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

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

head(Houses)

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

cor(Houses)

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

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

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

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

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

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

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

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

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

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

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

*Simple models are idea, than overaly complicated ones*

**Multicollinearity**

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

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

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

anova(HouseModel)

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

anova455(HouseModel)

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

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

*See below*

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

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

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

## [1] 0.588093

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

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

round(cor(Houses), 2)

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

1. Compute the Variance Inflation Factor (VIF).

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

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

## [1] 2.427732

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

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

summary(HouseModel)

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

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

## [1] 1.558118

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

## [1] 11.36012

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

## [1] 0.8423846

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

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

vif(HouseModel)

## Size Lot   
## 2.427732 2.427732

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

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

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

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

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

head(StateSAT)

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

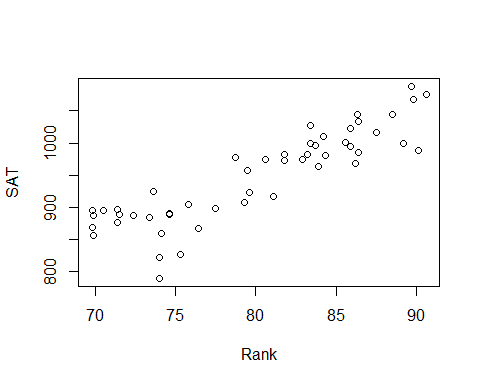
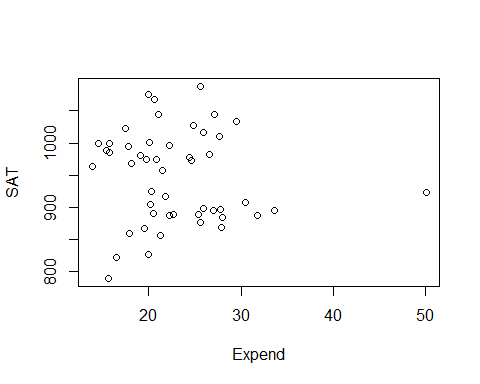
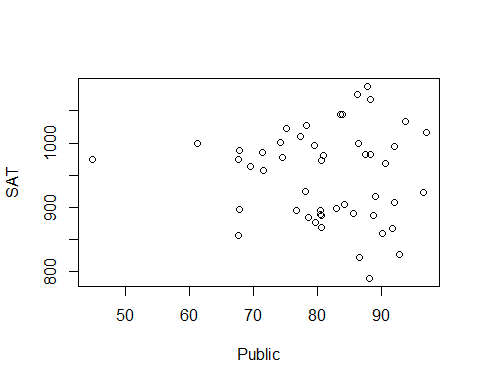
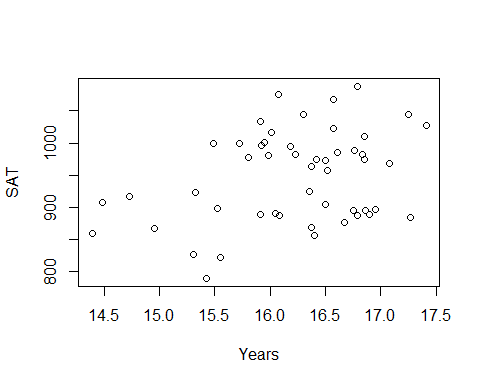
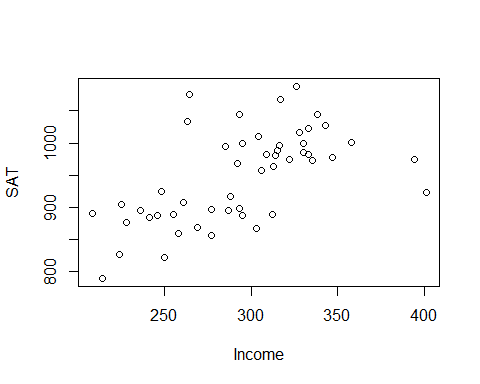
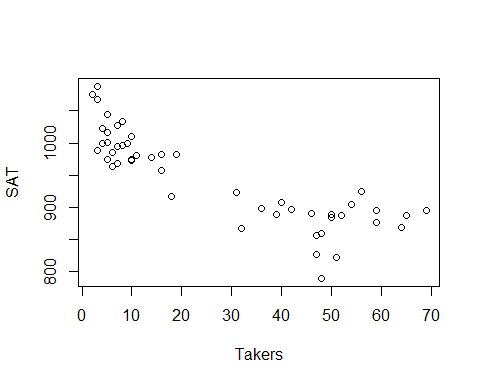
# Some states might have lower SAT takers ebcause the ACT mgiht be better

