   
## Bill 0.60812 0.13699 - -   
## Bob 0.31148 0.31148 0.60812 -   
## Bud 0.00229 3.8e-05 0.00079 0.00028  
##   
## P value adjustment method: none

# We see where the differences are in this table   
# We see what we saw with the confidence intervals, where one student has a sig difference, but others dont.   
# The table isn't filled out because where the dashes are, we already have those valeus; we have already compared bob and bill, but bill and bob would give us the same number

**Bonferroni Method in R**

pairwise.t.test(Exams4$Grade, Exams4$Student, p.adj = 'bonf')

##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: Exams4$Grade and Exams4$Student   
##   
## Barb Betsy Bill Bob   
## Betsy 0.53567 - - -   
## Bill 1.00000 1.00000 - -   
## Bob 1.00000 1.00000 1.00000 -   
## Bud 0.02293 0.00038 0.00789 0.00277  
##   
## P value adjustment method: bonferroni

# Similar, but the bonferronia adjustment   
# A little differently done by R;   
# When we made the C.I, we adj the sig level of test based on the number of groups we had; we lowered teh sig level, which made me needs a lower pvalue to find a difference  
# R's difference: it's calc the pvalues in the same way for the LSD method, then multiplying that by the number of comparisions; so rather than compariting the orignal pvalue to the smaller sig level, it's comparing an inflated pvalue to that same sig level   
# Mathmatically its teh same thing, and you can se it here   
  
# The 1st for the pvalues show it is mult the last case by 10 and if its 1, that's the upper limit because you cant have a prob higher than that   
  
# THis compares each of the pvalues to the same 0.05 level since they are all being multiplyed by 10  
  
# Bud is the big difference dude here again

*Tukey Function* Does the middle ground of making type one and type 2 errors;

diff = diff in mean scores for each pair - lower and upper = 95% conf int using the Tukey’s distribution - Trying to see if 0 in this distribution or not, if 0 is in it, we are predicting that there could be no difference in teh population, if 0 is not in it, then we’re - if there is no 0 invloved, then there may be a sig dif bt the two

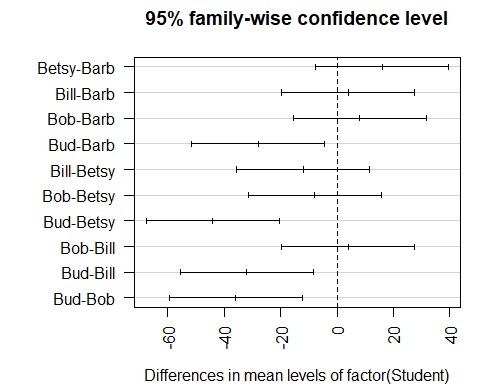
* we could look at the pvalue, which is being inflated; we compare then to the 0.05 level
* the results are the same

TukeyHSD(modS)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Grade ~ factor(Student), data = Exams4)  
##   
## $`factor(Student)`  
## diff lwr upr p adj  
## Betsy-Barb 16 -7.584413 39.584413 0.2720310  
## Bill-Barb 4 -19.584413 27.584413 0.9835150  
## Bob-Barb 8 -15.584413 31.584413 0.8295529  
## Bud-Barb -28 -51.584413 -4.415587 0.0166293  
## Bill-Betsy -12 -35.584413 11.584413 0.5360462  
## Bob-Betsy -8 -31.584413 15.584413 0.8295529  
## Bud-Betsy -44 -67.584413 -20.415587 0.0003116  
## Bob-Bill 4 -19.584413 27.584413 0.9835150  
## Bud-Bill -32 -55.584413 -8.415587 0.0060225  
## Bud-Bob -36 -59.584413 -12.415587 0.0021941

*Added Code* - Visual representation of this - We can plot all the hist on the same axis, and if any cross the vertial 0 line, then there is not sig dif there; - if they dont cross through it, then thats where we see those differences.

origpar = par()  
par(mar=c(4,7,3,1))  
hsd = TukeyHSD(modS)  
plot(hsd, las=2)



par(origpar)

## STOR 455 Class 36 Two Way ANOVA

library(readr)  
  
Exams4 <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/Exams4.csv")  
  
Diet <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/Diet.csv")  
  
Glue <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/Glue.csv")

*Diet Information* - Investigate the data to determine the impact of the type of diet and Sex on mean weight loss in separate models. - Evaluate the residuals for the appropriateness of using the ANOVA model in each model - Use pairwise methods to investigate differences between means if the overall ANOVA model is significant.

**Goal:** - Want to see the impact of each diet and how the diet and sex effects teh mean weight loss - First model: ANOVA MOdel by WEight loss by Sex - SEcond: Anova by WEight loss by Diet type

* Want to look at residual plot to check the conditions and if we do see there are sig differences by sex or diet, then we want to follow up and do a pairwise comparision like we did last time to see where those differences are.

head(Diet)

## # A tibble: 6 x 7  
## Person Sex Age Height Preweight Diet weight6weeks  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 25 NA 41 171 60 2 60   
## 2 26 NA 32 174 103 2 103   
## 3 1 0 22 159 58 1 54.2  
## 4 2 0 46 192 60 1 54   
## 5 3 0 55 170 64 1 63.3  
## 6 4 0 33 171 64 1 61.1

There is no weight change variable so we will need to make a variable called weight change and subtract teh differences in weight begining and afterwards.

* Then look back, make an anova model for weight change by diet and then follow through the process of checking the conditions and see wehre teh sig difference with teh pairwise analysis if tehy exists
* Repeat for Sex
* Then we will talk about how to do one test that does the two things above twice

**THE DIET INFORMATION IS NOT ALREADY WRITTEN IN THE RMARKDOWN FILE, YOU WILL NEED TO WRITE THIS FROM THE SLIDE** *Diet Data Conclusions* **Question1:** Are there dif in weight change depending on the diet you are on and if there are, which diet(s) seem to be sig different than each other?

*There are changes* - **Best diet:** The 3rd one – They lost about on average 2kg more and there are sig dif – Practically this doesnt mean its better. Maybe it’s more miserable to be on than the others – Maybe it’s not worth it, but from a stats standpoint, there is a difference there

**Diet Code** - First: Make a new var of the dataframe = weight change - Second: Make a model and factor the diet because we want to see the 3 separate diets - Third: Look at teh summary of that

Ho: Mean weight change for each diet is equal Ha: At least one mean weight change for a diet is differnt

Pvalue: small, tells us

Conclusion: We have sig eviduce that the diets are not equal in the population, so we have evidence to support the alternative

*We should also check that it fits the conditions that we need, so we look at the plots of residuals by fitted and QQNorm* - NOrmality: Seems pretty normal - Residuals by fitted - constant variance; the spread of each of the diets are the same; nothing here stands out; what would stand out is if one diet was tightly packed adn the others were spread out. –This looks different than normal because it’s clustered by diet –We could have also gotten the sd of teh diets and seen if one was bigger than the others; –**Rule of thumb: If the biggest one is more than double the smallest, there may be some issues**

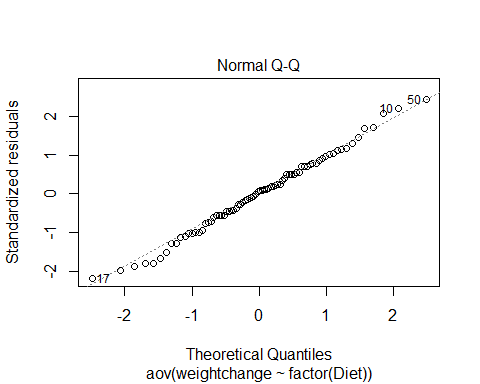
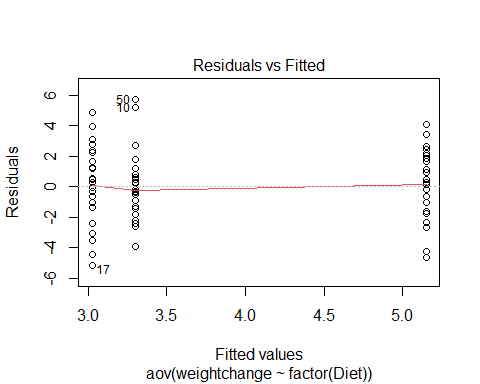
**To see where the differences are, look at the Tukey analysis** *Tukey’s* Gives us middle ground *Bonferroni* Increases chance of type 2 error, so we would need more evidence, which is why we don’t do that one. You can in the future if you ever need it though.

*Tukey’s output* - Inflated p-value: There are two places were the pvlaue is below a point 0.05 level and that’s wehre we are comparing diet 3:1 and 3:2; – Sig dif exist bt the mean weight change bt 3:1 and 3:2 –If you didn’t factor diet, Tukey would probably be mad at you and not run **Troubleshooting**

Diet$weightchange = Diet$Preweight - Diet$weight6weeks  
  
weightchange\_diet = aov(weightchange~factor(Diet), data = Diet)  
summary(weightchange\_diet)

## Df Sum Sq Mean Sq F value Pr(>F)   
## factor(Diet) 2 71.1 35.55 6.197 0.00323 \*\*  
## Residuals 75 430.2 5.74   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

plot(weightchange\_diet, 1:2)



TukeyHSD(weightchange\_diet)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = weightchange ~ factor(Diet), data = Diet)  
##   
## $`factor(Diet)`  
## diff lwr upr p adj  
## 2-1 -0.2740741 -1.8806155 1.332467 0.9124737  
## 3-1 1.8481481 0.2416067 3.454690 0.0201413  
## 3-2 2.1222222 0.5636481 3.680796 0.0047819

tapply(Diet$weightchange, Diet$Diet, mean)

## 1 2 3   
## 3.300000 3.025926 5.148148

**Additional Analysis for Sex** Not much is different; iot says two obs were delted, because there were 2 NA’s for sex that weren’t listed

Looking at teh summary: Sum of Squeares tells us little variability is explained by this model, but we have a high pvalue, so we don’t have evidence to say that Sex plays a role in the weight change

Ho: Sexes are all equal Ha: Sexes are not equal

**Looking at the plots: Does it fit the conditions?** - No need to do a Tukey follow up because we don’t see a difference between the groups, so there’s no reason to keep going with that.

*LAst two classes* We looked at the level combinations of the variables: ex: Student and grade; here there are more because sex and diet; many people of sex 1 on different diets, so there are more than 1 person in each categories.

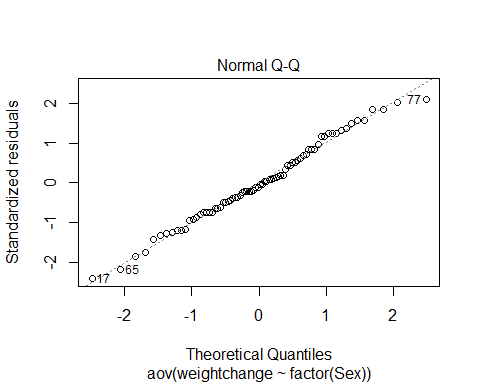
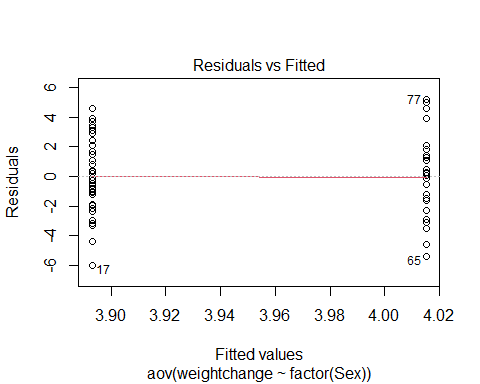
**HOw to tell if it is a good model?** - Look at teh summary to get an idea of how much variably is explained by this model compared to the total variablity(vs what is not being explained by the model is a better way to put it)

SSR, model 1: 71 w/ 430 left over in residual (unexplained) SSR, model 2: More left over unexplained, we are explaining very little of it by sex.

weightchange\_Sex <- aov(weightchange~factor(Sex), data = Diet)  
summary(weightchange\_Sex)

## Df Sum Sq Mean Sq F value Pr(>F)  
## factor(Sex) 1 0.3 0.278 0.044 0.835  
## Residuals 74 470.7 6.360   
## 2 observations deleted due to missingness

plot(weightchange\_Sex, 1:2)

 **Example: Exam Score** *We want to build a model with both Sex and the Diet* - We want to compare what is explained vs not explained - If there ar emore variables, then tehre will be less unexplained because they will explain some of the the differences that we see - When we build a bigger model, what you’re comparing (the variabilaity that is explained by that predictor too) is going to change

*We’ve shown (one-way):* *No significant differences between the exams* *Significant differences between the students*

**Question:** Can we use BOTH factors to help explain the variability in the exam scores?

**Repeat the models from last time** amodC is a different test now

**Simple Block Design**

**Define**: A simple block design has two factors with exactly one data value in each combination of the factors.

Assume: Factor A (Treatments) has K levels Factor B (Blocks) has J levels  n= K∙J data values

* Putting both model wants into one is a simpl block design; where we take 2 factors (Student and Exam) and we see which one is treatmetn or block and teh blocks (how we split them up) and treatment (apply different treatments)
* Treatmetns = levels (exams, student names, etc)
* We have one value for each person in our data

*thinking about what the model is doing* - Still thinking about the overall average of what is going one - we want to see how the different effects are on the model; what are the effects of it being this particualr student or this block that we are in (which exam or which student) - leaves us with what the differences are for each combination of the different levels of our treatment and block

**Two-way ANOVA: Main Effects Model** Y = mu + alphak + Bj+ E Y = GrandMean + effect for kth treatment + effect for jth block + Random Error

**Randomize Block - Calculations** 1. Find the mean for each treatment (row means), each block (column means), and grand mean. 2. Partition the SSTotal into three pieces: – SSTotal = SSA + SSB + SSE

SSTotal = sum(y-ybar)^2 = (n-1)s^2 SSA = row means SSB = column means SSE = SSTotal - SSA-SSB

**Randomized Block ANOVA Table** **Format:** Row(Header): [1,2],[1,3],[1,4],etc. Source: d.f, S.S., M.S, t.s., P-value Facotr A: K-1, SSA, SSA/(K-1), MSA/MSE Factor B: J-1, SSb, SSB/(J-1), MSB/MSE Error: (K-1)(J-1), SSE, SSE/(K-1)(J-1), [4,4], [4,5] Total: n-1, SSTotal

**Testing Two hypotheses** Ho: alpha1 = alpha 2=…= alphaK=0 Ha: Some alpha k != 0 Factor A: Difference in treatment means? Above, we are assuming that the factor a has no effect on the factor b

Ho: B1 = B 2=…= BJ=0 Ha: Some B j != 0 Facotr B: Difference in block means? THis second test is looking at the factor b and seeing if it has no effect on factor a; - we just swapped the two things we are testing, so we test both of the variables

* We get different results because the residuals left over will be smaller because we are combining to see which is being explained by each of these things at once

**Looking at the summary of amodA, B, and C** - WHen individually, there are no sig dif by exam, - When put together, there are sig difference

*Reason:* If you look at teh SSResid in teh Exam model, that is big, but if you look in the other it’s smaller because the var is being explained by Student - It’s not really testing the same thing; one is in a vaccum and one is in the model, but we are saying that if we are jsut looking at the unaccounted varibaility, then there are sig differences on the average scores by exam in the population.

*Conclusion* Evidence of a difference in means between exams. and Evidence of a difference in means between students.

amodA = aov(Grade~factor(Exam), data=Exams4)  
summary(amodA) #This SSR Residuals is bigger because there is alot that is not being explained by the predictor

## Df Sum Sq Mean Sq F value Pr(>F)  
## factor(Exam) 3 1030 343.3 1.056 0.395  
## Residuals 16 5200 325.0

amodB = aov(Grade~Student, data=Exams4)  
summary(amodB) #This SSR Res id smaller bc when we look at mean by student, a lot of the variability is explained; when we predict Bub's score with the other average that he did, we know more where it is coming from

## Df Sum Sq Mean Sq F value Pr(>F)   
## Student 4 4480 1120.0 9.6 0.000468 \*\*\*  
## Residuals 15 1750 116.7   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#amodC = combines amodA and amodB  
amodC = aov(Grade~factor(Exam)+Student, data=Exams4)  
summary(amodC)

## Df Sum Sq Mean Sq F value Pr(>F)   
## factor(Exam) 3 1030 343.3 5.722 0.0114 \*   
## Student 4 4480 1120.0 18.667 4.35e-05 \*\*\*  
## Residuals 12 720 60.0   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Following Up with Tukey** - This output is a series of the mult comparision with every poss comb of factor A and then every combo of factor B; - there are sig dif for each factor

*Looking at output by exam* Exam 2 and 1 have a low p value and 3 and 2 have a low p value; those are where we are saying *THERE IS SIG DIF IN THE POPULATION*; if its practical is another thing

*Looking at output by student* : simialr results,that bud has a lower average, bud and bob, bill, betsy and bard are all sig and none of the other values are because we don’t have a lot of added explained variability by having the exams

by adding the exam it didn;’t really show wehre to see diferences, but it worked the other way

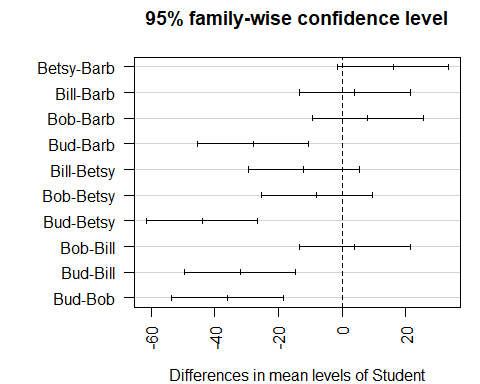
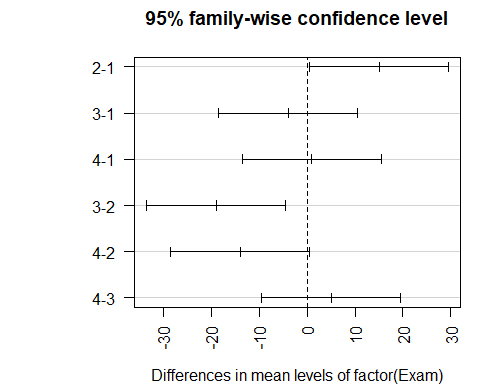
TukeyHSD(amodC)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Grade ~ factor(Exam) + Student, data = Exams4)  
##   
## $`factor(Exam)`  
## diff lwr upr p adj  
## 2-1 15 0.4554143 29.5445857 0.0425678  
## 3-1 -4 -18.5445857 10.5445857 0.8455825  
## 4-1 1 -13.5445857 15.5445857 0.9968212  
## 3-2 -19 -33.5445857 -4.4554143 0.0102092  
## 4-2 -14 -28.5445857 0.5445857 0.0605331  
## 4-3 5 -9.5445857 19.5445857 0.7409734  
##   
## $Student  
## diff lwr upr p adj  
## Betsy-Barb 16 -1.458285 33.458285 0.0782719  
## Bill-Barb 4 -13.458285 21.458285 0.9451946  
## Bob-Barb 8 -9.458285 25.458285 0.6039841  
## Bud-Barb -28 -45.458285 -10.541715 0.0019421  
## Bill-Betsy -12 -29.458285 5.458285 0.2467758  
## Bob-Betsy -8 -25.458285 9.458285 0.6039841  
## Bud-Betsy -44 -61.458285 -26.541715 0.0000293  
## Bob-Bill 4 -13.458285 21.458285 0.9451946  
## Bud-Bill -32 -49.458285 -14.541715 0.0006169  
## Bud-Bob -36 -53.458285 -18.541715 0.0002093

**Same visual from last time** - When they cross zero: means we predict there could be no difference betweent eh things - When they don’t cross zero: Visual to see wehre the differnces are that there could be a difference between tehse things

**We have been doing two way because we haev two predictors** *What about the conditions?*

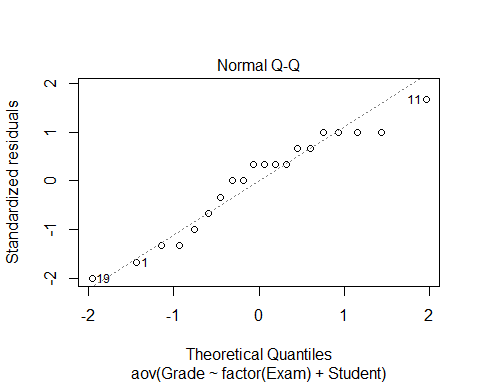
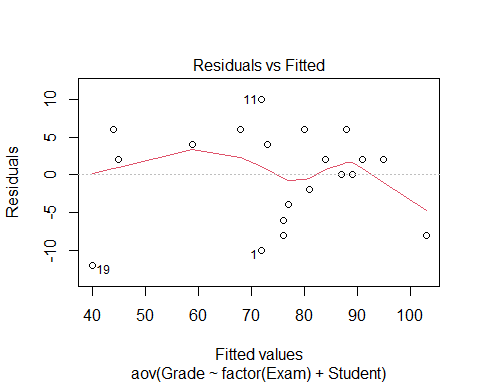
par(mar=c(4,7,3,1))  
hsd=TukeyHSD(amodC)  
plot(hsd,las=2)



par(mar=c(5,4,2,2))

*Looking at the conditions* - Constant variance: A little harder, but look at the fitted residuals, there are more columsn of data we haev to worry about here, overlal it looks good, but we can’t see much - NormalQQPlot of residuals : This is a small dataset, so nothing stands out as really bad.

plot(amodC, 1:2)

 **Looking back at the diet data** - GO in with a two way anova model Ho: Mean weights by diets are equal and themean weight changes are equal Ha: At least one is difference

Ho: Mean weight change by sex are equal Ho: Mean weight change by weight are not equal by sex

*Check COnditions* - None of them look sus

*Tukey tells us wehre changes are* - Tukey tells us that diets 3-1 and 3-2 are the different ones, - the dif bt sex, there ar eonly two here so we cant compare too mcuh - this test is giving us the same thing as before because its the same test

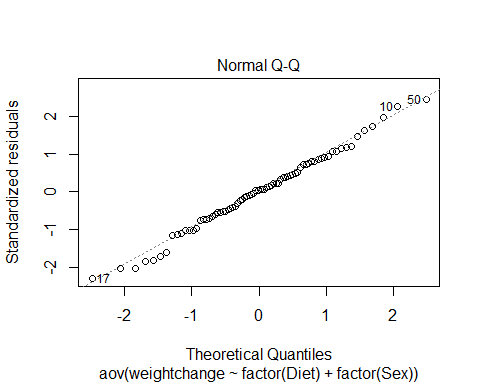
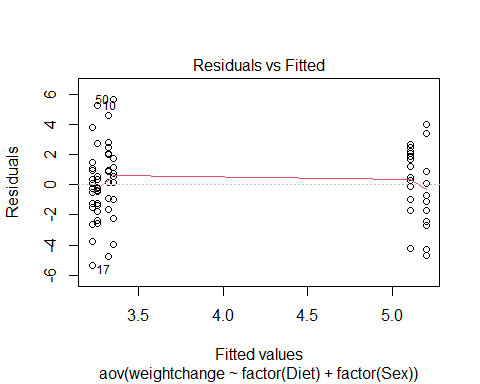
**Still need to think about: THe idea of interaction now** - We are saying that the diets there are sig differences there from weight changes and there are no sig difference by sex for weight change, but could there be interactions. I nother owords, woudl some diets work for some sexes thant he others and are there difference there?

