STAT645 - Homework 3

Salih Kilicli

9/22/2019

Problem 1: For the pollution data in "pollute data.csv" [Note: There is at least one missing value in these data. Just remove any records with at least one missing value prior to doing the following analyses.]:

(a) Based on the model

$$Mortality_{y_i} = \beta_0 + \beta_1 \times HCPot_i + \epsilon_i, \tag{1}$$

does there appear to be a significant linear relationship between HCPot and Mortality?

There is no apparent linear relationship between HCPot and Mortality, since the estimate of the coefficient β_1 appears to be insignificant (p-value is 0.161 which is pretty high). The summary of the model is given below:

```
setwd("~/Desktop/STAT645/Data")
pollute0 <- read.csv("pollute_data.csv", header=TRUE)
pollute = na.omit(pollute0)
attach(pollute)
model1 = lm(Mortality ~ HCPot)
summary(model1)
```

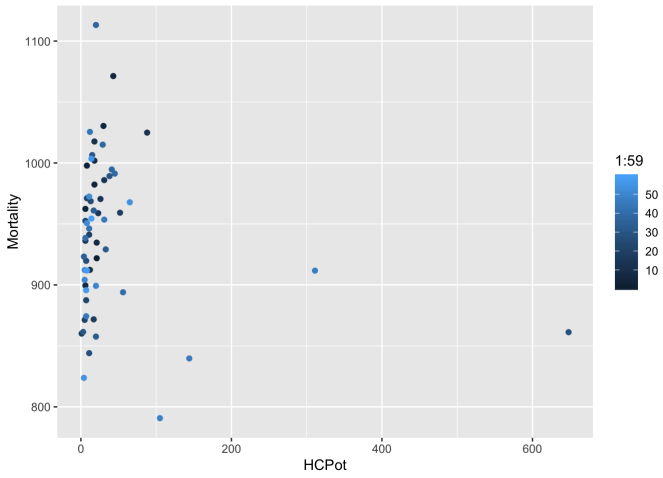
```
##
## Call:
## lm(formula = Mortality ~ HCPot)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -142.156  -42.699    4.474   41.206  169.686
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 945.96566    8.73448  108.30   <2e-16 ***
## HCPot        -0.12457    0.08771   -1.42    0.161
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61.88 on 57 degrees of freedom
## Multiple R-squared:  0.03418,    Adjusted R-squared:  0.01723
## F-statistic: 2.017 on 1 and 57 DF,  p-value: 0.161
```

```
#par(mfrow = c(2,2))
#plot(model1)
```

(b) Make a scatterplot of HCPot versus Mortality. Do you notice anything unusual that might have impacted your model in 1(a) [Hint: You should ;-)]. Dig into the data and provide an explanation for any unusual features you notice.

There appears to be 4 data points (rows 28, 46, 47, 48) (leverage points) that doesn't follow the trend of rest of the data. The scatter plot of HCPot vs Mortality is given below:

```
library(ggplot2)
#plot(Mortality, HCPot, pch=20, col = "blue", las=1)
#ggplot(pollute, aes(x=HCPot, y=Mortality)) + geom_point(color="blue")
ggplot(data = pollute) +
geom_point(mapping = aes(x = HCPot, y = Mortality, color = 1:59)) # I just like the ggplots better
```



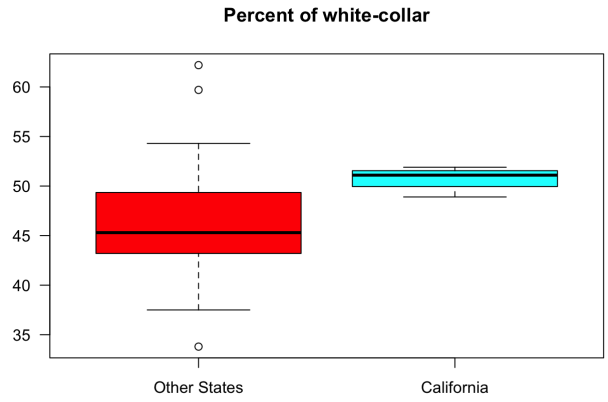

```
index = which(HCPot > 100) # also can be found by index = grep("CA", city)
cat("Data points given by the rows", index, "all have HCPot > 100 which clearly don't follow the pattern")
```

```
## Data points given by the rows 28 46 47 48 all have HCPot > 100 which clearly don't follow the pattern
```

(c) Compare the California records to all others, in terms of each of the following:

i. Percent of white-collar workers.

