STOR 455 Class 6 Outliers and Points of Influence

# message=FALSE, warning=FALSE suppress warnings and messages from appearing in knitted files  
  
library(readr)  
library(Stat2Data)

**Single Quantitative Predictor Model** Notation:  
- Y = Response variable - X = Predictor variable

Assume (for now) that both Y and X are quantitative variables.

**Simple Linear Model** X = Single quantitative predictor Y = Quantitative response

Find a line that best summarizes the trend in the data.

Y = Bo + B1X + E Response = intercept + Slope\*Predictor + Random Error

**Simple Linear Model- Conditions** **Model:** 1.Linearity: The means for Y vary as a linear function of X. **Error:** 2.Zero Mean: The distribution of the errors is centered at zero. 3.Constant variance: The variance for Y is the same at each X. (Homoscedasticity) 4.Independence: No relationships among errors. 5.Normality: Residuals are normally distributed (sometimes) At each X, the Y’s follow a normal distribution.

*Look at* What potent do these points have to influence our model?

**Types of “Unusual” Points in SLM** - Two Types - **Outlier:** A data point that is far from the regression line. –Points really above or below the regression line – Doesn’t always have a lot of influence on the model – Could be big enough that it has influence, but mostly depends on the value of the predictor – Data points that are closer to the edges of the predictor value (high or low) have a higher chance of having inlfuence in our model - **Influential point:** A data point that has a large effect on the regression fit. – can come from many things

**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

*This notebook covers the first two from above* - leverage = the potiential for a certain value to have infleuence on the model - Cook’s distance combines a lot of the things to do some calculations for us

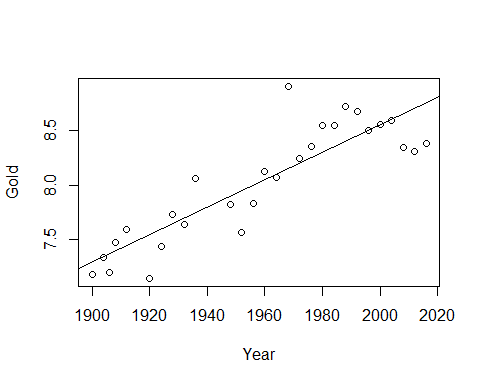
**Raw Residual** ei = yi - yhati *How can we tell if a residual is unusually large?*  CONTEXT! Example: Y = GPA ei = 2.6 is very large Y = SAT ei = 2.6 is very small

**Example: Men’s Olympic Long Jump**

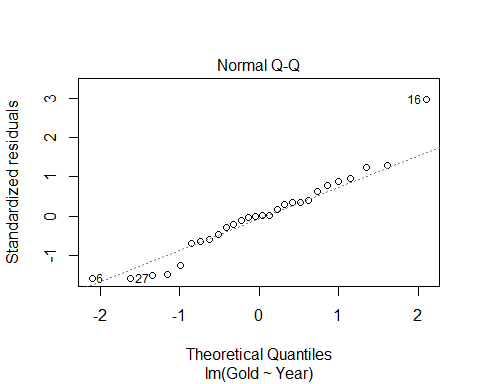
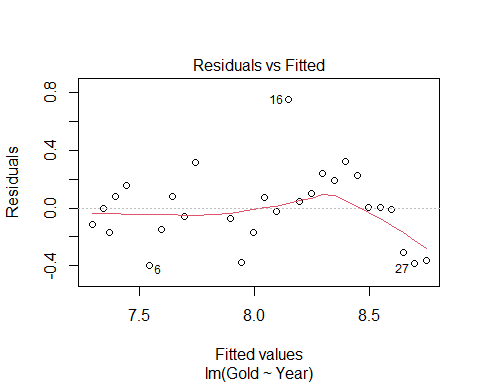
data("LongJumpOlympics2016")  
head(LongJumpOlympics2016)

## Year Gold  
## 1 1900 7.185  
## 2 1904 7.340  
## 3 1906 7.200  
## 4 1908 7.480  
## 5 1912 7.600  
## 6 1920 7.150

plot(Gold~Year, data=LongJumpOlympics2016) # Predict longjump distance by year   
GoldModel = lm(Gold~Year, data=LongJumpOlympics2016)  
abline(GoldModel) # Draw the line we made onto the plot



plot(GoldModel, 1:2) # To see the residual models, we see that there is one really big outlier



