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

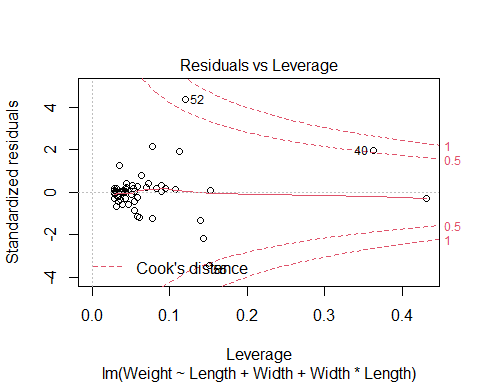
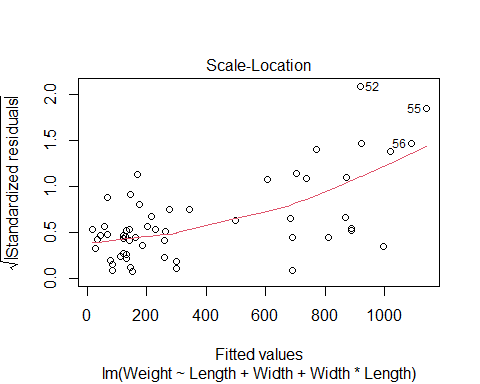
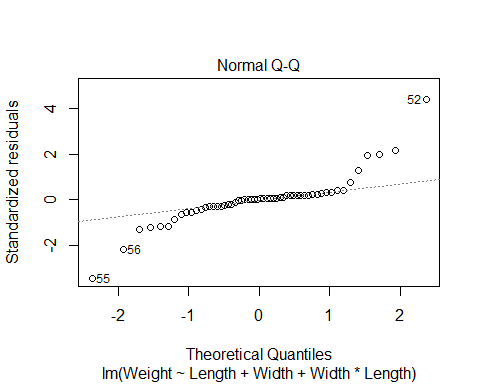
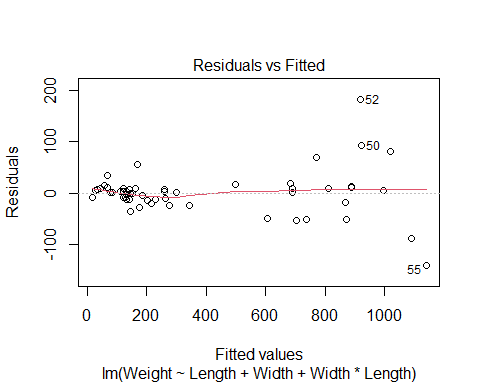
**Leverage in Multiple Regression: Perch** Perch\_lm = lm(Weight~Length+Width+Width\*Length, data=Perch)

* 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

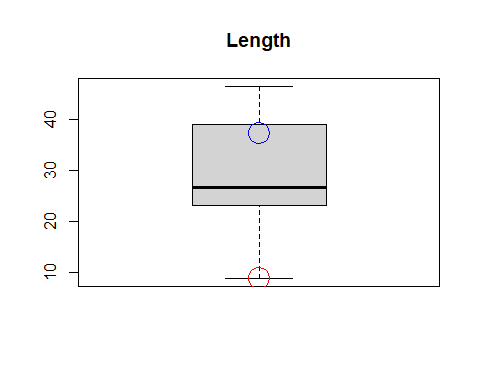
# Tells us which values in the dataset are over the highest boundary for leverage   
# Will give you waht is true   
Lev\_indices = which(hatvalues(Perch\_lm) >= 3\*(3+1)/56)  
  
#Two cases with high leverage  
# will tell you which of the perch values are potiential high leverage fish   
# What are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

