STOR 455 Class 7 Outliers and Points of Influence Again

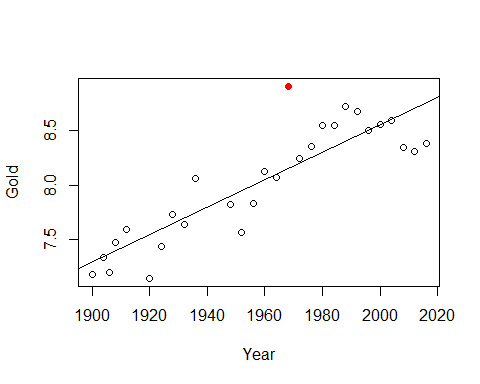
library(Stat2Data)  
  
data("LongJumpOlympics2016")  
data("PalmBeach")

**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.

**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

*Below:* We are making the same model as last time, in class 06

GoldModel=lm(Gold~Year, data=LongJumpOlympics2016)  
  
plot(Gold~Year, data=LongJumpOlympics2016)  
abline(GoldModel)  
points(LongJumpOlympics2016$Year[16], LongJumpOlympics2016$Gold[16], col="red", pch=16)



# Points just stands out a certain point   
# We need to give it the x and y coordinates, the the color and style   
  
summary(GoldModel) # The 0.2595 meaters is the standard error of the residual

##   
## 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

**Standardized Residuals** - Roughly equal to zscore - For residuals: mean=0 and std. dev. ≈𝜎 ̂\_𝜀 - Standardized Residual about = (yi-yhat)/std - Look for values: beyond +/-2 (mild) or beyond +/-3 - **Definition:** The standardized residuals are: – std. resi = ((yi-yhat)/(stdsqrt(1-hi))); where hi = leverage

*SQRT 1=leverae* - It wont tell us that it has influence, but it will tell us the potential for influence -

StanResidEst = GoldModel$resid/summary(GoldModel)$sigma # Pulls out the standard error of the residuals   
  
StanResidEst - rstandard(GoldModel) # What this leverage is

## 1 2 3 4 5   
## 0.0310213900 0.0014925005 0.0404484002 -0.0187751023 -0.0306759692   
## 6 7 8 9 10   
## 0.0624609303 0.0211917539 -0.0104655385 0.0062653144 -0.0314031236   
## 11 12 13 14 15   
## 0.0055095767 0.0275829381 0.0118383564 -0.0052073979 0.0019121259   
## 16 17 18 19 20   
## -0.0568637884 -0.0034991146 -0.0090387925 -0.0233983067 -0.0206451799   
## 21 22 23 24 25   
## -0.0385895552 -0.0299054544 -0.0004902075 -0.0005455954 0.0012967412   
## 26 27 28   
## 0.0655674849 0.0926377694 0.0982533934

**What is this leverage?** - Depending on where the value is, depending ont eh x coord, determines the leverage - Things that follow the mean along the predictor because it doesn’t have a lot of leverage - If it’s twoards the right or left of end the range of values, then even a small, very little values could have a lot fo influence on the model - Think of leverage as a seesaw. – if you have a long seesaw on one side, but short on the other side, it doesn’t matter who is on the short side, they won’t have as much leverage. – If you have a long seesaw on the other side, then if it’s an adult or child may impact the stronger amount of leverage. Think of if it was an adult of child by how far up or down the y axis it goes. If it’s an adult, it’s oging to be a bigger different from the axis, but if it’s a child it’s going to be smaller and have less influence. - **Bottom line** A lower weight, if its further away fromt eh balance point, can have more effect on the model than a higher weight that is closer to the balance point – **The balance point is the mean**

StanResid = GoldModel$resid/(summary(GoldModel)$sigma \* sqrt(1 - hatvalues(GoldModel)))  
  
StanResid - rstandard(GoldModel)

## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26   
## 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0   
## 27 28   
## 0 0

**Studentized Residuals** - If we took that point out of the model, how does that impact the variability of the residuals? - Definition: The studentized residuals are: stud.resi = *same as standardized, but without the one point* - If th epoint had a lot of influence on the varibaility, then the studentized value would be much bigger than the standarized value

* **The more different the standardized vs studentized values are, the more influence that point has on the model**

**Typical Leverage** - **For one point:** - For a simple linear model: hi = (1/n)+(((xi-xbar)2)/sum(xi-xbar)2) - **For all leverage in a model:** - sum(leverage) = sum(1/n) + ((sum(xi-xbar)2)/sum(xi-xbar)2) = 1+1 = 2 - The sum of the leverages of all of the points = 2 – this is usefulbecause we can think of what a typical leverage value is – If all teh leverages are 2, then 2/n = give you what the mean leverage is - *Look for:* – leverage > 2(2/n) – leverage hi > 3(2/n)

