STOR 455 - Class 20 – new predictors from old

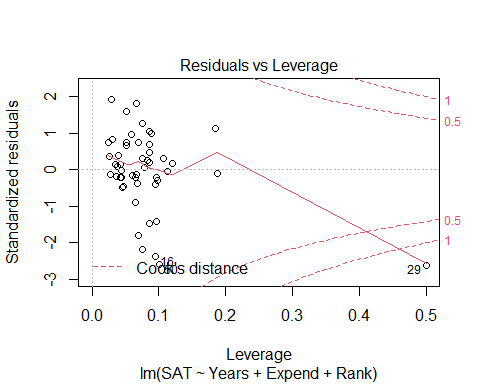
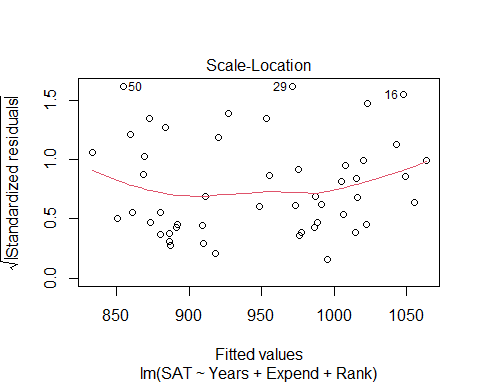
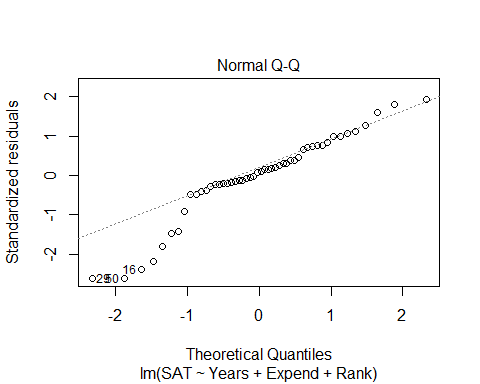
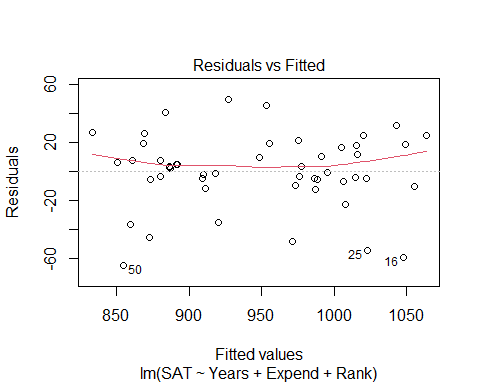
library(readr)  
library(leaps)  
  
StateSAT <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/StateSAT.csv")  
  
source("https://raw.githubusercontent.com/JA-McLean/STOR455/master/scripts/ShowSubsets.R")

