STOR 455 - Class 13 - R Model Section Methods

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

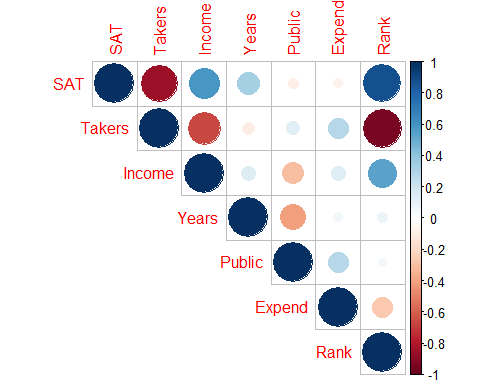
head(StateSAT)

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

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

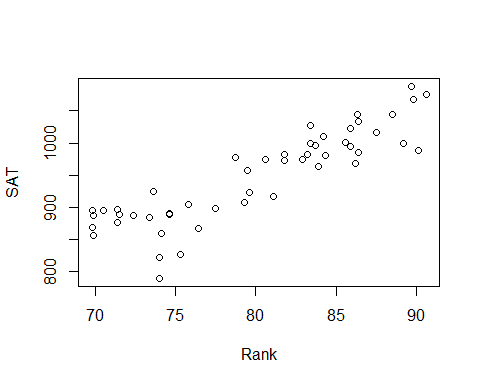
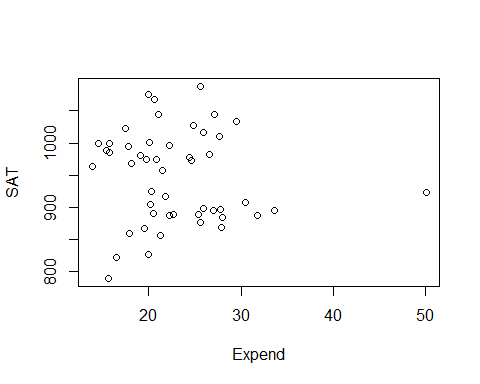
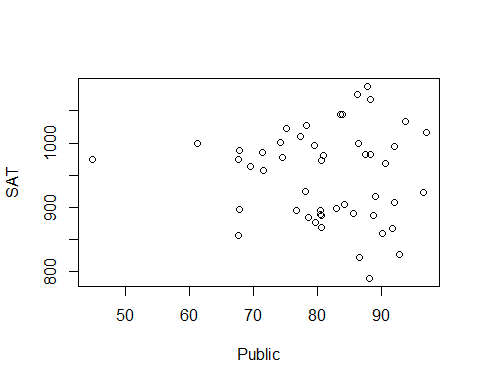
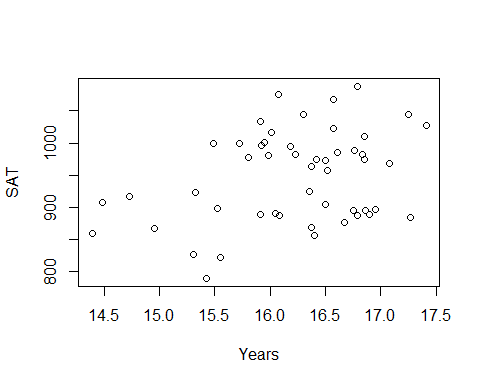
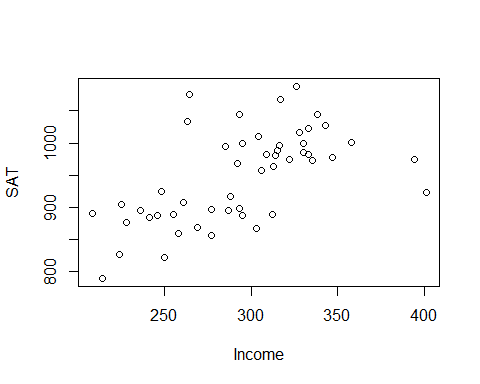
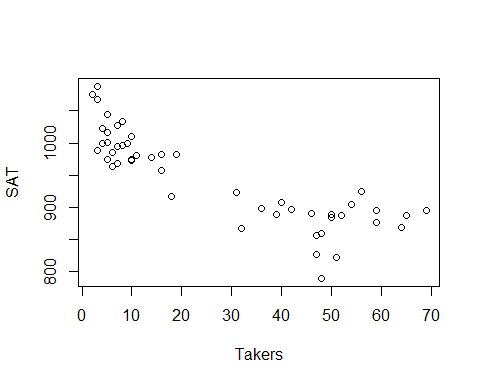
## SAT Takers Income Years Public Expend  
## SAT 1.00000000 -0.8578100 0.5844666 0.33096886 -0.08035688 -0.06287764  
## Takers -0.85780996 1.0000000 -0.6619351 -0.10154350 0.12355625 0.28363041  
## Income 0.58446657 -0.6619351 1.0000000 0.13476231 -0.30656703 0.13151942  
## Years 0.33096886 -0.1015435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