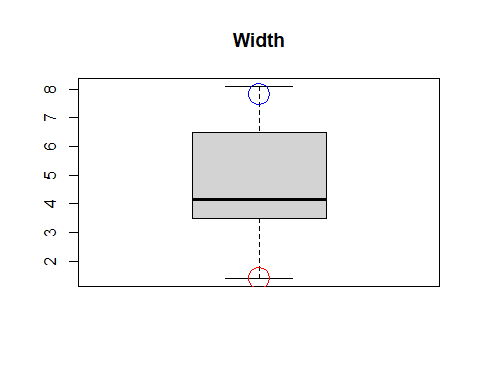
# Have fish 1 and 40   
# Fish 1 is kinda small   
# In teh slides we see how this are sorted differently

**Leverage in Multiple Regression: Perch**

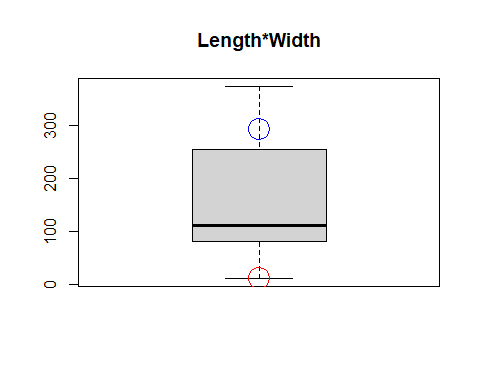
#boxplots for the three variables in the model  
#points() is used to show the values for cases 1 and 40 that have high leverage  
#cex=3 is the type of symbol to show in the plot  
  
# Tehse are box plots   
# first look at weight and length variable and made a box plot for each   
# Then drew some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

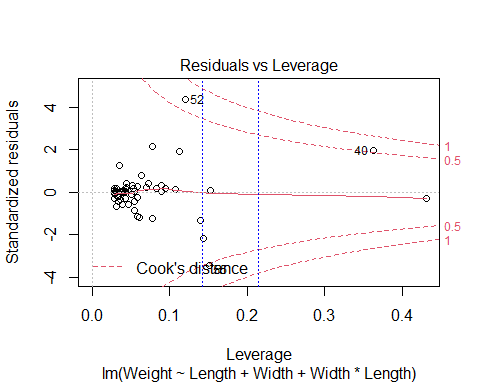
Cook’s Di = (((std.resi)^2)/(k+1))\*(hi/(1-hi)) Di increases with either poor fit (std.resi) and high leverage (hi).

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

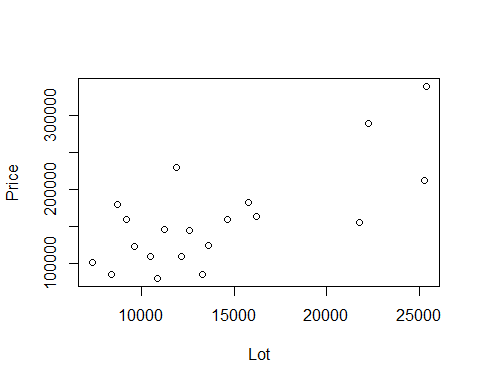
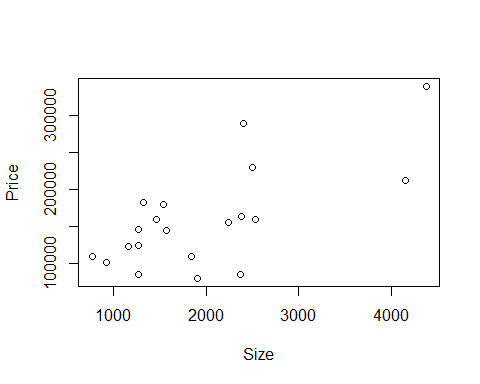
**Houses** Response variable: Y = House price Predictors: X1 = Size; X2 = Lot (size of the lot)

