Contents

[STOR 455 - Class 17 – Comparing two regression lines 1](#_Toc89874728)

[STOR 455 - Class Coding categorial variables 13](#_Toc89874729)

[STOR 455 - Class 19 – Testing a subset of predictors 25](#_Toc89874730)

[STOR 455 - Class 20 – new predictors from old 34](#_Toc89874731)

[STOR 455 - Class 21 - R Cross Validation 59](#_Toc89874732)

[STOR 455 Class 22 Influential Points in Multiple Regression 69](#_Toc89874733)

[STOR 455 Homework #4 97](#_Toc89874734)

## 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

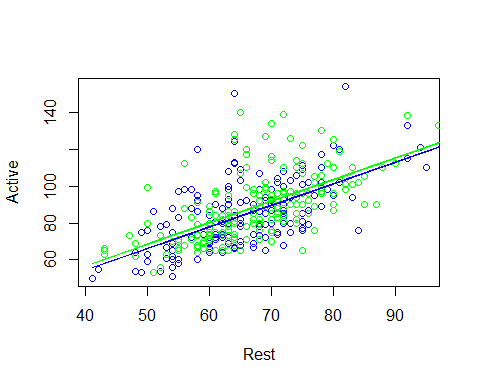
modelP2=lm(Active~Rest+Sex, data=Pulse) # model with quanti and categorical variables  
summary(modelP2)

##   
## 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)

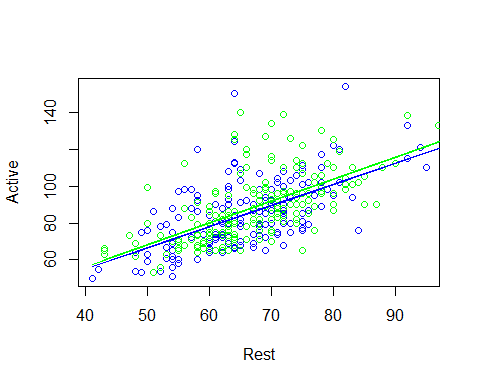
##   
## Call:  
## lm(formula = Active ~ Rest, data = Females)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.822 -9.088 -3.177 6.010 54.000   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.1527 7.0198 1.304 0.194   
## Rest 1.1823 0.1016 11.633 <2e-16 \*\*\*  
## ---  
## 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

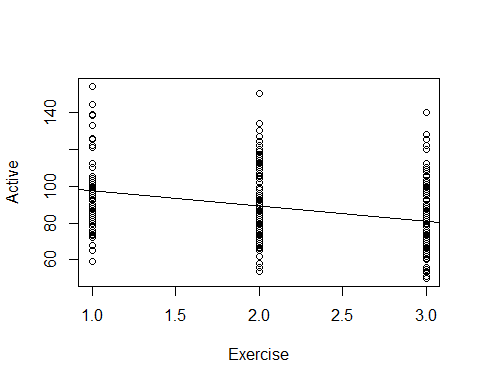
-Try a model to predict Y=Active pulse rates using X=Exercise. How should the coefficients be interpreted?

\_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

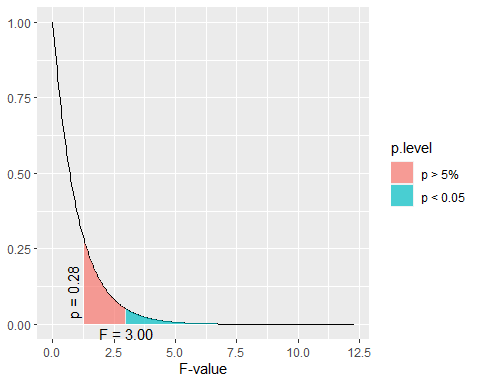
## [1] 1.274553

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

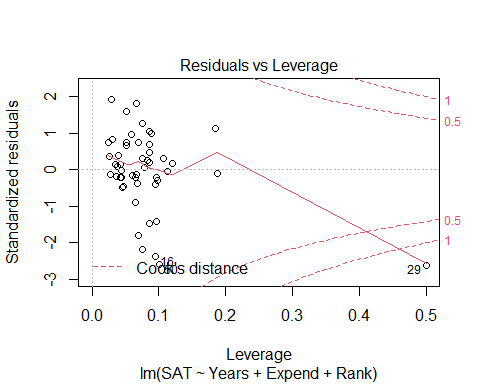
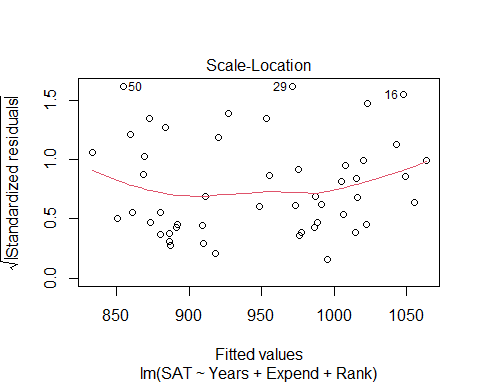
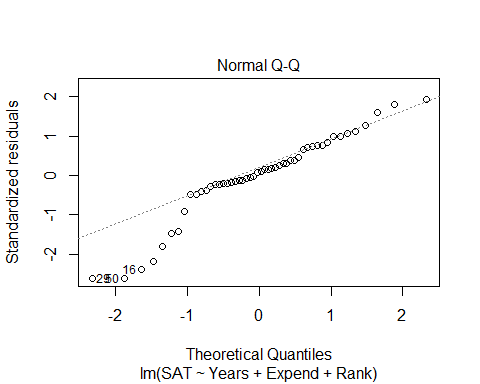
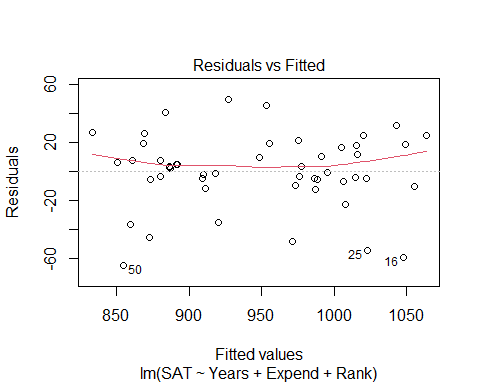
library(readr)  
library(leaps)  
  
StateSAT <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/StateSAT.csv")  
  
source("https://raw.githubusercontent.com/JA-McLean/STOR455/master/scripts/ShowSubsets.R")

**Example: State SAT** - Model #1: Y=SAT vs. X=Takers

mod = lm(SAT~Years+Expend+Rank, data=StateSAT)  
summary(mod)

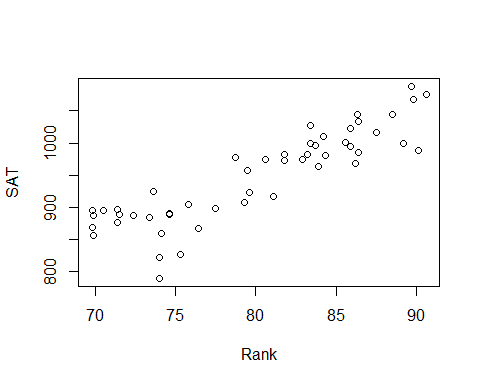
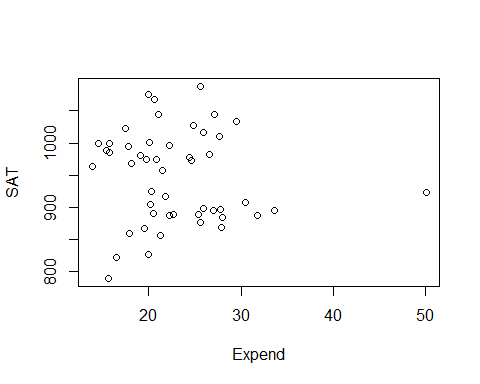
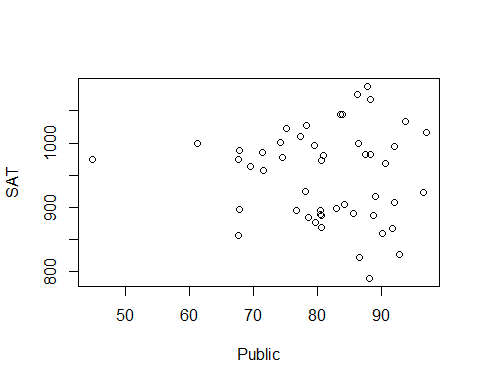
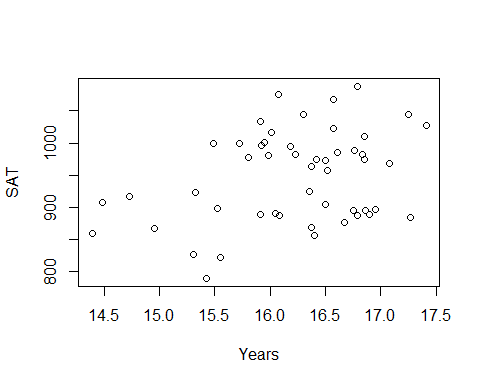
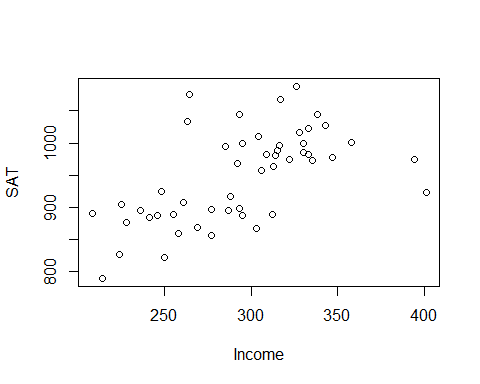
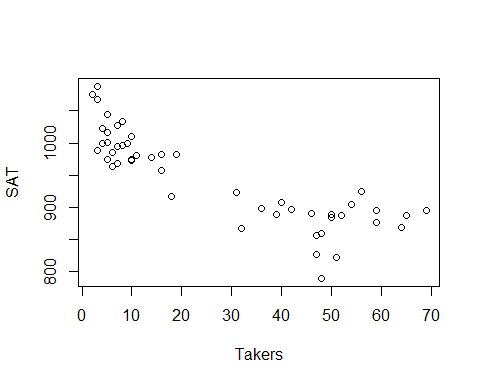
##   
## Call:  
## lm(formula = SAT ~ Years + Expend + Rank, data = StateSAT)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -64.802 -6.798 2.169 17.525 49.706   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -303.7243 97.8415 -3.104 0.00326 \*\*   
## Years 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])