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

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

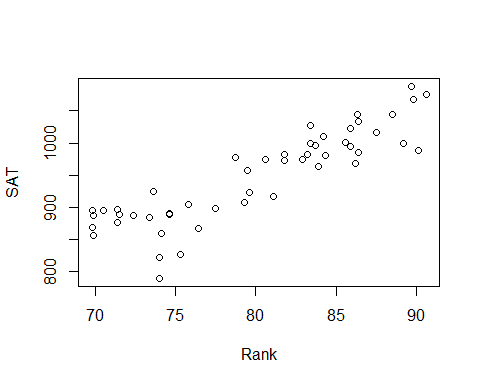
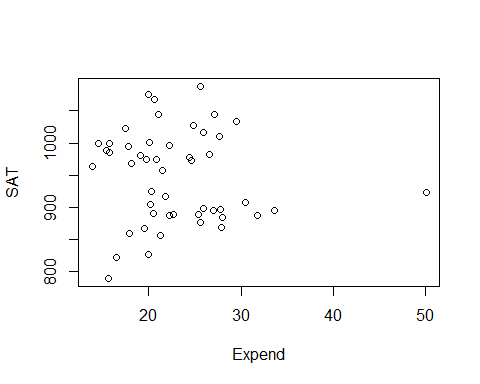
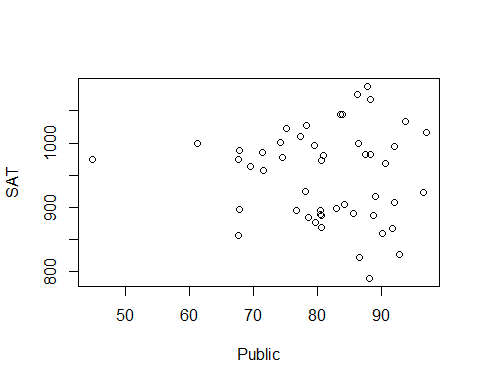
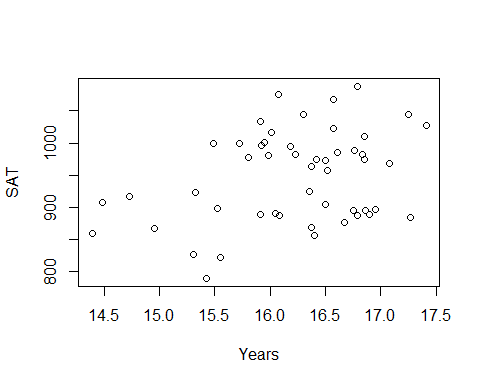
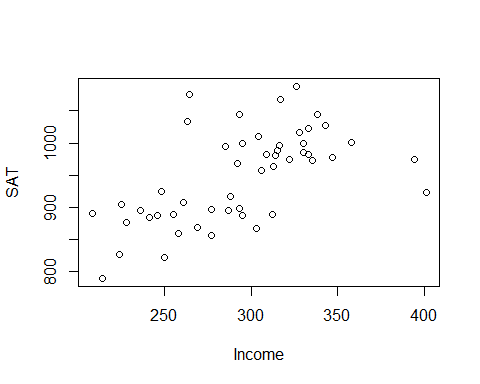
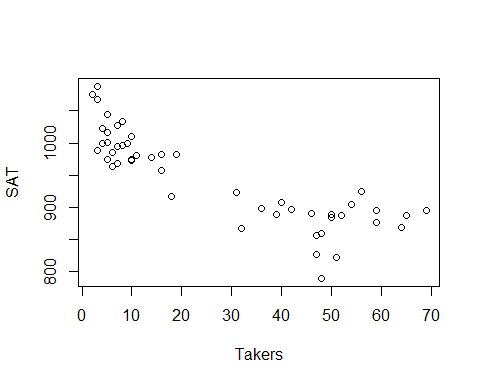
##   
## Call:  
## lm(formula = SAT ~ Years + Expend + Rank, data = StateSAT)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -64.802 -6.798 2.169 17.525 49.706   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -303.7243 97.8415 -3.104 0.00326 \*\*   
## Years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

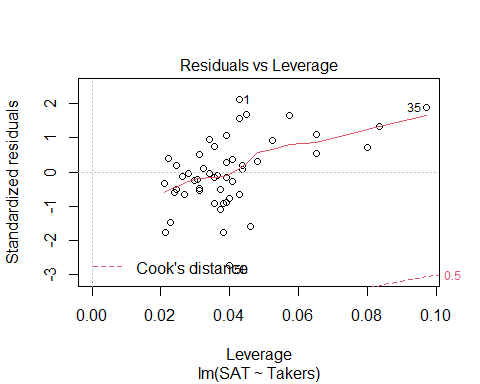
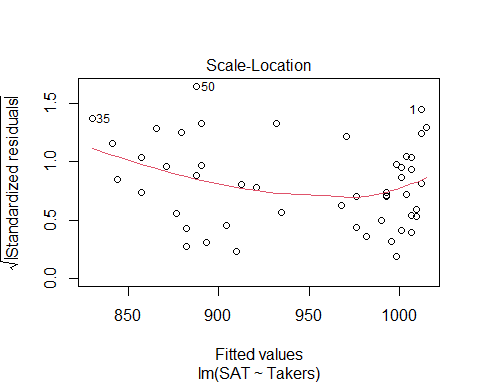
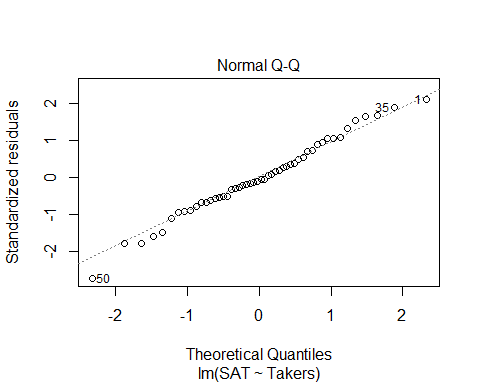
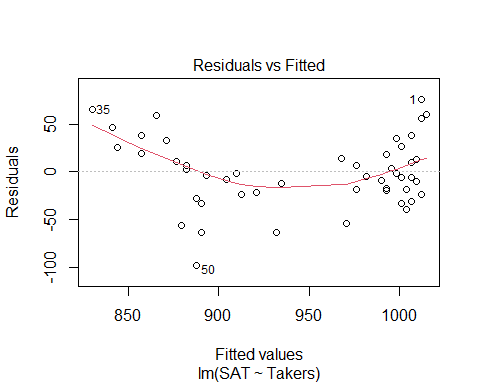