If we look at the dif of weight change by sex and diet, could we explain more variabiltiy that’s left unexplained here.

weightchange\_both <- aov(weightchange~factor(Diet)+factor(Sex), data = Diet)  
summary(weightchange\_both)

## Df Sum Sq Mean Sq F value Pr(>F)   
## factor(Diet) 2 60.5 30.264 5.312 0.00705 \*\*  
## factor(Sex) 1 0.2 0.169 0.030 0.86387   
## Residuals 72 410.2 5.698   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 2 observations deleted due to missingness

plot(weightchange\_both, 1:2)



TukeyHSD(weightchange\_both)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = weightchange ~ factor(Diet) + factor(Sex), data = Diet)  
##   
## $`factor(Diet)`  
## diff lwr upr p adj  
## 2-1 -0.032000 -1.6644380 1.600438 0.9987875  
## 3-1 1.848148 0.2455971 3.450699 0.0198145  
## 3-2 1.880148 0.2946561 3.465640 0.0160629  
##   
## $`factor(Sex)`  
## diff lwr upr p adj  
## 1-0 0.09502529 -1.006192 1.196243 0.8639058

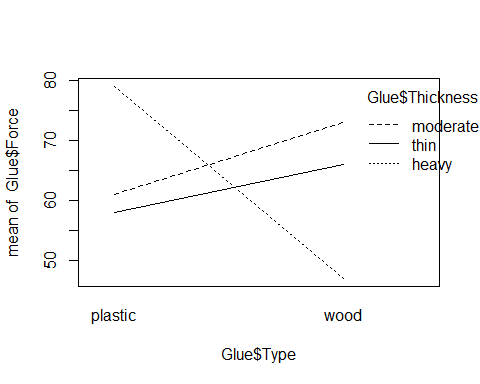
Glue\_model = aov(Force ~ Thickness + Type + Thickness\*Type, data=Glue)  
summary(Glue\_model)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Thickness 2 56 28 0.424 0.6725   
## Type 1 48 48 0.727 0.4265   
## Thickness:Type 2 1184 592 8.970 0.0157 \*  
## Residuals 6 396 66   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

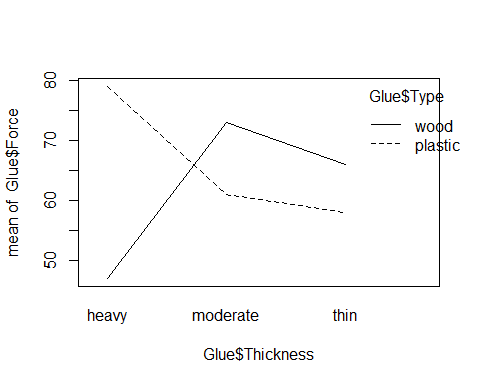
TukeyHSD(Glue\_model)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Force ~ Thickness + Type + Thickness \* Type, data = Glue)  
##   
## $Thickness  
## diff lwr upr p adj  
## moderate-heavy 4 -13.62589 21.62589 0.7743278  
## thin-heavy -1 -18.62589 16.62589 0.9834788  
## thin-moderate -5 -22.62589 12.62589 0.6768095  
##   
## $Type  
## diff lwr upr p adj  
## wood-plastic -4 -15.47703 7.477033 0.4264949  
##   
## $`Thickness:Type`  
## diff lwr upr p adj  
## moderate:plastic-heavy:plastic -18 -50.332428 14.3324278 0.3480798  
## thin:plastic-heavy:plastic -21 -53.332428 11.3324278 0.2331433  
## heavy:wood-heavy:plastic -32 -64.332428 0.3324278 0.0522177  
## moderate:wood-heavy:plastic -6 -38.332428 26.3324278 0.9689918  
## thin:wood-heavy:plastic -13 -45.332428 19.3324278 0.6259379  
## thin:plastic-moderate:plastic -3 -35.332428 29.3324278 0.9985907  
## heavy:wood-moderate:plastic -14 -46.332428 18.3324278 0.5637326  
## moderate:wood-moderate:plastic 12 -20.332428 44.3324278 0.6890028  
## thin:wood-moderate:plastic 5 -27.332428 37.3324278 0.9855637  
## heavy:wood-thin:plastic -11 -43.332428 21.3324278 0.7509689  
## moderate:wood-thin:plastic 15 -17.332428 47.3324278 0.5039318  
## thin:wood-thin:plastic 8 -24.332428 40.3324278 0.9074973  
## moderate:wood-heavy:wood 26 -6.332428 58.3324278 0.1170797  
## thin:wood-heavy:wood 19 -13.332428 51.3324278 0.3052653  
## thin:wood-moderate:wood -7 -39.332428 25.3324278 0.9431857

interaction.plot(Glue$Type, Glue$Thickness, Glue$Force)



interaction.plot(Glue$Thickness, Glue$Type, Glue$Force)



## STOR 455 - Class 37 – ANCOVA and Homogeneity of Variances

library(readr)  
  
Diet <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/Diet.csv")  
Glue <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/Glue.csv")  
Pulse <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/Pulse.csv")

**What’s an Interaction Effect?** **DEFINE:** An interaction effect occurs when a significant difference is present at a specific combination of factors.

*Example* Y=GPA Factor A = Year in School (FY, So, Jr, Se) Factor B = Major (Psych, Bio, Math)

FY is hard if alphaa1 < o (main effect) Bio is easy if B2 >0 (main effect) Jr in Math is hard if Y33 < 0 (interaction effect)

There may be times where one factor by itself are not sig; there may not be a sig diff by year or major; we may see for certain years for certain majors that is differnet

**Factorial Design** *In the past:* We hae only needed one data point that looked at each combo of the factor levels; for teh exam and student we needed one score for exam and student; here we can expand that if we only have C different values

*Future* Assume: Factor A has K levels, Factor B has J levels.

To estimate an interaction effect, we need more than one data value in each combination of factors.

Let nkj = sample size in (k,j)th cell

nkj = c

c = 1 = randomized block design c> 1 = blanaced factorial design

**Example: Glue Strength** Factor A: Thickness (thin, moderate, heavy) Factor B: Glue Type (plastic, wood) Factor B: Glue Type (plastic, wood)

Data: Plastic, Wood Thin: 52 64, 72 60 Moderate: 67 55, 78 68 Heavy: 86 72, 43 51

K = 3 j = 2 c = 2 n = 12

*Notes about the data* - They are in newtons - We could be using plastic or wood glue - How much glue we use is the other factor (Thin, Moderate, Heavy)

*Question of interest* Is there a dif in teh force reuired depending on the type of glue I use and/or a difference on how much glue I use? Is there an interaction? Is there some level of a combination that looks really different than the others and doesn’t follow the trend?

*Looking at the data* The thin and moderate force required, wood is higher until you get to heavy; a heavy amount of wood glue doesn’t seem to be as strong

*We’ll see:* there is some interaction going on between teh wood glue amount and the type of glue

# Want to try and predict the force requried to pull these things apart by the thickness and type of glue in the data and also that interaction  
Glue\_model = aov(Force ~ Thickness + Type + Thickness\*Type, data=Glue)  
summary(Glue\_model)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Thickness 2 56 28 0.424 0.6725   
## Type 1 48 48 0.727 0.4265   
## Thickness:Type 2 1184 592 8.970 0.0157 \*  
## Residuals 6 396 66   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# This gives more information than we wanted because it gives the first line as a test by thickness   
# When we look at the itneraction, there is a sig level there; the SSQ for this is higher; Factor A and B are not sig on their own, but the interation is

**Two-way ANOVA Table (with interaction)** Header: Source, df, SS, MS, ts, p-value Row1: FactorA, K-1, SSA, SSA/(K-1), MSA/MSE, BLANK Row2: FactorB, J-1, SSB, SSB/(J-1), MSB/MSE, BLANK Row3: AxB, (K-1)(J-1), SSAB, SSAB/df, MSAB/MSE, BLANK Row4: Error, JK(n-1), SSE, SE/df, BLANK, BLANK Row5: total, n-1, SSY, BLANK———>

Ho: All alphak = 0 Ha: Some alphak != 0

Ho: All Bj = 0 Ha: Some Bj != 0

Ho: All ykj = 0 Ha: Some ykj != 0

*In the table above, we do not have evidence to say that there is impact from the glue type of thickness. We don’t have evidence to reject the null, but in the third case, we do have evidence that there is a sig interaction between these variables*

TukeyHSD(Glue\_model)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Force ~ Thickness + Type + Thickness \* Type, data = Glue)  
##   
## $Thickness  
## diff lwr upr p adj  
## moderate-heavy 4 -13.62589 21.62589 0.7743278  
## thin-heavy -1 -18.62589 16.62589 0.9834788  
## thin-moderate -5 -22.62589 12.62589 0.6768095  
##   
## $Type  
## diff lwr upr p adj  
## wood-plastic -4 -15.47703 7.477033 0.4264949  
##   
## $`Thickness:Type`  
## diff lwr upr p adj  
## moderate:plastic-heavy:plastic -18 -50.332428 14.3324278 0.3480798  
## thin:plastic-heavy:plastic -21 -53.332428 11.3324278 0.2331433  
## heavy:wood-heavy:plastic -32 -64.332428 0.3324278 0.0522177  
## moderate:wood-heavy:plastic -6 -38.332428 26.3324278 0.9689918  
## thin:wood-heavy:plastic -13 -45.332428 19.3324278 0.6259379  
## thin:plastic-moderate:plastic -3 -35.332428 29.3324278 0.9985907  
## heavy:wood-moderate:plastic -14 -46.332428 18.3324278 0.5637326  
## moderate:wood-moderate:plastic 12 -20.332428 44.3324278 0.6890028  
## thin:wood-moderate:plastic 5 -27.332428 37.3324278 0.9855637  
## heavy:wood-thin:plastic -11 -43.332428 21.3324278 0.7509689  
## moderate:wood-thin:plastic 15 -17.332428 47.3324278 0.5039318  
## thin:wood-thin:plastic 8 -24.332428 40.3324278 0.9074973  
## moderate:wood-heavy:wood 26 -6.332428 58.3324278 0.1170797  
## thin:wood-heavy:wood 19 -13.332428 51.3324278 0.3052653  
## thin:wood-moderate:wood -7 -39.332428 25.3324278 0.9431857

**Interpreting Interaction** IF the two-way ANOVA indicates a significant interaction, plot the cell means vs. one factor with separate lines/symbols for the second factor

Called a Cell means plot or INteraction plot

Cell Means: Plastic, Wood Thin: 58.0, 66.0 Moderate: 61.0, 73.0 Heavy: 49.0, 47.0

interaction.plot(FactorA,FactorB,Response)

*Class NOtes* - Will look at teh itneractions between teh groups and plot them by the force required along one axis if plastic or wood, then other; see the difference through this tabl instead of having two values in each cell, they’re just averaged - why don’t we make this interaction plot factor A and factor B reponse, so type thickness force, but if we flip those teh axis will flip

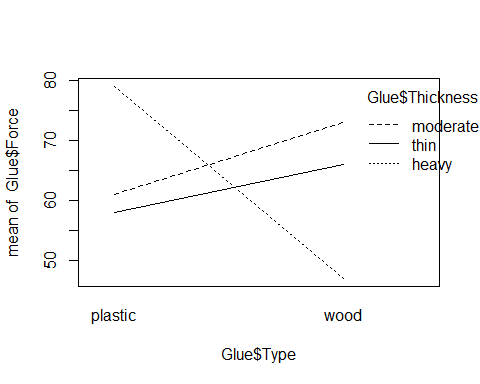
*See below for example on the flipping axises*

**Plot 1: interaction.plot(GlueThickness, Glue$Force)** - Looking at plastic and wood, then the force of the means for those - each one of the lines is the differnce amount of thickness, so wherever the poitns are that’s just the mean value - under 60, 62 or so and a bit higher. - we are looking at *are the trends as we go from plastic to wood similar for each level of thickness* – If they are similar, we expect lines t be similar in slope to eachother **The size of the data has a big impact here** - Small data is going to be skewed faster because it’s mean, so if there’s something out of wack, it’s going to show more easily since mean isn’t insoluated from outlies

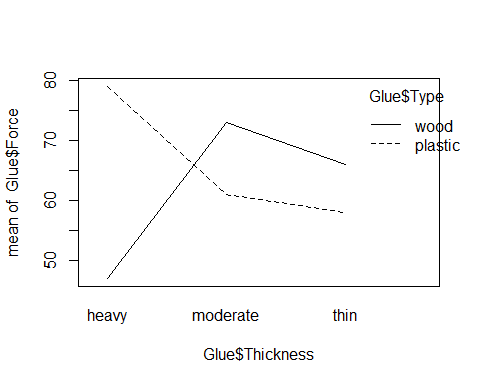
* The heavy amount of thickness is very different *There is some kind of interaction betcause heavy has a different interaction thatn the other lines*

**Plot 2: interaction.plot(GlueType, Glue$Force)** - Same relationship from a different perspective - Depending on what you’re looking at, it might look better or worse - Moderate to thin: – Those lines look similar in solpe for wood and plastic - - Heavy - Moderate the slope is different *Doesn’t show, but implied* - Going from Heavy to thin, it would be the very different relationship - Heavy to thin for one and thin to heavy to the other *Sometimes you have to assume there is no interaction to see if there is a relationship in the data*

interaction.plot(Glue$Type, Glue$Thickness, Glue$Force)



interaction.plot(Glue$Thickness, Glue$Type, Glue$Force)

 **Looking back at the Diet Data** *Make a 2-way anova model with interaction* - We made a thing by diet there were sig dif in the weight loss by which diet you were on - No sig dif by sex, - Made a two-way model to see if dif by diet and sex when in the same model, found similar results **One step further: 2-way with the interaction** *Question Interest: Are tehre any sig dif in weight loss by diet depending on which sex you are?* - Look at residuals and make these interaction plots to see where the sig difs are

*Diet Data* - Looking at the table, we have 3 different Hypo tests here

* Assuming there’s no effec ton the average weight change by diet, vs Ha: There is an effect; small p-value; – Type of diet shows sig there
* Sex has a high p-value, Ho: Assume there is no effect on average weight loss based on sex Ha: There is an effect on average weight loss based on sex

P-vale Conc: We dont have evidence to reject that

Interaction is below 0.05 level, so we are seeing a sig interaction between sex and diet, so tehre is some effect between what sex you are and which diet you are on

*Tukey* - It’s giving us 3 different outputs ’ 1. Dif by diet for all poss combos; 2. dif by sex for all combos; and 3. dif for interactions

The only real difference is in one situation; all other pvalues are huge, except for 3-0 and 1-0 there was a big difference.

**Question: What is 3-0 and what is 1-0?** *Anwser: Zero is for women, and 3 = diet 3 and 1 = diet 1; there is sig dif in teh 3rd and 1st diet for women, but we are not seeing this sig dif for men*

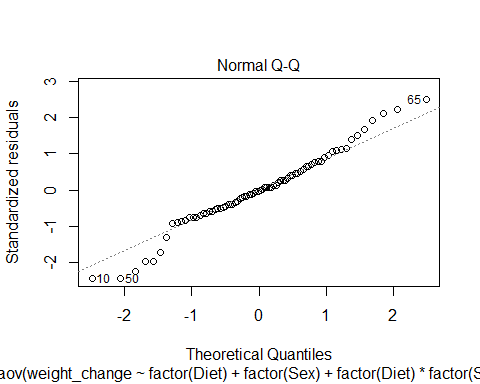
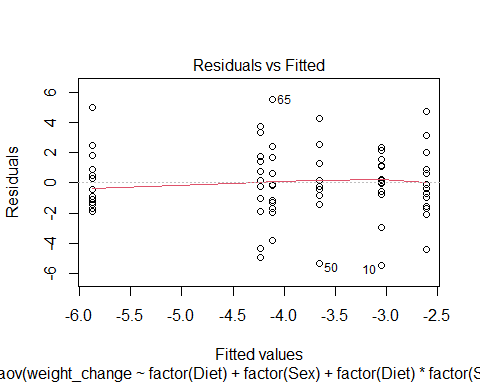
**Look at Resiudal analyss** - Nothing really stands out - Check for constant variance, they are spread similarly; one group is a little more compact, so may be an issue there, but not too bad - Normaility by QQPlot, some tail issues, but not too bad

**Interaction plot** - INteraction plot: sex on the horizontal axis: – For diet 3, it looks like we have a dif relationship than for diet 1 for each of teh sexes

Diet$weight\_change <- Diet$weight6weeks - Diet$Preweight  
  
mod1 <- aov(weight\_change ~   
 factor(Diet)+   
 factor(Sex)+  
 factor(Diet)\*factor(Sex),   
 data = Diet)  
  
summary(mod1)

## Df Sum Sq Mean Sq F value Pr(>F)   
## factor(Diet) 2 60.5 30.264 5.629 0.00541 \*\*  
## factor(Sex) 1 0.2 0.169 0.031 0.85991   
## factor(Diet):factor(Sex) 2 33.9 16.952 3.153 0.04884 \*   
## Residuals 70 376.3 5.376   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 2 observations deleted due to missingness

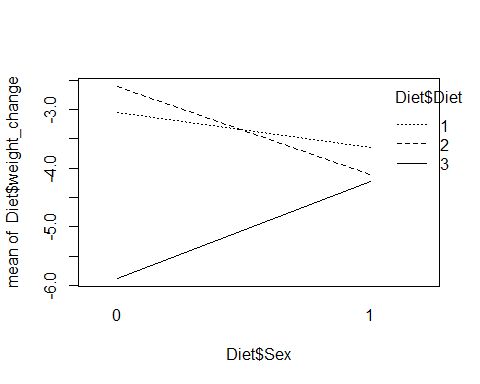
plot(mod1, 1:2)