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

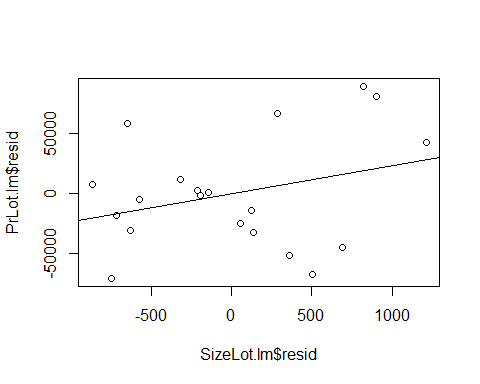
# based on teh slope it doesnt appear that size or ot is a good predictor forom price   
# looking at anova, we get that its fine   
# There is multicollinearity here

**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

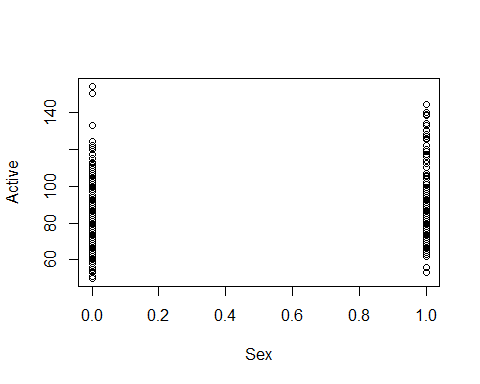
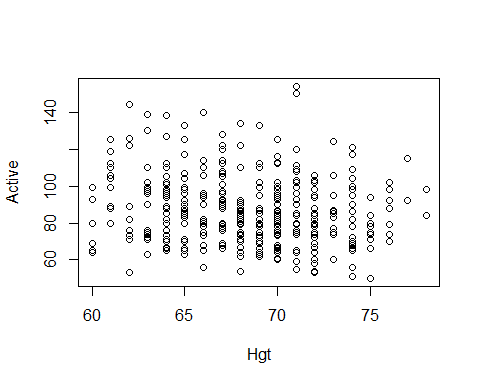
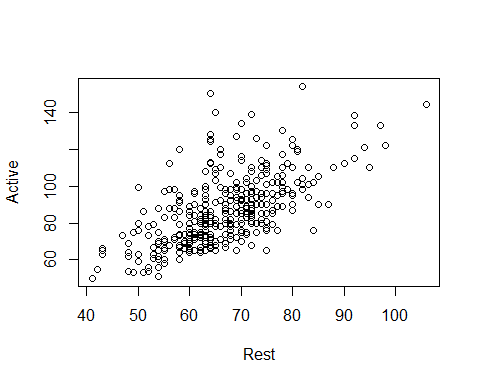
#Is there a relationship?   
# Equation of the line:   
# Looks familliar because the size estimate for the summary of the model above is the   
# New model fo teh residuals has an incet fo 0 and a slope of 23, and oges through gthe orignas, have teh same slope as size by lot   
# Do we see anythign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

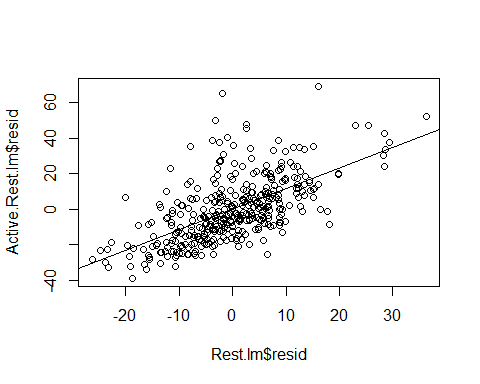
##   
## Call:  
## lm(formula = Active ~ Rest + Hgt + Sex, data = Pulse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.730 -9.381 -2.691 6.817 67.413   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.47044 20.40682 0.660 0.510   
## Rest 1.16249 0.07569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