# Here, we see the correlation between each of the variables in the dataset (each of teh numerical variables in teh dataset)   
# Expend is iffy   
# Years is good  
# Takers looks like the best pattern, but we dont have it because it doesn't look lienar

# 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

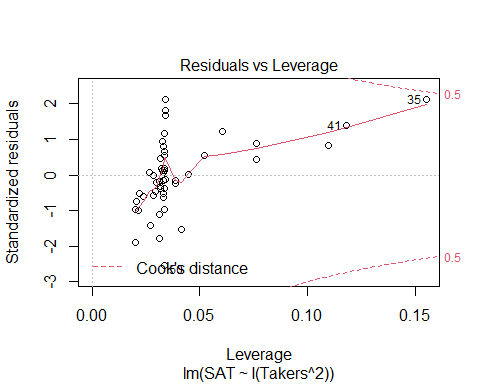
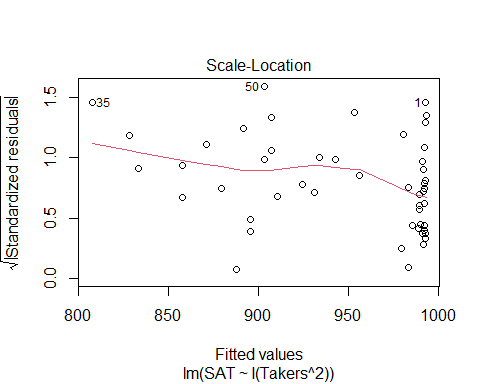
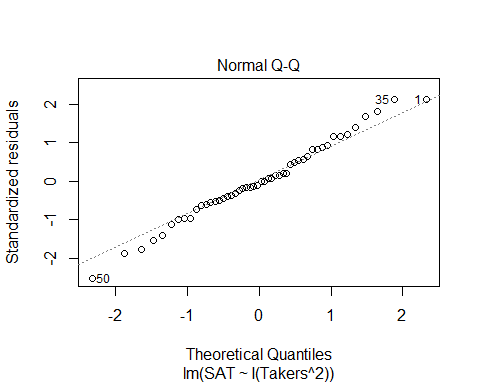
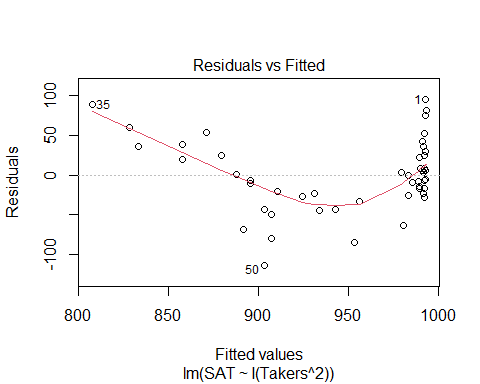
* **LINEAR** – 𝑌=𝛽\_𝑜+𝛽\_1 𝑋+𝜀
* **QUADRATIC** – 𝑌=𝛽\_𝑜+𝛽\_1 𝑋+𝛽\_2 𝑋^2+𝜀
* **CUBIC** –𝑌=𝛽\_𝑜+𝛽\_1 𝑋+𝛽\_2 𝑋^2+𝛽\_3 𝑋^3+𝜀

