S631 HW3

Shibi He

9/15/2019

Q1: Reading Assignment.

Q2: Find d_i and h_i in terms of x_i's and $\bar{x}$s such that, in simple linear regression,

$$\hat{\beta}_1 = \sum_{i=1}^{n} d_i y_i$$

and

$$\hat{\beta}_0 = \sum_{i=1}^{n} h_i y_i.$$

Answer:

$$\hat{\beta}_1 = \frac{\sum (x_i y_i - \bar{x}\bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - \bar{x}\sum y_i}{\sum (x_i - \bar{x})^2} = \sum \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} y_i = \sum_{i=1}^{n} d_i y_i,$$

where $d_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{1}{n}\sum y_i - \sum \frac{(x_i - \bar{x})\bar{x}}{\sum (x_i - \bar{x})^2} y_i = \sum \left[\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{\sum (x_i - \bar{x})^2} \right] y_i = \sum_{i=1}^{n} h_i y_i,$$

where $h_i = \frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{\sum (x_i - \bar{x})^2}$.

Q3: ALR 2.2 and 2.4.

Answer:

2.2

2.2.1 The key difference between points above this line and points below the line is that the rice price increased from 2003 to 2009 for the points above the $y = x$ line, while the rice price decreased for the points below the line.

2.2.2 Vilnius has the largest increase in the rice price. Mumbai has the largest decrease in rice price.

2.2.3 No, $\hat{\beta}_1 < 1$ does not necessarily mean the prices are lower in 2009 than 2003 because the price ($\hat{y}$) is also dependent on $\hat{\beta}_0$. If $\hat{\beta}_0$ is large enough, we would still have $\hat{y} > \hat{x}$, i.e. the rice price in 2009 is higher than the price in 2003.

2.2.4 One reason is that there are some outliers in the data, such as Vilnius and Mumbai. Another possible reason is that the relationship between the rice price in 2003 and 2009 is not likely to be linear, as the cities may be affected by the economic recession differently.