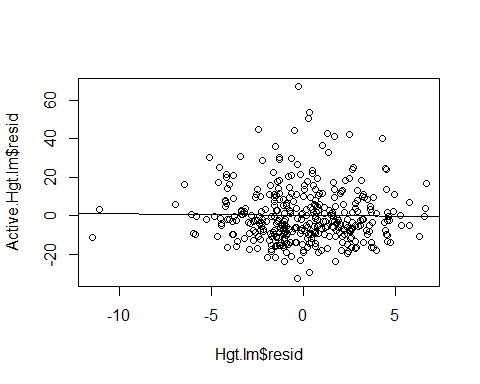
#Want to make a plot where actiec is predicted byt weverthing   
Active.Rest.lm = lm(Active~Hgt+Sex, data=Pulse)  
Rest.lm = lm(Rest~Hgt+Sex, data = Pulse)  
plot(Active.Rest.lm$resid~Rest.lm$resid)  
mod1 = lm(Active.Rest.lm$resid~Rest.lm$resid)  
abline(mod1)



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

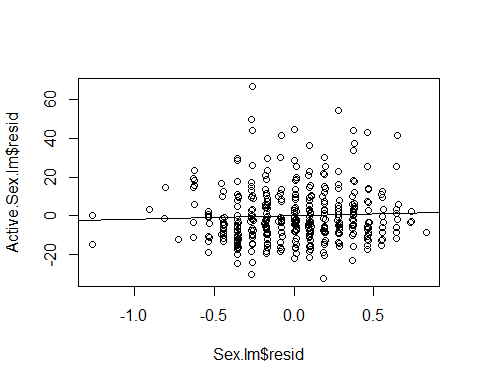
# Where active is predicted by rest and sex, not hgieht   
Active.Hgt.lm = lm(Active~Rest+Sex, data=Pulse)  
Hgt.lm = lm(Hgt~Rest+Sex, data = Pulse)  
plot(Active.Hgt.lm$resid~Hgt.lm$resid)  
mod2 = lm(Active.Hgt.lm$resid~Hgt.lm$resid)  
abline(mod2)



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)

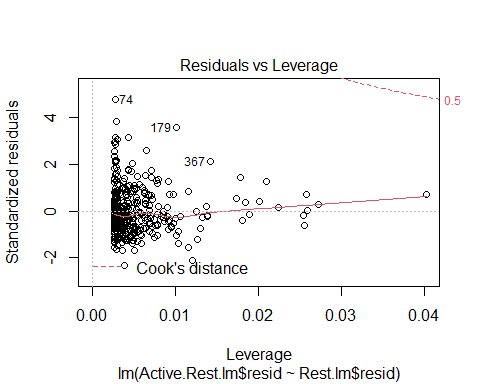


summary(mod3)

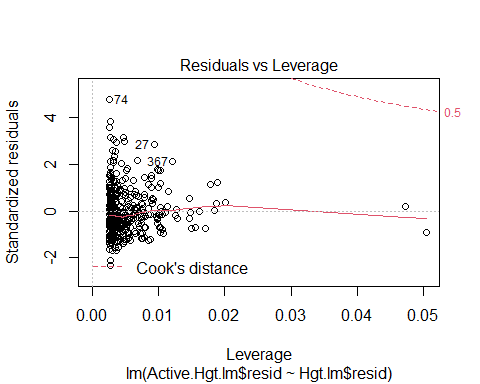
##   
## Call:  
## lm(formula = Active.Sex.lm$resid ~ Sex.lm$resid)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.730 -9.381 -2.691 6.817 67.413   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -9.785e-16 7.300e-01 0.000 1.000  
## Sex.lm$resid 1.909e+00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

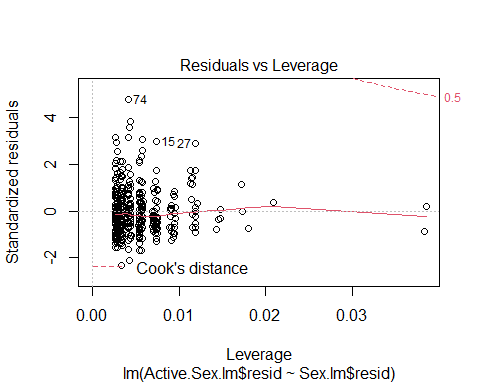
# Now we check each of these for leverage and influence   
plot(mod1, 5)



plot(mod2, 5)