**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)



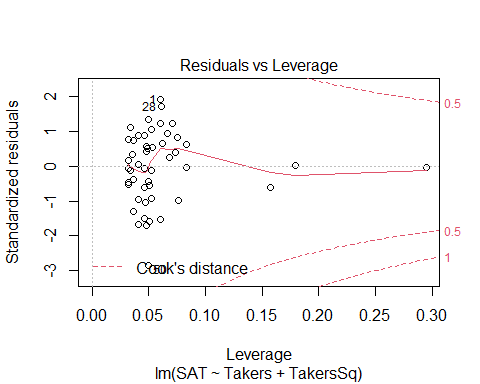
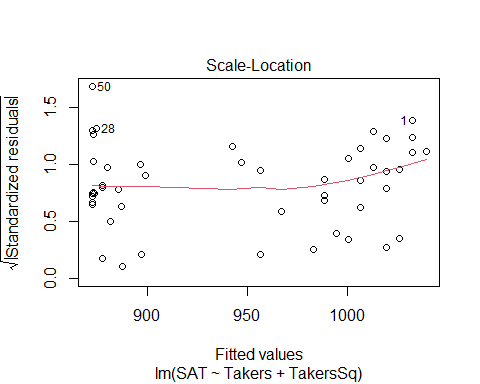
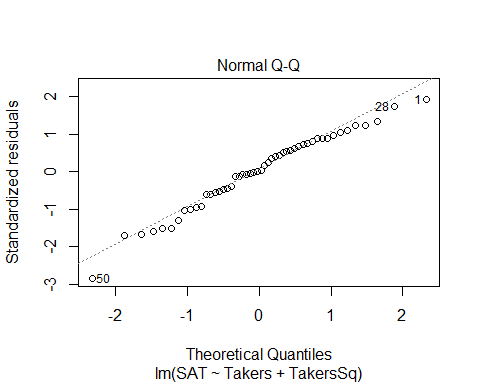
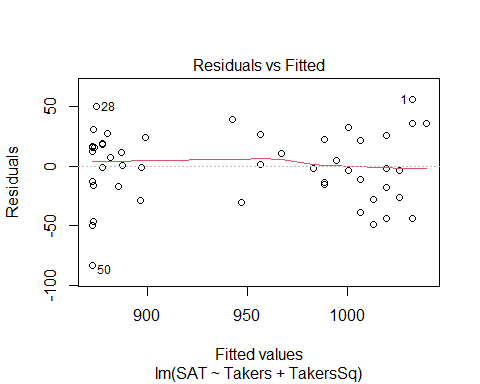
**Polynomial Regression in R** - We can add as many powers we want but then we might be over fitting, so that’s not always best

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)

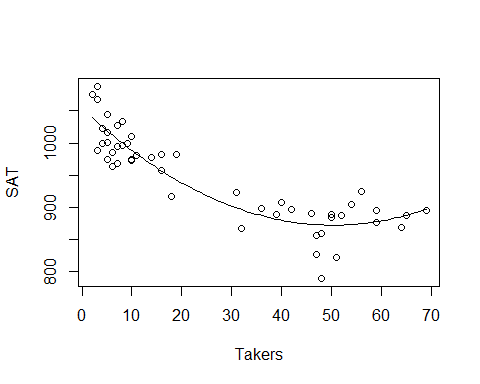
## Analysis of Variance Table  
##   
## Response: SAT  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Takers 1 181024 181024 202.089 < 2.2e-16 \*\*\*  
## TakersSq 1 22886 22886 25.549 6.992e-06 \*\*\*  
## Residuals 47 42101 896   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1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

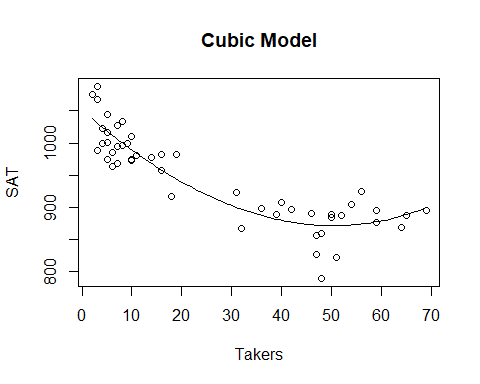
Would a Cubic work better?