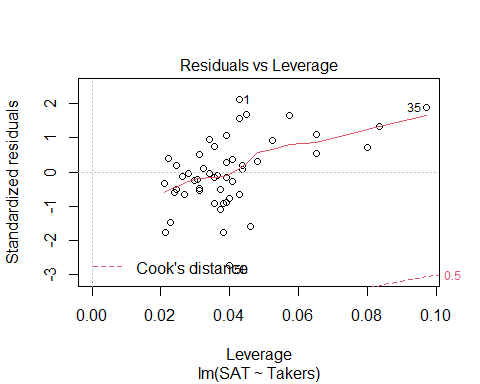
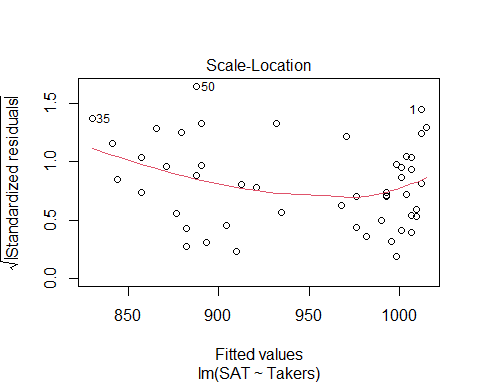
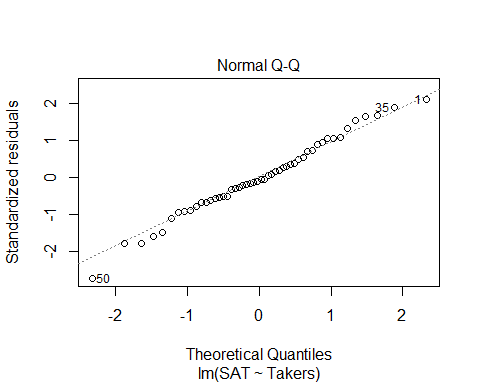
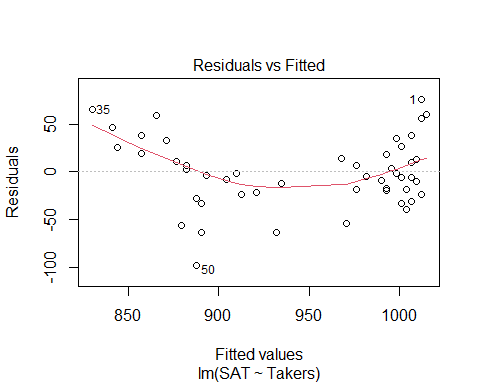


# Here, we see the correlation between each of the variables in the dataset (each of teh numerical variables in teh dataset)   
# Expend is iffy   
# Years is good  
# Takers looks like the best pattern, but we dont have it because it doesn't look lienar

# 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

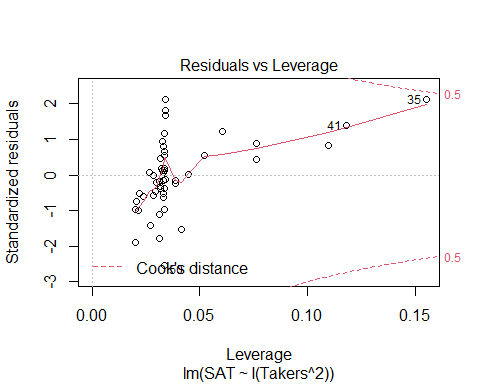
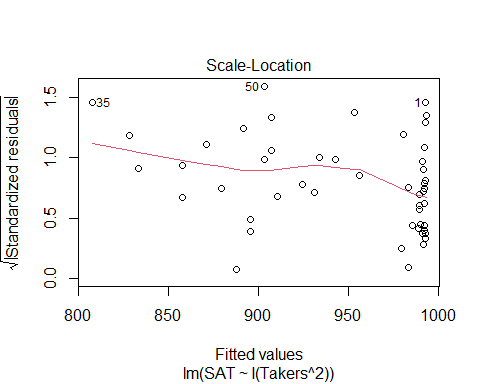
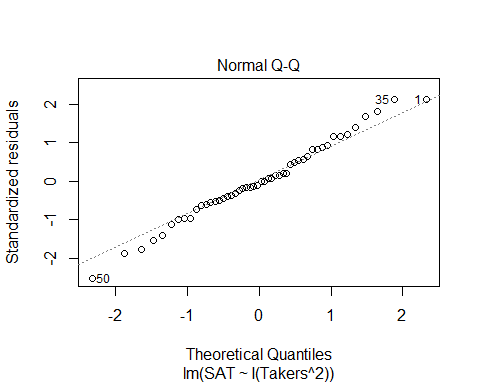
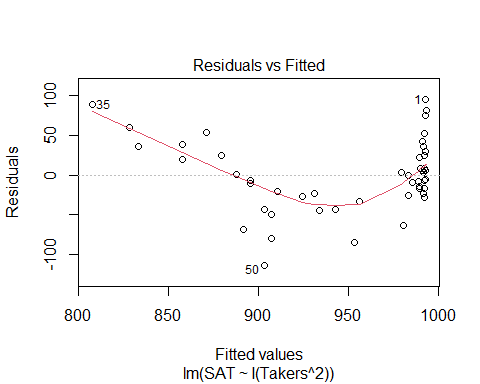
* **LINEAR** – 𝑌=𝛽\_𝑜+𝛽\_1 𝑋+𝜀
* **QUADRATIC** – 𝑌=𝛽\_𝑜+𝛽\_1 𝑋+𝛽\_2 𝑋^2+𝜀
* **CUBIC** –𝑌=𝛽\_𝑜+𝛽\_1 𝑋+𝛽\_2 𝑋^2+𝛽\_3 𝑋^3+𝜀

**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)



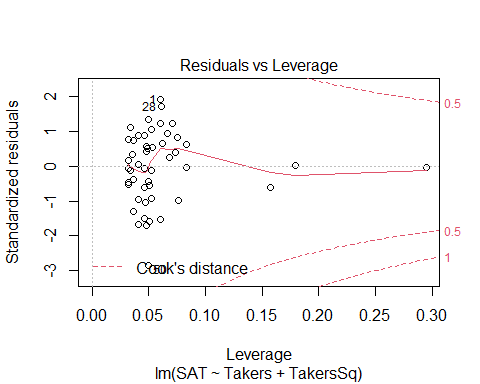
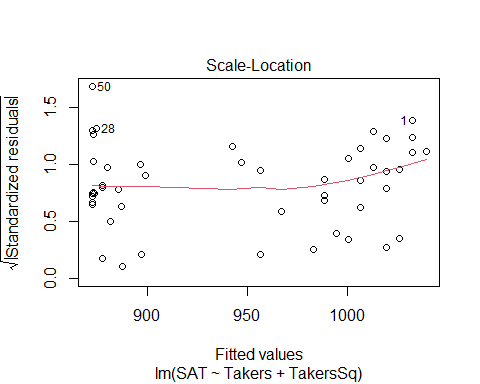
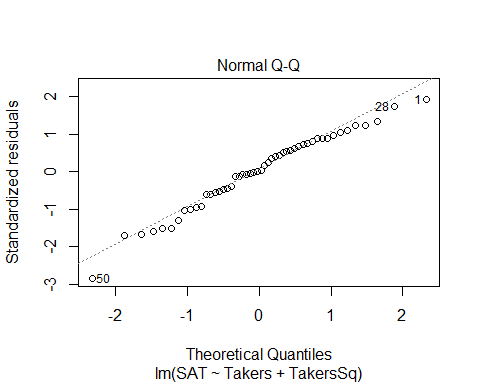
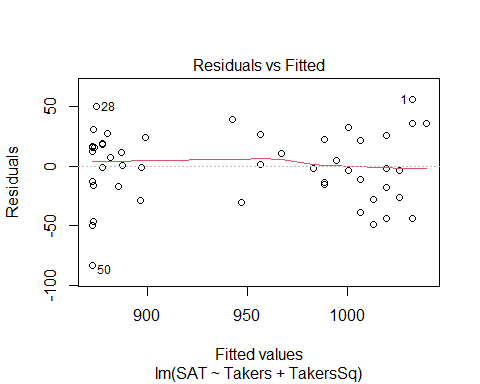
**Polynomial Regression in R** - We can add as many powers we want but then we might be over fitting, so that’s not always best