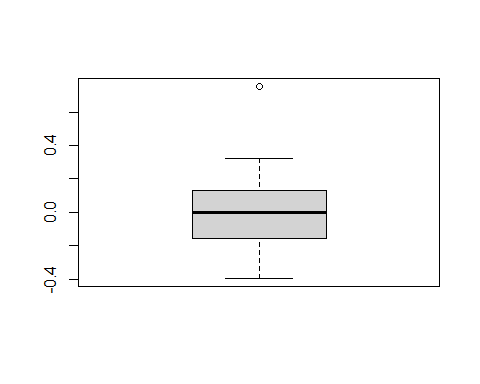
# Linearity looks like an issue because of the fitted plot. The red line goes down; the prediction curves   
# Point 16 is our outlier   
# R tells us which row this is because it looks like an ourlier to R   
# Looking at the normal QQ Plot, normal might be an issue   
# Constant variance doesn't appear to be an issue   
summary(GoldModel)

##   
## Call:  
## lm(formula = Gold ~ Year, data = LongJumpOlympics2016)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.39610 -0.15495 -0.00137 0.11606 0.75349   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -16.470194 2.666282 -6.177 1.56e-06 \*\*\*  
## Year 0.012508 0.001361 9.191 1.19e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2595 on 26 degrees of freedom  
## Multiple R-squared: 0.7646, Adjusted R-squared: 0.7556   
## F-statistic: 84.47 on 1 and 26 DF, p-value: 1.192e-09

# Look at estimate of the year: For every 1 year increase than that's how many meters we think the winning long distance jump is going to increase as well   
# Every 4 years it looks like its increasing by 5 cms

**What if we wanted to see what the data looked like without the outlier?**

boxplot(GoldModel$residuals)



# Outliers are more than 1.5 IQR’s beyond the Quartiles  
# Will give idea of how different that one value is from the others   
  
# Wee see there is an outlier, but how much of an outlier is this?   
# LOOK AT STANDARDIZED VALUES OF RESIDUALS INSTEAD  
  
max(GoldModel$residuals)

## [1] 0.7534932

which.max(GoldModel$residuals)

## 16   
## 16

**Standardized Residuals** - HOW TO TELL HOW MUCH OF AN OUTLIER THIS IS - ROughly equal to the acutal - predicted/stdeve; basically a zscore, but not exactly

* Fact: If X has mean mu and std. dev, then (𝑋−𝜇)/𝜎 has mean 0 and std. dev.=1.
* For residuals: mean=0 and std. dev. of errors
* Standardized Residuals about equals (yi - yhat)/stad dev of the population errors
* Look for values beyond +/-2 (mild) or beyond +/-3
* Once you have fit mymodel=lm(Y~X)
* Use: rstandard(mymodel)

*notes* - It will give us a thing centered at zero +/- unites; and that’s how many std they are away from teh average - THink about this as: - Once you are +/- 2 std away and its normally dist, then you’re in teh outer 5% of the data, so that’s starting to eb an outlier - IF you’re +/-3 away, then you’re into the .05 of the data and outliers, so it’s pretty extreme

*Below* - If we look at the standaized residual, it’s 2.96. so it’s 2.96 std above the line, which is an outlier

rstandard(GoldModel) # Put the model in

## 1 2 3 4 5 6   
## -0.45846037 -0.02447194 -0.69927805 0.34245659 0.62395356 -1.58872766   
## 7 8 9 10 11 12   
## -0.60354515 0.33349306 -0.22282237 1.24041457 -0.28038843 -1.47783119   
## 13 14 15 16 17 18   
## -0.65303672 0.28865616 -0.10391829 2.96025507 0.17095826 0.40756162   
## 19 20 21 22 23 24   
## 0.96124413 0.76569814 1.28443310 0.89029407 0.01303494 0.01295953   
## 25 26 27 28   
## -0.02754601 -1.24791995 -1.58338033 -1.51206192

# Will show standardized residuals   
# Taht's great when we have a small dataset, but if its not small, then don't use the above code   
  
which.max(GoldModel$residuals)

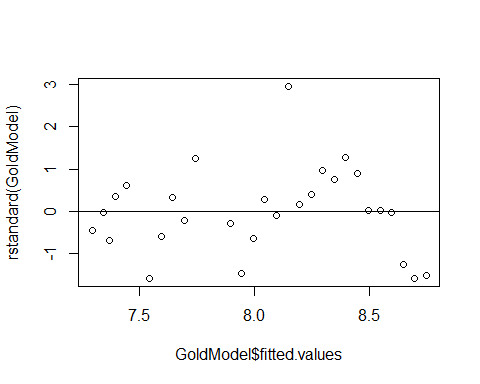
## 16   
## 16

rstandard(GoldModel)[16] # This will target the key point we ar elooking at

## 16   
## 2.960255

**We now know it’s an outlier, BUT WE DONT KNOW IF IT HAS ANY INFLUENCE ON TEH MODEL YET** - Can plot the rstand of the residuals by the fitted values - The plot is going to look identical to the others, other htan the axies, so but it’s a different measure of scale - How much infleunce is that really having?