TukeyHSD(mod1)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = weight\_change ~ factor(Diet) + factor(Sex) + factor(Diet) \* factor(Sex), data = Diet)  
##   
## $`factor(Diet)`  
## diff lwr upr p adj  
## 2-1 0.032000 -1.554658 1.6186577 0.9987150  
## 3-1 -1.848148 -3.405757 -0.2905393 0.0160200  
## 3-2 -1.880148 -3.421176 -0.3391199 0.0128543  
##   
## $`factor(Sex)`  
## diff lwr upr p adj  
## 1-0 -0.09502529 -1.16524 0.9751891 0.8599508  
##   
## $`factor(Diet):factor(Sex)`  
## diff lwr upr p adj  
## 2:0-1:0 0.4428571 -2.1250148 3.0107291 0.9958151  
## 3:0-1:0 -2.8300000 -5.3547114 -0.3052886 0.0191170  
## 1:1-1:0 -0.6000000 -3.4129628 2.2129628 0.9887997  
## 2:1-1:0 -1.0590909 -3.7964516 1.6782698 0.8656520  
## 3:1-1:0 -1.1833333 -3.8560592 1.4893925 0.7855223  
## 3:0-2:0 -3.2728571 -5.7975685 -0.7481458 0.0040103  
## 1:1-2:0 -1.0428571 -3.8558199 1.7701056 0.8852416  
## 2:1-2:0 -1.5019481 -4.2393087 1.2354126 0.5963201  
## 3:1-2:0 -1.6261905 -4.2989163 1.0465354 0.4833188  
## 1:1-3:0 2.2300000 -0.5436187 5.0036187 0.1863470  
## 2:1-3:0 1.7709091 -0.9260048 4.4678230 0.3965102  
## 3:1-3:0 1.6466667 -0.9846191 4.2779524 0.4513580  
## 2:1-1:1 -0.4590909 -3.4275816 2.5093998 0.9975014  
## 3:1-1:1 -0.5833333 -3.4923292 2.3256625 0.9915569  
## 3:1-2:1 -0.1242424 -2.9601974 2.7117126 0.9999949

interaction.plot(Diet$Sex, Diet$Diet, Diet$weight\_change)



**ANOVA via Dummy Regression** *Recall: For a single categorical factor with K levels* 1. Create K-1 indicator (dummy) predictors 2. Run regression with the dummy predictors 3. Constant estimates the mean of the reference group 4. Coefficients estimate how each other group differs 5. ANOVA tables match (depending…)

*NOtes* - We are really just looking at regression - If we wanted tot hink of these as regression models, we could do that, we just have to combine some ideas we are looking at *ANOVA MODEL FOR DIFFERENC EIN MEANS: ONE-WAY MODEL* - That’s the same as a regression model that is useing K-1 predictor variables, and each of those coeff would be some measure of how different that group is from the overall mean

*IF look at glue data, see below:* - For a math standpoint, the below models are the same, the output is just different - There is no sig evidence by dif by just thickness - we did have it in terms of an interaction - If we did this for 1,2, 2w/interaction we can build these eitherway with a linear or aov model and we are building teh same math thing

**How is this useful?** - Because this is all just regression, we can see some slightly different relationships and think about this kind of analysis of difference in means after considering teh variability of other quanatitive variables as well in the model, not just categorical - So far we have looked at how much variability is explained by the type of glue vs some other things - There might be quanatitive var we want to compare with as well

mod1 = aov(Force ~ Thickness, data=Glue)  
summary(mod1)

## Df Sum Sq Mean Sq F value Pr(>F)  
## Thickness 2 56 28.0 0.155 0.859  
## Residuals 9 1628 180.9

mod2 = lm(Force ~ Thickness, data=Glue)  
anova(mod2)

## Analysis of Variance Table  
##   
## Response: Force  
## Df Sum Sq Mean Sq F value Pr(>F)  
## Thickness 2 56 28.00 0.1548 0.8588  
## Residuals 9 1628 180.89

summary(mod2)

##   
## Call:  
## lm(formula = Force ~ Thickness, data = Glue)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.00 -10.50 0.50 9.25 23.00   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 63.000 6.725 9.368 6.14e-06 \*\*\*  
## Thicknessmoderate 4.000 9.510 0.421 0.684   
## Thicknessthin -1.000 9.510 -0.105 0.919   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 13.45 on 9 degrees of freedom  
## Multiple R-squared: 0.03325, Adjusted R-squared: -0.1816   
## F-statistic: 0.1548 on 2 and 9 DF, p-value: 0.8588

mod3 = aov(Force ~ Thickness + Type, data=Glue)  
summary(mod3)

## Df Sum Sq Mean Sq F value Pr(>F)  
## Thickness 2 56 28.0 0.142 0.870  
## Type 1 48 48.0 0.243 0.635  
## Residuals 8 1580 197.5

mod4 = lm(Force ~ Thickness + Type, data=Glue)  
anova(mod4)

## Analysis of Variance Table  
##   
## Response: Force  
## Df Sum Sq Mean Sq F value Pr(>F)  
## Thickness 2 56 28.0 0.1418 0.8700  
## Type 1 48 48.0 0.2430 0.6353  
## Residuals 8 1580 197.5

summary(mod4)

##   
## Call:  
## lm(formula = Force ~ Thickness + Type, data = Glue)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.00 -10.50 0.00 8.25 21.00   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 65.000 8.114 8.011 4.32e-05 \*\*\*  
## Thicknessmoderate 4.000 9.937 0.403 0.698   
## Thicknessthin -1.000 9.937 -0.101 0.922   
## Typewood -4.000 8.114 -0.493 0.635   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.05 on 8 degrees of freedom  
## Multiple R-squared: 0.06176, Adjusted R-squared: -0.2901   
## F-statistic: 0.1755 on 3 and 8 DF, p-value: 0.91

mod5 = aov(Force ~ Thickness + Type + Thickness\*Type, data=Glue)  
summary(mod5)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Thickness 2 56 28 0.424 0.6725   
## Type 1 48 48 0.727 0.4265   
## Thickness:Type 2 1184 592 8.970 0.0157 \*  
## Residuals 6 396 66   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

mod6 = lm(Force ~ Thickness + Type + Thickness\*Type, data=Glue)  
anova(mod6)

## Analysis of Variance Table  
##   
## Response: Force  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Thickness 2 56 28 0.4242 0.67247   
## Type 1 48 48 0.7273 0.42649   
## Thickness:Type 2 1184 592 8.9697 0.01574 \*  
## Residuals 6 396 66   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(mod6)

##   
## Call:  
## lm(formula = Force ~ Thickness + Type + Thickness \* Type, data = Glue)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7 -6 0 6 7   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 79.000 5.745 13.752 9.19e-06 \*\*\*  
## Thicknessmoderate -18.000 8.124 -2.216 0.06861 .   
## Thicknessthin -21.000 8.124 -2.585 0.04149 \*   
## Typewood -32.000 8.124 -3.939 0.00763 \*\*   
## Thicknessmoderate:Typewood 44.000 11.489 3.830 0.00866 \*\*   
## Thicknessthin:Typewood 40.000 11.489 3.482 0.01312 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.124 on 6 degrees of freedom  
## Multiple R-squared: 0.7648, Adjusted R-squared: 0.5689   
## F-statistic: 3.903 on 5 and 6 DF, p-value: 0.06386

*Go back to pulse data* - We looked at: How does exercise predict what your active heart rate was - Do we think the aveg heart rate is dif by exercise elvel

Ho: Do I see evidence of a difference? Ha: Do I not see evidence of a difference?

*Check the conditions for a linear model*

*ANOVA Table* - Tells us there is sig dif in teh average active heart rate based on what your exercise level is

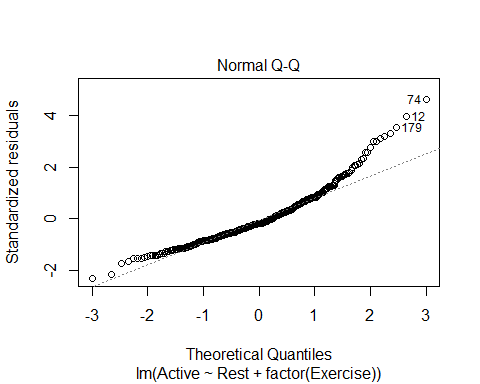
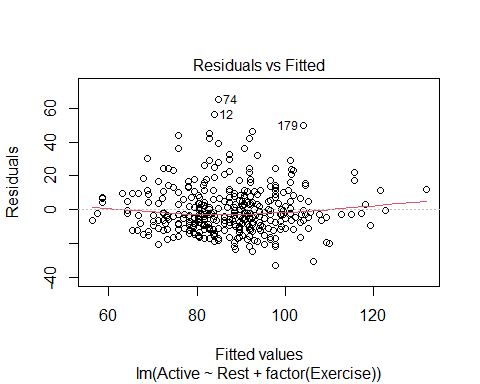
modp2 = lm(Active ~ Rest + factor(Exercise), data=Pulse)  
  
summary(modp2)

##   
## Call:  
## lm(formula = Active ~ Rest + factor(Exercise), data = Pulse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.653 -9.206 -2.629 7.231 65.073   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.25869 6.70517 1.381 0.168   
## Rest 1.15698 0.08611 13.436 <2e-16 \*\*\*  
## factor(Exercise)2 1.62128 2.15805 0.751 0.453   
## factor(Exercise)3 -0.51883 2.38266 -0.218 0.828   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.19 on 371 degrees of freedom  
## Multiple R-squared: 0.4043, Adjusted R-squared: 0.3995   
## F-statistic: 83.92 on 3 and 371 DF, p-value: < 2.2e-16

anova(modp2)

## Analysis of Variance Table  
##   
## Response: Active  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Rest 1 50342 50342 250.0302 <2e-16 \*\*\*  
## factor(Exercise) 2 351 175 0.8714 0.4192   
## Residuals 371 74699 201   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

plot(modp2, 1:2)



**Main Effects Two-way ANOVA via Dummy Regression** 1. Create indicator predictors for each factor 2. Run regression with the dummy predictors (leaving out one for each factor) 3. How to interpret the coefficients? 4. How to “recover” the two-way ANOVA table?

*NOtes* - Want to think about something a little differently *Do I still show these diferences after considering teh Varaiability explained by some over variable?* Is this sig AFTER looking at something else? - Look at the covar of the resting heart rate - We have that exercise level shows sig dif in resting heart rate on average; but does it show that after we consider the realtionship between your resting and active heart rate? *AKA, if we tell R that this is already a connection, will it still see a connection? I think that’s what this is saying*

modp2 <- lm(Active~Rest+factor(Exercise), data = Pulse)  
  
summary(modp2)

##   
## Call:  
## lm(formula = Active ~ Rest + factor(Exercise), data = Pulse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.653 -9.206 -2.629 7.231 65.073   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.25869 6.70517 1.381 0.168   
## Rest 1.15698 0.08611 13.436 <2e-16 \*\*\*  
## factor(Exercise)2 1.62128 2.15805 0.751 0.453   
## factor(Exercise)3 -0.51883 2.38266 -0.218 0.828   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.19 on 371 degrees of freedom  
## Multiple R-squared: 0.4043, Adjusted R-squared: 0.3995   
## F-statistic: 83.92 on 3 and 371 DF, p-value: < 2.2e-16

# People who exercise a small amount, we can get rid of the last two lines and it tells us our realtions bt rest and active heart reate follows that intercept and lope   
  
anova(modp2)

## Analysis of Variance Table  
##   
## Response: Active  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Rest 1 50342 50342 250.0302 <2e-16 \*\*\*  
## factor(Exercise) 2 351 175 0.8714 0.4192   
## Residuals 371 74699 201   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

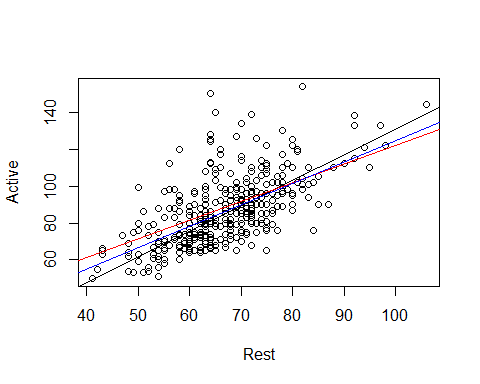
# How the active heart rate is adjected by the intercept   
# We are assuming that the resting heart rate follows teh same trend as the active, regardless of your exercise level

*Is it reasonable to make the assumption that resting heart rate follows the same trend as the active?* - What kind of relationships are there in the data if I subset by people who exercise a little, a moderate amount, and a lot?

**Two-way ANOVA w/Interaction via Dummy Regression** 1. Create indicator predictors for each factor 2. Run regression with the dummy predictors (leaving out one for each factor) 3. *To include interaction:* Use products of the (included) dummies. 4. How to interpret the coefficients? 5. How to “recover” the three ANOVA components?

*Reasonable to assume there is no interaction here. See the comments on the code below*

Exercise1 = subset(Pulse, Pulse$Exercise == 1)  
Exercise2 = subset(Pulse, Pulse$Exercise == 2)  
Exercise3 = subset(Pulse, Pulse$Exercise == 3)  
  
lme1 = lm(Active ~ Rest, data = Exercise1)  
lme2 = lm(Active ~ Rest, data = Exercise2)  
lme3 = lm(Active ~ Rest, data = Exercise3)  
  
plot(Active~Rest, data=Pulse)  
abline(lme1)  
abline(lme2, col='red')  
abline(lme3, col='blue')



# When we look at the three lines over the data, we are hoping that these will be similar in slope. IF they are much different in terms of slope, from a practical standpoint, there may be an interaction between these that we need to investigate.   
  
##Looking at the data, we are trying to see if they are similar enough with their slope, none look drastically different

*What if there are some issues? Lets try the above with a transformation*

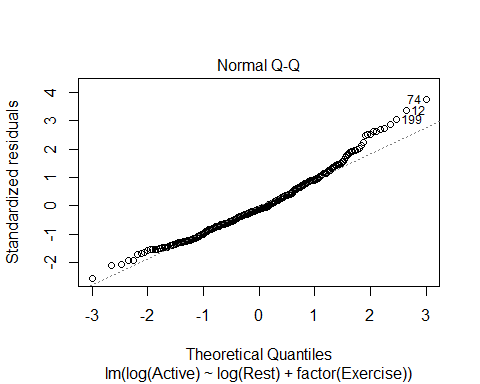
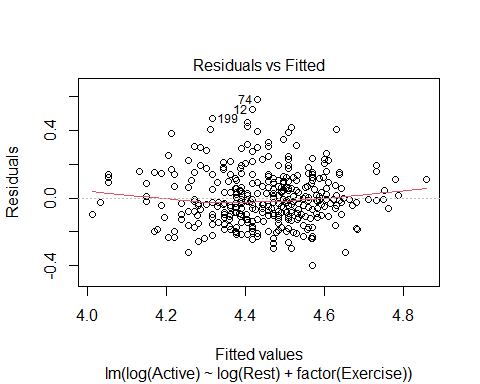
modp2log = lm(log(Active)~log(Rest) + factor(Exercise), data = Pulse)  
  
anova(modp2log)

## Analysis of Variance Table  
##   
## Response: log(Active)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## log(Rest) 1 6.7174 6.7174 277.991 <2e-16 \*\*\*  
## factor(Exercise) 2 0.0499 0.0250 1.033 0.3569   
## Residuals 371 8.9648 0.0242   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

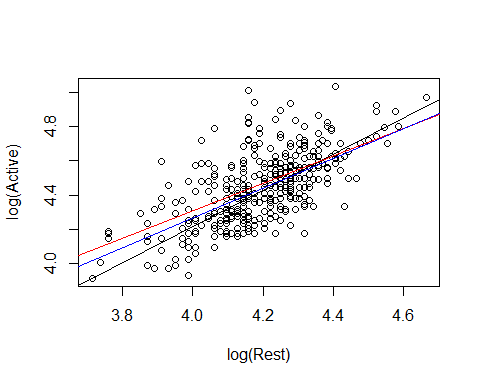
*Looking at Residuals* - There is some curve that is not ideal - Look at teh QQNorm, there are some big issues with that one tail on the untransformed thing - With the transformed data: the tail on QQNorm isn’t as bad and the little curve on the residual plot isn’t as bad.

*We also have to check and make sure that there is no interaction now that we have introduced the transformation* - Redo what you did before, but all in logs, See code below:

plot(modp2log, 1:2)



lme1.log = lm(log(Active) ~ log(Rest), data=Exercise1)  
lme2.log = lm(log(Active) ~ log(Rest), data=Exercise2)  
lme3.log = lm(log(Active) ~ log(Rest), data=Exercise3)  
  
plot(log(Active) ~ log(Rest), data=Pulse)  
abline(lme1.log)  
abline(lme2.log, col='red')  
abline(lme3.log, col='blue')



# Doesn't show anything idfferent than before   
# IT helped with teh conditions, but it doesn't effect the interaction to consider

**WHy do we do this?** - We want ot know for a hypo test: Is there a dif in teh active heart rates of people at each of the three exercise levels on average

Ho: Mean grousp are qual Ha: At least one mean group is difference

After accounting for the variability in their active heart rate; so basically doing a nested test to see if exercise variabile signifigant after considering active heart rate?

*look at ANOVA* FOcus on secon d line, We are thinking about after the var is explained in active heart rate by rest, do we see sig dif in the mean active heart rate by exercise level?

The pvalue is high, so we dont see a sig dif

anova(modp2log)

## Analysis of Variance Table  
##   
## Response: log(Active)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## log(Rest) 1 6.7174 6.7174 277.991 <2e-16 \*\*\*  
## factor(Exercise) 2 0.0499 0.0250 1.033 0.3569   
## Residuals 371 8.9648 0.0242   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

*How does this relate to the homework?* - Last questionon teh HW, are there dif in the mean price of the 6 models of car after taking into account another varibale?

* Account for the variability in mileage and then see if your car has sig difference in prices
* By chance you might have had cars of a certain model with high or low mileage
* if one car has only been out for a few years, then it might not have high mileage; so it might not be as useful.
* If you have cars that are different ages, then you might have a wide spread with some cars vs the other, then you can see after you take into account the variability ofX affects the price of car, are there any differences left by the model or is X really explaining that difference?

**Analysis of Covariance (ANCOVA)** **Basic idea:** If we can use dummy predictors to convert an ANOVA for means into a regression model, why not also include quantitative predictors?

**Analysis of Covariance (ANCOVA)** WeightLoss = β0 + β1Age + β2Height + β3Diet1 + β4Diet2 + ε Age and Height = Covariates Diet1 and Diet2 = Factor

* If want to see if sig dif by diet after taking into account other variables, we cna see that as well; see below
* We dont know how they chose the groups for these people, so we don’t know if its skewed. This will keep that in mind.
* If they didn’t, we could test to see if someone’s age explains how much weight they loss and if there is a sig diff bt the diet they were on after considering their age or height

Diet$weightchange = Diet$weight6weeks - Diet$Preweight  
  
mod7 = lm(weightchange ~ factor(Diet), data=Diet)  
anova(mod7)

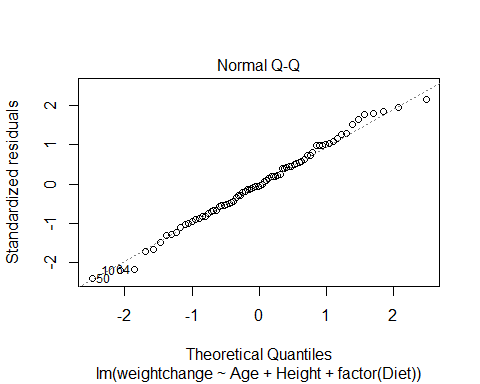
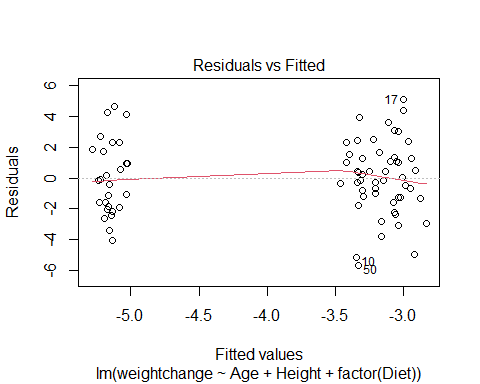
## Analysis of Variance Table  
##   
## Response: weightchange  
## Df Sum Sq Mean Sq F value Pr(>F)   
## factor(Diet) 2 71.09 35.547 6.1974 0.003229 \*\*  
## Residuals 75 430.18 5.736   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**What does this tell us?** - After explaining the variabilit yin weight change by age and height, neither were sig, we still saw that there were sig dif bt the weight change by which diet you were on - There could be other variables that did not show this and owuld not show any difference diet after thinking about that.