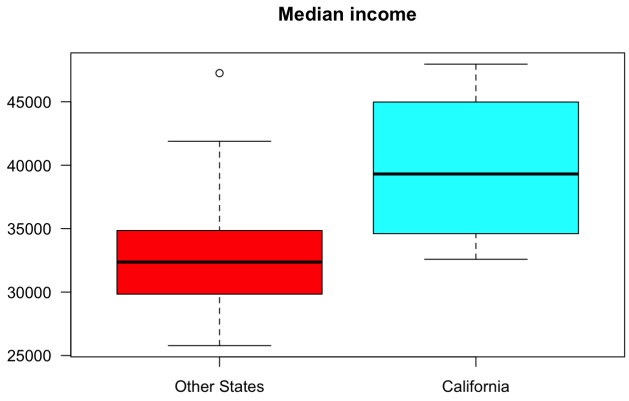
```
index = grep("CA", city)
#index = which(HCPot > 100)
boxplot(X.WC[-index],X.WC[index], xaxt = "n", col=rainbow(2), las=1, main="Percent of white-collar")
axis(1, at = 1:2, labels = c("Other States", "California"))
```



Percent of white-coller has much more variabilty in other states, whereas California has lesser variability with higher median and mean compared to other states. Also, other states have 3 outliers.

ii. Median income.

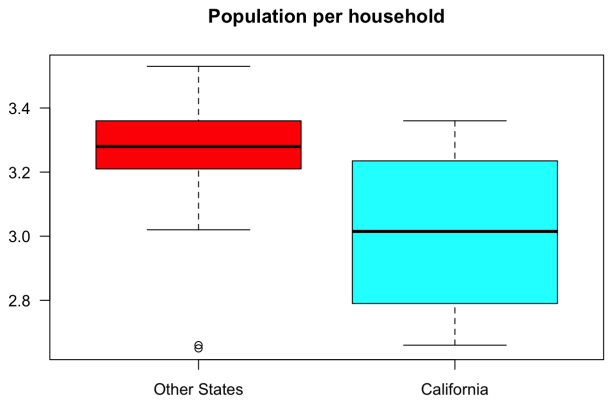
```
boxplot(income[-index],income[index], xaxt = "n", col=rainbow(2), las=1, main="Median income")
axis(1, at = 1:2, labels = c("Other States", "California"))
```



Median income in California has a bigger variability than other states both of which almost normally distributed. The median and mean of median income in California is nearly %30 more than those of in other states. There is one outlier with high median income in the other states.

iii. Population per household.

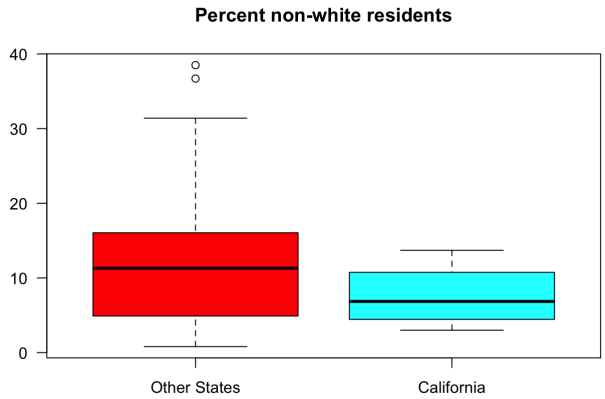
```
boxplot(pop.house[-index],pop.house[index], xaxt = "n", col=rainbow(2), las=1, main="Population per household")
axis(1, at = 1:2, labels = c("Other States", "California"))
```



Population per household in California has a lesser median whereas its variability is higher. Also, there appears to be 2 (extreme small values) outliers in the other states when it comes to population per household.

iv. Percent non-white residents.