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)

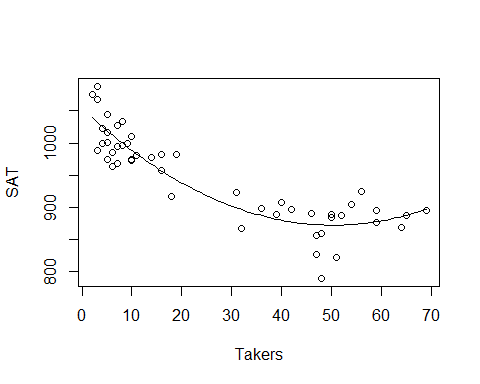
## Analysis of Variance Table  
##   
## Response: SAT  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Takers 1 181024 181024 202.089 < 2.2e-16 \*\*\*  
## TakersSq 1 22886 22886 25.549 6.992e-06 \*\*\*  
## Residuals 47 42101 896   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 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

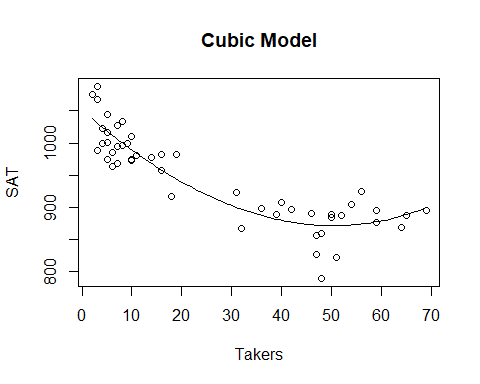
Would a Cubic work better?

#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)

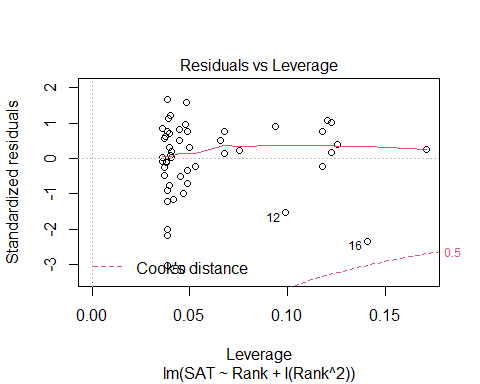
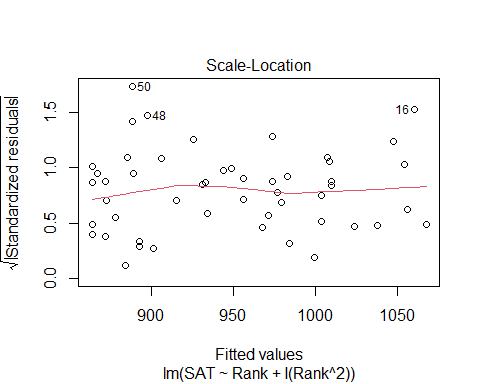
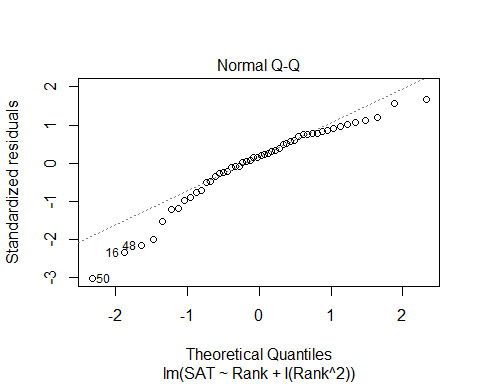
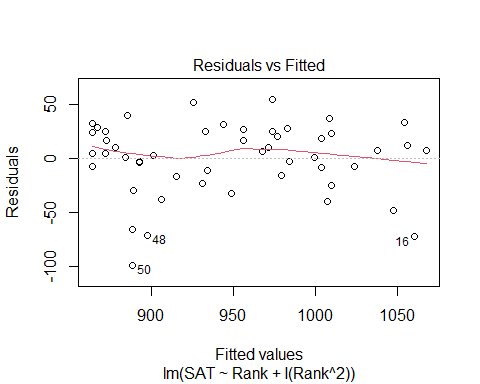
## Takers I(Takers^2) I(Takers^3)   
## 147.2369 678.9666 222.6922

# shows that there is a high multicollinearity with takers

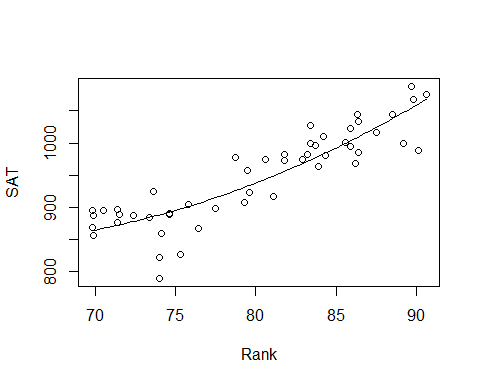
modSATquad4 = lm(SAT~ Rank+ I(Rank^2), data=StateSAT)  
summary(modSATquad4)

##   
## 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

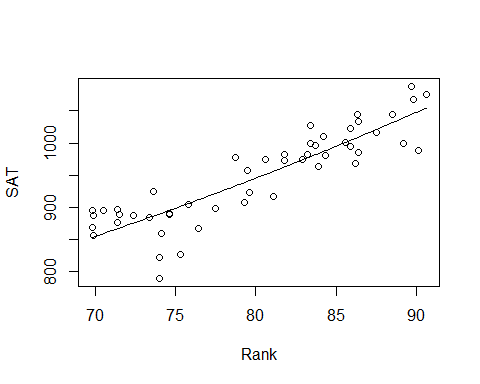
# Model looks pretty good, rank and rank^@ are not sig; btut ehre is a hug amount of multicollinearity   
# P value is close to 0, but the individual tests say things different because of the multicollinearity   
  
plot(modSATquad4)



# 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)

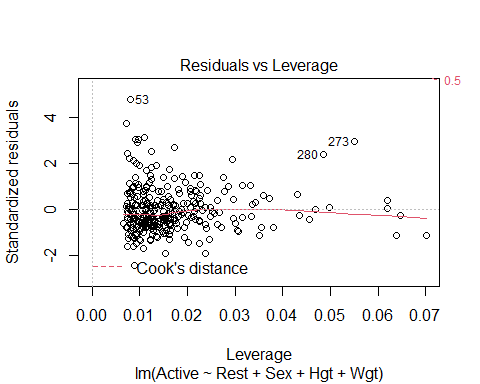
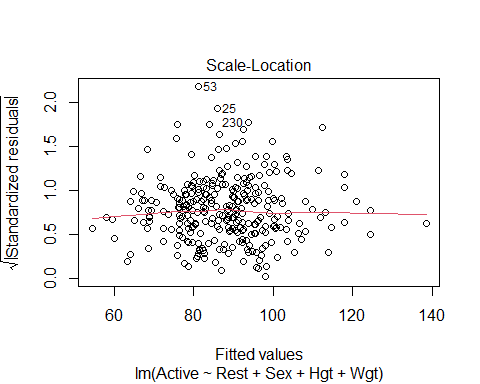
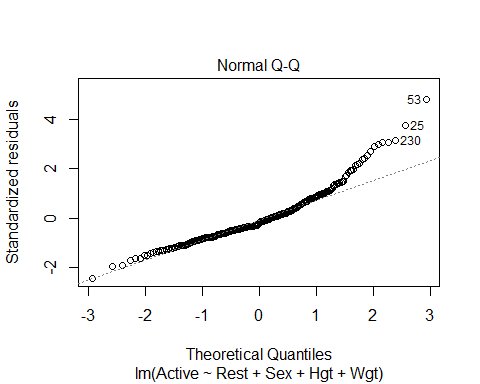
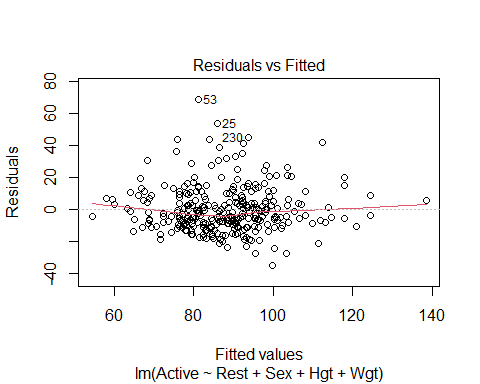
set.seed(12345) # Only set the seed to get the same results at the end, if you want actual random, then don't set seed  
# Want to do it randomly because we dont want to have a certain connection with the different rows  
rows <- sample(nrow(Pulse)) # Counts rows in pulse and takes 374 values without 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

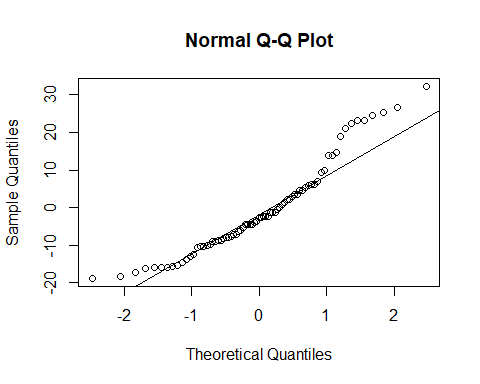
#predict active heart rates for data in the holdout sample with the model made from the training data  
fitActive=predict(PulseTrainMod,newdata=PulseHoldout)  
# Just all teh predctions   
  