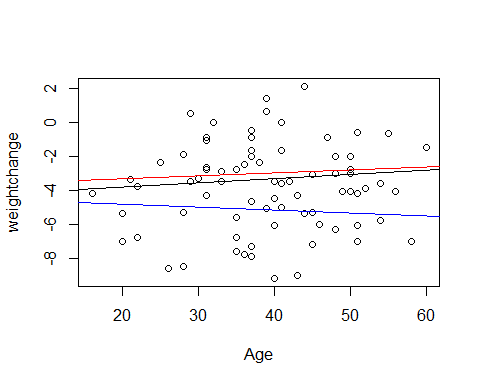
mod8 = lm(weightchange ~ Age + Height + factor(Diet), data=Diet)  
anova(mod8)

## Analysis of Variance Table  
##   
## Response: weightchange  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Age 1 1.69 1.689 0.2870 0.593795   
## Height 1 6.23 6.227 1.0580 0.307068   
## factor(Diet) 2 63.68 31.838 5.4092 0.006448 \*\*  
## Residuals 73 429.68 5.886   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

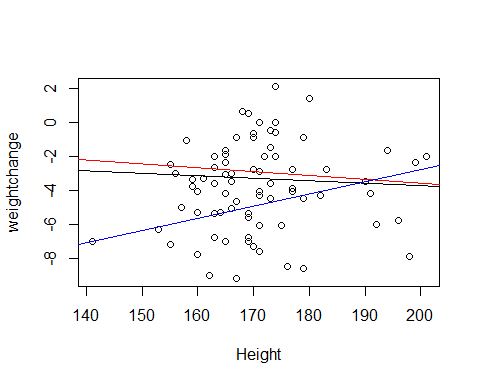
plot(mod8, 1:2)



Diet1 = subset(Diet, Diet==1)  
Diet2 = subset(Diet, Diet==2)  
Diet3 = subset(Diet, Diet==3)  
  
lme1.diet = lm(weightchange ~ Age, data=Diet1)  
lme2.diet = lm(weightchange ~ Age, data=Diet2)  
lme3.diet = lm(weightchange ~ Age, data=Diet3)  
  
plot(weightchange ~ Age, data=Diet)  
abline(lme1.diet)  
abline(lme2.diet, col='red')  
abline(lme3.diet, col='blue')



lme4.diet = lm(weightchange ~ Height, data=Diet1)  
lme5.diet = lm(weightchange ~ Height, data=Diet2)  
lme6.diet = lm(weightchange ~ Height, data=Diet3)  
  
plot(weightchange ~ Height, data=Diet)  
abline(lme4.diet)  
abline(lme5.diet, col='red')  
abline(lme6.diet, col='blue')



## STOR 455 Class 38 Homogeneity of Variances and Contrasts

library(readr)  
library(car)  
library(Stat2Data)  
  
Exams4 <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/Exams4.csv")

*Goal: Trying to see if there is a dif in means bt groups* - Look how we can test variability between means

**ANOVA for Difference in K Means** Data: Samples from K different groups Summary statistics: n1, ybar1, s1 for each group Combine all

Test: Ho: mu1 = mu2 = muk Ha: Some mui != muj

**Checking Conditions for ANOVA** E~N(0, Std of Errors) Check with residuals

* Zero Mean: Always holds for sample residuals
* COnstant Variance: Plot residuals vs. fits and/or compare std. dev.’s of groups (Check if some group si is more than twice another)
* Normaility: Histogram/normal plot of residuals
* Independence:Pay attention to data collection

**Example: Five Students** *Question of intereset: Is there a significant difference in average grade among the five students?* - Test tells us if there is sig, but it doesn’t always mean useful - Found there are sig dif by student, and the variability are suspect to assume there weas constant variance here (Betsy’s scores were all sim to each other while Bud was all close) - Laready groups so that the first student is grouped with their scores, etc, etc

**ANOVA for Grades vs. Students** *Can we assume equal variances?*

tapply(Exams4$Grade, Exams4$Student, median)

## Barb Betsy Bill Bob Bud   
## 75.5 91.5 77.5 82.5 48.5

medians = rep(tapply(Exams4$Grade, Exams4$Student, median), each=4)  
medians # Median scores for each student; some skew, which makes it different than means

## Barb Barb Barb Barb Betsy Betsy Betsy Betsy Bill Bill Bill Bill Bob   
## 75.5 75.5 75.5 75.5 91.5 91.5 91.5 91.5 77.5 77.5 77.5 77.5 82.5   
## Bob Bob Bob Bud Bud Bud Bud   
## 82.5 82.5 82.5 48.5 48.5 48.5 48.5

# Want to keep track of not comparing this with an anova analysis   
# Want to see how far away each score is from the median value   
#Analysis of the variances of these absolute devations

* Need to take the abso score - media to make a new array with all the values
* It’s going to work nicely here, but it’s not going to look unless we make the data look like this data

**Levene’s Test for Equality of Variances**

Test: Ho: variance1 = variance2 =…=variancek H1: some variancei!=variancej

(y-ybar) = (ybark-ybar) + (y-ybark) where (y-ybark) conducts analysis of variance on teh collection of absolute deviations abs(y-ytildek); where ytildek = median of group k

LeveneData = abs(Exams4$Grade - medians)  
LeveneData

## Barb Barb Barb Barb Betsy Betsy Betsy Betsy Bill Bill Bill Bill Bob   
## 13.5 11.5 1.5 1.5 2.5 3.5 5.5 2.5 9.5 15.5 4.5 4.5 3.5   
## Bob Bob Bob Bud Bud Bud Bud   
## 14.5 12.5 3.5 1.5 14.5 20.5 1.5

Levene\_aov = aov(LeveneData~Exams4$Student)  
summary(Levene\_aov)

## Df Sum Sq Mean Sq F value Pr(>F)  
## Exams4$Student 4 88.8 22.20 0.567 0.69  
## Residuals 15 587.0 39.13

leveneTest(Exams4$Grade, factor(Exams4$Student))

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 4 0.5673 0.6903  
## 15

**Levene’s Test for Grades versus Students** Ho: variance(barb) = variance2 = variance3 = etc Ha: variancei != variancej; for at least one pair of students (i,j)

*Conclusion*: There is not statistically significant evidence to suggest that students’ grades have difference variances.

medians=rep(tapply(Exams4$Grade, Exams4$Student,median),each=4) # Each = 4 because we want barbs 4 times, betsy 4 times, etc.   
LeveneData=abs(Exams4$Grade-medians) #They are lined up perfectly right in this dataset; so all the grades will be subed by each student's median score.   
LeveneData

## Barb Barb Barb Barb Betsy Betsy Betsy Betsy Bill Bill Bill Bill Bob   
## 13.5 11.5 1.5 1.5 2.5 3.5 5.5 2.5 9.5 15.5 4.5 4.5 3.5   
## Bob Bob Bob Bud Bud Bud Bud   
## 14.5 12.5 3.5 1.5 14.5 20.5 1.5

Levene\_aov = aov(LeveneData~Exams4$Student) # Do we see any sig dif in the subbed median by actual score?   
#Testing: Build model assume the var are equal in each group and trying to see if we have edv that there are some dif here; same as like anova model, but at var   
# most of the sum of squares are in resdu, which assum there are equal variance; little are in sumsquares of the model itself (grand variance)   
# Go in with teh same ideas, but with variances and assume that htey are equal, then the p value is big, we dont have evidence to support Ha; even though the rule of variance not being double another, but the test it didnt show sign  
# Comes down to sample size; we would have a lot of dif to have a sig dif   
# tehere is dif here, but its not big enough to reject Ho  
summary(Levene\_aov)

## Df Sum Sq Mean Sq F value Pr(>F)  
## Exams4$Student 4 88.8 22.20 0.567 0.69  
## Residuals 15 587.0 39.13

library(car)   
leveneTest(Exams4$Grade, Exams4$Student) # It wants grade and student (what we are splitting grade up with); it shouldnt really matter if you facotr it

## Warning in leveneTest.default(Exams4$Grade, Exams4$Student): Exams4$Student  
## coerced to factor.

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 4 0.5673 0.6903  
## 15

**Cancer Survival with Ascorbate Supplement**  - Use this when you have issues with constant variance and how to use transofmraitons to help with that

**INformation about the data:** *In the 1970’s doctors wondered if giving terminal cancer patients a supplement of ascorbate would prolong their lives. They designed an experiment to compare cancer patients who received ascorbate to cancer patients who did not receive the supplement. The result of that experiment was that, in fact, ascorbate did seem to prolong the lives of these patients. But then a second question arose. Was the effect of the ascorbate different when different organs were affected by the cancer? The researchers took a second look at the data. This time they concentrated only on those patients who received the ascorbate and divided the data up by which organ was affected by the cancer. They had 5 different organs represented among the patients (all of whom only had one organ affected): Stomach, bronchus, colon, ovary, and breast.*

* Looking at treatmetn for cancer; is the outcome dife for the type of cancer the people have?

data("CancerSurvival")  
head(CancerSurvival)

## Survival Organ  
## 1 124 Stomach  
## 2 42 Stomach  
## 3 25 Stomach  
## 4 45 Stomach  
## 5 412 Stomach  
## 6 51 Stomach

**Cancer Survival with Ascorbate Supplement** Test:

Ho: mu1 = mu2 = …= muk Ha: Some mui != muj

* Want to see if the avg # days are equal depn on type of cancer or if no dif in treatment vs cancer type

Cancermod = aov(Survival~Organ, data=CancerSurvival)  
summary(Cancermod) # Sig evid to say surivial time changes depending on the organ that has cancer

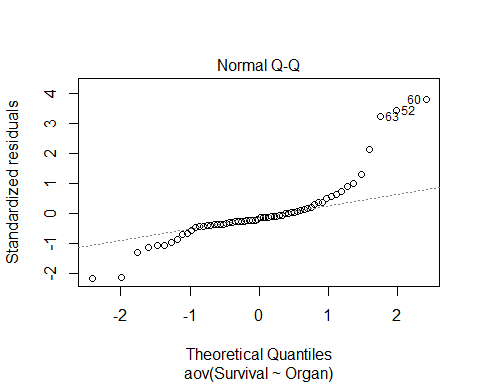
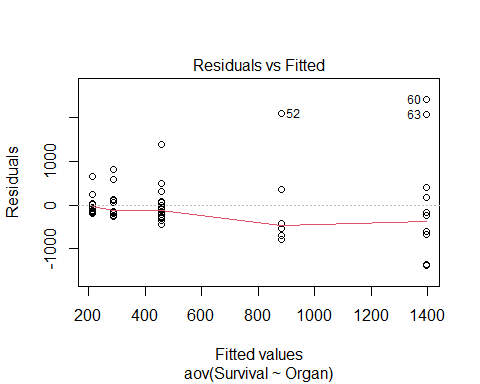
## Df Sum Sq Mean Sq F value Pr(>F)   
## Organ 4 11535761 2883940 6.433 0.000229 \*\*\*  
## Residuals 59 26448144 448274   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# THIS IS ASSUME CONDITIONS ARE MET

**Cancer Survival with Ascorbate Supplement**  *Check conditions* - See there is a trend in variability increases - The organ that has the lowest survial rate = more compact - as survival rate for organ increase as - clear pattern in plot

* Normal resiudal s
* off because the tails are wonky

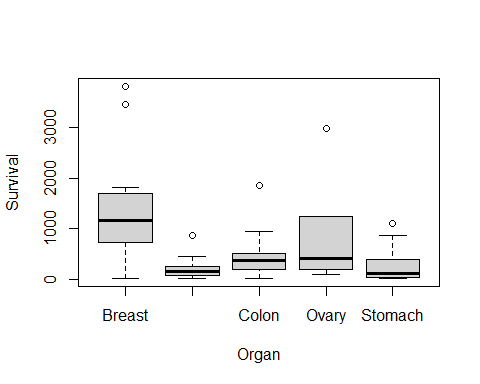
plot(Cancermod, 1:2)



**Cancer Survival with Ascorbate Supplement**  Ho: variance1 = variance2=…=variancek Ha: Some variancei != variancej

* Looking at visuals of survival by organ

boxplot(Survival~Organ, data=CancerSurvival)



tapply(CancerSurvival$Survival, CancerSurvival$Organ, sd)

## Breast Bronchus Colon Ovary Stomach   
## 1238.9667 209.8586 427.1686 1098.5788 346.3096

leveneTest(CancerSurvival$Survival, CancerSurvival$Organ)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)   
## group 4 4.4524 0.003271 \*\*  
## 59   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

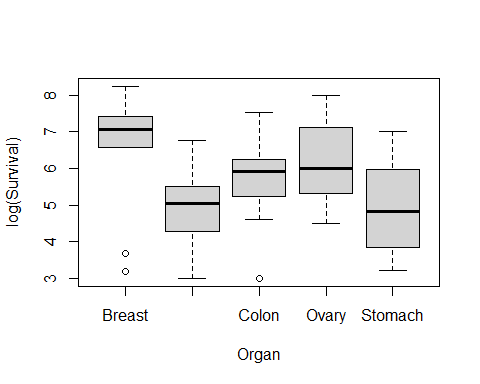
# Constant variance is a huge issue here   
  
#Levenetest gives a really small pvalue

* are these sim by chance or
* we have evid to say that the var is dif than the others
* we cant really use the results because the conditions aren’t good

**Would a transformation help?** - Yes, if you can make it look more normal then you will be good

LeveneTest Ho: variance1=varaince2=…varicnek Ha: some variancei != variancej

boxplot(log(Survival)~Organ, data=CancerSurvival)



tapply(log(CancerSurvival$Survival), CancerSurvival$Organ, sd)

## Breast Bronchus Colon Ovary Stomach   
## 1.6477550 0.9534041 0.9974766 1.2569313 1.2502073

leveneTest(log(CancerSurvival$Survival), CancerSurvival$Organ)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 4 0.6685 0.6164  
## 59

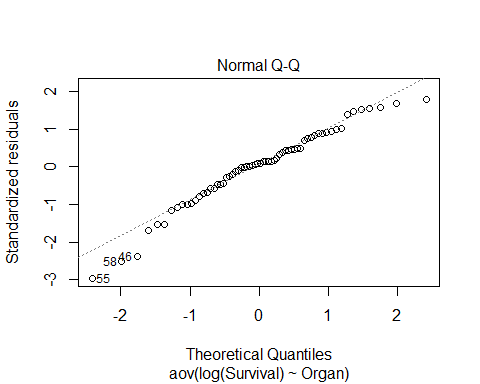
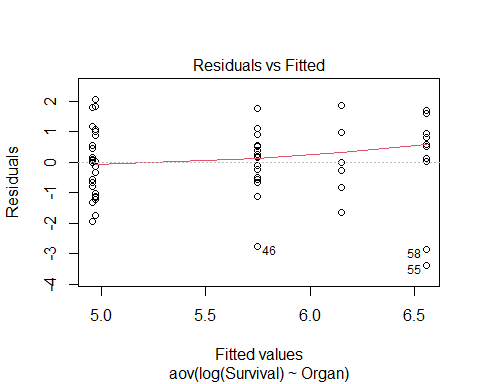
* above looks a lot better
* the sd are much closer and they are squished better
* the levenetest is also better with teh pvalue is high
* now we dont see there is dif
* we like higher pvalue in this

Cancermod\_log = aov(log(Survival)~Organ, data=CancerSurvival)  
summary(Cancermod\_log)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Organ 4 24.49 6.122 4.286 0.00412 \*\*  
## Residuals 59 84.27 1.428   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

*Cehck conditions* - not as big of an issue - not perfect for QQNorm, but its beter

plot(Cancermod\_log, 1:2)



TukeyHSD(Cancermod\_log) # shows us that Bronchus-Breast and stomach-breast are sig

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = log(Survival) ~ Organ, data = CancerSurvival)  
##   
## $Organ  
## diff lwr upr p adj  
## Bronchus-Breast -1.60543320 -2.906741 -0.3041254 0.0083352  
## Colon-Breast -0.80948110 -2.110789 0.4918267 0.4119156  
## Ovary-Breast -0.40798703 -2.114754 1.2987803 0.9615409  
## Stomach-Breast -1.59068365 -2.968399 -0.2129685 0.0158132  
## Colon-Bronchus 0.79595210 -0.357534 1.9494382 0.3072938  
## Ovary-Bronchus 1.19744617 -0.399483 2.7943753 0.2296079  
## Stomach-Bronchus 0.01474955 -1.224293 1.2537924 0.9999997  
## Ovary-Colon 0.40149407 -1.195435 1.9984232 0.9540004  
## Stomach-Colon -0.78120255 -2.020245 0.4578403 0.3981146  
## Stomach-Ovary -1.18269662 -2.842480 0.4770864 0.2763506

*Below shows the difference between small level what is “sig” and big sample level what is “sig”* - As sample size increases, if its sig doesn’t mean its pracitcal to us - we want ot make sure its true roughly, but we might have practical difference - we dont always use it because it can be sus

n=10 # 10 groups, and taking 10 random groups; of 4 groups; and binding them todaya dn making a df;   
  
#Name the data you randomly call   
x1 <- rnorm(n)  
x2 <- rnorm(n)  
x3 <- rnorm(n)  
x4 <- rnorm(n)  
  
x = c(rbind(x1, x2, x3, x4))  
data <- data.frame(x)  
  
data$group[ 1:n ] = "A"  
data$group[( n + 1):(2\*n)] = "B"  
data$group[(2\*n + 1):(3\*n)] = "C"  
data$group[(3\*n + 1):(4\*n)] = "D"  
  
#Lavene test  
# Should all have the same sd from population; by chance what the vlav test would look like   
leveneTest(data$x, factor(data$group))

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 3 0.673 0.5742  
## 36

# Get sds  
tapply(data$x, data$group, sd)

## A B C D   
## 1.2604865 0.9596752 1.1035292 0.6933545

# D is almost double and levne pval = 0.2313  
# If we run a few times, then we eventually get something closer to 0.05  
# Gives idea of the dif that we would need to see for levene test to not like the dat aor say its sig

n=400 # 400 groups, and taking 10 random groups; of 4 groups; and binding them todaya dn making a df;   
  
#Name the data you randomly call   
x1 <- rnorm(n)  
x2 <- rnorm(n)  
x3 <- rnorm(n)  
x4 <- rnorm(n)  
  
x = c(rbind(x1, x2, x3, x4))  
data <- data.frame(x)  
  
data$group[ 1:n ] = "A"  
data$group[( n + 1):(2\*n)] = "B"  
data$group[(2\*n + 1):(3\*n)] = "C"  
data$group[(3\*n + 1):(4\*n)] = "D"  
  
#Lavene test  
# Should all have the same sd from population; by chance what the vlav test would look like   
leveneTest(data$x, factor(data$group))

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 3 0.6395 0.5896  
## 1596

# Get sds  
tapply(data$x, data$group, sd)

## A B C D   
## 0.9938886 1.0174135 1.0623579 1.0265391

# Unit 7: Conclusion

## Simple Linear Regression

### Building a Model & Assessing Conditions

Building a Model and Assessing Conditions

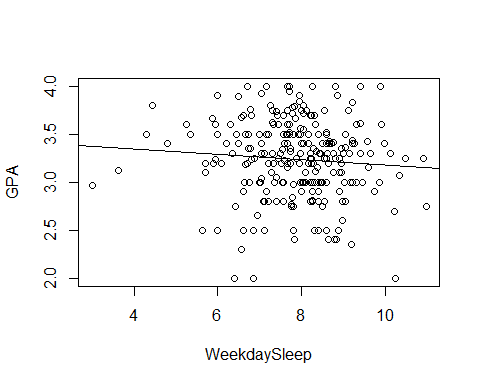
R Markdown

library(readr)  
  
Sleep <- read.csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

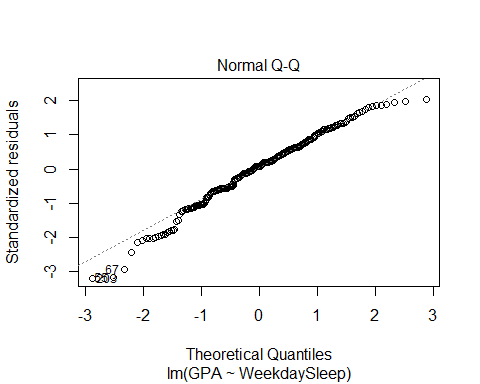
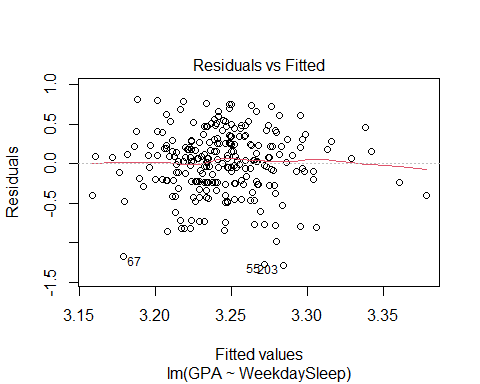
1. Build a linear model predicting GPA based on average hours of sleep during weekdays. Comment on how well your data appear to fit the conditions for a linear model.

Looking at original plot of the data with the linear model, the data does not seem to follow a linear pattern as described by the model line. The residuals vs. fitted plot shows a relatively linear pattern (looking at the plotted line as compared to the origin line) and the variance seems relatively constant across the residual data, with possible slight tapering in towards the right-hand side. The normal Q-Q line shows some skew on the left hand side of the plot, showing possible deviation from a normal distribution pattern in the residuals. Left-skew is also observed in the histogram for the residual data, also implying a deviation from normal distribution.