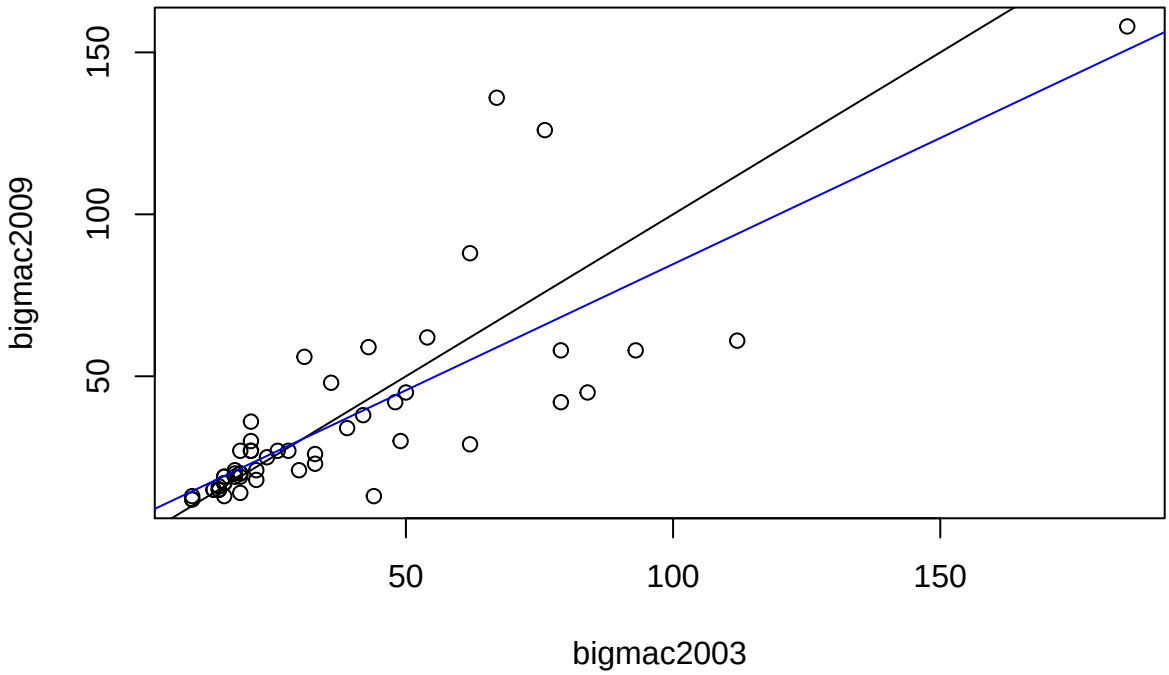
2.4

2.4.1 Draw the plot of y=bigmac2009 versus x=bigmac2003.

```
library(alr4)
```

```
data(UBSprices)
plot(bigmac2009 ~ bigmac2003, UBSprices)
abline(0, 1)

m1 <- lm(bigmac2009 ~ bigmac2003, data=UBSprices)
abline(m1, col= "blue")
```




```
#text(bigmac2009 ~ bigmac2003, labels=rownames(UBSprices),data=UBSprices, cex=0.9, font=2)
```

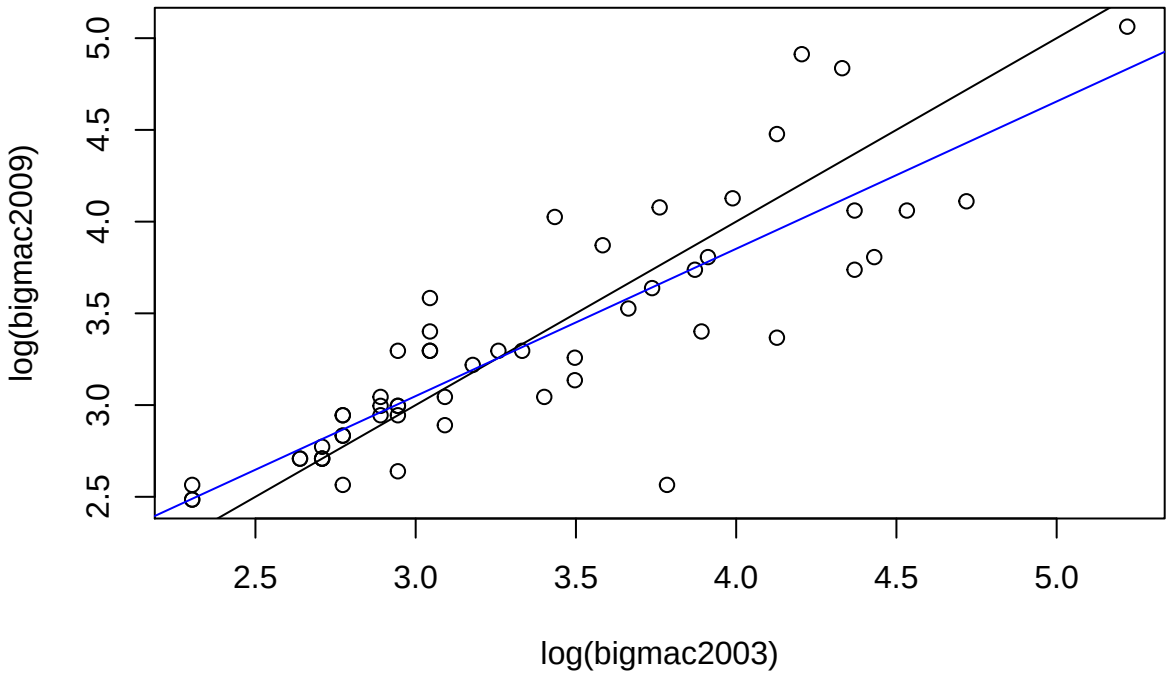
Nairobi is one the most unusual cities as its price for a Big Mac hamburger is much higher than all other cities, both in 2003 and 2009. Other unusual cases include Jakarta, Caracas, and Mumbai, as Jakarta and Caracas had the largest increase in the price for a Big Mac hamburger from 2003 to 2009, while Mumbai had the largest decrease in the price.