#Actual active heart rates in holdout sample minus their predicted values  
# how far off are we?   
holdoutresid=PulseHoldout$Active - 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

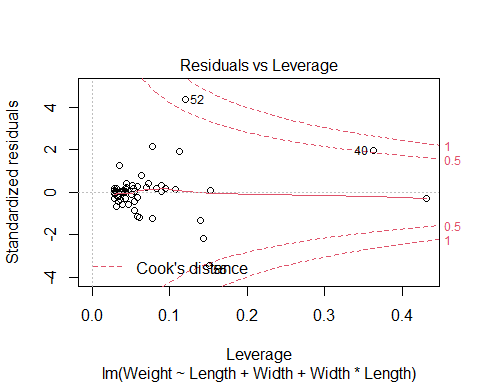
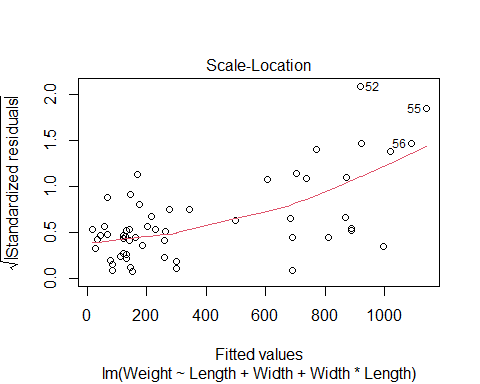
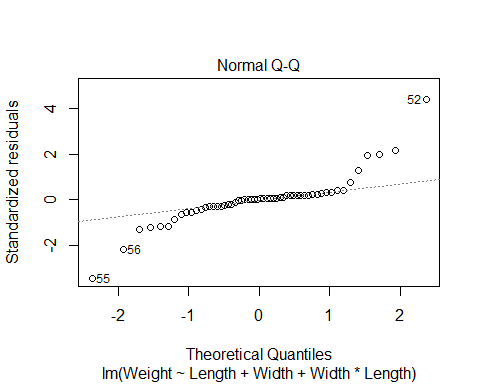
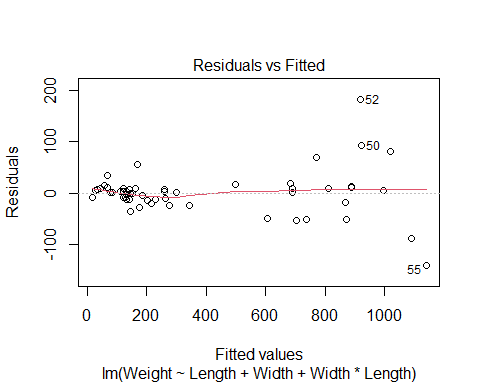
**Leverage in Multiple Regression: Perch** Perch\_lm = lm(Weight~Length+Width+Width\*Length, data=Perch)

* 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

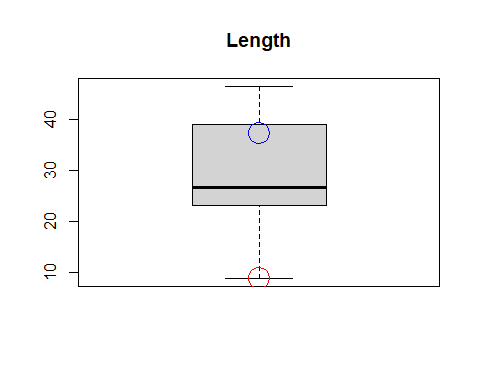
# Tells us which values in the dataset are over the highest boundary for leverage   
# Will give you waht is true   
Lev\_indices = which(hatvalues(Perch\_lm) >= 3\*(3+1)/56)  
  
#Two cases with high leverage  
# will tell you which of the perch values are potiential high leverage fish   
# What are 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

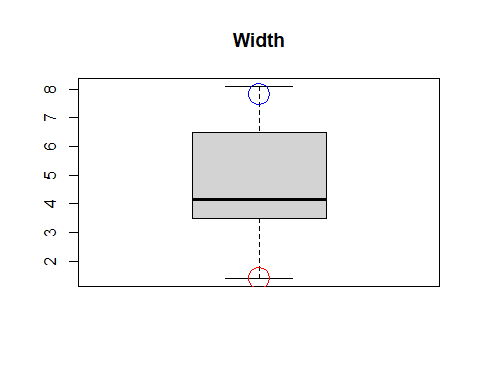
# Have fish 1 and 40   
# Fish 1 is kinda small   
# In teh slides we see how this are sorted differently

**Leverage in Multiple Regression: Perch**

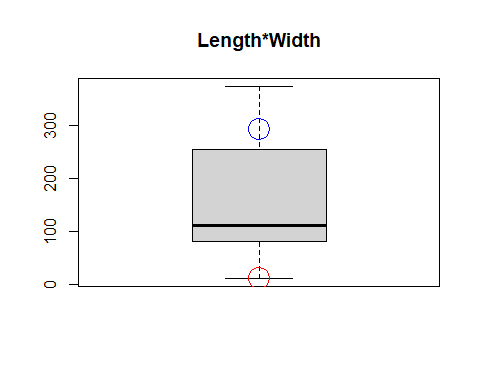
#boxplots for the three variables in the model  
#points() is used to show the values for cases 1 and 40 that have high leverage  
#cex=3 is the type of symbol to show in the plot  
  
# Tehse are box plots   
# first look at weight and length variable and made a box plot for each   
# Then drew some 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?

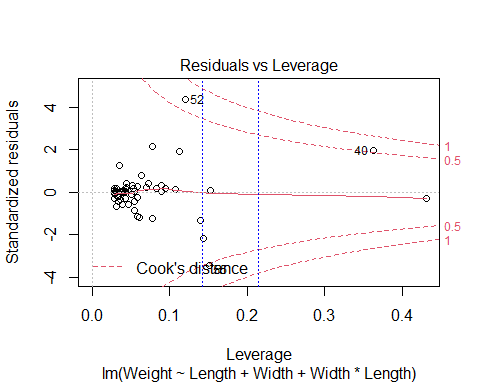
Cook’s Di = (((std.resi)^2)/(k+1))\*(hi/(1-hi)) Di increases with either poor fit (std.resi) and high leverage (hi).

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

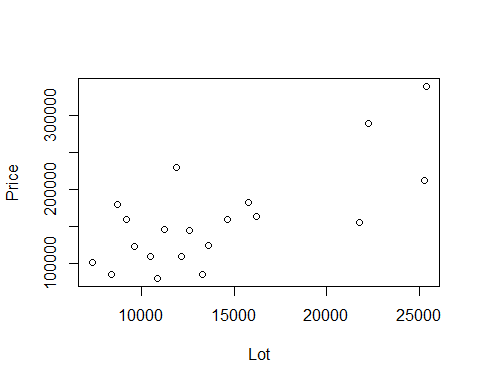
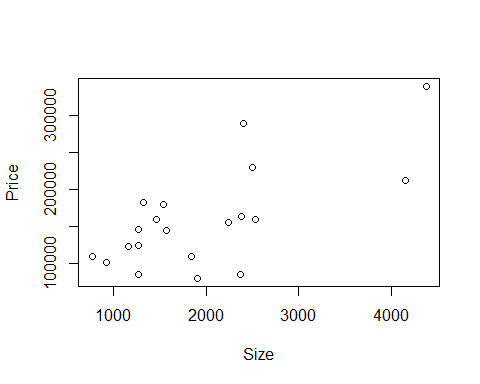
**Houses** Response variable: Y = House price Predictors: X1 = Size; X2 = Lot (size of the lot)

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

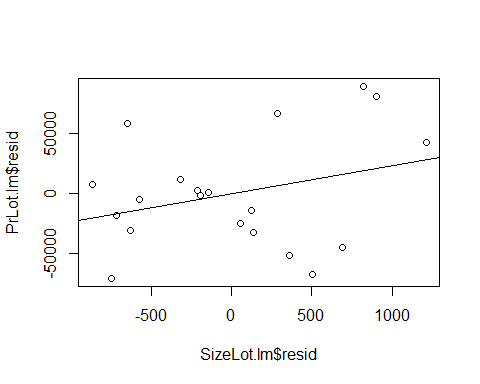
# based on teh slope it doesnt appear that size or ot is a good predictor forom price   
# looking at anova, we get that its fine   
# There is multicollinearity here

