Assignment 6 CS 4070

Nirmal Roy, 4724429

January 15, 2018

Question 1

A model for the regression of education on income(X_1), young(X_2) and urban((X_2)) is given by

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i \tag{1}$$

Assessing Leverage: Hat-Values

Figure 1 plots the hat values for all the states. The average hat value $\bar{h} = \frac{k+1}{n} = \frac{3+1}{51} = 0.078$. We can observe that AK has the highest hat value $= 0.531$ where as VT, Nm and UT falls in the range $2\bar{h}$ and $3\bar{h}$.

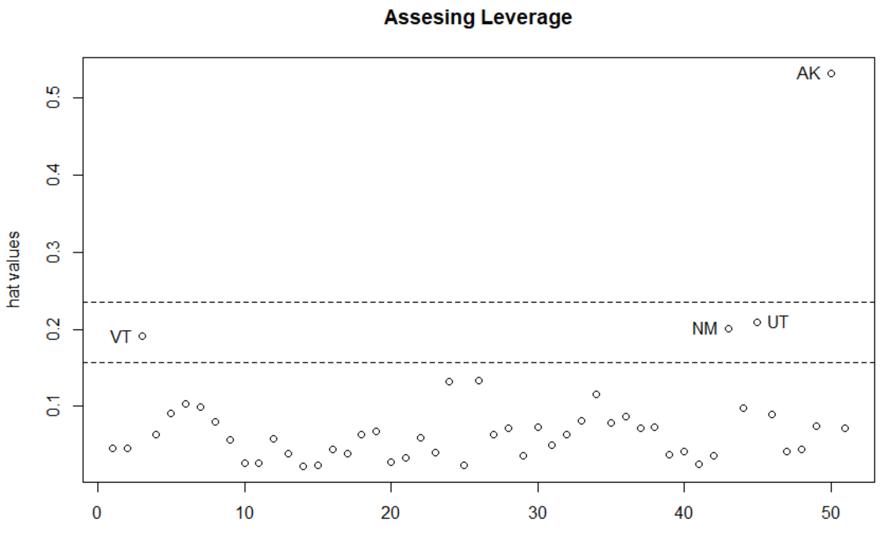


Figure 1: Plot of hat values for Anscombe data. The dashed lines are drawn at $2\bar{h}$ and $3\bar{h}$. AK has the highest hat value meaning it has highest influence

Detecting Outliers: Studentized Residuals

To identify an outlying observation, we need an index of the unusualness of Y given the Xs. It is provided by *studentized residuals*. Figure 2 shows a plot of the Studentized Residual values of different states. State CT = 2.51 is clearly an outlier as it lies well beyond the margin of $|E_i^*| \leq 2$

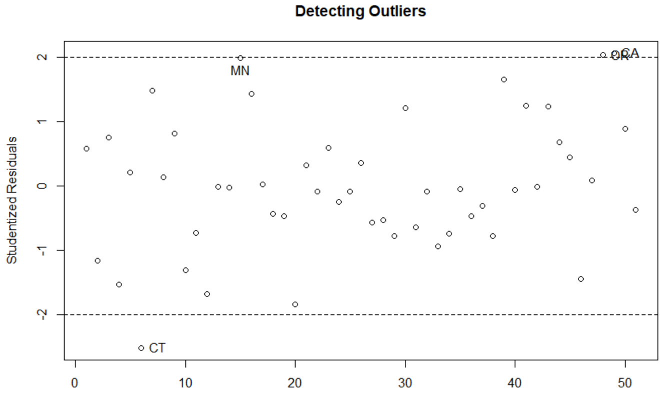


Figure 2: Plot of Studentized Residuals for Anscombe data. The dashed lines are drawn at -2 and 2.Only Ct lies significantly out side the cut off range

Influence on Coefficients and Standard Errors

Influence on regression coefficient combines leverage and discrepancy of data. A measure of calculating influence on coefficients is Cook's D which is shown in the first plot of Figure 3 Although the all the values are within the cut off value of $\frac{2}{\sqrt{n}} = 0.28$, AK stands off with the highest Cook's D of 0.225 . To test the influence plot on Standard Errors, a measure known as COVRATIO is used. Even in this case AK has the highest value of 2.17. In figure 4 we show an influence plot with hat value on x axis and studentized residuals on y axis and the size of the dots are proportional to cook D. We see that AK has high leverage and big influence of coefficients whereas CT is an outlier and also with relatively big influence on the coefficients.