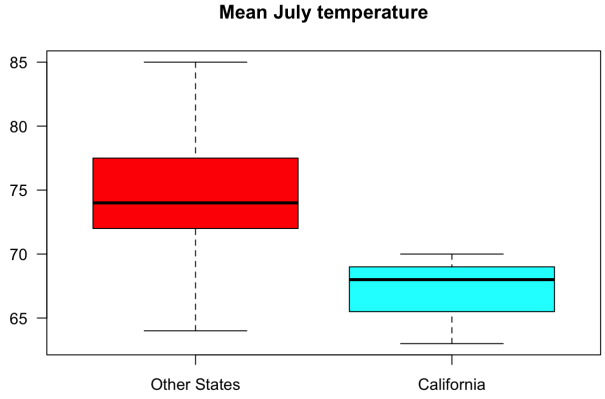
```
boxplot(X.NonWhite[-index],X.NonWhite[index], xaxt = "n", col=rainbow(2), las=1, main="Percent non-white resident
s")
axis(1, at = 1:2, labels = c("Other States", "California"))
```



Percent of non-white residents in California has lesser variability and lesser median compared to other states. Other states have 2 really high outliers.

v. Mean July temperature.

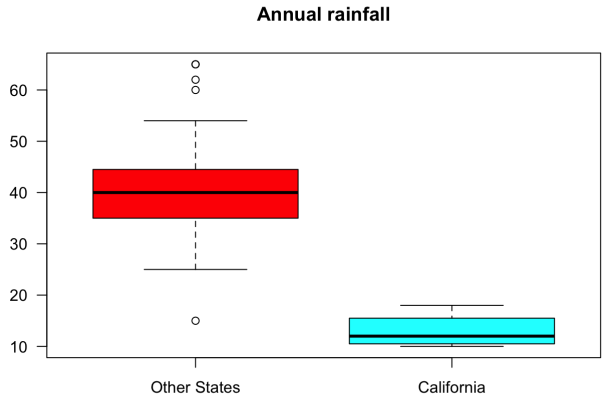
```
boxplot(JulyTemp[-index],JulyTemp[index], xaxt = "n", col=rainbow(2), las=1, main="Mean July temperature")
axis(1, at = 1:2, labels = c("Other States", "California"))
```



Mean July temperature in California appears to be around 5 degrees less than that of in other states. There appears no outlier in either dataset and variability in other states is slightly more.

vi. Annual rainfall.

```
boxplot(Rain[-index],Rain[index], xaxt = "n", col=rainbow(2), las=1, main="Annual rainfall")
axis(1, at = 1:2, labels = c("Other States", "California"))
```



Annual rainfall in California seems to be much lesser than those of in other states, which explains water resourse problems in the state. Also, variability in other states are more than the one in California, and there appears to be 4 outliers in the other states.

(d) Write down the model for Mortality as a function of log(HCPot), as well as all variables from 1(c). Note: By "write down the model," I mean for you to write down an equation analogous to equation (1) above.

$$Mortality_i = \beta_0 + \beta_1 log(HCPot_i) + \beta_2 X.WC_i + \beta_3 income_i + \beta_4 pop.house_i + \beta_5 X.NonWhite_i + \beta_6 JulyTemp_i + \beta_7 Rain_i \quad (2)$$

(e) Interpret all coefficients in the model from 1(d).

According to the summary of model, the most effective coefficients are Intercept, pop.house and log(HCPot) whereas pop variable has the least effect on Mortality whose coefficient estimate is almost zero. On the other hand, X.WC and JulyTemp have negative effects on the response variable. Finally, Intercept, X.WC, X.NonWhite and Rain appear to be statistically significant in terms of explaining a linear relationship with Mortality.

```
model2 = lm(Mortality ~  log(HCPot) + X.WC + income + pop.house + X.NonWhite + JulyTemp  +  Rain)
summary(model2)
```

```
##
## Call:
## lm(formula = Mortality ~ log(HCPot) + X.WC + income + pop.house +
##     X.NonWhite + JulyTemp + Rain)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -87.32 -21.85  -3.04  25.78 113.38
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 858.639846 222.144989   3.865 0.000315 ***
## log(HCPot)   16.996502   7.574546   2.244 0.029208 *
## X.WC         -2.397372   1.215760  -1.972 0.054055 .
## income       -0.001247   0.001419  -0.879 0.383750
## pop.house    40.649881  35.965769   1.130 0.263663
## X.NonWhite    3.174328   1.033974   3.070 0.003426 **
## JulyTemp     -0.862296   1.973816  -0.437 0.664052
## Rain          2.131673   0.619165   3.443 0.001158 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.54 on 51 degrees of freedom
## Multiple R-squared:  0.6106, Adjusted R-squared:  0.5571
## F-statistic: 11.42 on 7 and 51 DF,  p-value: 1.297e-08
```