**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

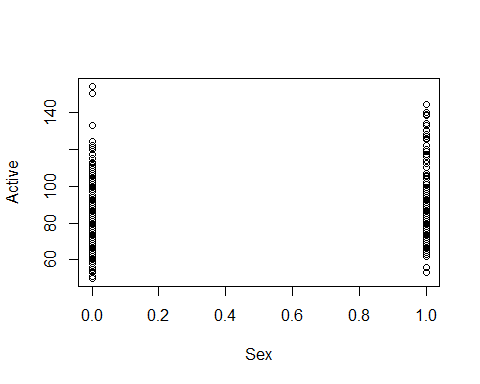
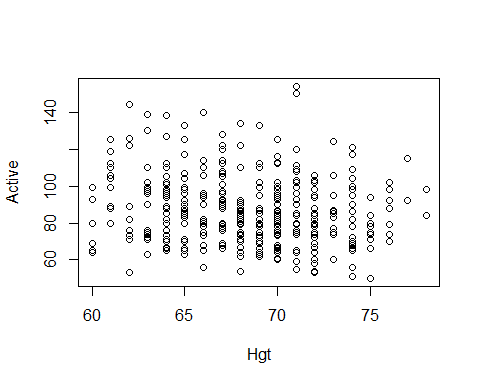
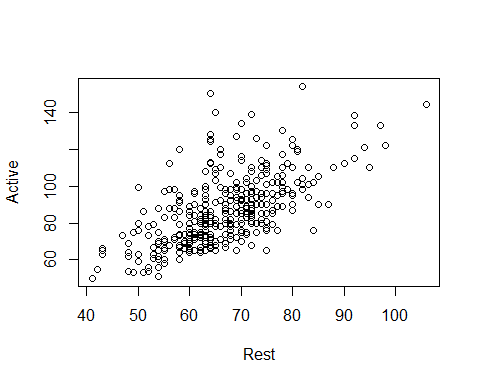
#Is there a relationship?   
# Equation of the line:   
# Looks familliar because the size estimate for the summary of the model above is the   
# New model fo teh residuals has an incet fo 0 and a slope of 23, and oges through gthe orignas, have teh same slope as size by lot   
# Do we see anythign 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)

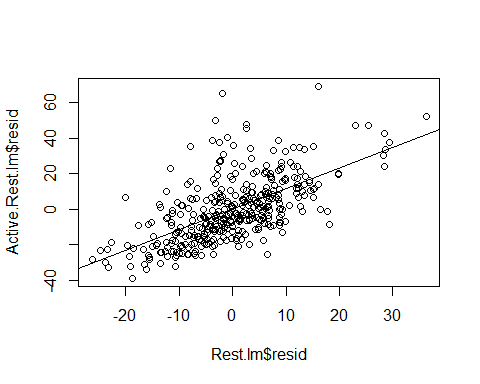
##   
## Call:  
## lm(formula = Active ~ Rest + Hgt + Sex, data = Pulse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.730 -9.381 -2.691 6.817 67.413   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.47044 20.40682 0.660 0.510   
## Rest 1.16249 0.07569 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

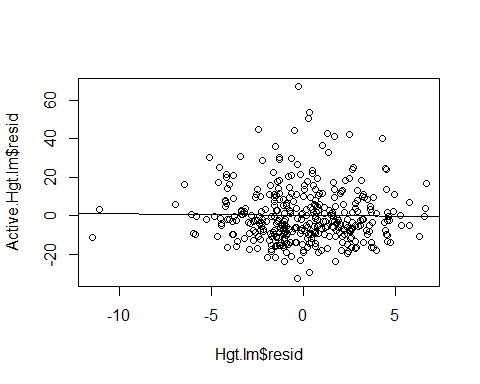
#Want to make a plot where actiec is predicted byt weverthing   
Active.Rest.lm = lm(Active~Hgt+Sex, data=Pulse)  
Rest.lm = lm(Rest~Hgt+Sex, data = Pulse)  
plot(Active.Rest.lm$resid~Rest.lm$resid)  
mod1 = lm(Active.Rest.lm$resid~Rest.lm$resid)  
abline(mod1)



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

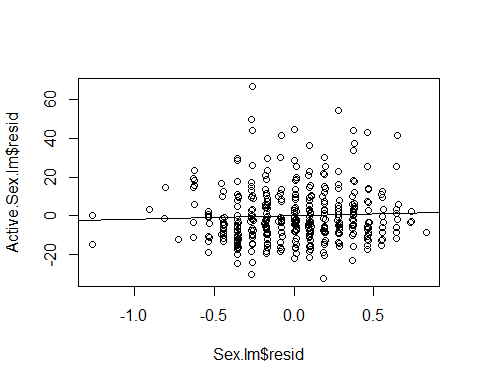
# Where active is predicted by rest and sex, not hgieht   
Active.Hgt.lm = lm(Active~Rest+Sex, data=Pulse)  
Hgt.lm = lm(Hgt~Rest+Sex, data = Pulse)  
plot(Active.Hgt.lm$resid~Hgt.lm$resid)  
mod2 = lm(Active.Hgt.lm$resid~Hgt.lm$resid)  
abline(mod2)



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)

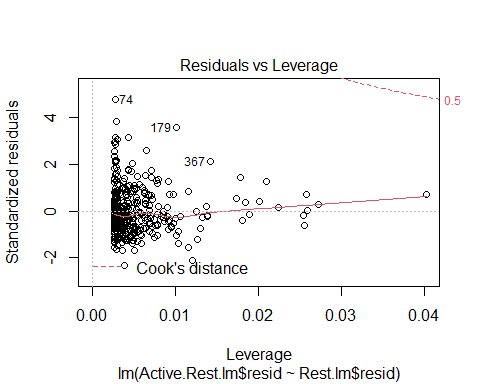


summary(mod3)

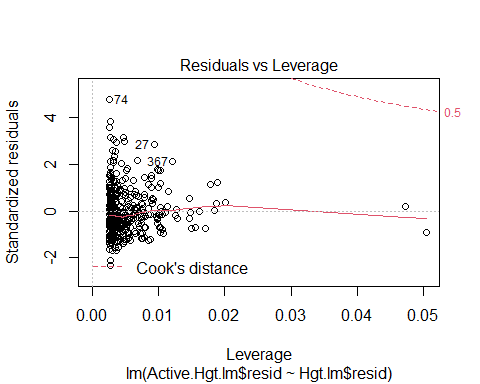
##   
## Call:  
## lm(formula = Active.Sex.lm$resid ~ Sex.lm$resid)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.730 -9.381 -2.691 6.817 67.413   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -9.785e-16 7.300e-01 0.000 1.000  
## Sex.lm$resid 1.909e+00 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

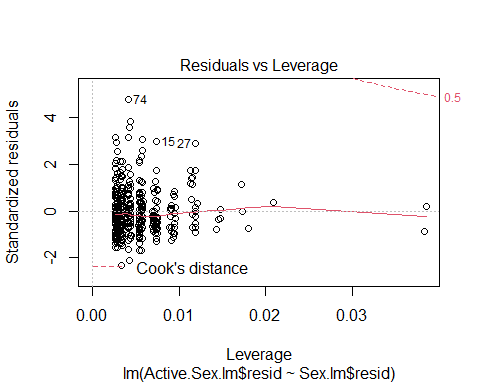
# Now we check each of these for leverage and influence   
plot(mod1, 5)



plot(mod2, 5)

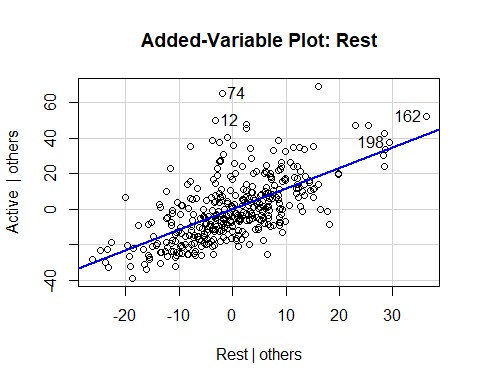


plot(mod3, 5)

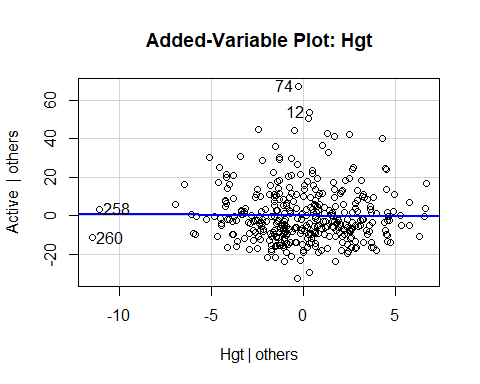


**Added Variable Plots** HOW TO DO THE ADDED VARIBALE PLOT EASILY AND QUICKLY IN R

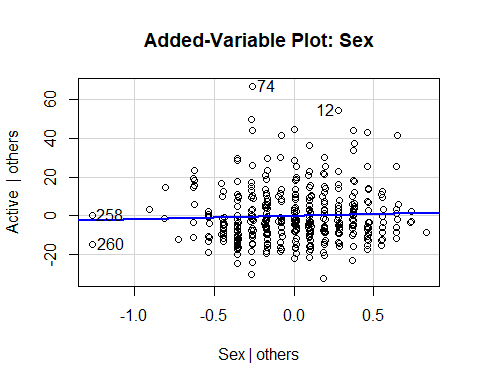
library(car)  
  
# Give the lienar model and what you want to look at as the added variabe   
# So it looks at the thing in pulse.lm, but takes out rest in the first one and then compares what the residual plots would look like between those two   
avPlot(Pulse.lm, "Rest") # need to give the string name of the variable in the lm model you're referencing from the first argument



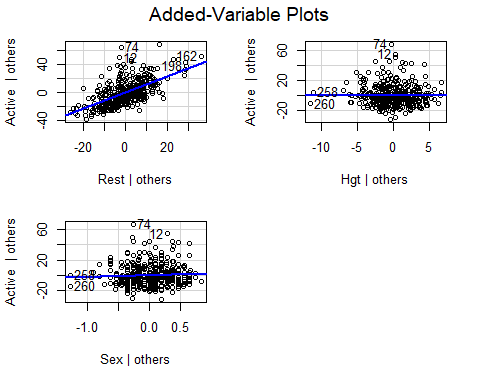
avPlot(Pulse.lm, "Hgt")



avPlot(Pulse.lm, "Sex")