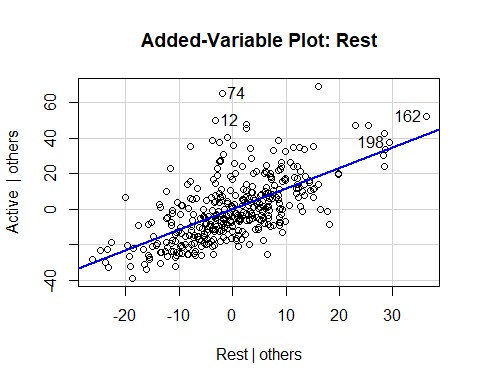


plot(mod3, 5)

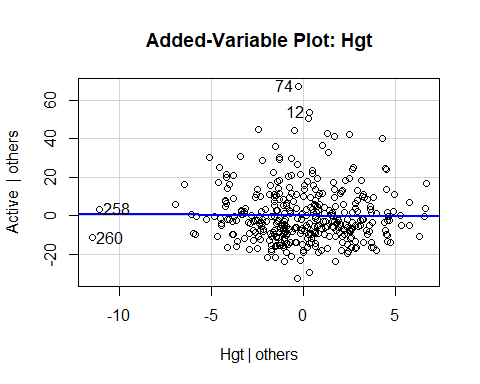


**Added Variable Plots** HOW TO DO THE ADDED VARIBALE PLOT EASILY AND QUICKLY IN R

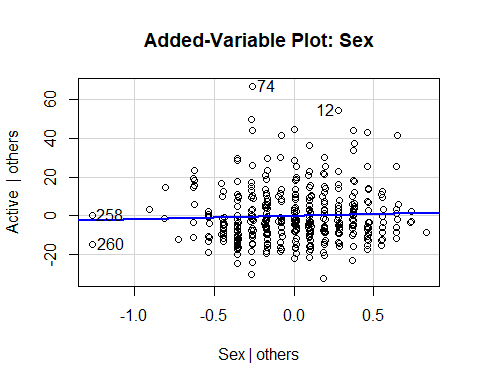
library(car)  
  
# Give the lienar model and what you want to look at as the added variabe   
# So it looks at the thing in pulse.lm, but takes out rest in the first one and then compares what the residual plots would look like between those two   
avPlot(Pulse.lm, "Rest") # need to give the string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



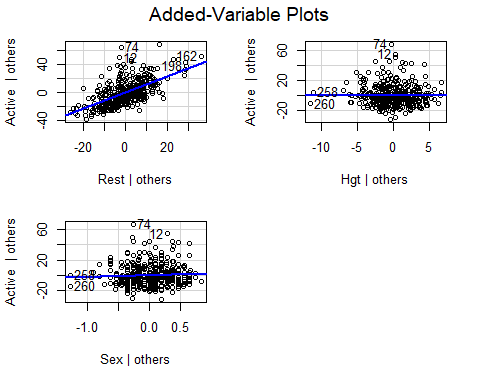
avPlot(Pulse.lm, "Hgt")



avPlot(Pulse.lm, "Sex")



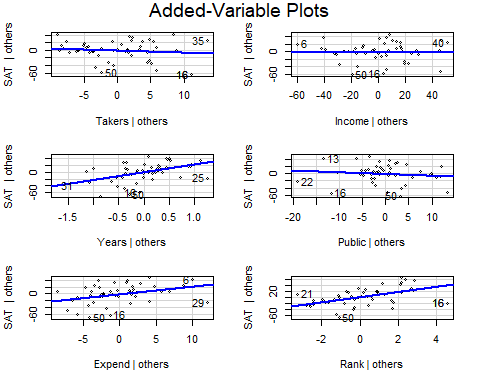
avPlots(Pulse.lm, ~.) # This is an avplot for all of the variables in the Pulse.lm linear regression model