*Notes* - How different is this point from the average predictive value in compraision to all those other values - Xi - xbar = predictor where the xi = a specific predictor value, and xbar = the mean of the values - Xi-xbar squared = in a horizontal way, how different is that point regardless if above or below mean - The sum(xi-xbar)^2 = how different is this point proportionally to the rest of the data - scared accoriding the the sample of the data with the 1/n - One point will haev mor einfluence if the data is small - values that are double or triple the average leverage are the ones we have to worry about because they may have influence in the model – **We are still talking about potenital here because we are only looking at the predictors** –Because we are only looking at years in this model, vs the entire context of the data –It tells us the data points we should look at more closely

# average leverage : 2/28   
# IF poitns are bigger than this, might have a lot of influence on the model   
  
2\*(2/28)

## [1] 0.1428571

3\*(2/28)

## [1] 0.2142857

# Above are just the cut off points for influence, we don't really care too much about them   
  
# The below code will show the ordered data because of the way the data is organized in the dataset   
# We can see how the leverages are going to be similar there where the points that are small, very different for the mean, they wil start with a higher leverage, and by the time we get the middle values, its going down to a leverage of nearly zero,   
# Get further waya from average year, we will get higher leverages reported   
  
hatvalues(GoldModel)

## 1 2 3 4 5 6 7   
## 0.13075010 0.11825693 0.11234035 0.10664378 0.09591065 0.07708445 0.06899139   
## 8 9 10 11 12 13 14   
## 0.06177835 0.05544533 0.04999234 0.03891349 0.03698058 0.03592770 0.03575484   
## 15 16 17 18 19 20 21   
## 0.03646200 0.03804918 0.04051638 0.04386361 0.04809086 0.05319813 0.05918543   
## 22 23 24 25 26 27 28   
## 0.06605275 0.07380009 0.08242745 0.09193483 0.10232224 0.11358967 0.12573712

# Iif you look at the 16th value, it has a leverage of 0.038; based onteh creitera foor what tis a big lleverage value, this point doesn't appear to be high leverage   
# the 16the point has to be a REALLY big outlier for it to have influence, because its potential to have influence is low

**Wanted to look at the leverages better** - Sort them to have a better idea of where the potenitals are to gave influence - Sort the hat values - sort will by default go from acending order/increasing order - mostly interested in the biggest leverages - Want to change it to decreasing = TRUE so we see the biggest leverages first instead of the smallest ones first

2\*(2/28)

## [1] 0.1428571

3\*(2/28)

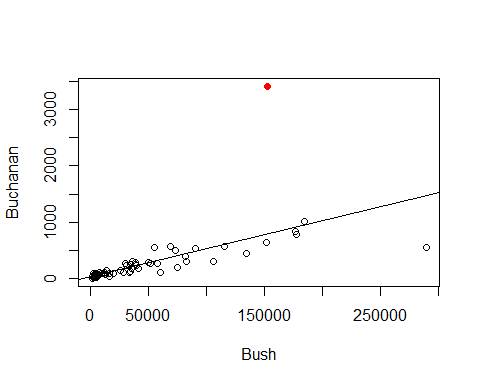
## [1] 0.2142857

sort(hatvalues(GoldModel), decreasing = TRUE)

## 1 28 2 27 3 4 26   
## 0.13075010 0.12573712 0.11825693 0.11358967 0.11234035 0.10664378 0.10232224   
## 5 25 24 6 23 7 22   
## 0.09591065 0.09193483 0.08242745 0.07708445 0.07380009 0.06899139 0.06605275   
## 8 21 9 20 10 19 18   
## 0.06177835 0.05918543 0.05544533 0.05319813 0.04999234 0.04809086 0.04386361   
## 17 11 16 12 15 13 14   
## 0.04051638 0.03891349 0.03804918 0.03698058 0.03646200 0.03592770 0.03575484

# The one with the biggests index is 0.1307, even the smallest and biggest predicotr value aresnt so far from teh mean predictor compared to the other values that hey have potiental for influence.   
# They would need to be big outliers to have influence   
# If we had more datapoints, then we could also just look at some of the values   
# The head function can be added to the sorting hatvalues

ElectionModel=lm(Buchanan~Bush,data=PalmBeach)  
plot(Buchanan~Bush,data=PalmBeach)  
abline(ElectionModel)  
points(PalmBeach$Bush[50], PalmBeach$Buchanan[50], col="red", pch=16)



# Red is palm beach

#Average leverage = 2/67  
  
#Potiential for influence   
2\*(2/67)

## [1] 0.05970149

3\*(2/67)

## [1] 0.08955224

head(sort(hatvalues(ElectionModel), decreasing=TRUE), n=10)

## 13 52 6 29 50 16 48   
## 0.29747301 0.10761608 0.09859725 0.09820784 0.07085197 0.07007421 0.05365982   
## 5 36 53   
## 0.03899504 0.03331541 0.02511932

# We see that point 13 probably has potiential leverage   
  
# We want to see the first 6 with the highest potienital for influence so we can look at them closer  
# We want all columns, so just leave it with the ,  
# This tells us teh counties in the order of teh output, not the leverages though   
# Want to add the leverages to it though   
PalmBeach[c(6,13,16,29,50,52),]