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

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

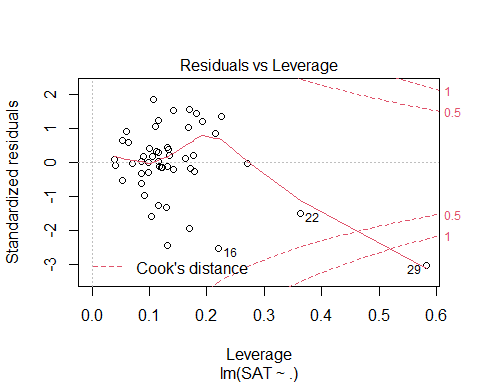
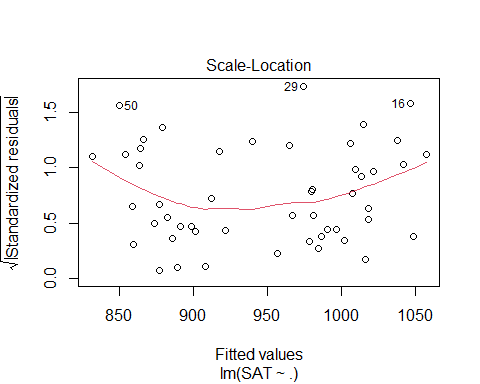
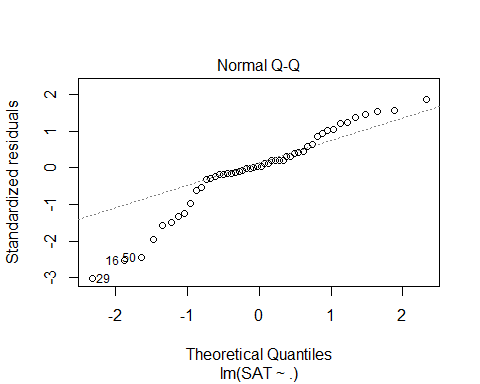
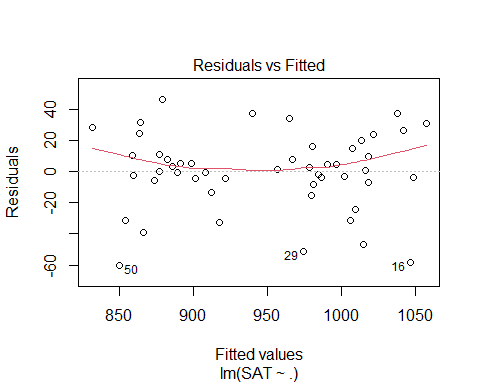
*What determines “best”?*

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



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

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



summary(SAT\_Model)

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

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

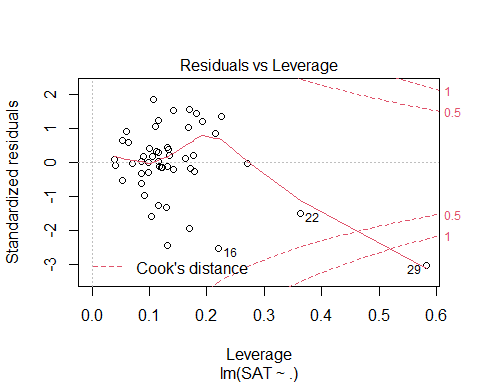
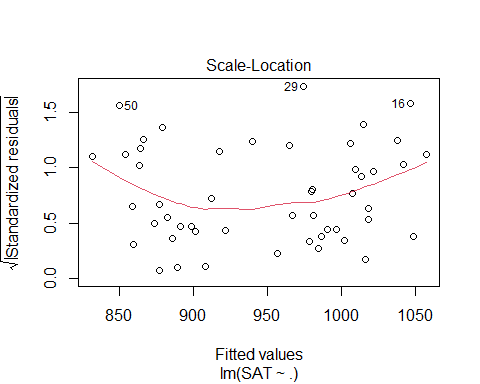
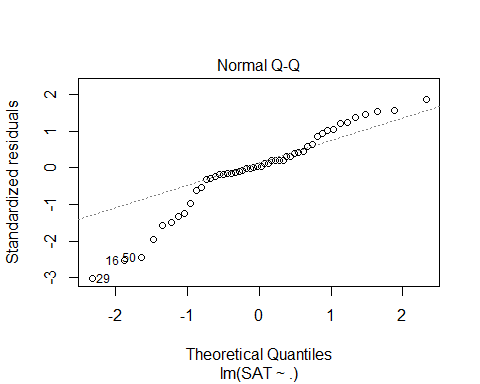
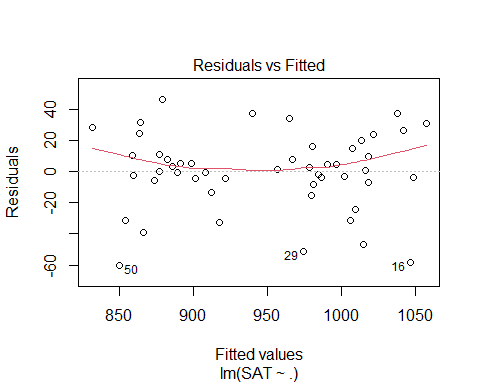
vif(SAT\_Model)