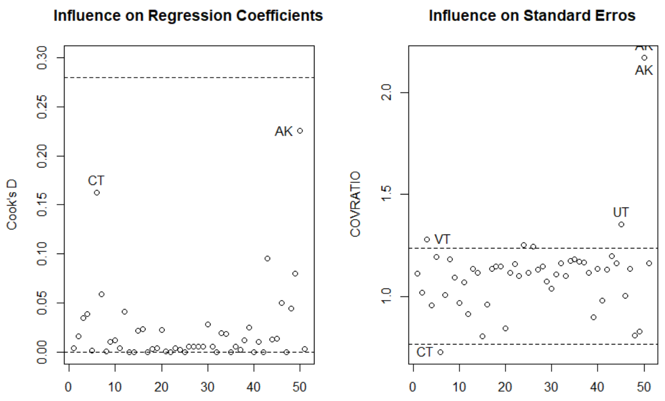


Figure 3: Left: Plot of Cook's D for Anscombe data. The dashed lines are drawn at $\frac{2}{\sqrt{n}}$ and $-\frac{2}{\sqrt{n}}$. Right: Plot of COVRATIO. The dashed line are drawn at $|\text{COVRATIO}_i - 1| > \frac{3(k+1)}{n}$

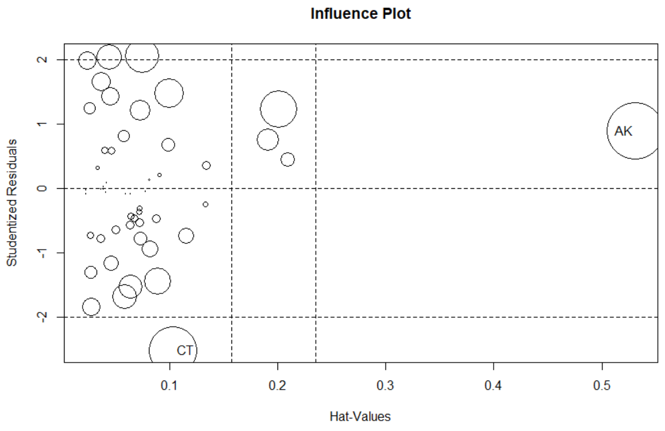


Figure 4: Influence Plot

Joint influence on coefficients and their standard error

Subsets of observation can be jointly influential. In figure 5, from the avPlots function in R, we see that AK and CT is jointly influential on the regression coefficients and standard error. Out of these, CT tend to decrease the slope of income and urban but it tends to increase the slope of young. On the other hand, AK remains above the regression line for all the three plots but it's not discrepant enough. Hence, we might conclude that CT being highly discrepant yet not high leverage but AK having high leverage but nit being discrepant enough, the joint influence is not that much on the final regression line.

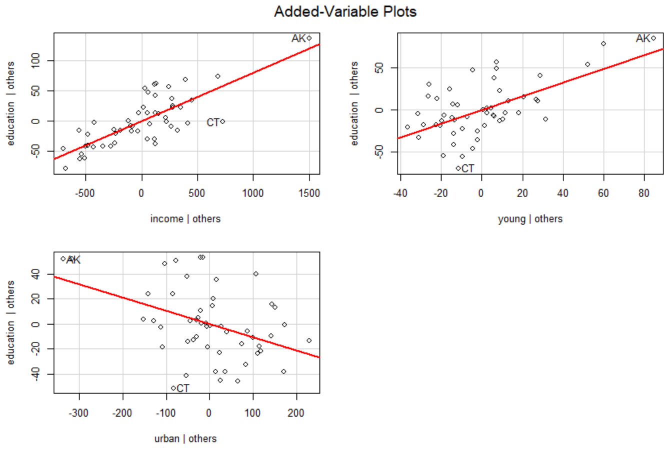


Figure 5: Added Variable Plots

Unusual State

AK stands out in the most of the diagnostics. So we regress education on the explanatory variables using equation 1 once with AK and once without AK and compare the coefficients and standard error in table 1

Table 1

	Full		Without AK	
	Estimate	**Standard Error**	**Estimate**	**Standard Error**
Intercept	-286.8	64.92	- 242.19	82.2
Income	0.08065	0.093	0.074	0.011
Young	0.8173	0.16	0.71	0.2
Urban	-0.1058	0.034	-0.087	0.041

The multiple correlation coefficient for the full model is 0.69 whereas for the reduced model is .57. The standard error for the full model is 26.09 whereas that for the reduced model is 26.75.

As we see, though the fit has decreased in terms of correlation coefficient and standard error without AK but it exerts a really high influence on the regression as is apparent from the changes in the regression coefficient. Although CT's removal would have improved the regression fit but it's not as influential as AK.

Question 2

As seen from figure 6, the residual errors are normal for the reduced model. None of the points lie beyond the 95% confidence envelop for the QQ plot. Hence, we don't need transformation for this data. Although, the density plot suggests we can make the data more symmetric with a power transformation but it is not required.