(f) Using both a likelihood ratio test and an F test, test the null hypothesis that all coefficients other than that for log(HCPot) equal 0, in the model from 1(d). Test at α = 0.05.

The null hypothesis is given by $H_0 : A\beta = c$ where

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}_{6\times 8}, \quad \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_7 \end{bmatrix}_{8\times 1}, \quad c = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}_{6\times 1}$$

```
model0 = lm(Mortality ~ log(HCPot)) # Null Hypothesis, Alternative Hypothesis is given by model2 in (d)
# LR Test
df = length(coef(model2))-length(coef(model0))
lambda = -2*(as.numeric(logLik(model0))-as.numeric(logLik(model2)))
p = pchisq(lambda, df, lower.tail=FALSE) # or calculate by 1-pchisq(lambda,df) In this example p=q
q = 1 - pchisq(lambda, df)
cat('p-value obtained from the likelihood ratio test is', p)
```

```
## p-value obtained from the likelihood ratio test is 5.312965e-10
```

```
# F Test
anova(model0, model2)
```

```
## Analysis of Variance Table
##
## Model 1: Mortality ~ log(HCPot)
## Model 2: Mortality ~ log(HCPot) + X.WC + income + pop.house + X.NonWhite +
##     JulyTemp + Rain
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     57 222440
## 2     51  88008  6    134432 12.984 7.533e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at either of the tests, we see that the p-values for the tests are much smaller than $\alpha = 0.05$, therefore, we reject the null hypothesis.

(g) Using both a likelihood ratio test and an F test, test the null hypothesis that the coefficients for percent white collar and percent non-white sum to zero. Test at α = 0.05.

In this case the Null hypothesis is given by: $H_0 : \beta_2 + \beta_5 = 0$ or in other words $H_0 : \beta_5 = -\beta_2$, and the model in this case would be

$$Mortality_{y_i} = \beta_0 + \beta_1 log(HCPot_i) + \beta_2(X.WC_i - X.NonWhite_i) + \beta_3 pop_i + \beta_4 pop.house_i + \beta_6 JulyTemp_i + \beta_7 Rain_i \quad (3)$$

```
model3 = lm(Mortality ~  log(HCPot) + I(X.WC-X.NonWhite) + income + pop.house + JulyTemp  +  Rain)
# LR Test
df = length(coef(model3))-length(coef(model0))
lambda = -2*(as.numeric(logLik(model0))-as.numeric(logLik(model3)))
p = pchisq(lambda, df, lower.tail=FALSE) # or calculate by 1-pchisq(lambda,df) In this example p=q
q = 1 - pchisq(lambda, df)
cat('p-value obtained from the likelihood ratio test is', p)
```

```
## p-value obtained from the likelihood ratio test is 1.695343e-10
```

```
# F Test
anova(model0, model3)
```

```
## Analysis of Variance Table
##
## Model 1: Mortality ~ log(HCPot)
## Model 2: Mortality ~ log(HCPot) + I(X.WC - X.NonWhite) + income + pop.house +
##     JulyTemp + Rain
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     57 222440
## 2     52  88394  5    134045 15.771 1.971e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at either of the tests, we see that the p-values for the tests are much smaller than $\alpha = 0.05$, therefore, again we reject the null hypothesis.