## County Buchanan Bush  
## 6 BROWARD 789 177279  
## 13 DADE 561 289456  
## 16 DUVAL 650 152082  
## 29 HILLSBOROUGH 836 176967  
## 50 PALM BEACH 3407 152846  
## 52 PINELLAS 1010 184312

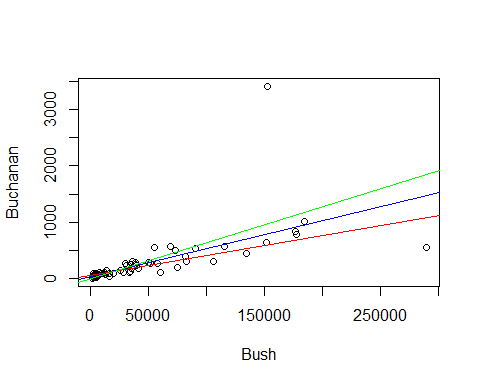
# Defines new variable inside of the dataframe  
PalmBeach$Leverage = hatvalues(ElectionModel)  
# This is putting the leverage into the dataframe   
  
# Now run this again: PalmBeach[c(6,13,16,29,50,52),]  
PalmBeach[c(6,13,16,29,50,52),]

## County Buchanan Bush Leverage  
## 6 BROWARD 789 177279 0.09859725  
## 13 DADE 561 289456 0.29747301  
## 16 DUVAL 650 152082 0.07007421  
## 29 HILLSBOROUGH 836 176967 0.09820784  
## 50 PALM BEACH 3407 152846 0.07085197  
## 52 PINELLAS 1010 184312 0.10761608

# This will now include the palmbeach levearge

*Want to visualize the impact dave county has vs palm beach* - Dave has a smaller outlier, it’s leverage is high so it could havea similar impact on the model

plot(Buchanan~Bush,data=PalmBeach)  
abline(ElectionModel, col="blue")  
# The above is the model we have now   
  
# Below, make new model with no palmbeach   
NoPalmBeach=subset(PalmBeach,County!="PALM BEACH")  
ElectionModel\_noPB=lm(Buchanan~Bush,data=NoPalmBeach)  
abline(ElectionModel\_noPB, col="red")  
  
# Below, make a new model with no dade  
NoDade = subset(PalmBeach,County!="DADE")  
ElectionModel\_noD=lm(Buchanan~Bush,data=NoDade)  
abline(ElectionModel\_noD, col="green")



# Then we compare the lines that we have written to see how the slope is different.   
# Red = no palm beach   
# Green = no dade   
# Blue = with both   
# The change in slope appears the same, so it tells me that the impact these datapoint shave on our model are about similar   
# Even though one is a bigger outlier, it has a lower potenital to have infleunce,   
# The other is smaller outlier, but has a high leverage, it has abotu the same typ eof influence

*NOtes* - FOr high leverage, doesnt need to be a big outlier to have an impact on the model - for high outlier, it doesnt have to have that much leverage - We are looking at cook’s distance, which combines teh two above things and quantify the value instead of the other vlaues - Cook’s will tell you if the pointhas influence on the model - Brings in stanard resid; the leverage of the point; and how many predictors there are - Cooks’ distance is acombien of leverage, student, adn standard

**Cook’s Distance** - How much would the fit change if one data value were omitted? - 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.

*Cook’s Di:* = (((std.resi)^2)/(k+1))\*(hi/(1-hi))

head(sort(cooks.distance(ElectionModel), decreasing=TRUE), n=5)

## 50 13 48 36 6   
## 2.231935359 1.981365681 0.016228381 0.009781056 0.007928585

# Two ar emuch bigger adn the rest of smaller, those two points rae the main that are havign influence on our model   
# 50 = =palm beach   
# 13 = Dade   
# They're not exactly the same, but have about teh same impact ont eh model we are making   
  
#We've seen Cook's distance between with teh plots for the model   
plot(ElectionModel,5)  
# Will show you what point shave leverage   
# LOOK AT WHAT THE THING IS   
# HOw big of an outlier is this value   
# Mode values have standardiex values within +/- 2  
# Th etwo that have big differences are point 50 and 13  
# Point 50 = std score of 8   
# Point 13, Dade, has a -3.0 std value; but it is high leverage   
# Points low leverage = closer to the left side of the graph   
# Further right = different data, more leverage   
  
# O.5  
# Gives us a cook's distance of 0.5  
# OUtler line: 1 = cook's distance of 1   
# Above these lines, we can claim that by putting the leverage and output together it's having an influence on the model   
# POint 50, has small leverage, but a big outlier that it's goig outside of the curve and having influence on the model   
# Dade county has a huge leverage, but it's not an outlier, but it's enough to put it outside of the curve and have an high impact on teh model   
  
# Think about: WHY DO THESE POINTS HAVE IMPACT ON THE MDOEL>   
# DO WE WANT THEM TO IMPACT THE MODEL?   
# sometimes might not want to inlcude some things int eh model, but it's good to look at it anyways so you know what is going on   
abline(v = 4/67, col="blue", lty=3)  
abline(v = 6/67, col="blue", lty=3)