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

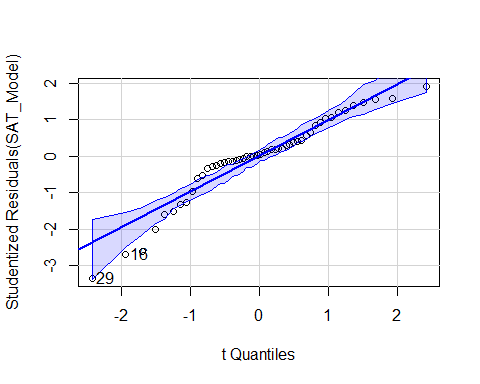
# looking at the VIF for this model, we see that Takers and Rank have really high inflation rates, which means they should probably be expcluded from the final regression model   
# It teslls us wif we make a mdoel with takers as the repsonse, we will get a really high r squared value; almost all the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

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

plot(SAT\_Model)



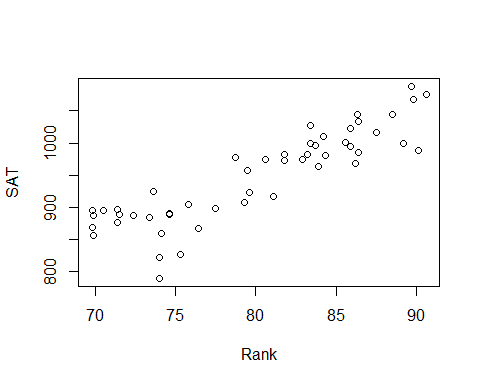
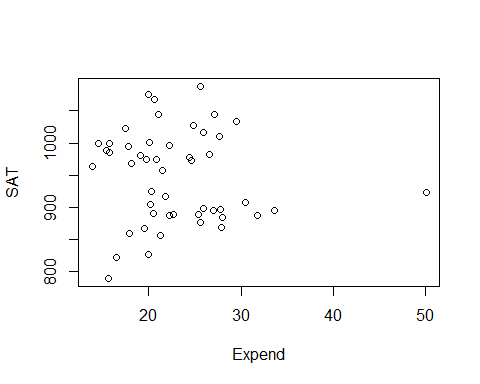
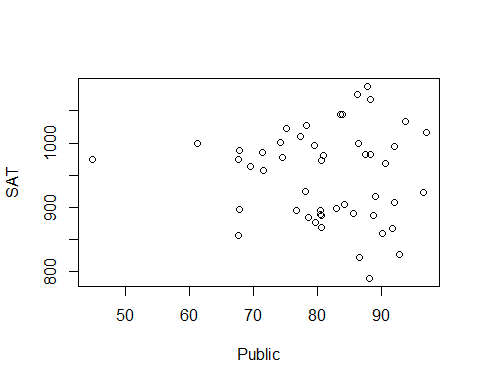
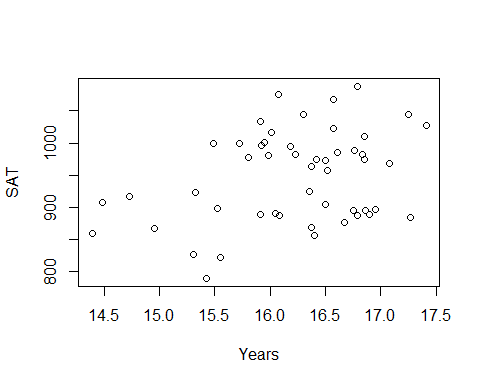
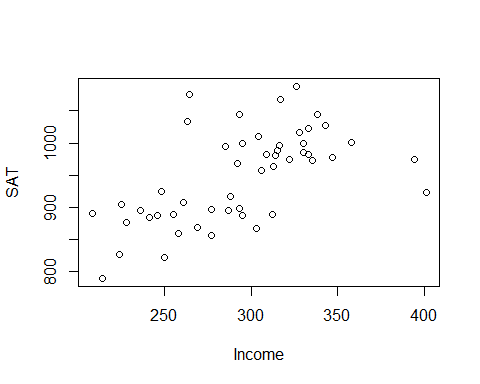
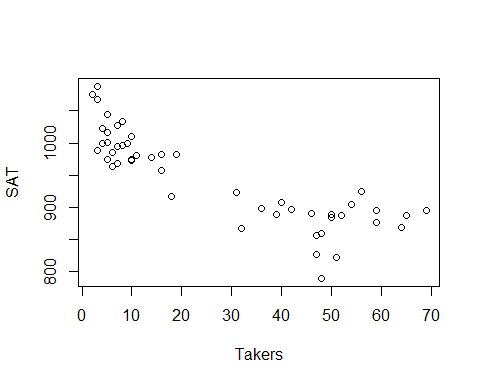
# We can look at these plots to see if the ocnditions are met   
# The model doesn't appear to fit a line well   
# Transfomration might be useful   
# Or, dont include all the variables in teh model   
# Maybe tranfomraiotn or may less predicotrs   
qqPlot(SAT\_Model)



## [1] 16 29

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

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



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

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

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