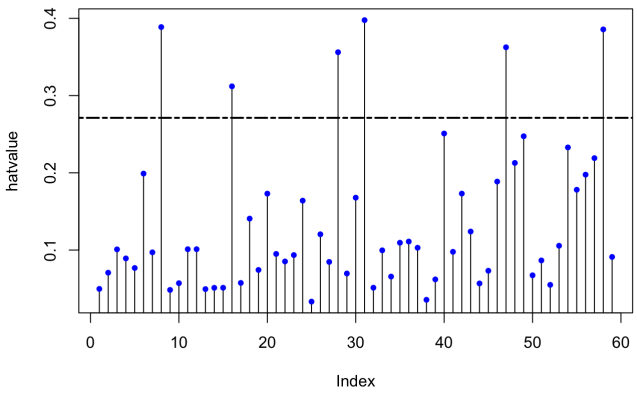
(h) Report a 95% confidence interval for the sum of the coefficients for percent white collar and percent non-white.

```
sigmahatsq=(summary(model2)$sigma)^2            # estimate of error variance
v=c(0,0,1,0,0,1,0,0)                            # v vector in the equation v'betahat=0
M=summary(model2)$cov.unscaled                  # gives the inv(X'X) - covariance matrix for model2
varhat=sigmahatsq*(t(v)%*%M%*%v)                # varhat = sigmahat^2 v' inv(X'X) v
se=sqrt(as.numeric(varhat))                     # se = sqrt(varhat(v' betahat))
betahat=model2$coefficients                     # betahat = estimates for betas
ci=as.numeric(t(v)%*%betahat)+c(-1,1)*qt(1-0.05/2,51)
names(ci)=c("Lower Bound","Upper Bound")        # ci = v' betahat +- t_{\alpha/2,n-p}xse(v' betahat)
ci
```

```
## Lower Bound Upper Bound
##   -1.230628    2.784540
```

(i) For the model that includes log(HCPot), as well as all variables from 1(c), identify any potential leverage or inluential points from this data.

```
# you can also use plot(model2) to show diagnostics plots for each assumption
# plot of hat-values to identify leverage points
hatvalue=hatvalues(model2)
plot(hatvalue, type="h")
points(hatvalue, pch=20, col="blue", bg=2)
abline(h=(2*8)/59, lty=6, lwd=2)
```



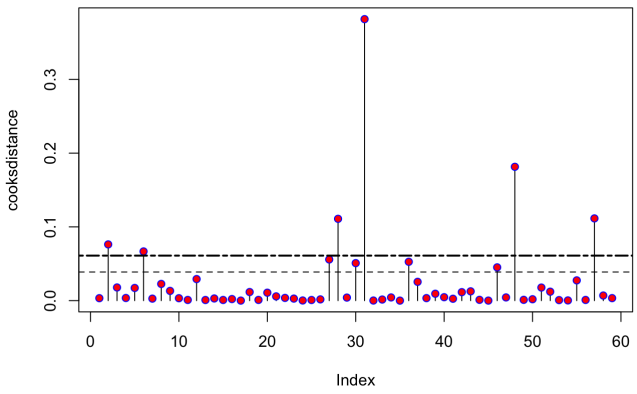

```
leverage=which(hatvalue>(2*8)/59)
cat("Potential leverage points are the data points given by", leverage)
```

```
## Potential leverage points are the data points given by 8 16 28 31 47 58
```

```
# plot with cooks distance to identify influential observation
cooksdistance=cooks.distance(model2)
summary(cooksdistance)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0000001 0.0012667 0.0041097 0.0245592 0.0177771 0.3817230
```

```
IQR= 0.0165529-0.0017265                        # check the D_i values above Q3+1.5IQR and Q3+3IQR
l1= 0.0165529 + 1.5*IQR
l2= 0.0165529 + 3*IQR
plot(cooksdistance, type="h")
points(cooksdistance, pch=21, col="blue", bg=2)
abline(h=l1, lty=2)                                 # gives a threshold for influential points
abline(h=l2, lty=6, lwd=2)                          # gives a threshold for HIGHLY influential poi
nts
```




```
influential1=which(cooksdistance>l1)
influential2=which(cooksdistance>l2)

cat("Influential points are the data points given by", influential1)
```

```
## Influential points are the data points given by 2 6 27 28 30 31 36 46 48 57
```

```
cat("Highly influential points are the data points given by", influential2)
```

```
## Highly influential points are the data points given by 2 6 28 31 48 57
```