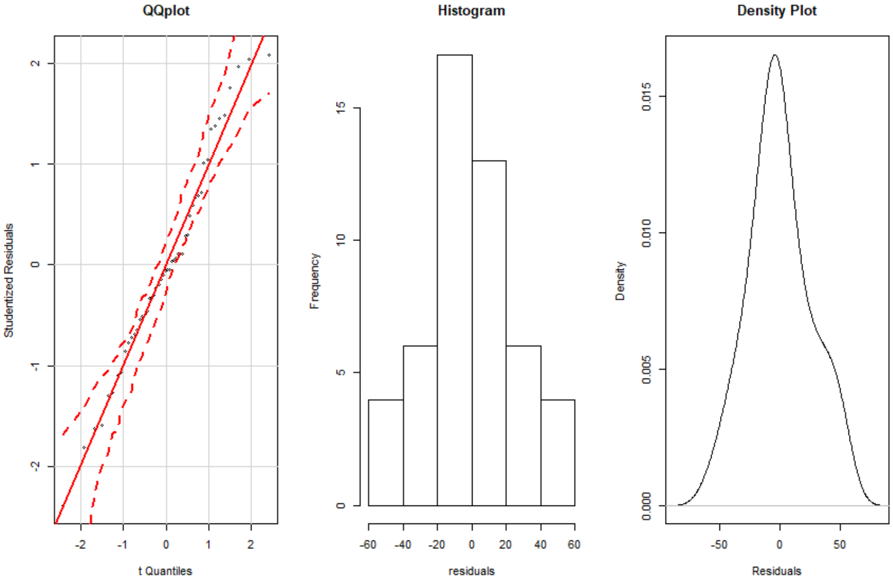
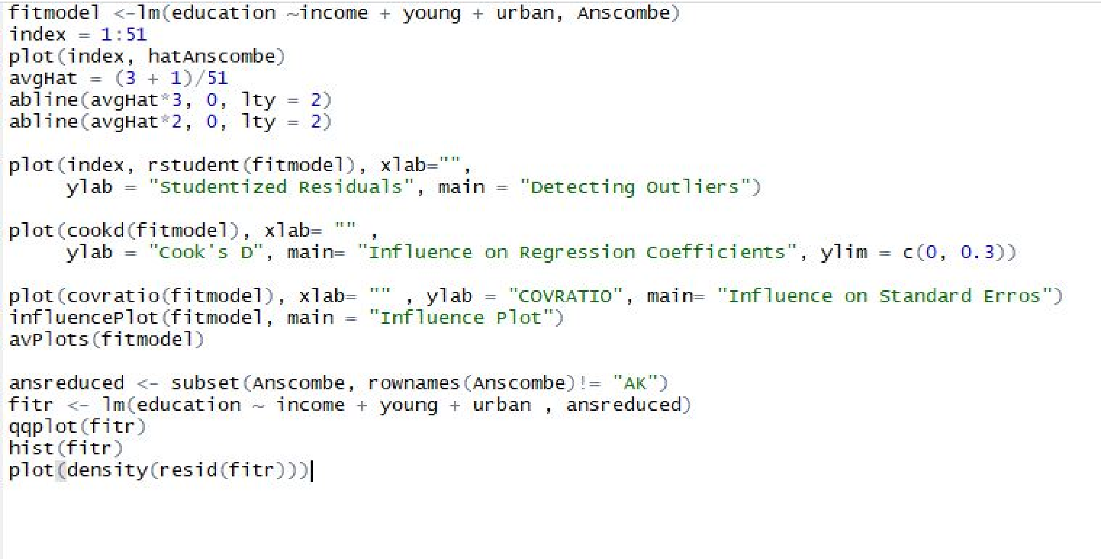


Figure 6: Checkin normality of residual errors

Appendix



```
fitmodel <-lm(education ~income + young + urban, Anscombe)
index = 1:51
plot(index, hatAnscombe)
avgHat = (3 + 1)/51
abline(avgHat*3, 0, lty = 2)
abline(avgHat*2, 0, lty = 2)

plot(index, rstudent(fitmodel), xlab="",
     ylab = "Studentized Residuals", main = "Detecting Outliers")

plot(cookd(fitmodel), xlab= "" ,
     ylab = "Cook's D", main= "Influence on Regression Coefficients", ylim = c(0, 0.3))

plot(covratio(fitmodel), xlab= "" , ylab = "COVRATIO", main= "Influence on Standard Erros")
influencePlot(fitmodel, main = "Influence Plot")
avPlots(fitmodel)

ansreduced <- subset(Anscombe, rownames(Anscombe)!= "AK")
fitr <- lm(education ~ income + young + urban , ansreduced)
qqplot(fitr)
hist(fitr)
plot(density(resid(fitr)))
```

Figure 7: Relevant Codes