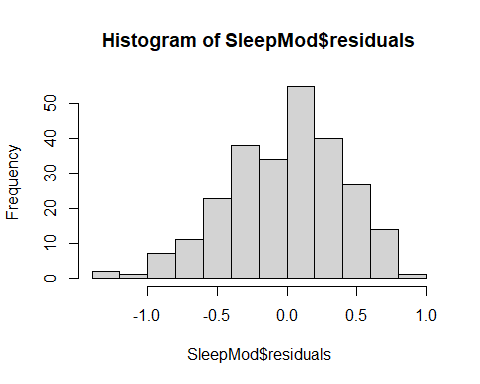
SleepMod = lm(GPA~WeekdaySleep, data=Sleep)  
plot(GPA~WeekdaySleep, data=Sleep)  
abline(SleepMod)



plot(SleepMod, 1:2)



hist(SleepMod$residuals)



### Transformations

R Notebook   
library(readr)   
  
sleep <- read\_csv("https://raw.githubusercontent.com/JA-  
McLean/STOR455/master/data/SleepStudy.csv")   
## Rows: 253 Columns: 27   
## -- Column specification --------------------------------------------------  
------   
## Delimiter: ","   
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse   
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA,   
ClassesMissed, ...   
##   
## i Use `spec()` to retrieve the full column specification for this data.   
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this   
message.   
#Transformations   
#Building a Model with a single predictors   
head(sleep)   
## # A tibble: 6 x 27   
## Gender ClassYear LarkOwl NumEarlyClass EarlyClass GPA ClassesMissed   
## <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <dbl>   
## 1 0 4 Neither 0 0 3.6 0   
## 2 0 4 Neither 2 1 3.24 0   
## 3 0 4 Owl 0 0 2.97 12   
## 4 0 1 Lark 5 1 3.76 0   
## 5 0 4 Owl 0 0 3.2 4   
## 6 1 4 Neither 0 0 3.5 0   
## # ... with 20 more variables: CognitionZscore <dbl>, PoorSleepQuality   
<dbl>,   
## # DepressionScore <dbl>, AnxietyScore <dbl>, StressScore <dbl>,   
## # DepressionStatus <chr>, AnxietyStatus <chr>, Stress <chr>, DASScore   
<dbl>,   
## # Happiness <dbl>, AlcoholUse <chr>, Drinks <dbl>, WeekdayBed <dbl>,   
## # WeekdayRise <dbl>, WeekdaySleep <dbl>, WeekendBed <dbl>, WeekendRise   
<dbl>,   
## # WeekendSleep <dbl>, AverageSleep <dbl>, AllNighter <dbl>   
1. Using CognitionZscore as a predictor for GPA, calculate the least squares regression   
line that best fits your data and plot this.

mod1 = lm(GPA ~ CognitionZscore, data = sleep)   
plot(GPA ~ CognitionZscore, data = sleep)   
abline(mod1)   
  
plot(mod1)

2. Experiment with some transformation(s) to attempt to find one that seems to do a   
better job of satisfying the linear model conditions. Plot this as well.   
mod2 = lm(log(GPA) ~ log(CognitionZscore), data = sleep)

### Outliers & Influential Points

Group 2 Questions

library(readr)  
sleep = read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

## Rows: 253 Columns: 27

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse  
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA, ClassesMissed, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

1. Start by building a model that predicts GPA based on the number of classes a student has missed, ClassesMissed

mod = lm(GPA~ClassesMissed, data = sleep)

1. Find and display the top 10 standardized residuals (be sure to look at absolute values when sorting). Of these, which data points may be considered influential?

Any values greater than 2 / less than -2 are considered to be influential. In our case, the top 9 data points have high standardized residuals.

head(sort(abs(rstandard(mod)), decreasing = TRUE), 10)

## 55 67 203 126 89 224 34 72   
## 3.198291 3.198291 3.140456 2.273283 2.250347 2.077591 2.037753 2.021602   
## 169 50   
## 2.021602 1.998586

1. Find and display the top 10 studentized residuals (be sure to look at absolute values when sorting). Of these, which data points may be considered influential?

Values we consider to be high: Any values greater than 2 / less than -2 In our case, all of the top 10 data points have high standardized residuals.

head(sort(abs(rstudent(mod)), decreasing = TRUE), 10)

## 55 67 203 126 89 224 34 72   
## 3.259012 3.259012 3.197646 2.292473 2.268865 2.091510 2.050724 2.034199   
## 169 50   
## 2.034199 2.010663

1. Based on only the information from the previous two questions, do we have any indications about data points that may exert significant influence on the model?

For question 4, I think the answer should be we have no evidence that the data points may have significant influence because standardized residuals and studentized residuals only tell us how far away each point deviates from the mean, they don’t tell us how influential each point is.

1. Briefly explain the difference between influence and leverage.

Influence is the effect of a single data point on the regression line, and represents how well the data point matches the of the rest of the points. Leverage represents how much influence a data point has on the regression line. It explains how far away the data point is from the other points (explains how much of an outlier it is)

For question 5, a better explanation of leverage is the “potential” to exert influence.

1. Briefly explain the difference between standardized and studentized residuals.

Standardized: Residuals of the data are standardized If a rstandard value is greater than 2 (or less than -2) it is a mild outlier. If the value is greater than 3 (or less than -3), it is a strong case of being an outlier. Studentized: An outlier is removed from the data, then standardized.

For question 6, the studentized residual is calculated by removing the data point whose studentized value you want to calculate, not removing an outlier.

1. Find and display the top 10 points with highest leverage. Do any points have significantly large leverage?

None of these data points appear to have high leverage.

2/12 # Average leverage (2/n)

## [1] 0.1666667

2 \* 2/12 # Points we might need to worry about

## [1] 0.3333333

3 \* 2/12 # Points we might need to worry about

## [1] 0.5

head(sort(hatvalues(mod), decreasing=TRUE), 10)

## 20 26 85 163 65 204 148   
## 0.12357263 0.12357263 0.06578311 0.06578311 0.05649286 0.05649286 0.04795851   
## 3 27 66   
## 0.04018004 0.02689076 0.02689076

1. Calculate the Cook’s Distance for the points in our data set. How many of these points may be considered influential outliers?

From class, we said to study any case with Di > 0.5; worry if Di > 1.0. In our data set, none of the cook’s distances exceed 0.5, so we have 0 influential outliers in our dataset.

head(sort(cooks.distance(mod), decreasing=TRUE),10)

## 65 26 55 67 163 245 34   
## 0.09610117 0.02662156 0.02314754 0.02314754 0.02079628 0.02033074 0.01966539   
## 203 89 126   
## 0.01965083 0.01476538 0.01341317

### Inference for Regression Slope/Correlation

STOR 455 Homework #9 Part 1

Inference for Regression Slope/Correlation

library(readr)  
SleepStudy <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

## Rows: 253 Columns: 27

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse  
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA, ClassesMissed, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

1. Create a model to predict GPA from the average amount of sleep and use the test for correlation and test for slope to test the strength of the linear relationship between your variables.

test for corelation: H0: p = 0 Hα: p != 0

test for slope: H0:β1 = 0 Hα:β1 != 0 According to the test for slope, our p-value is 0.337 while the p-value of the correlation test is 0.3368; Both of which do not give us enough evidence to reject the null hypothesis, which means that there isnt a strong relationship between GPA and AverageSleep.

mod1 = lm(GPA~AverageSleep, data = SleepStudy)  
cor.test(SleepStudy$GPA,SleepStudy$AverageSleep)

##   
## Pearson's product-moment correlation  
##   
## data: SleepStudy$GPA and SleepStudy$AverageSleep  
## t = -0.96238, df = 251, p-value = 0.3368  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.18259574 0.06316715  
## sample estimates:  
## cor   
## -0.06063317

summary(mod1)

##   
## Call:  
## lm(formula = GPA ~ AverageSleep, data = SleepStudy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.29018 -0.24395 0.03294 0.26317 0.80407   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.44618 0.21183 16.269 <2e-16 \*\*\*  
## AverageSleep -0.02541 0.02640 -0.962 0.337   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4043 on 251 degrees of freedom  
## Multiple R-squared: 0.003676, Adjusted R-squared: -0.000293   
## F-statistic: 0.9262 on 1 and 251 DF, p-value: 0.3368

### ANOVA for simple linear regression

Group\_Review

library(readr)  
sleep = read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

## Rows: 253 Columns: 27

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse  
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA, ClassesMissed, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

head(sleep)

## # A tibble: 6 x 27  
## Gender ClassYear LarkOwl NumEarlyClass EarlyClass GPA ClassesMissed  
## <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 0 4 Neither 0 0 3.6 0  
## 2 0 4 Neither 2 1 3.24 0  
## 3 0 4 Owl 0 0 2.97 12  
## 4 0 1 Lark 5 1 3.76 0  
## 5 0 4 Owl 0 0 3.2 4  
## 6 1 4 Neither 0 0 3.5 0  
## # ... with 20 more variables: CognitionZscore <dbl>, PoorSleepQuality <dbl>,  
## # DepressionScore <dbl>, AnxietyScore <dbl>, StressScore <dbl>,  
## # DepressionStatus <chr>, AnxietyStatus <chr>, Stress <chr>, DASScore <dbl>,  
## # Happiness <dbl>, AlcoholUse <chr>, Drinks <dbl>, WeekdayBed <dbl>,  
## # WeekdayRise <dbl>, WeekdaySleep <dbl>, WeekendBed <dbl>, WeekendRise <dbl>,  
## # WeekendSleep <dbl>, AverageSleep <dbl>, AllNighter <dbl>

#Topic 4: ANOVA for simple linear regression  
#Question:Test the strength of the linear relationship between \_\_GPA\_\_ and \_\_AverageSleep\_\_ using ANOVA for regression. Include hypotheses and your conclusions in the context of the problem.

#Solution:  
mod1 = lm(GPA~AverageSleep, data = sleep)  
anova(mod1)

## Analysis of Variance Table  
##   
## Response: GPA  
## Df Sum Sq Mean Sq F value Pr(>F)  
## AverageSleep 1 0.151 0.15142 0.9262 0.3368  
## Residuals 251 41.037 0.16349

#Null Hypothesis: The coefficient for AverageSleep is 0.  
#Alternative Hypothesis: The coefficient for AverageSleep is not 0.  
#Conclusion: According to the table, p-value is 0.3368, which is a really large number. Therefore, we fail to reject the null hypothesis and come to the conclusion that the coefficient for AverageSleep is 0.

#Topic 17: Test of significance for nested models

#Question\_1:Construct a logistic model to predict if the student has an early class using Stress, AnxietyStatus, and the interaction between them as the predictor variables.

#Solution\_1:  
mod2 = glm(EarlyClass~Stress+AnxietyStatus+Stress\*AnxietyStatus,family = binomial, data = sleep)  
summary(mod2)

##   
## Call:  
## glm(formula = EarlyClass ~ Stress + AnxietyStatus + Stress \*   
## AnxietyStatus, family = binomial, data = sleep)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6651 -1.4944 0.8906 0.8906 1.7941   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.51083 0.36515 1.399 0.1618   
## Stressnormal 0.58779 0.59628 0.986 0.3243   
## AnxietyStatusnormal -0.04082 0.67700 -0.060 0.9519   
## AnxietyStatussevere 0.47000 0.76920 0.611 0.5412   
## Stressnormal:AnxietyStatusnormal -0.33774 0.84118 -0.402 0.6881   
## Stressnormal:AnxietyStatussevere -2.95491 1.43662 -2.057 0.0397 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 322.99 on 252 degrees of freedom  
## Residual deviance: 317.01 on 247 degrees of freedom  
## AIC: 329.01  
##   
## Number of Fisher Scoring iterations: 4

#Question2:Conduct a drop in deviance hypothesis test to determine the effectiveness of the interaction terms in the model constructed in the previous question. Cite your hypotheses, p-value, and conclusion in context.

#Solution2:  
reduced = glm(EarlyClass~Stress+AnxietyStatus,family = binomial, data = sleep)  
anova(mod2, reduced,test = "Chisq")

## Analysis of Deviance Table  
##   
## Model 1: EarlyClass ~ Stress + AnxietyStatus + Stress \* AnxietyStatus  
## Model 2: EarlyClass ~ Stress + AnxietyStatus  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 247 317.01   
## 2 249 322.16 -2 -5.1531 0.07604 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Null Hypothesis:The coefficient for the interaction terms is 0.  
#Alternative Hypothesis:The coefficient for the interaction terms is not 0.  
#Conclusion: According to the table, p-value is 0.07604, which is greater than 0.05. Therefore, we fail to reject the null hypothesis and come to the conclusion that the coefficient for interaction terms is 0.

### Confidence and Prediction Intervals

Homework 9 - Topic 5

Rhea Bhagia, Meghna Sharma, Olivia Voss, Connor Hammond

11/28/2021

SleepStudy <- read.csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

*Question*: Construct 90% confidence and prediction intervals to assess the interval for average hours of sleep for all days for a first year and fourth year student. Interpret your findings in context.

SleepStudy\_mod = lm(AverageSleep ~ factor(ClassYear), data=SleepStudy)  
  
first\_year = data.frame(ClassYear = 1)  
fourth\_year = data.frame(ClassYear = 4)  
   
predict.lm(SleepStudy\_mod, first\_year, interval = "confidence", level = 0.90 )

## fit lwr upr  
## 1 7.925319 7.691953 8.158685

predict.lm(SleepStudy\_mod, first\_year, interval = "prediction", level = 0.90 )

## fit lwr upr  
## 1 7.925319 6.30851 9.542129

predict.lm(SleepStudy\_mod, fourth\_year, interval = "confidence", level = 0.90 )

## fit lwr upr  
## 1 7.95 7.738091 8.161909

predict.lm(SleepStudy\_mod, fourth\_year, interval = "prediction", level = 0.90 )

## fit lwr upr  
## 1 7.95 6.336148 9.563852

We are 90% confident that a first year student will sleep on average 7.691953 to 8.158685 hours every night.

We can predict 90% of the time that a first year student will sleep on average between 6.30851 and 9.542129 hours a night.

We are 90% confident that a fourth year student will sleep on average 7.738091 to 8.161909 hours every night.

We can predict 90% of the time that a first year student will sleep on average between 6.336148 and 9.563852 hours a night.

## Multiple Regression

### Building a Model & Assessing Conditions

STOR 455 Homework #9

Due 11/29

library(readr)  
library(Stat2Data)  
"SleepStudy" <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

## Rows: 253 Columns: 27

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse  
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA, ClassesMissed, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

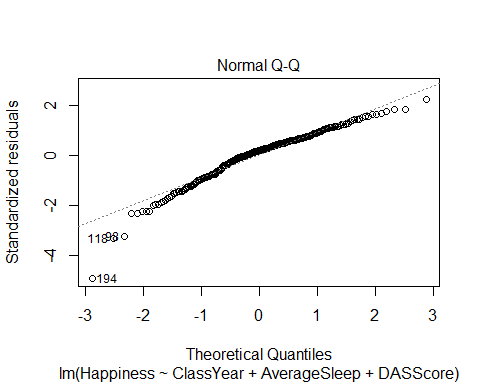
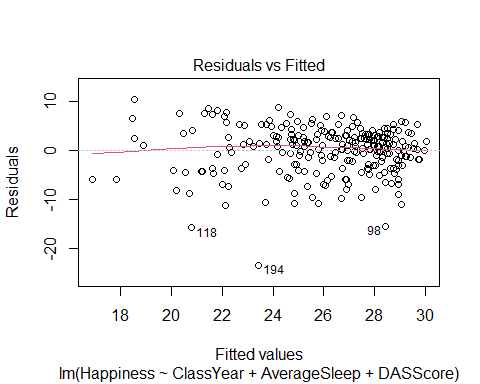
1. Multiple Regression: Building a Model & Assessing Conditions
2. Construct a linear model with *Happiness* as the response and *ClassYear* , *AverageSleep* , *DASScore* as the predictors. Include the output for the summary of the model.

mod1 = lm(Happiness~ClassYear + AverageSleep + DASScore, data=SleepStudy)  
summary(mod1)

##   
## Call:  
## lm(formula = Happiness ~ ClassYear + AverageSleep + DASScore,   
## data = SleepStudy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23.4348 -2.8943 0.8034 3.0193 10.4424   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 25.34113 2.68833 9.426 < 2e-16 \*\*\*  
## ClassYear 0.39920 0.29229 1.366 0.173   
## AverageSleep 0.37165 0.31511 1.179 0.239   
## DASScore -0.15870 0.01839 -8.630 7.37e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.81 on 249 degrees of freedom  
## Multiple R-squared: 0.2412, Adjusted R-squared: 0.2321   
## F-statistic: 26.39 on 3 and 249 DF, p-value: 7.432e-15

1. Asses the conditions for linearity of the model you constructed in question 1

plot(mod1, 1:2)

 Looking at our residuals vs. fitted plot, we can see that there seem to be some issues with constant variance, as the data is more compact towards the larger fitted values. Again looking at the residuals vs. fitted plot, there are not issues with linearity as there is not a clear, defined non-linear pattern. Looking at the normal quantile plot, there seem to be issues of normality with the tails of the data, especially the lower tail, as it does not seem to quite fit a normal distribution.

### Methods for Choosing Predictors

R Notebook Group Project 3 - Topic 7: Methods for Choosing Predictors

library(readr)  
  
SleepData = read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

## Rows: 253 Columns: 27

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse  
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA, ClassesMissed, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

head(SleepData)

## # A tibble: 6 x 27  
## Gender ClassYear LarkOwl NumEarlyClass EarlyClass GPA ClassesMissed  
## <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 0 4 Neither 0 0 3.6 0  
## 2 0 4 Neither 2 1 3.24 0  
## 3 0 4 Owl 0 0 2.97 12  
## 4 0 1 Lark 5 1 3.76 0  
## 5 0 4 Owl 0 0 3.2 4  
## 6 1 4 Neither 0 0 3.5 0  
## # ... with 20 more variables: CognitionZscore <dbl>, PoorSleepQuality <dbl>,  
## # DepressionScore <dbl>, AnxietyScore <dbl>, StressScore <dbl>,  
## # DepressionStatus <chr>, AnxietyStatus <chr>, Stress <chr>, DASScore <dbl>,  
## # Happiness <dbl>, AlcoholUse <chr>, Drinks <dbl>, WeekdayBed <dbl>,  
## # WeekdayRise <dbl>, WeekdaySleep <dbl>, WeekendBed <dbl>, WeekendRise <dbl>,  
## # WeekendSleep <dbl>, AverageSleep <dbl>, AllNighter <dbl>

library(leaps)  
  
source("https://raw.githubusercontent.com/JA-McLean/STOR455/master/scripts/ShowSubsets.R")

Find the best multiple regression model predicting GPA using the following four method selections: all subsets, forward regression, backward regression, and stepwise regression.

