Chapter 3 Exercise 3

Allen Church

September 25, 2019

Load necessary data and view

load("C:/Users/WB537822/Desktop/stats1/Ch3\_Exercise3\_Height\_and\_Wages\_UK.RData")  
  
data <- dta  
  
View(data)

3a. Estimate model where height at 33 explains income at age 33. Explain beta hat 1 and beta hat 0

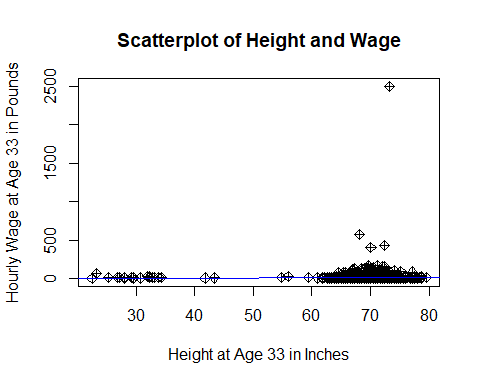
#Create variables for income and height  
income <- data$gwage33  
height <- data$height33  
  
#Create OLS model and summarize  
OLSresults <- lm(income ~ height)  
summary(OLSresults)

##   
## Call:  
## lm(formula = income ~ height)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.68 -4.96 -3.11 -0.68 2488.68   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -6.5994 12.2143 -0.540 0.589  
## height 0.2447 0.1757 1.393 0.164  
##   
## Residual standard error: 44.71 on 3694 degrees of freedom  
## Multiple R-squared: 0.0005252, Adjusted R-squared: 0.0002546   
## F-statistic: 1.941 on 1 and 3694 DF, p-value: 0.1636

The value for beta hat 1 (0.2447) estimates that for each inch in additional height, the hourly wage increases by 0.2447 pounds. The value for beta hat 0 (-6.5994) estimates an hourly wage of -6.5994 for someone with 0 inches of height.

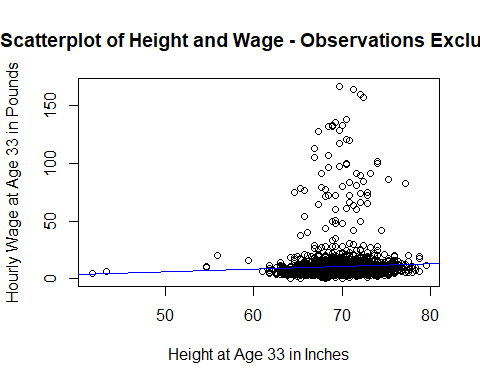
3b. Scatterplot of height and income, identify outliers

plot(height, income, main="Scatterplot of Height and Wage",  
 xlab="Height at Age 33 in Inches", ylab = "Hourly Wage at Age 33 in Pounds", pch=9)  
  
abline(lm(income ~ height, data=data), col='blue')

 The scatterplot above contains outliers where at least 8 entries have a height below 30 inches which - while possible - could be outliers. Additionally, there is an outlier of an individual with a wage of approximately 2500 pounds per hour.

3c. Create scatterplot, but exclude observations with wages per hour more than 400 pounds and height less than 40 inches.

#Select rows that match the above specifications  
newdata <- subset(data, gwage33 < 400 & height33 > 40)  
  
plot(newdata$height33, newdata$gwage33, main="Scatterplot of Height and Wage - Observations Excluded",  
 xlab="Height at Age 33 in Inches", ylab = "Hourly Wage at Age 33 in Pounds")  
  
abline(lm(income ~ height, data=data), col='blue')



The plot above seems to be a more reasonable basis for statistical analysis because we are now able to see more variance in observations. It also shows a stronger positive correlation between height and wage.

3d. Reestimate the model from part a but exclude 4 outliers with very high wages and outliers with height below 40 inches

#Subset newdata and exclude 4 highest wages and height below 40 inches  
newdata2 <- subset(newdata, gwage33 < 156 & height33 > 40)  
  
#Create new OLS model with data that excludes outliers, and summarize results  
OLSresults2 <- lm(newdata2$gwage33 ~ newdata2$height33)  
summary(OLSresults2)

##   
## Call:  
## lm(formula = newdata2$gwage33 ~ newdata2$height33)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.349 -3.698 -1.867 0.498 128.649   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.95343 4.56052 -1.525 0.127420   
## newdata2$height33 0.23136 0.06543 3.536 0.000411 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.94 on 3663 degrees of freedom  
## Multiple R-squared: 0.003402, Adjusted R-squared: 0.00313   
## F-statistic: 12.5 on 1 and 3663 DF, p-value: 0.0004112

Having dropped the outliers, the new beta hat 1 is 0.23136 vs. the old value of 0.2447. Dropping the outliers weakens the positive correlation between height and wages. However, an interesting observation is that standard error for beta hat 1 with outliers removed is 0.06543, whereas with outliers standard error for beta hat 1 was 0.1757.

3e. What happens when sample size is smaller? Reestimate bivariate ols model (newdata23), but limit analysis to first 800 observations. Which changes more from the results with the full sample: the estimated coefficient on height or estimated standard error of coefficient on height?

#Create small sample by selecting first 800 rows of the previously subsetted data and verify the number of rows  
small\_sample <- newdata2[1:800,]  
nrow(small\_sample)

## [1] 800

#Reestimate bivariate ols model with small sample and summarize results  
OLSresults3<- lm(small\_sample$gwage33 ~ small\_sample$height33)  
summary(OLSresults3)

##   
## Call:  
## lm(formula = small\_sample$gwage33 ~ small\_sample$height33)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.224 -3.351 -1.616 0.626 111.070   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.0164 8.0151 -0.751 0.4531   
## small\_sample$height33 0.2109 0.1155 1.827 0.0681 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.86 on 798 degrees of freedom  
## Multiple R-squared: 0.004164, Adjusted R-squared: 0.002916   
## F-statistic: 3.337 on 1 and 798 DF, p-value: 0.06811

The re-restimated bivariate OLS model above shows that the estimated standard error of the coefficient on height changed more when reducing the original data to 800 observations. Namely, reducing the number of observations increased the estimated standard error from 0.06543 with the full set to 0.1155 with the limited set.