#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)

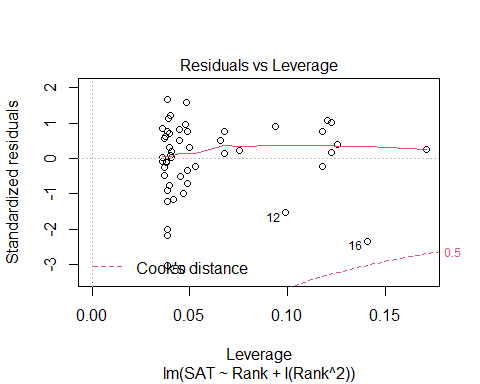
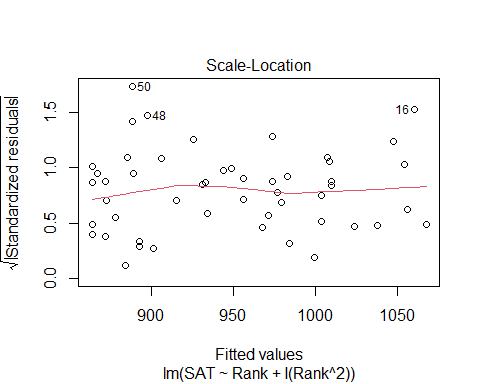
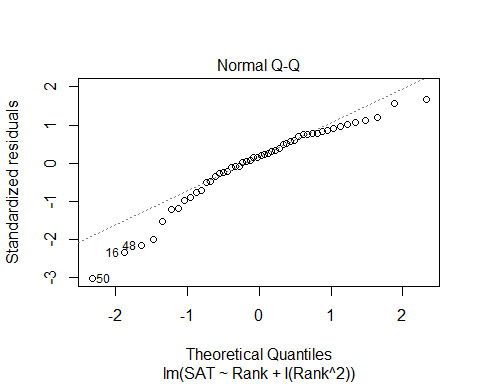
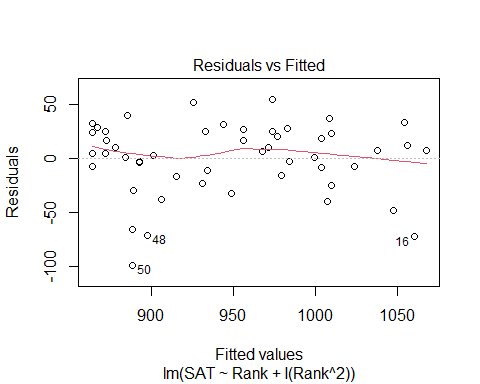
## Takers I(Takers^2) I(Takers^3)   
## 147.2369 678.9666 222.6922

# shows that there is a high multicollinearity with takers

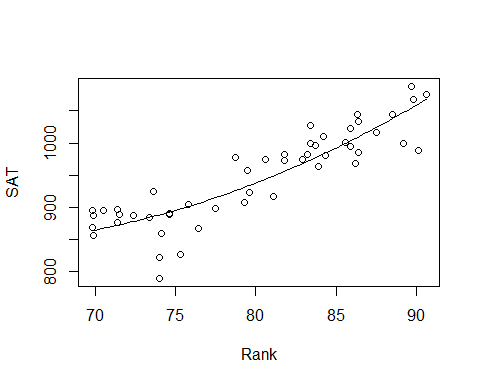
modSATquad4 = lm(SAT~ Rank+ I(Rank^2), data=StateSAT)  
summary(modSATquad4)

##   
## 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

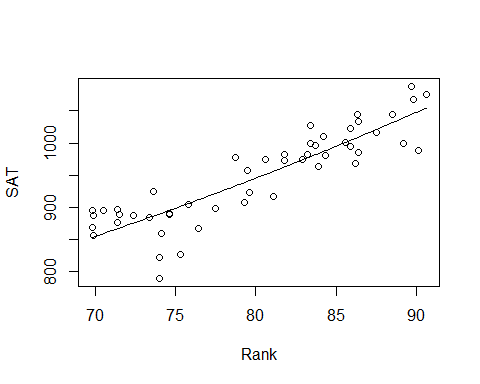
# Model looks pretty good, rank and rank^@ are not sig; btut ehre is a hug amount of multicollinearity   
# P value is close to 0, but the individual tests say things different because of the multicollinearity   
  
plot(modSATquad4)



# 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