1. All Subsets Method

all = regsubsets(GPA~., data=SleepData)

## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =  
## force.in, : 1 linear dependencies found

## Reordering variables and trying again:

ShowSubsets(all)

## Gender ClassYear LarkOwlNeither LarkOwlOwl NumEarlyClass EarlyClass  
## 1 ( 1 )   
## 2 ( 1 )   
## 3 ( 1 )   
## 4 ( 1 )   
## 5 ( 1 ) \*   
## 6 ( 1 ) \*   
## 7 ( 1 ) \* \*   
## 8 ( 1 ) \* \*   
## 9 ( 1 ) \* \*   
## ClassesMissed CognitionZscore PoorSleepQuality DepressionScore  
## 1 ( 1 )   
## 2 ( 1 ) \*   
## 3 ( 1 ) \*   
## 4 ( 1 ) \*   
## 5 ( 1 ) \*   
## 6 ( 1 ) \*   
## 7 ( 1 ) \*   
## 8 ( 1 ) \*   
## 9 ( 1 ) \* \*  
## AnxietyScore StressScore DepressionStatusnormal DepressionStatussevere  
## 1 ( 1 )   
## 2 ( 1 )   
## 3 ( 1 ) \*   
## 4 ( 1 ) \*   
## 5 ( 1 ) \*   
## 6 ( 1 ) \*   
## 7 ( 1 ) \*   
## 8 ( 1 ) \*   
## 9 ( 1 ) \*   
## AnxietyStatusnormal AnxietyStatussevere Stressnormal DASScore  
## 1 ( 1 )   
## 2 ( 1 )   
## 3 ( 1 )   
## 4 ( 1 ) \*  
## 5 ( 1 ) \*  
## 6 ( 1 ) \*  
## 7 ( 1 ) \*  
## 8 ( 1 ) \*  
## 9 ( 1 ) \*  
## Happiness AlcoholUseHeavy AlcoholUseLight AlcoholUseModerate Drinks  
## 1 ( 1 ) \*  
## 2 ( 1 ) \*  
## 3 ( 1 ) \*  
## 4 ( 1 ) \*  
## 5 ( 1 ) \*  
## 6 ( 1 ) \*  
## 7 ( 1 ) \*  
## 8 ( 1 ) \*  
## 9 ( 1 ) \*  
## WeekdayBed WeekdayRise WeekdaySleep WeekendBed WeekendRise  
## 1 ( 1 )   
## 2 ( 1 )   
## 3 ( 1 )   
## 4 ( 1 )   
## 5 ( 1 )   
## 6 ( 1 ) \*   
## 7 ( 1 ) \*   
## 8 ( 1 ) \* \*   
## 9 ( 1 ) \* \*  
## WeekendSleep AverageSleep AllNighter Rsq adjRsq Cp  
## 1 ( 1 ) 7.25 6.88 36.44  
## 2 ( 1 ) 12.43 11.73 22.49  
## 3 ( 1 ) 16.16 15.15 13.03  
## 4 ( 1 ) 19.71 18.41 4.11  
## 5 ( 1 ) 23.07 21.52 -4.25  
## 6 ( 1 ) 23.92 22.06 -4.85  
## 7 ( 1 ) 24.41 22.25 -4.37  
## 8 ( 1 ) 24.78 22.31 -3.50  
## 9 ( 1 ) 25.18 22.41 -2.73

ShowSubsetsModel = lm(GPA~Drinks+StressScore, data=SleepData)  
summary(ShowSubsetsModel)

##   
## Call:  
## lm(formula = GPA ~ Drinks + StressScore, data = SleepData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.29843 -0.24307 0.03535 0.23565 0.89868   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.298430 0.052105 63.30 < 2e-16 \*\*\*  
## Drinks -0.024851 0.005946 -4.18 4.04e-05 \*\*\*  
## StressScore 0.008848 0.003062 2.89 0.00419 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3845 on 250 degrees of freedom  
## Multiple R-squared: 0.1025, Adjusted R-squared: 0.09533   
## F-statistic: 14.28 on 2 and 250 DF, p-value: 1.346e-06

1. Forwards Selection Method

full = lm(GPA~., data=SleepData)  
none = lm(GPA~1, data=SleepData)  
MSE = (summary(full)$sigma)^2  
  
step(none, scope=list(upper=full), scale=MSE, direction="forward", trace=FALSE)

##   
## Call:  
## lm(formula = GPA ~ Drinks + CognitionZscore + StressScore + DASScore +   
## ClassYear + WeekdayRise, data = SleepData)  
##   
## Coefficients:  
## (Intercept) Drinks CognitionZscore StressScore   
## 3.79938 -0.01700 0.13410 0.03478   
## DASScore ClassYear WeekdayRise   
## -0.01288 -0.07212 -0.04114

ForwardModel = lm(GPA ~ Drinks + CognitionZscore + StressScore + DASScore +   
 ClassYear + WeekdayRise, data = SleepData)  
summary(ForwardModel)

##   
## Call:  
## lm(formula = GPA ~ Drinks + CognitionZscore + StressScore + DASScore +   
## ClassYear + WeekdayRise, data = SleepData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.21505 -0.19065 0.01833 0.24037 0.84845   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.799377 0.223741 16.981 < 2e-16 \*\*\*  
## Drinks -0.017005 0.005692 -2.987 0.003098 \*\*   
## CognitionZscore 0.134102 0.032478 4.129 4.99e-05 \*\*\*  
## StressScore 0.034780 0.007691 4.522 9.53e-06 \*\*\*  
## DASScore -0.012883 0.003648 -3.531 0.000494 \*\*\*  
## ClassYear -0.072116 0.021940 -3.287 0.001160 \*\*   
## WeekdayRise -0.041143 0.024927 -1.651 0.100113   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3569 on 246 degrees of freedom  
## Multiple R-squared: 0.2392, Adjusted R-squared: 0.2206   
## F-statistic: 12.89 on 6 and 246 DF, p-value: 1.17e-12

1. Backwards Regression Method

step(full, scale=MSE, trace=FALSE)

##   
## Call:  
## lm(formula = GPA ~ Gender + ClassYear + CognitionZscore + DepressionScore +   
## AnxietyScore + StressScore + Drinks + WeekdaySleep, data = SleepData)  
##   
## Coefficients:  
## (Intercept) Gender ClassYear CognitionZscore   
## 3.712435 -0.086172 -0.074506 0.128105   
## DepressionScore AnxietyScore StressScore Drinks   
## -0.009911 -0.018460 0.022651 -0.014895   
## WeekdaySleep   
## -0.029240

BackwardsModel = lm(GPA ~ Gender + ClassYear + CognitionZscore + DepressionScore +   
 AnxietyScore + StressScore + Drinks + WeekdaySleep, data = SleepData)  
summary(BackwardsModel)

##   
## Call:  
## lm(formula = GPA ~ Gender + ClassYear + CognitionZscore + DepressionScore +   
## AnxietyScore + StressScore + Drinks + WeekdaySleep, data = SleepData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.20958 -0.18108 0.02274 0.23235 0.79364   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.712435 0.175763 21.122 < 2e-16 \*\*\*  
## Gender -0.086172 0.052490 -1.642 0.101944   
## ClassYear -0.074506 0.022046 -3.380 0.000845 \*\*\*  
## CognitionZscore 0.128105 0.032871 3.897 0.000126 \*\*\*  
## DepressionScore -0.009911 0.004927 -2.012 0.045354 \*   
## AnxietyScore -0.018460 0.006133 -3.010 0.002890 \*\*   
## StressScore 0.022651 0.004556 4.972 1.25e-06 \*\*\*  
## Drinks -0.014895 0.006041 -2.466 0.014364 \*   
## WeekdaySleep -0.029240 0.019654 -1.488 0.138103   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3568 on 244 degrees of freedom  
## Multiple R-squared: 0.2457, Adjusted R-squared: 0.221   
## F-statistic: 9.937 on 8 and 244 DF, p-value: 5.801e-12

1. Stepwise Regression Method

step(none, scope=list(upper=full), scale=MSE, trace=FALSE)

##   
## Call:  
## lm(formula = GPA ~ Drinks + CognitionZscore + StressScore + DASScore +   
## ClassYear + WeekdayRise, data = SleepData)  
##   
## Coefficients:  
## (Intercept) Drinks CognitionZscore StressScore   
## 3.79938 -0.01700 0.13410 0.03478   
## DASScore ClassYear WeekdayRise   
## -0.01288 -0.07212 -0.04114

StepwiseModel = lm(GPA ~ Drinks + CognitionZscore + StressScore + DASScore +   
 ClassYear + WeekdayRise, data = SleepData)  
summary(StepwiseModel)

##   
## Call:  
## lm(formula = GPA ~ Drinks + CognitionZscore + StressScore + DASScore +   
## ClassYear + WeekdayRise, data = SleepData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.21505 -0.19065 0.01833 0.24037 0.84845   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.799377 0.223741 16.981 < 2e-16 \*\*\*  
## Drinks -0.017005 0.005692 -2.987 0.003098 \*\*   
## CognitionZscore 0.134102 0.032478 4.129 4.99e-05 \*\*\*  
## StressScore 0.034780 0.007691 4.522 9.53e-06 \*\*\*  
## DASScore -0.012883 0.003648 -3.531 0.000494 \*\*\*  
## ClassYear -0.072116 0.021940 -3.287 0.001160 \*\*   
## WeekdayRise -0.041143 0.024927 -1.651 0.100113   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3569 on 246 degrees of freedom  
## Multiple R-squared: 0.2392, Adjusted R-squared: 0.2206   
## F-statistic: 12.89 on 6 and 246 DF, p-value: 1.17e-12

### Interaction Terms

R Notebook

Group Members: Katherine Bacon Kay Youngstrom Ali Floyd Bethany Newcomb

library(readr)  
sleep <- read.csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

1. Interaction Terms

Create a model that predicts GPA based on Average Sleep.

mod1 = lm(GPA~ AverageSleep, data = sleep)  
summary(mod1)

##   
## Call:  
## lm(formula = GPA ~ AverageSleep, data = sleep)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.29018 -0.24395 0.03294 0.26317 0.80407   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.44618 0.21183 16.269 <2e-16 \*\*\*  
## AverageSleep -0.02541 0.02640 -0.962 0.337   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4043 on 251 degrees of freedom  
## Multiple R-squared: 0.003676, Adjusted R-squared: -0.000293   
## F-statistic: 0.9262 on 1 and 251 DF, p-value: 0.3368

Now create a model precting GPA based on Average Sleep, Classes Missed, and an interaction between these two variables.

mod2 = lm(GPA~ AverageSleep + ClassesMissed + AverageSleep\*ClassesMissed, data = sleep)  
summary(mod2)

##   
## Call:  
## lm(formula = GPA ~ AverageSleep + ClassesMissed + AverageSleep \*   
## ClassesMissed, data = sleep)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.33450 -0.23587 0.04836 0.23665 0.77052   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.5748832 0.2608322 13.706 <2e-16 \*\*\*  
## AverageSleep -0.0350662 0.0323790 -1.083 0.280   
## ClassesMissed -0.0310116 0.0454508 -0.682 0.496   
## AverageSleep:ClassesMissed 0.0009665 0.0057040 0.169 0.866   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3987 on 249 degrees of freedom  
## Multiple R-squared: 0.0388, Adjusted R-squared: 0.02722   
## F-statistic: 3.35 on 3 and 249 DF, p-value: 0.01968

Does the ClassesMissed term improve our model? Use Anova or another test we have learned in class. No need for a formal hypothesis.

anova(mod1, mod2)

## Analysis of Variance Table  
##   
## Model 1: GPA ~ AverageSleep  
## Model 2: GPA ~ AverageSleep + ClassesMissed + AverageSleep \* ClassesMissed  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 251 41.037   
## 2 249 39.590 2 1.4466 4.5492 0.01147 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The P-value is smaller than .05 which shows that it is significant. This means that at least one of the coefficients of ClassesMissed is non-zero and therefore makes our model better.

### Polynomial Models

STOR 455: Homework 9 Topic 9

1. Fit a quadratic model using *Happiness* to predict GPA and construct a scatterplot of the data with the quadratic fit.

library(readr)  
Sleep = read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

## Rows: 253 Columns: 27

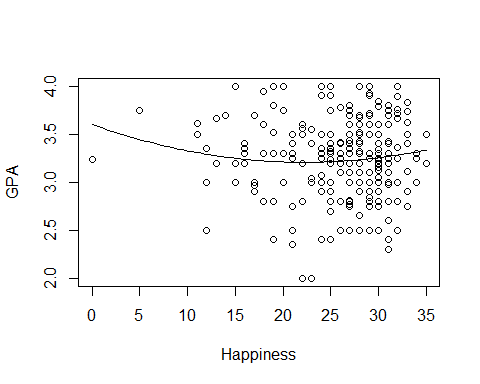
## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse  
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA, ClassesMissed, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

mod1 = lm(GPA~Happiness+I(Happiness^2), data=Sleep)  
summary(mod1)

##   
## Call:  
## lm(formula = GPA ~ Happiness + I(Happiness^2), data = Sleep)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.20924 -0.24386 0.03966 0.27400 0.78894   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.6035793 0.2724150 13.228 <2e-16 \*\*\*  
## Happiness -0.0353191 0.0240513 -1.468 0.143   
## I(Happiness^2) 0.0007902 0.0005248 1.506 0.133   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4041 on 250 degrees of freedom  
## Multiple R-squared: 0.008996, Adjusted R-squared: 0.001068   
## F-statistic: 1.135 on 2 and 250 DF, p-value: 0.3232

plot(GPA~Happiness, data=Sleep)  
curve(summary(mod1)$coef[3,1]\*x^2 + summary(mod1)$coef[2,1]\*x + summary(mod1)$coef[1,1], add=TRUE)



1. Determine if the fit would improve if a cubic term was included.

mod2 = lm(GPA~Happiness+I(Happiness^2)+I(Happiness^3), data=Sleep)  
anova(mod1, mod2)

## Analysis of Variance Table  
##   
## Model 1: GPA ~ Happiness + I(Happiness^2)  
## Model 2: GPA ~ Happiness + I(Happiness^2) + I(Happiness^3)  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 250 40.818   
## 2 249 40.721 1 0.096534 0.5903 0.443

Doing a nested F-test, we have a p-value greater than .05 so adding the cubic term would not significantly improve the model.

### Second Order Models

R Notebook

This is the “Iris” dataset. Originally published at UCI Machine Learning Repository: Iris Data Set, this small dataset from 1936 is often used for testing out machine learning algorithms and visualizations (for example, Scatter Plot). Each row of the table represents an iris flower, including its species and dimensions of its botanical parts, sepal and petal, in centimeters.

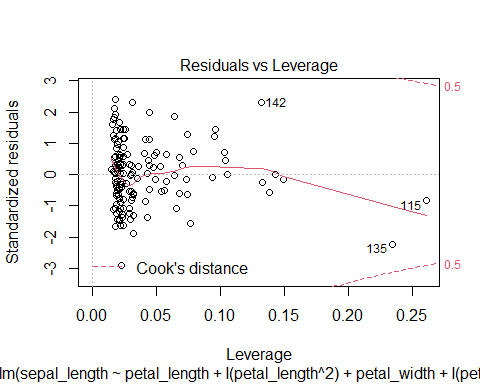
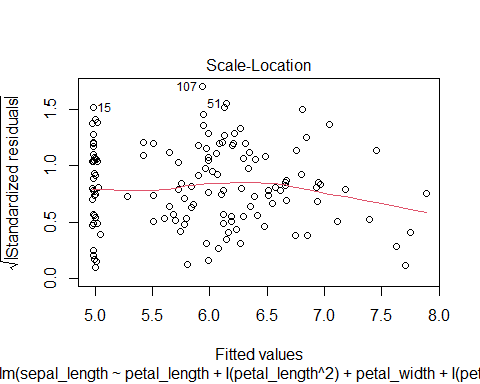
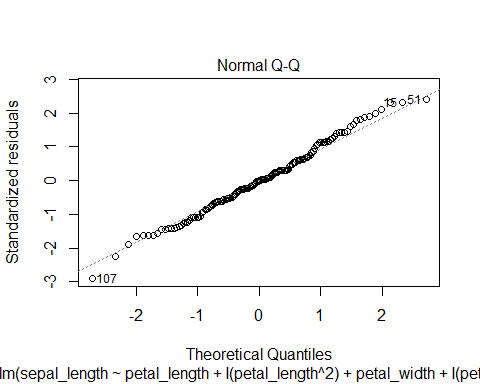
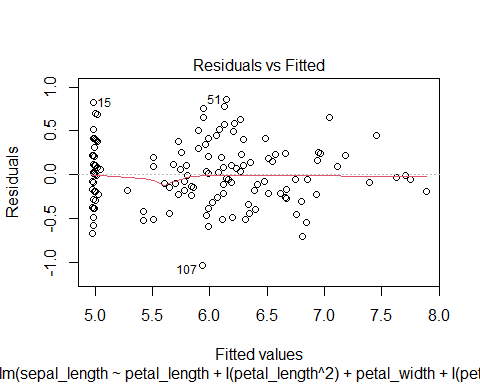
Iris = readr::read\_csv(  
 "https://gist.githubusercontent.com/curran/a08a1080b88344b0c8a7/raw/0e7a9b0a5d22642a06d3d5b9bcbad9890c8ee534/iris.csv", show\_col\_types = FALSE)

Q10: Fit a complete second order model for predicting a sepal\_length based on petal\_length and petal\_width and examine the residuals. Examine the conditions for your model. Discuss any issues that you see in the conditions.

Iris10 = lm(sepal\_length~petal\_length + I(petal\_length^2) + petal\_width + I(petal\_width^2)   
 + petal\_length\*petal\_width, data = Iris)  
summary(Iris10)

##   
## Call:  
## lm(formula = sepal\_length ~ petal\_length + I(petal\_length^2) +   
## petal\_width + I(petal\_width^2) + petal\_length \* petal\_width,   
## data = Iris)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.03541 -0.22444 -0.00044 0.21695 0.85811   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.06985 0.23422 21.646 <2e-16 \*\*\*  
## petal\_length -0.17961 0.30746 -0.584 0.560   
## I(petal\_length^2) 0.07061 0.09779 0.722 0.471   
## petal\_width 0.00520 0.68842 0.008 0.994   
## I(petal\_width^2) -0.27946 0.51092 -0.547 0.585   
## petal\_length:petal\_width 0.13631 0.44288 0.308 0.759   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3613 on 144 degrees of freedom  
## Multiple R-squared: 0.816, Adjusted R-squared: 0.8096   
## F-statistic: 127.7 on 5 and 144 DF, p-value: < 2.2e-16

plot(Iris10)



Overall, the residuals are constantly distributed with no obvious pattern. When fitted value is equal to 5.0, the residuals seem to be more compact.

Linear condition is satisfied since the fitted line is generally linear.

According to the qqplot, the points are generally lying on the fitted line, indicating satisfied normality condition.

### Multicollinearity & VIF

R Notebook

library(readr)  
Sleep = read.csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv  
")  
library(Stat2Data)  
data("CountyHealth") ##from Stat2Data

1. Create a linear model using GPA as the response and AlcoholUse, DepressionStatus, Happiness and Drinks as predictors.

mod1 = lm(GPA~AlcoholUse+Drinks+DepressionStatus+Happiness, data=Sleep)  
summary(mod1)

##   
## Call:  
## lm(formula = GPA ~ AlcoholUse + Drinks + DepressionStatus + Happiness,   
## data = Sleep)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.32735 -0.23308 0.03205 0.23060 0.85236   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.442522 0.148267 23.218 < 2e-16 \*\*\*  
## AlcoholUseHeavy 0.227005 0.152703 1.487 0.138   
## AlcoholUseLight 0.114326 0.087170 1.312 0.191   
## AlcoholUseModerate 0.148325 0.097688 1.518 0.130   
## Drinks -0.035619 0.008522 -4.180 4.07e-05 \*\*\*  
## DepressionStatusnormal -0.060470 0.076218 -0.793 0.428   
## DepressionStatussevere -0.194575 0.144806 -1.344 0.180   
## Happiness -0.002487 0.005029 -0.494 0.621   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.391 on 245 degrees of freedom  
## Multiple R-squared: 0.09052, Adjusted R-squared: 0.06454   
## F-statistic: 3.484 on 7 and 245 DF, p-value: 0.001405

1. Compute the VIF values for the model

library(car)

## Loading required package: carData

vif(mod1)

## GVIF Df GVIF^(1/(2\*Df))  
## AlcoholUse 2.029196 3 1.125176  
## Drinks 2.007391 1 1.416824  
## DepressionStatus 1.279483 2 1.063552  
## Happiness 1.255598 1 1.120535

Since the vif values are less than 5, this suggests there is not multicollinearity between the predictors, where multicollinearity is when one or more of the predictors is strongly correlated with some combination of the other predictors in the set.

##Using CountyHealth dataset 3. Construct a linear model using the number of Doctors in a county, MDs, as the response and number of hospitals in the county, Hospitals, and the number of beds in the county, Beds, as predictors

mod2 = lm(MDs~Hospitals+Beds, data=CountyHealth)  
summary(mod2)

##   
## Call:  
## lm(formula = MDs ~ Hospitals + Beds, data = CountyHealth)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1238.57 -291.40 79.93 200.92 1511.16   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -472.0761 135.3677 -3.487 0.00103 \*\*   
## Hospitals 117.3967 55.1110 2.130 0.03810 \*   
## Beds 1.2600 0.1435 8.778 1.07e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 494.8 on 50 degrees of freedom  
## Multiple R-squared: 0.9304, Adjusted R-squared: 0.9276   
## F-statistic: 334.1 on 2 and 50 DF, p-value: < 2.2e-16

1. Create a correlation matrix to look for multicollinearity between the Hospitals and Bed predictors

cor(CountyHealth[,3:4])

## Hospitals Beds  
## Hospitals 1.0000000 0.9094098  
## Beds 0.9094098 1.0000000

The correlation is 0.9 which is close to 1 so they are highly correlated and therefore explain the same kind of variability i.e. show multicollinearity.

1. Compute the VIF values for the model

library(car)  
vif(mod2)

## Hospitals Beds   
## 5.781221 5.781221

Since the vif values are greater than 5, this suggests multicollinearity. So you should consider removing one of the variables from the model.

### Nested F-Test

None

### Cross Validation Correlation

STOR 455 Cross Validation Correlation Practice

The data in FirstYearGPA contains information on 219 college students.