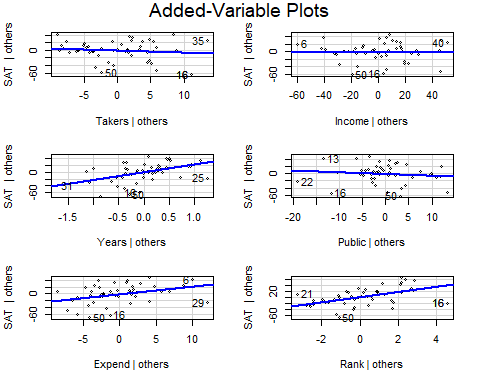
avPlots(Pulse.lm, ~.) # This is an avplot for all of the variables in the Pulse.lm linear regression model



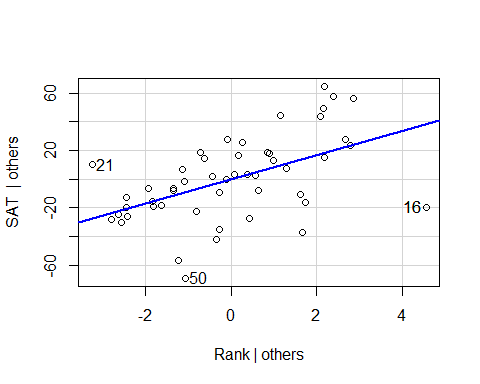
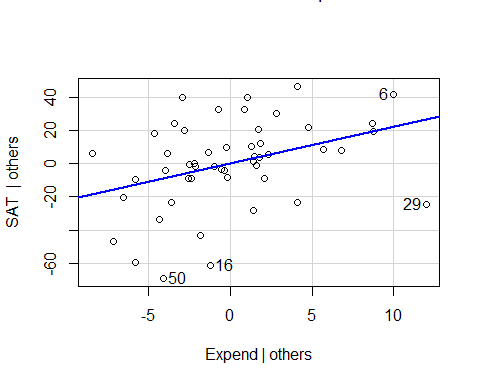
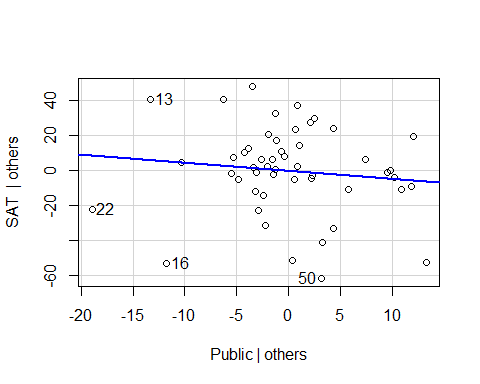
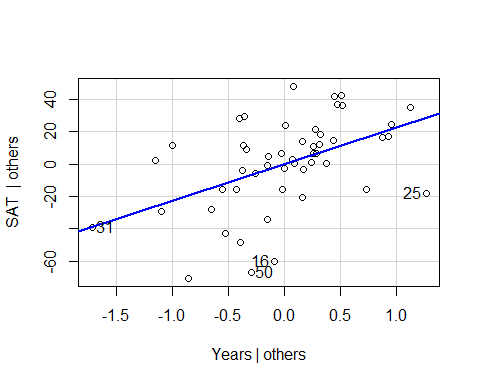
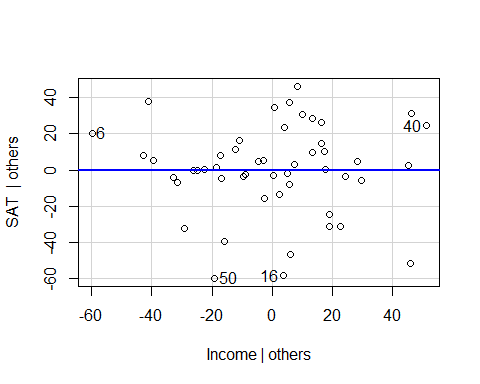
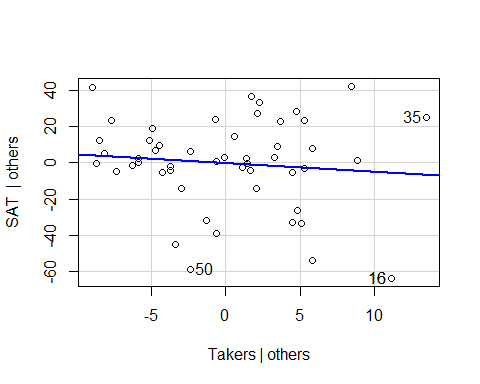
# This iwll make an avPlot for all the variables in teh pulse.lm linear regression model so you don't have to   
# THIS IS PLURARL   
# Its not super great in how it does it   
# YOu have to do it by all the predictors  
# issues: it's really hard to read it's really jumbled, you can read it   
# All 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)

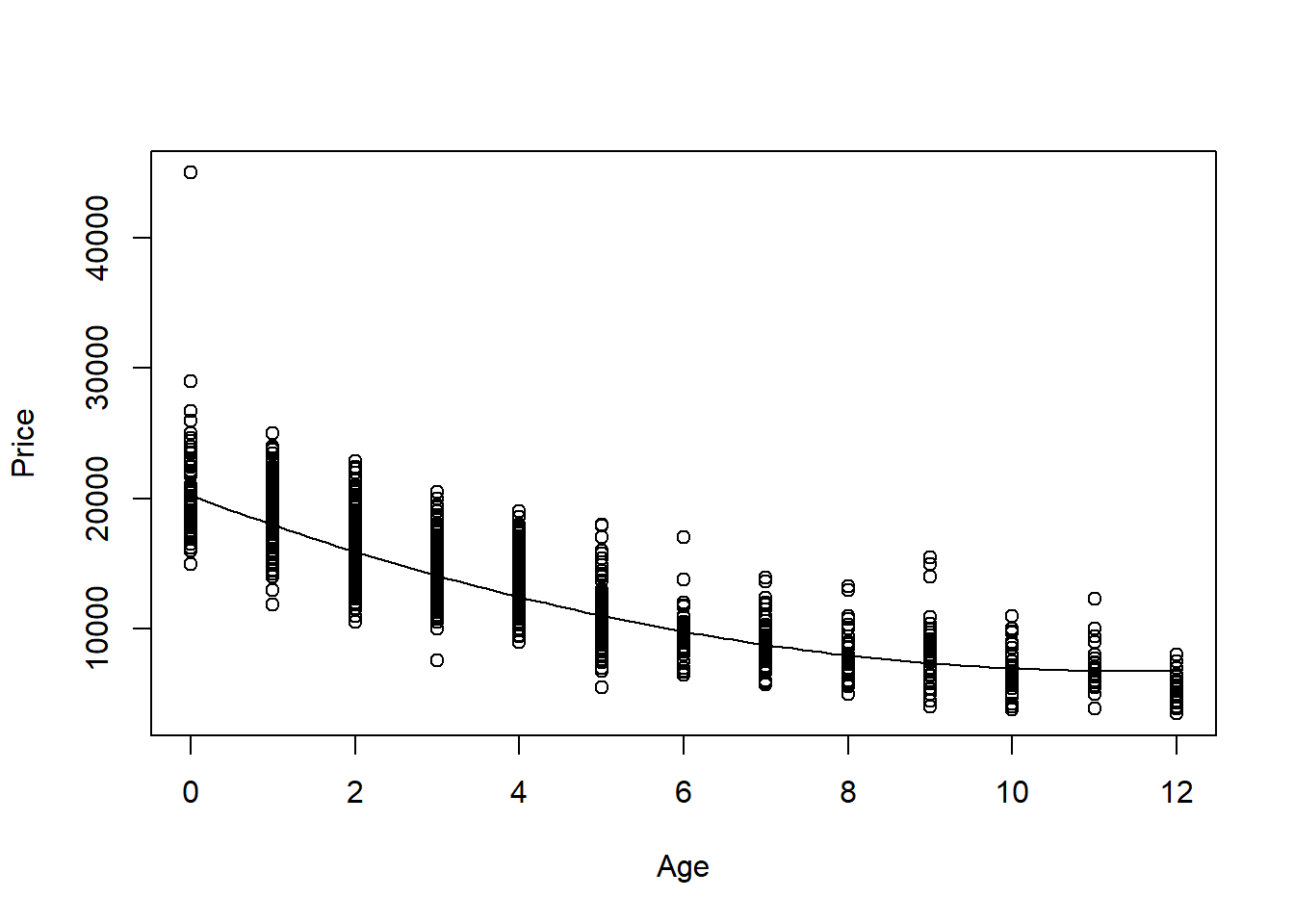
plot(Price~Age, data=CombinedStates)

a = summary(modq10)$coef[3]