plot(rstandard(GoldModel)~GoldModel$fitted.values)  
abline(0,0)

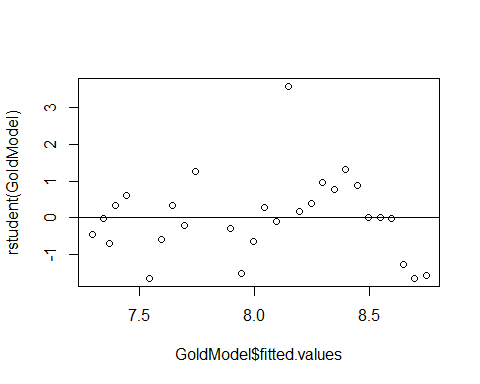
 **THING ABOUT THE STUDNTIZED RESIDUALS TO SEE HOW MUCH INFLUENCE SOMETHING HAS** - Standard = Outliers - Student = Influence (uses a different standard deviation)

**Studentized Residual** - Takes the single data out of the dataset, make a new regression line and get a new std of resid and seeing how far away the new line is from the outlier point in terms of the new standard residuals - If we take out an oulter, the std of the residuals is going to go down because we are removing an extreme case, so now its going to take more std to get to the outlier, and its going to give us a value that is bigger - When we see a studentized residual that is larger thant the stadnard residual, then it tells us that by removing this point, we are reallying changing the varibaility of the model and condenseing it more. - uses a different standard deviation than standariza

* **Concern:** An unusual value may exert great influence on the fit
* Its residual might be underestimated because the model “moves” a lot to fit it and/or
* The standard error of regression may be inflated due to the outlier error
* **Studentize:** Fit the model without that case, then find new 𝑦−𝑦hat and 𝜎\_𝜀 (of the population) to standardize. (R does this for every point)

**Influence** The effect of a single data point on the regression line depends on: - how well it matches the “trend” of the rest of the points - how “unusual” is its predictor value

plot(rstudent(GoldModel)~GoldModel$fitted.values)  
abline(0,0)



rstudent(GoldModel)[16]

## 16   
## 3.565083

# When we took the outlier from teh thing, its more away, which is bigger   
# There is some kind of influence here, but is it really noticable influence?

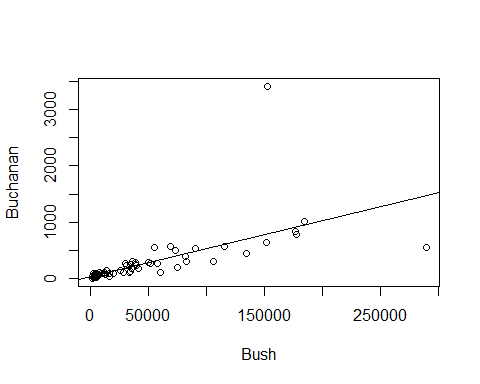
# plot(IceModel3) # From the previous notes   
  
# max(rstandard(IceModel3))  
# max(rstudent(IceModel3))  
# When we look at the values, themore different they are the mode influence they have in the model   
# The more close they are, then they have less influence on the model   
# No real bounds on what is a big or little influence on set number ot look at

**Dataset: PalmBeach** - County vote counts in Florida (n=67) for George Bush and Pat Buchanan in 2000. - Model: Use Bush votes to predict Buchanan votes.

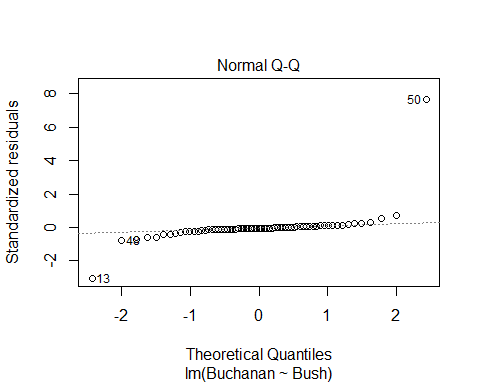
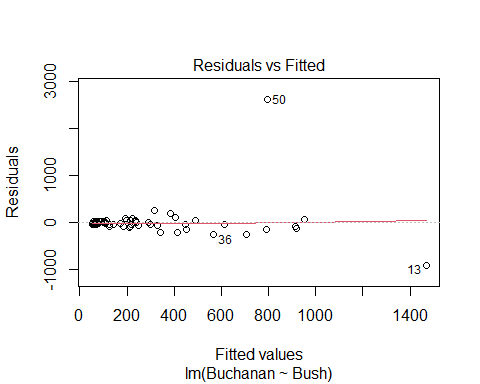
data(PalmBeach)  
head(PalmBeach)