Variable Description HSGPA | High school GPA SATV | Verbal/critical reading SAT score SATM | Math SAT score Male | 1 for male, 0 for female HU | Number of credit hours earned in humanities courses in high school SS | Number of credit hours earned in social science courses in high school FirstGen | 1 if the student is the first in her or his family to attend college White | 1 for white students, 0 for others CollegeBound | 1 if attended a high school where of students intend to go on to college

library(Stat2Data)  
data("FirstYearGPA")

#Creating training sample  
GPATrain = data.frame(FirstYearGPA[c(0:150),])  
  
#Creating holdout sample  
GPAHoldout = data.frame(FirstYearGPA[c(150:219),])

1. Use the training sample to fit a multiple regression to predict GPA using HSGPA, HU, and White. Compute the predicted GPA for each case in the holdout sample using this model, then compute the residuals for each of the holdout cases.

GPAModel = lm(GPA~HSGPA + HU + factor(White), data=GPATrain)  
fitGPA = predict(GPAModel, newdata=GPAHoldout)  
GPAResid = GPAHoldout$GPA-fitGPA  
GPAResid

## 150 151 152 153 154   
## -0.6289464013 0.1774394377 0.7093971450 0.7199727824 -0.5703145067   
## 155 156 157 158 159   
## -0.4140782175 -0.0382432643 -0.2047372186 -0.6550454987 0.3798684276   
## 160 161 162 163 164   
## 0.4580618768 0.1985923190 0.2664629939 -0.2747119553 0.2351734504   
## 165 166 167 168 169   
## -0.7751471876 -0.2536553431 -0.0018953044 -0.3524895565 -0.3600627205   
## 170 171 172 173 174   
## 0.0141157790 0.7683507724 -0.7868886116 0.2257948302 0.6026659164   
## 175 176 177 178 179   
## -0.0881825223 -0.6084714179 0.4681877699 0.2216631324 -0.0818348277   
## 180 181 182 183 184   
## 0.1975249313 -0.3461618179 -0.6202984124 0.2034320264 -0.4424330626   
## 185 186 187 188 189   
## 0.1059798830 -0.4148954015 -0.4579081385 -0.3659294161 -0.3708417600   
## 190 191 192 193 194   
## -0.0751703326 0.1907910433 -0.5770129223 -0.1300272176 0.0181296694   
## 195 196 197 198 199   
## 0.4429532055 -0.5626864554 0.3596999810 0.2423307665 -0.2842874743   
## 200 201 202 203 204   
## -0.1221447143 0.1743707542 -0.6322232584 -0.5838442333 0.7272796483   
## 205 206 207 208 209   
## -0.7945570625 0.1053929468 -0.1978247748 -0.4516368696 0.4624243698   
## 210 211 212 213 214   
## 0.4690425865 0.4176008337 -0.2202564063 0.0842287668 0.0529403002   
## 215 216 217 218 219   
## 0.0408244100 -0.2425121782 -0.2950094662 -0.0006948518 -0.1901644095

1. Compute the mean and standard deviation for the residuals. Is the mean reasonably close to zero? Is the standard deviation reasonably close to the standard deviation of the error term from the fit to the training sample?

mean(GPAResid)

## [1] -0.06760761

sd(GPAResid)

## [1] 0.4092978

summary(GPAModel)

##   
## Call:  
## lm(formula = GPA ~ HSGPA + HU + factor(White), data = GPATrain)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.09844 -0.23079 0.03517 0.23600 0.82933   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.147478 0.311524 3.683 0.000323 \*\*\*  
## HSGPA 0.466053 0.088393 5.273 4.75e-07 \*\*\*  
## HU 0.015328 0.004091 3.747 0.000257 \*\*\*  
## factor(White)1 0.199174 0.076152 2.615 0.009846 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3773 on 146 degrees of freedom  
## Multiple R-squared: 0.2842, Adjusted R-squared: 0.2695   
## F-statistic: 19.32 on 3 and 146 DF, p-value: 1.319e-10

The mean of -0.06760761 is reasonably close to 0. The standard deviation of the residuals of 0.4092978 is quite close to the standard deviation from our model of 0.3773. This points to our model being a good predictor of GPA in general, not just for our specific training sample.

1. Compute the cross-validation correlation between the actual and predicted GPA values for the cases in the holdout sample.

crosscor = cor(GPAHoldout$GPA, fitGPA)  
crosscor

## [1] 0.5795404

1. Square the cross-validation correlation and subtract from for the training sample to compute the shrinkage. Does it look like the training model works reasonably well for the holdout sample or has there been a considerable drop in the amount of variability explained?

crosscor^2

## [1] 0.3358671

summary(GPAModel)$r.squared - crosscor^2

## [1] -0.05166974

The shrinkage here is actually a negative value, which means that my model created from the training data actually predicts GPA from the holdout model better than the original training data. The values in the holdout model are predicted around 5% better than the training model values, which is not a considerable problem as it is less than 10% and does not point to a significant difference in the effectiveness of model prediction between two datasets.

## Logistic Regression

### Building a Model with a single predictors

Logistic Regression

### Building a Model with a single predictors

library(readr)  
sleep = read.csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

1. Construct a model predicting *AllNighter* by *ClassesMissed*. Include a summary of this model. Use this model to perform a hypothesis test to determine if there is significant evidence of a relationship between *AllNighter* and *ClassesMissed*.

allnighter = glm(AllNighter ~ ClassesMissed, family = binomial, data = sleep)  
summary(allnighter)

##   
## Call:  
## glm(formula = AllNighter ~ ClassesMissed, family = binomial,   
## data = sleep)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.1200 -0.5224 -0.4986 -0.4758 2.1140   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.12138 0.23121 -9.175 <2e-16 \*\*\*  
## ClassesMissed 0.09924 0.04587 2.164 0.0305 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 199.69 on 252 degrees of freedom  
## Residual deviance: 195.45 on 251 degrees of freedom  
## AIC: 199.45  
##   
## Number of Fisher Scoring iterations: 4

Hypothesis test

anova(allnighter, test = "Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: AllNighter  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 252 199.69   
## ClassesMissed 1 4.244 251 195.44 0.03939 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Null hypothesis: The slope to the relationship between *AllNighter* and *ClassesMissed* is zero.

Alternative hypothesis: The slope to the relationship between *AllNighter* and *ClassesMissed* is nonzero.

1. For a student that has missed 9 classes, what is the probability the model predicts that the student has pulled an all nighter?

sleepmod = glm(AllNighter~ClassesMissed, data=sleep, family=binomial)  
predict(sleepmod, data.frame(ClassesMissed = 9), type="response")

## 1   
## 0.2264939

### Building a Model with multiple predictors

Group Project #3

library(readr)  
SleepStudy <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

## Rows: 253 Columns: 27

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse  
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA, ClassesMissed, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

head(SleepStudy)

## # A tibble: 6 x 27  
## Gender ClassYear LarkOwl NumEarlyClass EarlyClass GPA ClassesMissed  
## <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 0 4 Neither 0 0 3.6 0  
## 2 0 4 Neither 2 1 3.24 0  
## 3 0 4 Owl 0 0 2.97 12  
## 4 0 1 Lark 5 1 3.76 0  
## 5 0 4 Owl 0 0 3.2 4  
## 6 1 4 Neither 0 0 3.5 0  
## # ... with 20 more variables: CognitionZscore <dbl>, PoorSleepQuality <dbl>,  
## # DepressionScore <dbl>, AnxietyScore <dbl>, StressScore <dbl>,  
## # DepressionStatus <chr>, AnxietyStatus <chr>, Stress <chr>, DASScore <dbl>,  
## # Happiness <dbl>, AlcoholUse <chr>, Drinks <dbl>, WeekdayBed <dbl>,  
## # WeekdayRise <dbl>, WeekdaySleep <dbl>, WeekendBed <dbl>, WeekendRise <dbl>,  
## # WeekendSleep <dbl>, AverageSleep <dbl>, AllNighter <dbl>

1. Construct a logistic model to predict the *Stress* of a student using *GPA*, *AverageSleep*, and *ClassYear* as the predictor variable.

mod1 = glm(factor(Stress) ~ GPA + AverageSleep + ClassYear, data = SleepStudy, family = binomial)  
summary(mod1)

##   
## Call:  
## glm(formula = factor(Stress) ~ GPA + AverageSleep + ClassYear,   
## family = binomial, data = SleepStudy)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.1843 0.4386 0.6356 0.7592 1.0841   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 5.64974 2.10371 2.686 0.00724 \*\*  
## GPA -1.15547 0.42717 -2.705 0.00683 \*\*  
## AverageSleep -0.01023 0.16400 -0.062 0.95026   
## ClassYear -0.20183 0.14690 -1.374 0.16946   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 267.47 on 252 degrees of freedom  
## Residual deviance: 258.64 on 249 degrees of freedom  
## AIC: 266.64  
##   
## Number of Fisher Scoring iterations: 4

1. For a sophomore student that has a 3.50 GPA, gets an average of 8 hours of sleep a night, what does your model predict is their stress level? Interpret.

'The student has a 0.75 chance of having a high level of stress.'

## [1] "The student has a 0.75 chance of having a high level of stress."

student = data.frame(GPA = 3.50, AverageSleep = 8, ClassYear = 2)  
predict(mod1, student, type="response")

## 1   
## 0.7540053

1. Construct a second logistic model to predict the *Stress* of a student using *GPA*, *AverageSleep*, and *ClassYear* as well as the interaction between *GPA* and *ClassYear* as the predictor variable.

mod2 = glm(factor(Stress) ~ GPA + AverageSleep + ClassYear + GPA\*ClassYear, data = SleepStudy, family = binomial)  
summary(mod2)

##   
## Call:  
## glm(formula = factor(Stress) ~ GPA + AverageSleep + ClassYear +   
## GPA \* ClassYear, family = binomial, data = SleepStudy)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.2103 0.4257 0.6382 0.7649 1.0378   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 6.72521 3.89996 1.724 0.0846 .  
## GPA -1.47051 1.05151 -1.398 0.1620   
## AverageSleep -0.00903 0.16423 -0.055 0.9562   
## ClassYear -0.63892 1.33515 -0.479 0.6323   
## GPA:ClassYear 0.12824 0.38933 0.329 0.7419   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 267.47 on 252 degrees of freedom  
## Residual deviance: 258.53 on 248 degrees of freedom  
## AIC: 268.53  
##   
## Number of Fisher Scoring iterations: 4

### Tests of significance for a model (single or multiple predictors)

STOR 455 Homework #9 Part 1 #16

library(readr)  
Sleep <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

## Rows: 253 Columns: 27

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse  
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA, ClassesMissed, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

Tests of significance for a model (single or multiple predictors)

1. Construct a logistic model to predict *Stress* using *WeekdaySleep* and *WeekendSleep* as predictor variables. Perform a hypothesis test to determine if there is significant evidence of a relationship between *Stress*, *WeekdaySleep*, and *WeekendSleep*. Cite your hypotheses, p-value, and conclusion in context.

H0: B1 = 0 HA: B1 != 0 The p-value estimated p-value for WeekdaySleep is 0.362 and the p-value for WeekendSleep is 0.186. This concludes that both tests are statistically insignificant at any level less than or equal to 0.1, because their p-value are greater than 0.1, and you should fail to reject the null.

mod2 = glm(factor(Stress)~WeekdaySleep+WeekendSleep, data = Sleep, family = binomial)  
summary(mod2)

##   
## Call:  
## glm(formula = factor(Stress) ~ WeekdaySleep + WeekendSleep, family = binomial,   
## data = Sleep)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.0400 0.5892 0.6761 0.7319 1.0052   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 1.5883 1.3020 1.220 0.223  
## WeekdaySleep 0.1173 0.1288 0.911 0.362  
## WeekendSleep -0.1507 0.1139 -1.324 0.186  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 267.47 on 252 degrees of freedom  
## Residual deviance: 265.14 on 250 degrees of freedom  
## AIC: 271.14  
##   
## Number of Fisher Scoring iterations: 4

I think that unless specifically stated, we tend to use the 95 percent confidence interval which means that alpha, or the p-value to determine whether or not the null should be rejected would be .05. This would mean that the test is not statistically significant. I think that for the interpretation it could be put into context, like the slope would be zero or non-zero. I also think that the conclusion would need to be in context by saying that the slope for the logit model predicting stress by weekend and weekday sleep is zero, and that there is not statistically significant evidence from the model to say that there is a non-zero slope for stress with those variables. - Group 11: Haley Hawkins, Rafel Al Ghrary, Alyssa Warnock, Gloria Su.

### Test of significance for nested models

Group\_Review

library(readr)  
sleep = read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

## Rows: 253 Columns: 27

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse  
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA, ClassesMissed, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

head(sleep)

## # A tibble: 6 x 27  
## Gender ClassYear LarkOwl NumEarlyClass EarlyClass GPA ClassesMissed  
## <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 0 4 Neither 0 0 3.6 0  
## 2 0 4 Neither 2 1 3.24 0  
## 3 0 4 Owl 0 0 2.97 12  
## 4 0 1 Lark 5 1 3.76 0  
## 5 0 4 Owl 0 0 3.2 4  
## 6 1 4 Neither 0 0 3.5 0  
## # ... with 20 more variables: CognitionZscore <dbl>, PoorSleepQuality <dbl>,  
## # DepressionScore <dbl>, AnxietyScore <dbl>, StressScore <dbl>,  
## # DepressionStatus <chr>, AnxietyStatus <chr>, Stress <chr>, DASScore <dbl>,  
## # Happiness <dbl>, AlcoholUse <chr>, Drinks <dbl>, WeekdayBed <dbl>,  
## # WeekdayRise <dbl>, WeekdaySleep <dbl>, WeekendBed <dbl>, WeekendRise <dbl>,  
## # WeekendSleep <dbl>, AverageSleep <dbl>, AllNighter <dbl>

#Topic 4: ANOVA for simple linear regression  
#Question:Test the strength of the linear relationship between \_\_GPA\_\_ and \_\_AverageSleep\_\_ using ANOVA for regression. Include hypotheses and your conclusions in the context of the problem.

#Solution:  
mod1 = lm(GPA~AverageSleep, data = sleep)  
anova(mod1)

## Analysis of Variance Table  
##   
## Response: GPA  
## Df Sum Sq Mean Sq F value Pr(>F)  
## AverageSleep 1 0.151 0.15142 0.9262 0.3368  
## Residuals 251 41.037 0.16349

#Null Hypothesis: The coefficient for AverageSleep is 0.  
#Alternative Hypothesis: The coefficient for AverageSleep is not 0.  
#Conclusion: According to the table, p-value is 0.3368, which is a really large number. Therefore, we fail to reject the null hypothesis and come to the conclusion that the coefficient for AverageSleep is 0.

#Topic 17: Test of significance for nested models

#Question\_1:Construct a logistic model to predict if the student has an early class using Stress, AnxietyStatus, and the interaction between them as the predictor variables.

#Solution\_1:  
mod2 = glm(EarlyClass~Stress+AnxietyStatus+Stress\*AnxietyStatus,family = binomial, data = sleep)  
summary(mod2)

##   
## Call:  
## glm(formula = EarlyClass ~ Stress + AnxietyStatus + Stress \*   
## AnxietyStatus, family = binomial, data = sleep)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6651 -1.4944 0.8906 0.8906 1.7941   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.51083 0.36515 1.399 0.1618   
## Stressnormal 0.58779 0.59628 0.986 0.3243   
## AnxietyStatusnormal -0.04082 0.67700 -0.060 0.9519   
## AnxietyStatussevere 0.47000 0.76920 0.611 0.5412   
## Stressnormal:AnxietyStatusnormal -0.33774 0.84118 -0.402 0.6881   
## Stressnormal:AnxietyStatussevere -2.95491 1.43662 -2.057 0.0397 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 322.99 on 252 degrees of freedom  
## Residual deviance: 317.01 on 247 degrees of freedom  
## AIC: 329.01  
##   
## Number of Fisher Scoring iterations: 4

#Question2:Conduct a drop in deviance hypothesis test to determine the effectiveness of the interaction terms in the model constructed in the previous question. Cite your hypotheses, p-value, and conclusion in context.

#Solution2:  
reduced = glm(EarlyClass~Stress+AnxietyStatus,family = binomial, data = sleep)  
anova(mod2, reduced,test = "Chisq")

## Analysis of Deviance Table  
##   
## Model 1: EarlyClass ~ Stress + AnxietyStatus + Stress \* AnxietyStatus  
## Model 2: EarlyClass ~ Stress + AnxietyStatus  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 247 317.01   
## 2 249 322.16 -2 -5.1531 0.07604 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Null Hypothesis:The coefficient for the interaction terms is 0.  
#Alternative Hypothesis:The coefficient for the interaction terms is not 0.  
#Conclusion: According to the table, p-value is 0.07604, which is greater than 0.05. Therefore, we fail to reject the null hypothesis and come to the conclusion that the coefficient for interaction terms is 0.

## Methods for Choosing Predictors

### Assessing linearity of the logit model

Assessing linearity of the logit model

1). Build a logistic model that predicts whether a student has pulled a “AllNighter”, based on their “WeekdaySleep”. Comment on the linearity of the model using a emplogitplot.

library(readr)  
library(Stat2Data)  
Sleep <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

## Rows: 253 Columns: 27

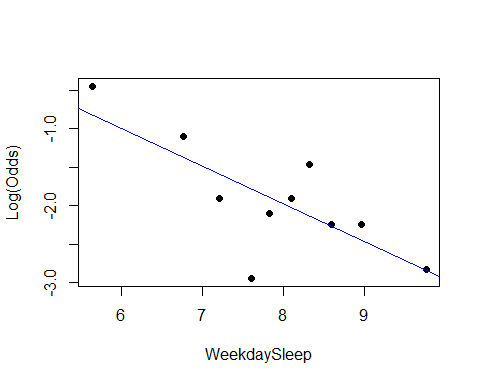
## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse  
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA, ClassesMissed, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

All\_Nighter = glm(AllNighter~WeekdaySleep, family=binomial, data=Sleep)  
summary(All\_Nighter)

##   
## Call:  
## glm(formula = AllNighter ~ WeekdaySleep, family = binomial, data = Sleep)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.5137 -0.5438 -0.4543 -0.3618 2.5528   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.4767 1.1789 2.101 0.035656 \*   
## WeekdaySleep -0.5713 0.1579 -3.617 0.000298 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 199.69 on 252 degrees of freedom  
## Residual deviance: 185.78 on 251 degrees of freedom  
## AIC: 189.78  
##   
## Number of Fisher Scoring iterations: 5

emplogitplot1(AllNighter~WeekdaySleep, data=Sleep, ngroups=10)

 As we can see in our summary output, our WeekdaySleep term looks significant, and looking at the plot it seems as if the points follow a linear pattern, as there is no clear curve.

### Confidence interval for Odds Ratio

Stor455-Hw9

library(readr)  
library(Stat2Data)

data("Election16")  
head(Election16)

## State Abr Income HS BA Adv Dem.Rep TrumpWin  
## 1 Alabama AL 43623 84.3 23.5 8.7 -17 1  
## 2 Alaska AK 72515 92.1 28.0 10.1 -17 1  
## 3 Arizona AZ 50255 86.0 27.5 10.2 -1 1  
## 4 Arkansas AR 41371 84.8 21.1 7.5 -7 1  
## 5 California CA 61818 81.8 31.4 11.6 16 0  
## 6 Colorado CO 60629 90.7 38.1 14.0 -1 0

1. Run a logistic regression model to predict *TrumpWin* for each state using the per capita *Income* of the state. Print a summary of the model.
2. Find a 95% confidence interval for the odds ratio using the model constructed in question 1. (Explain if you wanna)

Answers

1. Run a logistic regression model to predict *TrumpWin* for each state using the per capita *Income* of the state. Print a summary of the model.

Election\_logitmod = glm(TrumpWin~Income,   
 data = Election16, family = binomial)  
summary(Election\_logitmod)

##   
## Call:  
## glm(formula = TrumpWin ~ Income, family = binomial, data = Election16)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.2049 -0.7510 0.4074 0.6566 2.5000   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.118e+01 3.076e+00 3.635 0.000277 \*\*\*  
## Income -1.967e-04 5.582e-05 -3.523 0.000426 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 67.301 on 49 degrees of freedom  
## Residual deviance: 45.923 on 48 degrees of freedom  
## AIC: 49.923  
##   
## Number of Fisher Scoring iterations: 5

1. Find a 95% confidence interval for the odds ratio using the model constructed in question 1. (Explain if you wanna)

exp(confint.default(Election\_logitmod))

## 2.5 % 97.5 %  
## (Intercept) 173.032045 2.980697e+07  
## Income 0.999694 9.999127e-01

We are 95% confident that for every increase of one dollar, the person’s odds that Trump Wins changes by a factor between 0.999694 and 9.999127e-01.