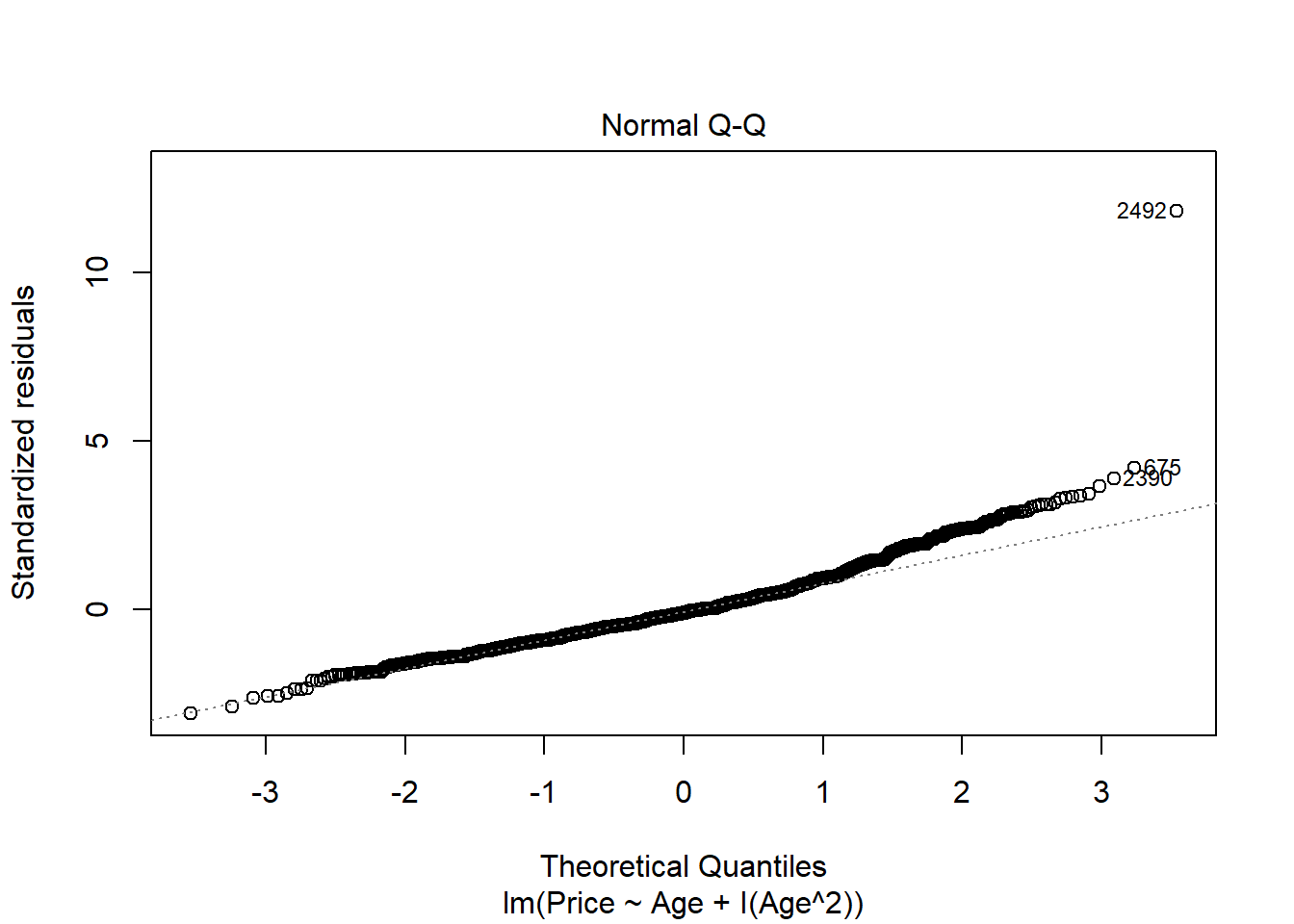
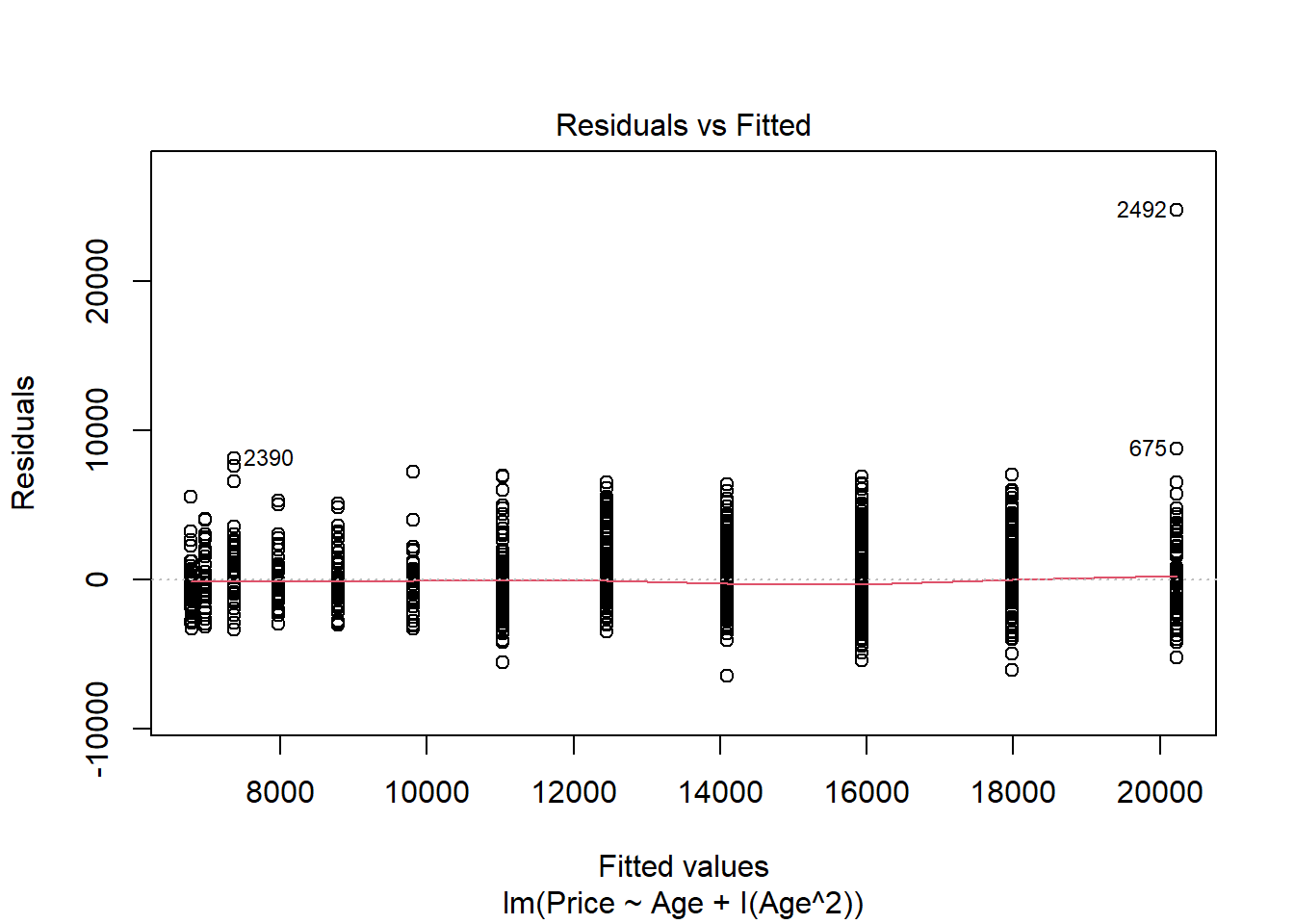
b = summary(modq10)$coef[2]

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))

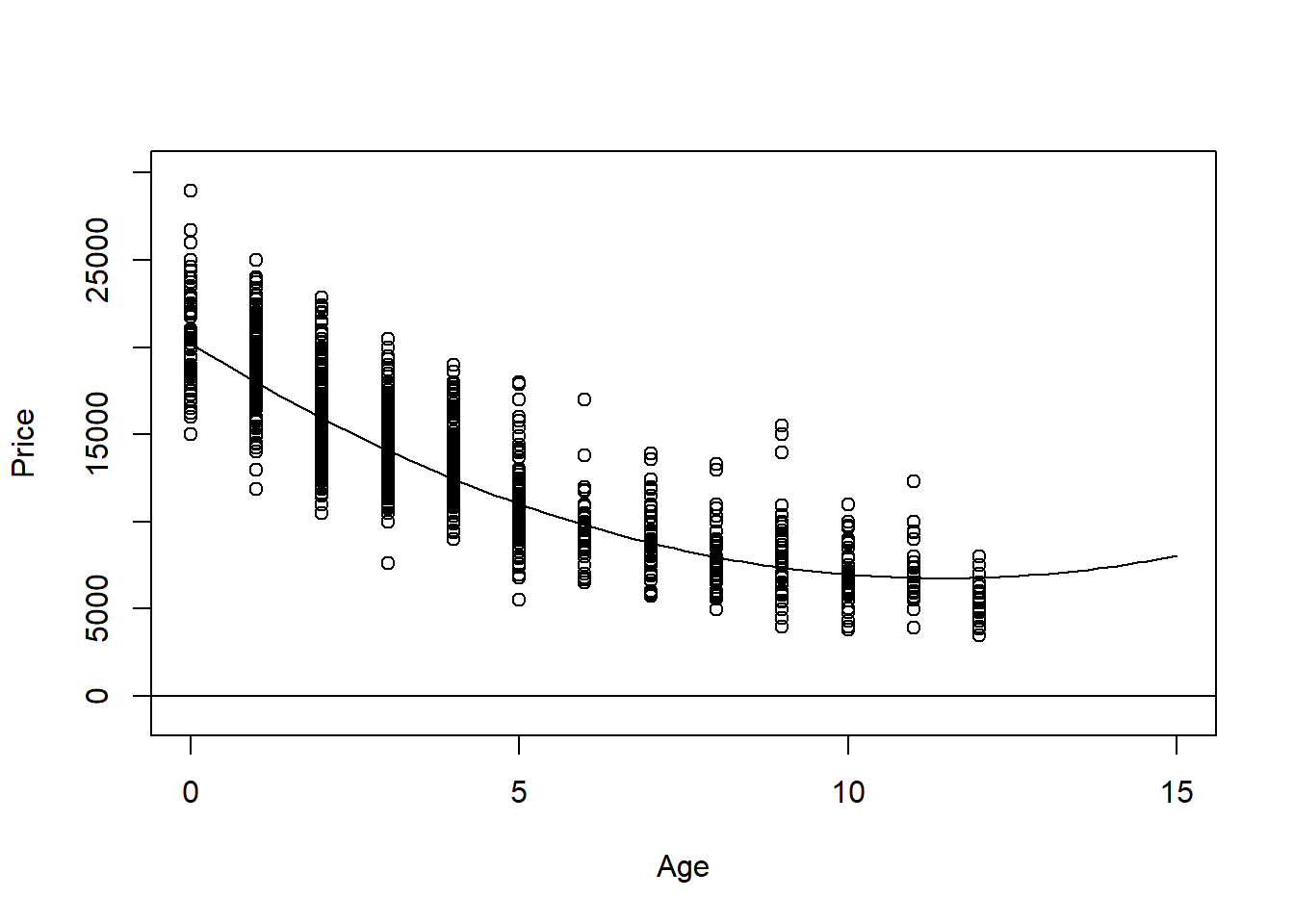
## [1] 11.44833+8.10717i 11.44833-8.10717i