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



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

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

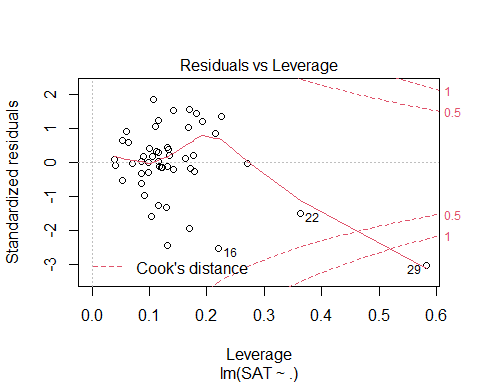
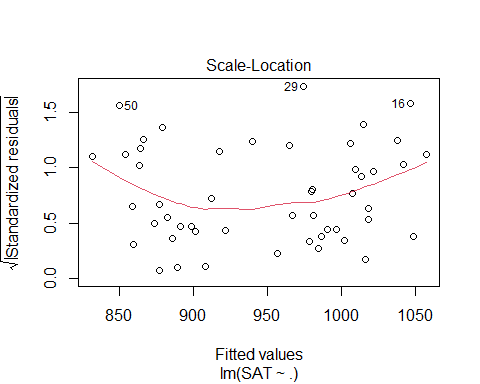
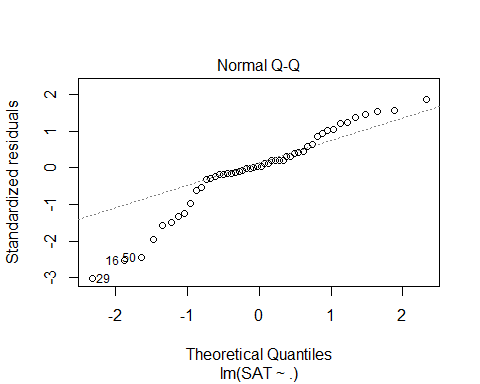
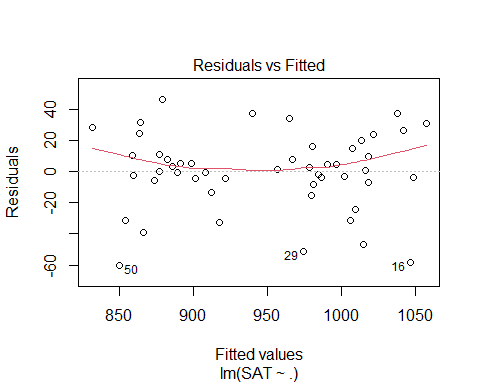


# Plot the data against each of the predictors int eh dataframe   
# Excludes state (Because we would have to factor state, and that would be a lot of information to process)  
  
# Rank adn takers have a recise pattern wtih SAT scores; its appears to have a curved realtionship there   
# Might not have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

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

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

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



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

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

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

1. Look for large adjusted R2

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

1. Look at individual t-tests

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

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

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

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

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

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

# IMPORTANT  
ShowSubsets(all)

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

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

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

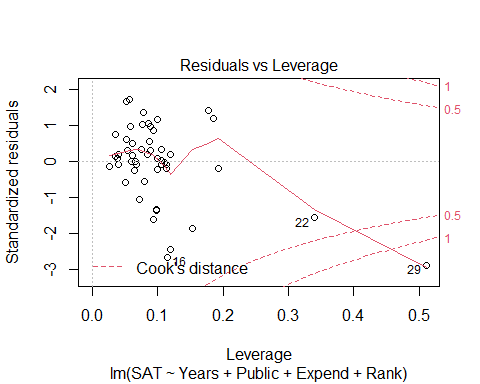
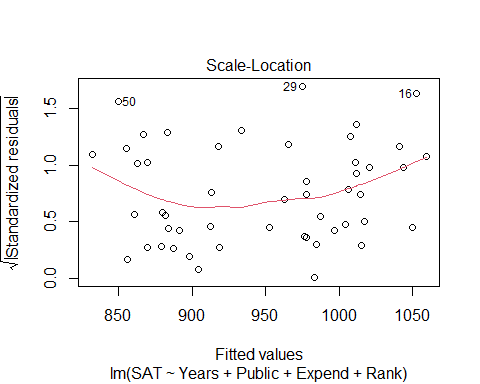
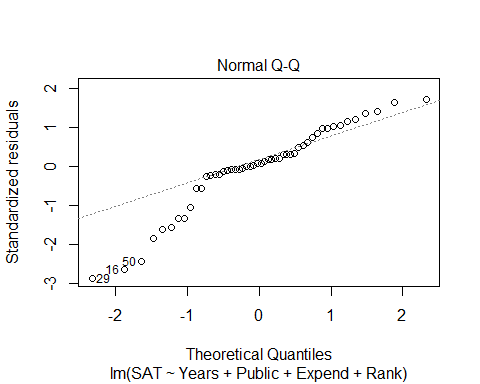
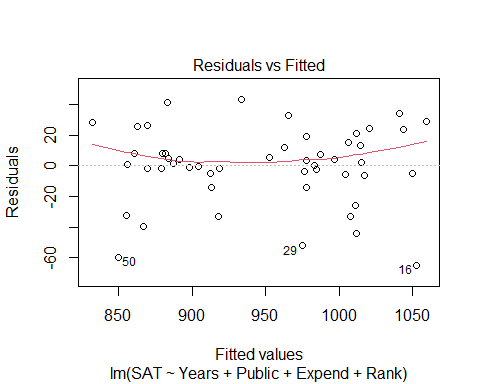
**Notes on Cp** - Cp depends on the larger pool of predictors as well as the set being considered. - For full model Cp = k+1 - For a “good” set of predictor, Cp should be small. - Like Adj R2, Cp weighs both the effectiveness of the model (SSEm) and the # of predictors (m).

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

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

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

plot(modSAT3)



vif(modSAT3)

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

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

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

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

summary(modSAT1)

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

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

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

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

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

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

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

**How to do backwards elimination in R**

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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