2.4.2 One reason that fitting simple linear regression is not appropriate is that there are some outliers in the data, and the linear regression model will be heavily affected by outliers. Another possible reason is that the true relationship between the prices in 2003 and 2009 is not linear.

2.4.3 Plot log(bigmac2009) versus log(bigmac2003).

```
plot(log(bigmac2009) ~ log(bigmac2003), UBSprices)
abline(0, 1)

m1 <- lm(log(bigmac2009) ~ log(bigmac2003), data=UBSprices)
abline(m1, col= "blue")
```



After taking logs, the relationship seems to be more linear. Therefore, it is more appropriate to use a linear regression to summarize the relationship.

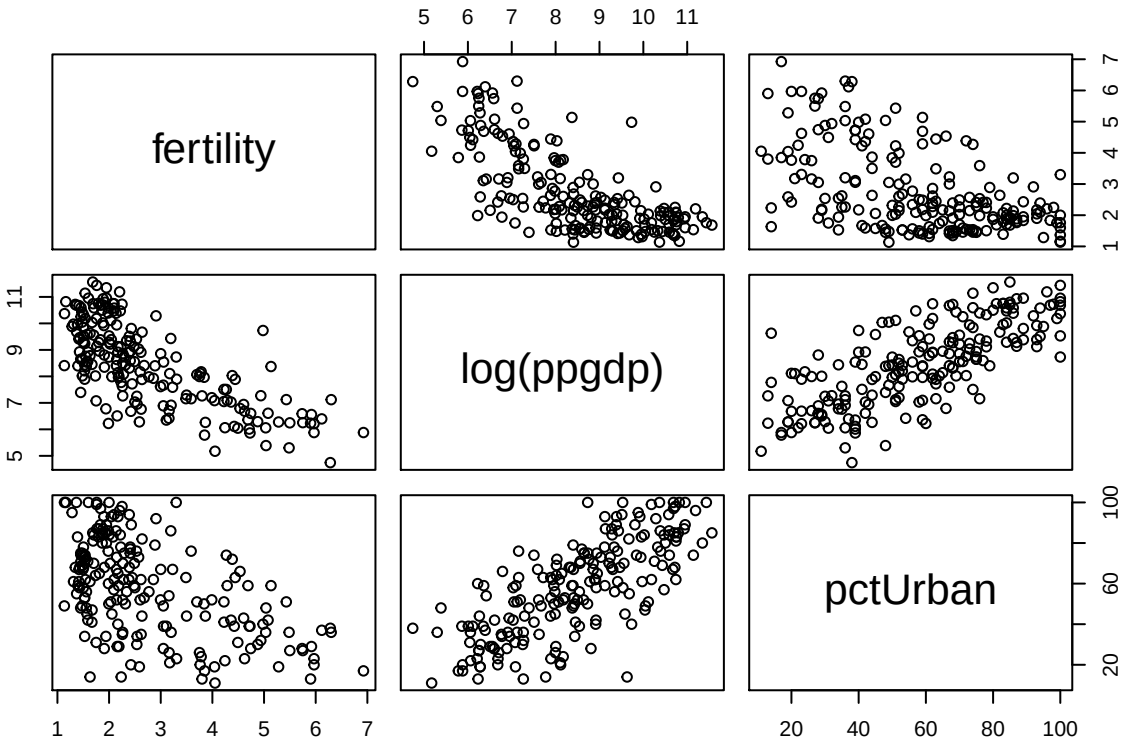
Q4: ALR 3.2.

Answer:

3.2.1 Examine the scatterplot matrix.

```
data(UN11)
```

```
pairs(~fertility + log(ppgdp) + pctUrban , data=UN11)
```



The scatterplot matrix suggests that *fertility* seems to be negatively correlated with *log(ppgdp)* and *pctUrban*, respectively. Moreover, *log(ppgdp)* and *pctUrban* are positively correlated with each other.