# 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**

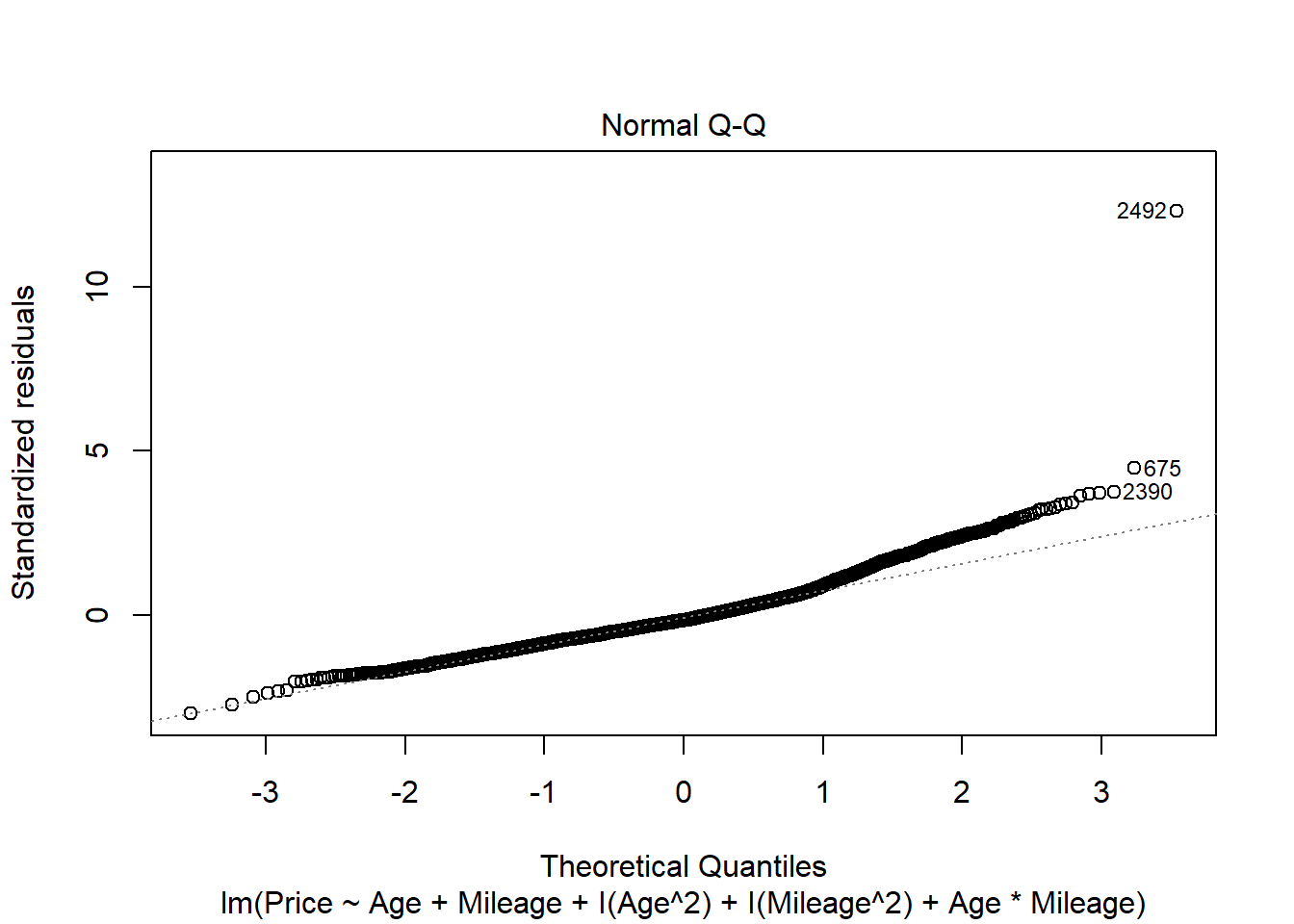
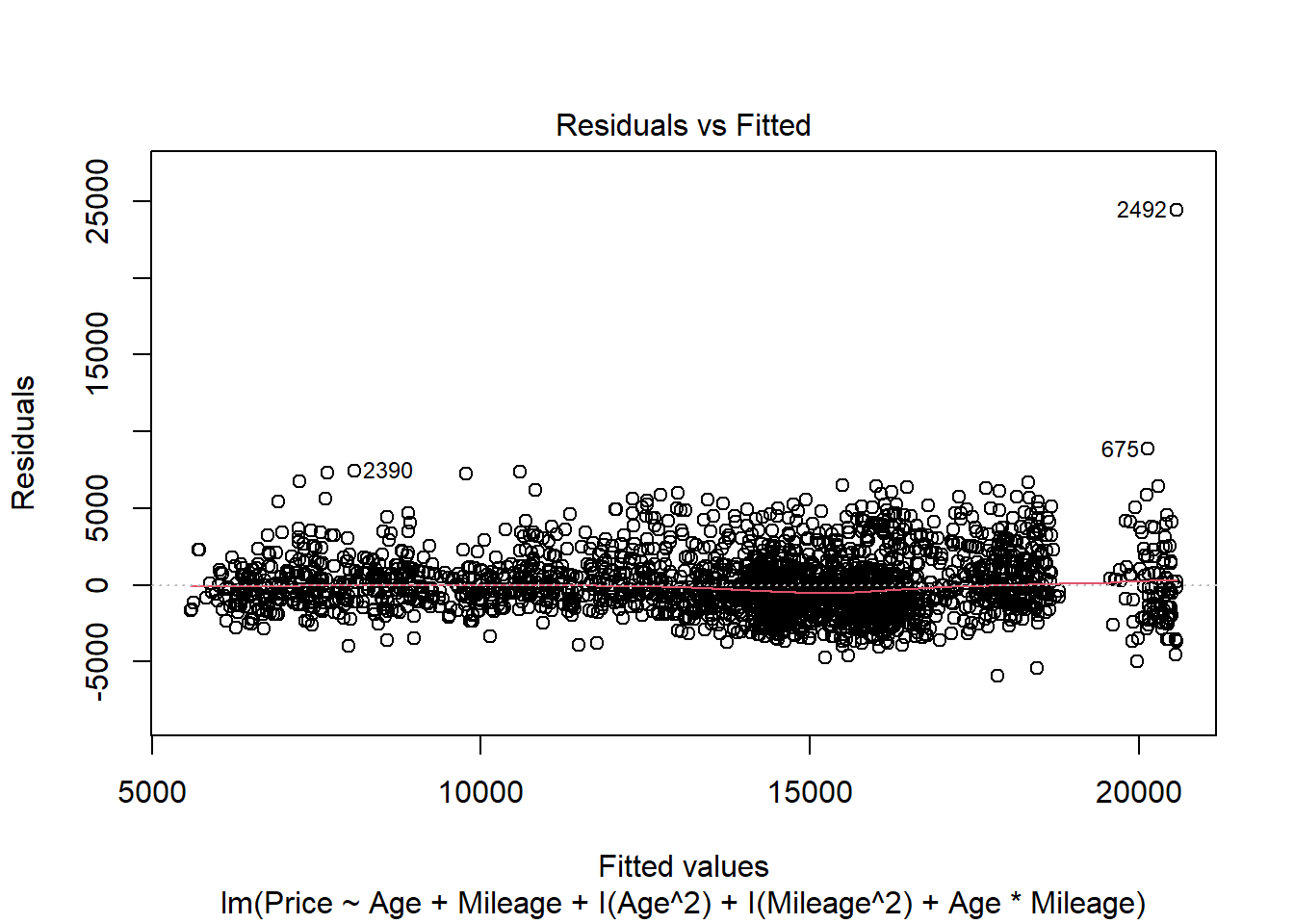
For this section you should again use the dataset with cars from three states that you used for models 5 and 6.

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