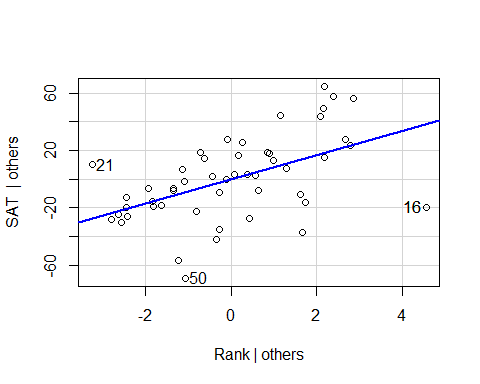
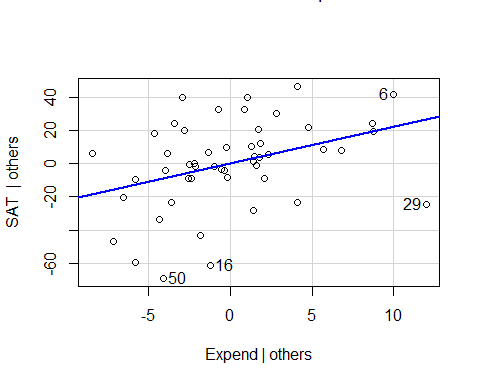
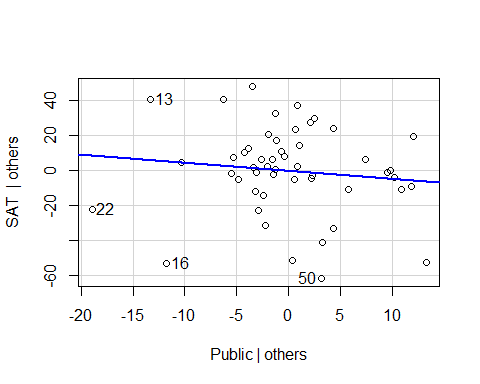
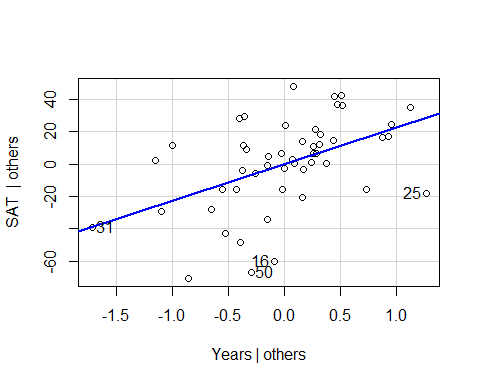
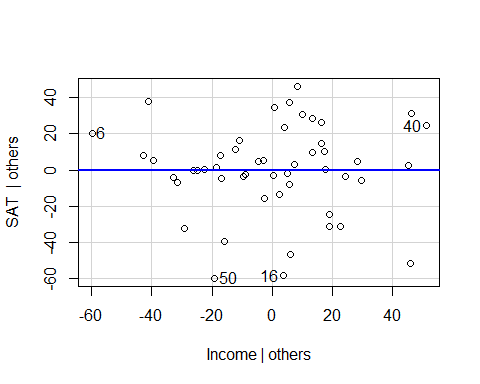
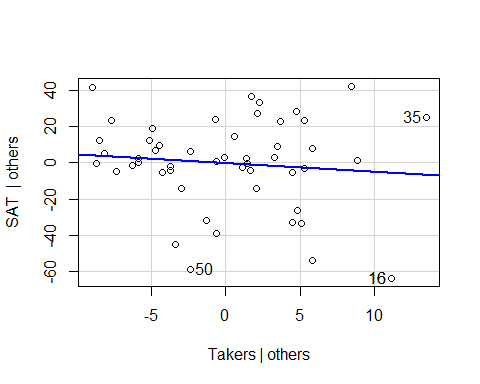
# This iwll make an avPlot for all the variables in teh pulse.lm linear regression model so you don't have to   
# THIS IS PLURARL   
# Its not super great in how it does it   
# YOu have to do it by all the predictors  
# issues: it's really hard to read it's really jumbled, you can read it   
# All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