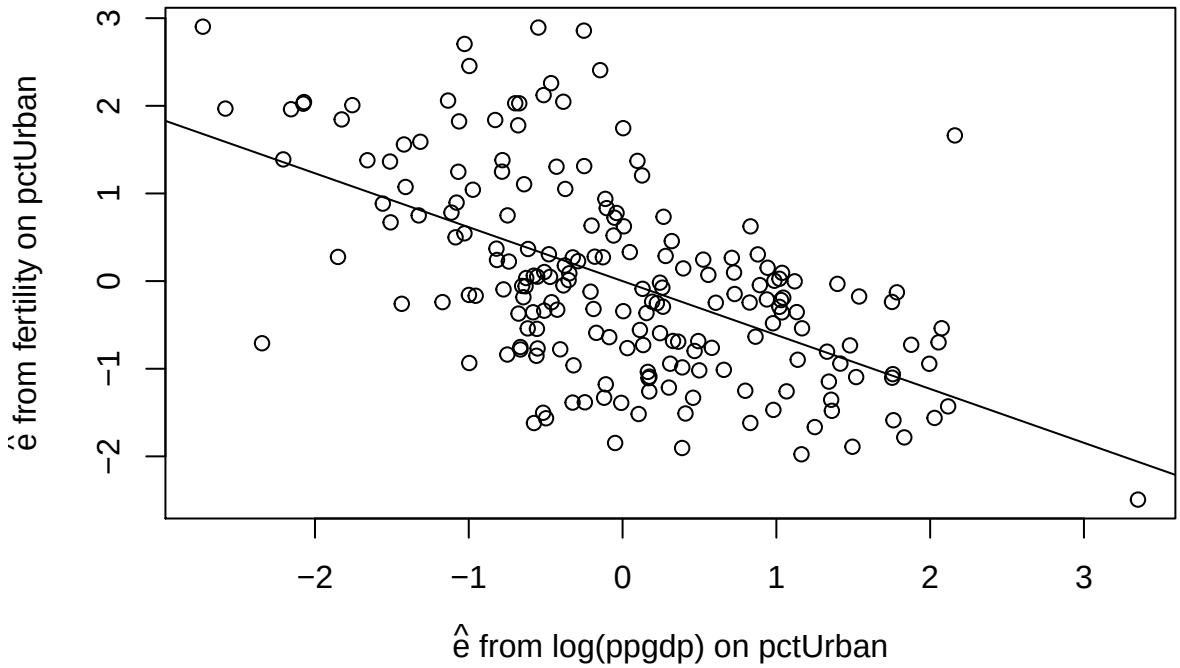
3.2.3 Obtain the added-variable plots for both predictors.

Added-variable plot for log(ppgdp):

```
r1 <- residuals(lm(fertility ~ pctUrban, UN11))
r2 <- residuals(lm(log(ppgdp) ~ pctUrban, UN11))
m2 <- lm(r1 ~ r2)
m2
```

```
##
## Call:
## lm(formula = r1 ~ r2)
##
## Coefficients:
## (Intercept)           r2
##  -1.986e-16   -6.151e-01
```

```
plot(r1 ~ r2,
xlab=expression(paste(hat(e), " from log(ppgdp) on pctUrban")),
ylab=expression(paste(hat(e), " from fertility on pctUrban")))
abline(m2)
```