### Interpreting Odds Ratio

R Notebook

Group Members: Katherine Bacon Kay Youngstrom Ali Floyd Bethany Newcomb

library(readr)  
sleep <- read.csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

Create a logistic model that predicts gender based on GPA.

mod1 = glm(Gender~GPA, data=sleep)

how does your model predict the odds of being male or female will change based on GPA?

summary(mod1)

##   
## Call:  
## glm(formula = Gender ~ GPA, data = sleep)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.7730 -0.3865 -0.2675 0.5243 0.8217   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.36768 0.24322 5.623 4.98e-08 \*\*\*  
## GPA -0.29734 0.07441 -3.996 8.47e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.2280315)  
##   
## Null deviance: 60.877 on 252 degrees of freedom  
## Residual deviance: 57.236 on 251 degrees of freedom  
## AIC: 347.97  
##   
## Number of Fisher Scoring iterations: 2

exp(summary(mod1)$coef[2])

## [1] 0.7427898

For every 1 point increase in GPA, the odds of being male increase by a factor of .74.

### Predicting Probabilities

Homework 9

Alyssa Warnock, Helen Johnston, Ashley Wade, Hunter Barbee

Sleep<-read.csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")  
head(Sleep)

## Gender ClassYear LarkOwl NumEarlyClass EarlyClass GPA ClassesMissed  
## 1 0 4 Neither 0 0 3.60 0  
## 2 0 4 Neither 2 1 3.24 0  
## 3 0 4 Owl 0 0 2.97 12  
## 4 0 1 Lark 5 1 3.76 0  
## 5 0 4 Owl 0 0 3.20 4  
## 6 1 4 Neither 0 0 3.50 0  
## CognitionZscore PoorSleepQuality DepressionScore AnxietyScore StressScore  
## 1 -0.26 4 4 3 8  
## 2 1.39 6 1 0 3  
## 3 0.38 18 18 18 9  
## 4 1.39 9 1 4 6  
## 5 1.22 9 7 25 14  
## 6 -0.04 6 14 8 28  
## DepressionStatus AnxietyStatus Stress DASScore Happiness AlcoholUse Drinks  
## 1 normal normal normal 15 28 Moderate 10  
## 2 normal normal normal 4 25 Moderate 6  
## 3 moderate severe normal 45 17 Light 3  
## 4 normal normal normal 11 32 Light 2  
## 5 normal severe normal 46 15 Moderate 4  
## 6 moderate moderate high 50 22 Abstain 0  
## WeekdayBed WeekdayRise WeekdaySleep WeekendBed WeekendRise WeekendSleep  
## 1 25.75 8.70 7.70 25.75 9.50 5.88  
## 2 25.70 8.20 6.80 26.00 10.00 7.25  
## 3 27.44 6.55 3.00 28.00 12.59 10.09  
## 4 23.50 7.17 6.77 27.00 8.00 7.25  
## 5 25.90 8.67 6.09 23.75 9.50 7.00  
## 6 23.80 8.95 9.05 26.00 10.75 9.00  
## AverageSleep AllNighter  
## 1 7.18 0  
## 2 6.93 0  
## 3 5.02 0  
## 4 6.90 0  
## 5 6.35 0  
## 6 9.04 0

1. For an female who has a stress score of 15, what does your model predict is the probability they have pulled an all nighter?

mod1=glm(AllNighter~Gender+AnxietyScore,family="binomial", data = Sleep)  
summary(mod1)

##   
## Call:  
## glm(formula = AllNighter ~ Gender + AnxietyScore, family = "binomial",   
## data = Sleep)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.9740 -0.6274 -0.4097 -0.3440 2.4140   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.85800 0.40034 -7.139 9.4e-13 \*\*\*  
## Gender 1.27236 0.39675 3.207 0.00134 \*\*   
## AnxietyScore 0.06036 0.03428 1.761 0.07824 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 199.69 on 252 degrees of freedom  
## Residual deviance: 187.23 on 250 degrees of freedom  
## AIC: 193.23  
##   
## Number of Fisher Scoring iterations: 5

newx=data.frame(Gender=0,AnxietyScore=15)  
predict(mod1,newx,type="response")

## 1   
## 0.1242689

For a female with an anxiety score of 15, the model predicts the probability of having had an all-nighter this semester is 0.1242689

## ANOVA for Means

### One Way ANOVA

R Notebook

This is the “Iris” dataset. Originally published at UCI Machine Learning Repository: Iris Data Set, this small dataset from 1936 is often used for testing out machine learning algorithms and visualizations (for example, Scatter Plot). Each row of the table represents an iris flower, including its species and dimensions of its botanical parts, sepal and petal, in centimeters.

Iris = readr::read\_csv(  
 "https://gist.githubusercontent.com/curran/a08a1080b88344b0c8a7/raw/0e7a9b0a5d22642a06d3d5b9bcbad9890c8ee534/iris.csv", show\_col\_types = FALSE)

Q23: Construct a one-way ANOVA model for the sepal length for each species of flower, using the species type as the predictor. Include the output showing the ANOVA table. Comment on what this output tells you about the sepal length across different spieces.

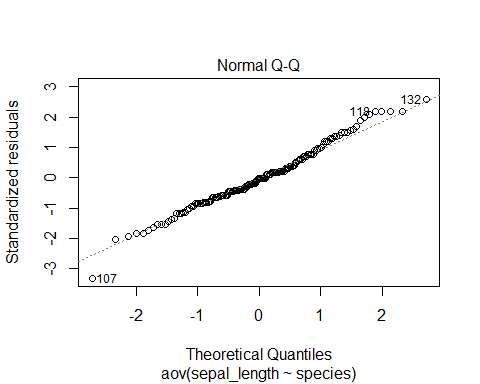
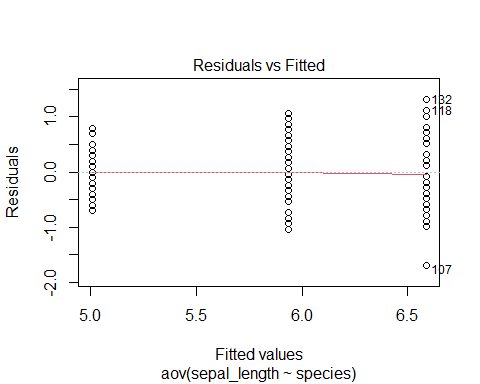
Iris23 = aov(sepal\_length~species, data = Iris)  
summary(Iris23)

## Df Sum Sq Mean Sq F value Pr(>F)   
## species 2 63.21 31.606 119.3 <2e-16 \*\*\*  
## Residuals 147 38.96 0.265   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

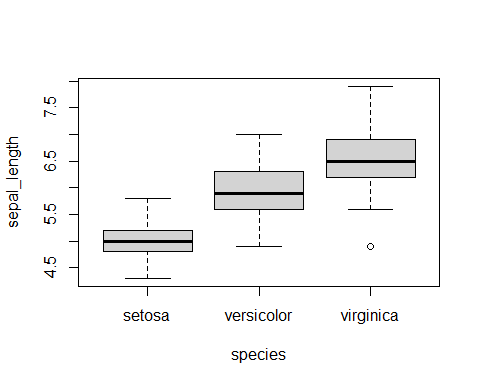
round(tapply(Iris$sepal\_length,Iris$species,sd),2)

## setosa versicolor virginica   
## 0.35 0.52 0.64

plot(Iris23, 1:2)



boxplot(sepal\_length~species, data = Iris)

 Since p-value is less than 0.05, we conclude that species is a significant predictor for sepal length. The spread of residuals is slightly increasing at each specie, however, there’s no specie that has a standard deviation of more than double of the other. From the qqplot, we can see that the points are generally following the fitted line as well. Therefore, there’s no big issue with constant variance and normalityï¼Œ which means species is an useful predictor in predicting sepal length.

### Two Way ANOVA

R Notebook

library(readr)  
  
Sleepstudy = read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

## Rows: 253 Columns: 27

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse  
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA, ClassesMissed, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

Sleepstudy

## # A tibble: 253 x 27  
## Gender ClassYear LarkOwl NumEarlyClass EarlyClass GPA ClassesMissed  
## <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 0 4 Neither 0 0 3.6 0  
## 2 0 4 Neither 2 1 3.24 0  
## 3 0 4 Owl 0 0 2.97 12  
## 4 0 1 Lark 5 1 3.76 0  
## 5 0 4 Owl 0 0 3.2 4  
## 6 1 4 Neither 0 0 3.5 0  
## 7 1 2 Lark 2 1 3.35 2  
## 8 0 2 Lark 0 0 3 0  
## 9 0 1 Neither 2 1 4 0  
## 10 0 4 Neither 2 1 2.9 0  
## # ... with 243 more rows, and 20 more variables: CognitionZscore <dbl>,  
## # PoorSleepQuality <dbl>, DepressionScore <dbl>, AnxietyScore <dbl>,  
## # StressScore <dbl>, DepressionStatus <chr>, AnxietyStatus <chr>,  
## # Stress <chr>, DASScore <dbl>, Happiness <dbl>, AlcoholUse <chr>,  
## # Drinks <dbl>, WeekdayBed <dbl>, WeekdayRise <dbl>, WeekdaySleep <dbl>,  
## # WeekendBed <dbl>, WeekendRise <dbl>, WeekendSleep <dbl>,  
## # AverageSleep <dbl>, AllNighter <dbl>

summary(Sleepstudy)

## Gender ClassYear LarkOwl NumEarlyClass   
## Min. :0.0000 Min. :1.000 Length:253 Min. :0.000   
## 1st Qu.:0.0000 1st Qu.:2.000 Class :character 1st Qu.:0.000   
## Median :0.0000 Median :2.000 Mode :character Median :2.000   
## Mean :0.4032 Mean :2.478 Mean :1.735   
## 3rd Qu.:1.0000 3rd Qu.:3.000 3rd Qu.:3.000   
## Max. :1.0000 Max. :4.000 Max. :5.000   
## EarlyClass GPA ClassesMissed CognitionZscore   
## Min. :0.000 Min. :2.000 Min. : 0.000 Min. :-1.62e+00   
## 1st Qu.:0.000 1st Qu.:3.000 1st Qu.: 0.000 1st Qu.:-4.80e-01   
## Median :1.000 Median :3.300 Median : 1.000 Median :-1.00e-02   
## Mean :0.664 Mean :3.244 Mean : 2.209 Mean :-3.95e-05   
## 3rd Qu.:1.000 3rd Qu.:3.500 3rd Qu.: 3.000 3rd Qu.: 4.40e-01   
## Max. :1.000 Max. :4.000 Max. :20.000 Max. : 1.96e+00   
## PoorSleepQuality DepressionScore AnxietyScore StressScore   
## Min. : 1.000 Min. : 0.000 Min. : 0.000 Min. : 0.000   
## 1st Qu.: 4.000 1st Qu.: 1.000 1st Qu.: 1.000 1st Qu.: 3.000   
## Median : 6.000 Median : 3.000 Median : 4.000 Median : 8.000   
## Mean : 6.257 Mean : 5.202 Mean : 5.372 Mean : 9.466   
## 3rd Qu.: 8.000 3rd Qu.: 7.000 3rd Qu.: 8.000 3rd Qu.:14.000   
## Max. :18.000 Max. :35.000 Max. :26.000 Max. :37.000   
## DepressionStatus AnxietyStatus Stress DASScore   
## Length:253 Length:253 Length:253 Min. : 0.00   
## Class :character Class :character Class :character 1st Qu.: 7.00   
## Mode :character Mode :character Mode :character Median :16.00   
## Mean :20.04   
## 3rd Qu.:28.00   
## Max. :82.00   
## Happiness AlcoholUse Drinks WeekdayBed   
## Min. : 0.00 Length:253 Min. : 0.000 Min. :21.80   
## 1st Qu.:24.00 Class :character 1st Qu.: 3.000 1st Qu.:24.20   
## Median :28.00 Mode :character Median : 5.000 Median :24.80   
## Mean :26.11 Mean : 5.569 Mean :24.85   
## 3rd Qu.:30.00 3rd Qu.: 8.000 3rd Qu.:25.50   
## Max. :35.00 Max. :24.000 Max. :29.10   
## WeekdayRise WeekdaySleep WeekendBed WeekendRise   
## Min. : 5.500 Min. : 3.000 Min. :21.50 Min. : 5.25   
## 1st Qu.: 8.000 1st Qu.: 7.200 1st Qu.:24.88 1st Qu.: 9.25   
## Median : 8.500 Median : 7.950 Median :25.50 Median :10.25   
## Mean : 8.586 Mean : 7.866 Mean :25.58 Mean :10.20   
## 3rd Qu.: 9.150 3rd Qu.: 8.600 3rd Qu.:26.25 3rd Qu.:11.00   
## Max. :12.020 Max. :10.970 Max. :30.25 Max. :15.00   
## WeekendSleep AverageSleep AllNighter   
## Min. : 4.000 Min. : 4.950 Min. :0.0000   
## 1st Qu.: 7.250 1st Qu.: 7.430 1st Qu.:0.0000   
## Median : 8.250 Median : 8.000 Median :0.0000   
## Mean : 8.217 Mean : 7.966 Mean :0.1344   
## 3rd Qu.: 9.250 3rd Qu.: 8.590 3rd Qu.:0.0000   
## Max. :12.750 Max. :10.620 Max. :1.0000

1. Construct a two-way ANOVA model for the mean GPA in the dataset, *GPA*, using the *LarkOwl* and *Stress* as the predictors. Include the output showing the ANOVA table. Comment on what this output tells you about the GPA across college stduents. Provide formal hypotheses, p-values, and conclusions.

amodA = aov(GPA~LarkOwl+Stress, data=Sleepstudy)  
summary(amodA)

## Df Sum Sq Mean Sq F value Pr(>F)   
## LarkOwl 2 0.43 0.2150 1.347 0.262   
## Stress 1 1.02 1.0215 6.401 0.012 \*  
## Residuals 249 39.74 0.1596   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Hypotheses:

H0: μ1 = μ2 = μ3 (1= Lark, 2 = Owl, 3= Neither) (Mean GPA across LarkOwl status are all same) Ha: Some μi != μk (There is at least one mean GPA of one LarkOwl status that is different from the mean GPA of another LarkOwl status)

P-value: 0.262

Conclusion: Because p-value is not small enough (0.262), we fail to reject the null hypothesis. Mean GPA across LarkOwl statuses are all same.

Hypotheses:

H0: α1 = α2 = 0 (1 = Normal, 2 = High) (The affect for both stress scores are zeros) Ha: One of αi != 0 (There is at least one stress score that does not have zero for the affect)

p-value: 0.012

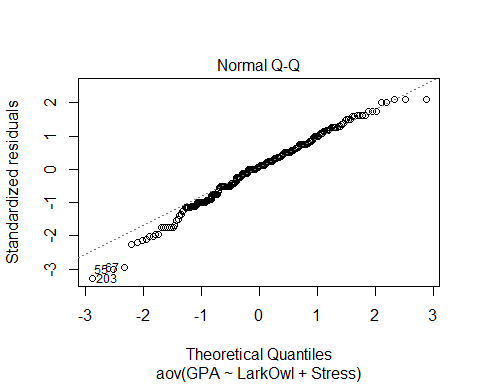
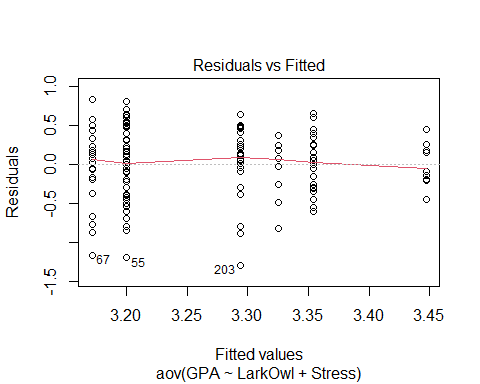
Conclusion: Because our p-value is small enough (0.012), we reject the null hypothesis. There is at least one stress score that does not have zero for the affect.

1. Construct residual plots and comment on the conditions of equality of variances and normality of residuals for the model that you created.

amodA

## Call:  
## aov(formula = GPA ~ LarkOwl + Stress, data = Sleepstudy)  
##   
## Terms:  
## LarkOwl Stress Residuals  
## Sum of Squares 0.43001 1.02147 39.73688  
## Deg. of Freedom 2 1 249  
##   
## Residual standard error: 0.399482  
## Estimated effects may be unbalanced

plot(amodA, 1:2)



tapply(Sleepstudy$GPA, Sleepstudy$LarkOwl, sd)

## Lark Neither Owl   
## 0.3972581 0.3935868 0.4413586

tapply(Sleepstudy$GPA, Sleepstudy$Stress, sd)

## high normal   
## 0.3513277 0.4123180

Variances: In general, we have constant variances, but as fitted values increase, the variances (vertical distances) are decreasing. In addition, the standard deviations for different combinations of LarkOwl and Stress predictors are roughly similar.

Normality: The normal Q-Q line shows some skewness on the left tail.

### Two Way ANOVA with Interaction

### Tukey HSD

library(readr)  
data=read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

## Rows: 253 Columns: 27

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse  
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA, ClassesMissed, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

1. Construct an ANOVA model for the mean AverageSleep value using LarkOwl and DepressionStatus as factors along with the interaction term and show the summary output. Is there evidence for a significant difference in mean AverageSleep values for any of the predictors?

amod1= aov(GPA~factor(ClassYear)+DepressionStatus+factor(ClassYear)\*DepressionStatus,data=data)  
summary(amod1)

## Df Sum Sq Mean Sq F value Pr(>F)   
## factor(ClassYear) 3 5.13 1.7109 11.606 3.93e-07 \*\*\*  
## DepressionStatus 2 0.22 0.1105 0.750 0.474   
## factor(ClassYear):DepressionStatus 6 0.31 0.0514 0.349 0.910   
## Residuals 241 35.53 0.1474   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

There is evidence for a significant difference in the mean AverageSleep value between class years but not for depression status or the interaction term.

1. Perform pairwise comparisons using Tukey HSD methods to show where any significant differences may occur among factors that showed a significant difference in part a).

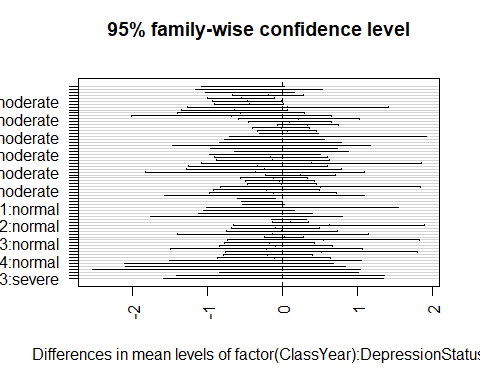
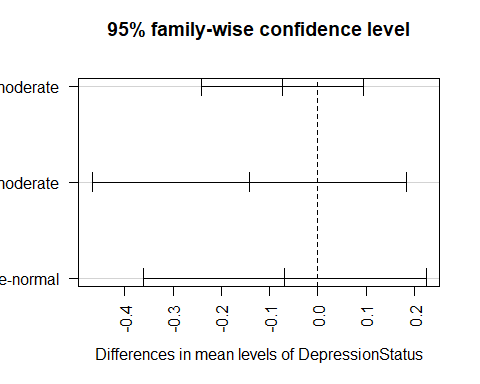
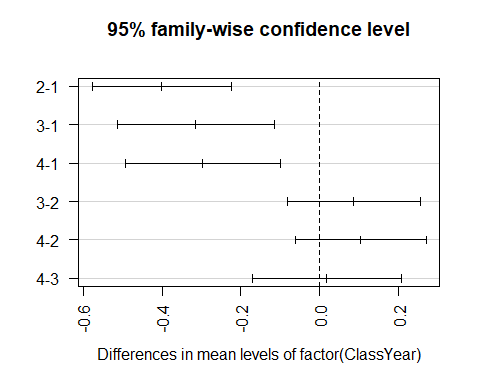
hsd = TukeyHSD(amod1)  
hsd$`factor(ClassYear)`[,4][hsd$`factor(ClassYear)`[,4]<.05] #significant differences in means between class groups

## 2-1 3-1 4-1   
## 9.727889e-08 3.285752e-04 6.721431e-04

There appear to be significant differences in mean AverageSleep values between class years 1:2, 1:3, and 1:4.

1. Produce 95% family-wise confidence level plots for the ANOVA model created in part a).

hsd=TukeyHSD(amod1)  
 plot(hsd,las=2)



### Levene’s Test

- Levene test tries to see if difference in group means when comparing things

- Useful, but may have some issues with

- asses how likely we get these diff by chance if we assume the two groups are similar

- TUkey looks at the group means to see wehre the differences are

- Leven only looks at evidence

- Only looked at levene in a one way ANOVA

- Tests to see if htere are difference in the variances,

More noets on Levene's Test

- Basically does an anova of differences rather than anova of abs values

- main point: see if looking bt the categories of teh variables are the abs deviances

- just testing to see if the variances are difference

- Tests for constant variance

- Use to check conditions