## County Buchanan Bush  
## 1 ALACHUA 262 34062  
## 2 BAKER 73 5610  
## 3 BAY 248 38637  
## 4 BRADFORD 65 5413  
## 5 BREVARD 570 115185  
## 6 BROWARD 789 177279

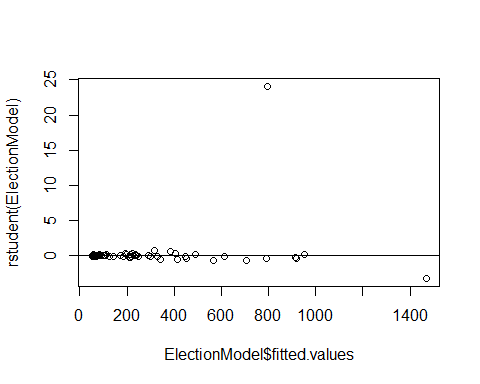
ElectionModel = lm(Buchanan~Bush, data=PalmBeach)  
# Bush = republican   
# Buchana - the other person   
  
plot(Buchanan~Bush, data=PalmBeach) # Look at the residuals and check the conditions   
abline(ElectionModel)

 **Example: Palm Beach Butterfly Ballot** - Palm Beach is the outlier

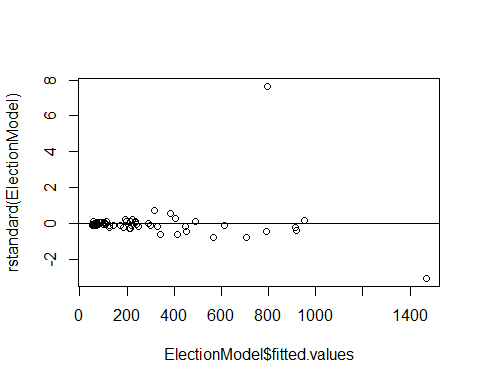
plot(ElectionModel, 1:2)



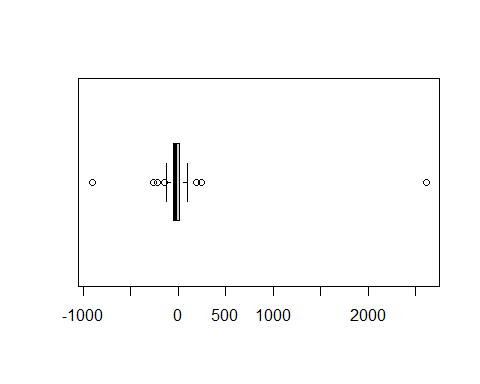
plot(rstudent(ElectionModel)~ElectionModel$fitted.values)  
abline(0,0)



plot(rstandard(ElectionModel)~ElectionModel$fitted.values)  
abline(0,0)



boxplot(ElectionModel$residuals, horizontal=TRUE) # Look at the outliers, there looks like ther eare a lot of outliers, but it's okay we are only looking at one of them



**What to do with an extreme residual?** - Try a transformation (Loging works really well, log(Data)) - Redo the analysis with the point omitted

*Below, we redo with the point omitted*

newdata = subset(PalmBeach, County!="PALM BEACH")  
  
ElectionModel\_noPB = lm(Buchanan~Bush, data=newdata)  
  
summary(ElectionModel)

##   
## Call:  
## lm(formula = Buchanan ~ Bush, data = PalmBeach)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -907.50 -46.10 -29.19 12.26 2610.19   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.529e+01 5.448e+01 0.831 0.409   
## Bush 4.917e-03 7.644e-04 6.432 1.73e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 353.9 on 65 degrees of freedom  
## Multiple R-squared: 0.3889, Adjusted R-squared: 0.3795   
## F-statistic: 41.37 on 1 and 65 DF, p-value: 1.727e-08

summary(ElectionModel\_noPB)