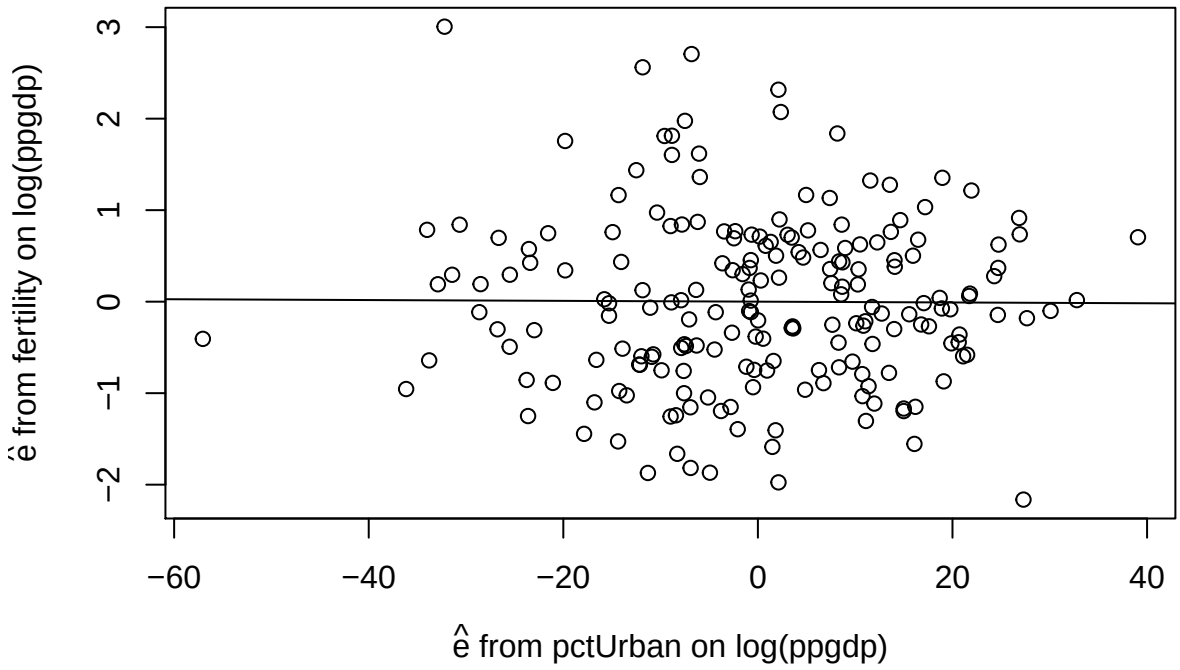


Added-variable plot for pctUrban:

```
r3 <- residuals(lm(fertility ~ log(ppgdp), UN11))
r4 <- residuals(lm(pctUrban ~ log(ppgdp), UN11))
m3 <- lm(r3~ r4)
m3
```

```
##
## Call:
## lm(formula = r3 ~ r4)
##
## Coefficients:
## (Intercept)           r4
##   6.313e-17   -4.393e-04
```

```
plot(r3 ~ r4,
xlab=expression(paste(hat(e), " from pctUrban on log(ppgdp)")),
ylab=expression(paste(hat(e), " from fertility on log(ppgdp)")))
abline(m3)
```



The added-variable plots show that *log(ppgdp)* is useful after adjusting for *pctUrban*, and *pctUrban* is also useful after adjusting for *log(ppgdp)*. Specifically, the added-variable plots show that there is a relatively strong negative relationship between *fertility* and *log(ppgdp)*, after adjusting for *pctUrban*. The negative relationship between *fertility* and *pctUrban* is relatively weak, after adjusting for *log(ppgdp)*.

```
m4 <- lm(fertility ~ log(ppgdp)+pctUrban, UN11)
m4
```

```
##
## Call:
## lm(formula = fertility ~ log(ppgdp) + pctUrban, data = UN11)
##
## Coefficients:
## (Intercept)   log(ppgdp)     pctUrban
##   7.9932699   -0.6151425   -0.0004393
```

The estimated mean function with both predictors is:

$$\hat{E}(fertility|log(ppgdp), pctUrban) = 7.99 - 0.62log(ppgdp) - 0.0004pctUrban.$$

The estimated coefficients are exactly the same as the coefficients obtained in the added-variable plots.

3.2.4 The estimated coefficient for *log(ppgdp)* is -0.62, which is exactly the same as the estimated slope in the added-variable plots.

3.2.5 Show that the residuals in the added-variable plot are identical to the residuals from te mean function with both predictors.

```
# residual from the regression with both predictors
R1 <- residuals(lm(fertility ~ log(ppgdp)+pctUrban, UN11))

# residual from added-varaible plot
r3 <- residuals(lm(fertility ~ log(ppgdp), UN11))
r4 <- residuals(lm(pctUrban ~ log(ppgdp), UN11))
R2 <- residuals(lm(r3~ r4))
```

```r
# Check the residuals are identical
all.equal(R1, R2)
```

```
## [1] TRUE
```

The results suggest that the residuals in the added-variable plot are identical to the residuals from the mean function with both predictors.

Q5: I did questions 1-4 and did not do question 5.