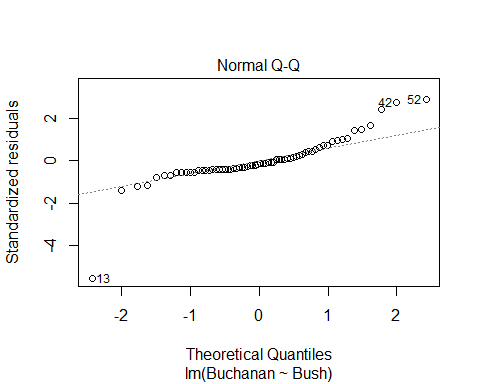
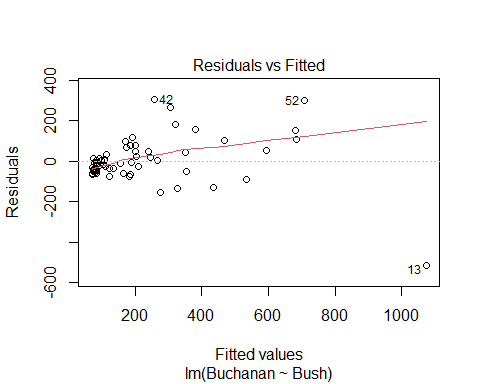
##   
## Call:  
## lm(formula = Buchanan ~ Bush, data = newdata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -512.43 -47.97 -17.09 41.78 305.45   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.557e+01 1.733e+01 3.784 0.000343 \*\*\*  
## Bush 3.482e-03 2.501e-04 13.923 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 112.5 on 64 degrees of freedom  
## Multiple R-squared: 0.7518, Adjusted R-squared: 0.7479   
## F-statistic: 193.8 on 1 and 64 DF, p-value: < 2.2e-16

# Compare teh summary of old vs new model   
# We see that the new model residuals std error is really high; it s std value   
# The origanl value is a lot smaller for std

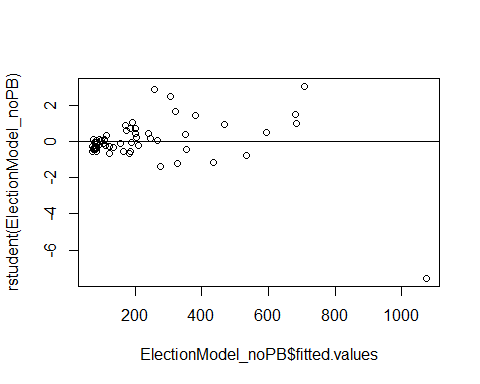
**Model with/without Palm Beach** - If compare the rstudent and rstandard of the with palm beach, we see that there is a really big jump between tehse numbers; tells us we have a value that is taking the whole regression and dragging it - So all our other predictiosn are going to be higher becuase of this vlaue

* If we look at the intercept and slope
* the slope is more dramatic when we take out the point, with the intercept goes down
* the residual standarad errors are very different between teh two
* the linearirity is really good without the palm beach
* the normal residuals look pretty bad when you take out palm beach

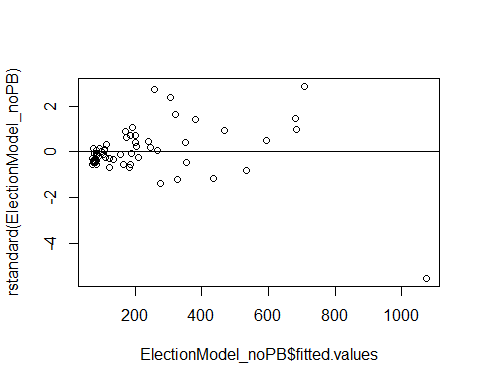
plot(ElectionModel\_noPB, 1:2)



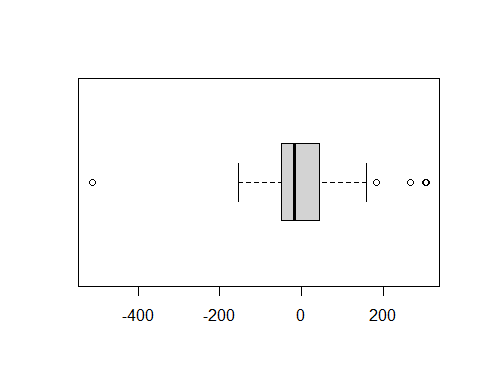
plot(rstudent(ElectionModel\_noPB)~ElectionModel\_noPB$fitted.values)  
abline(0,0)



plot(rstandard(ElectionModel\_noPB)~ElectionModel\_noPB$fitted.values)  
abline(0,0)



boxplot(ElectionModel\_noPB$residuals, horizontal=TRUE)



**BE careful with the student and standard** - look at how differen tbetween student and standard, you cna’t just know based on the one value