# Biostats - MLR: Outliers, Leverage and Influential Points

Rob Leonard (robleonard@tamu.edu (mailto:robleonard@tamu.edu))

# 1) Pollution and Mortality

**1a) Read in Data, does there appear to be a significant linear relationship between HCPot and Mortality?**

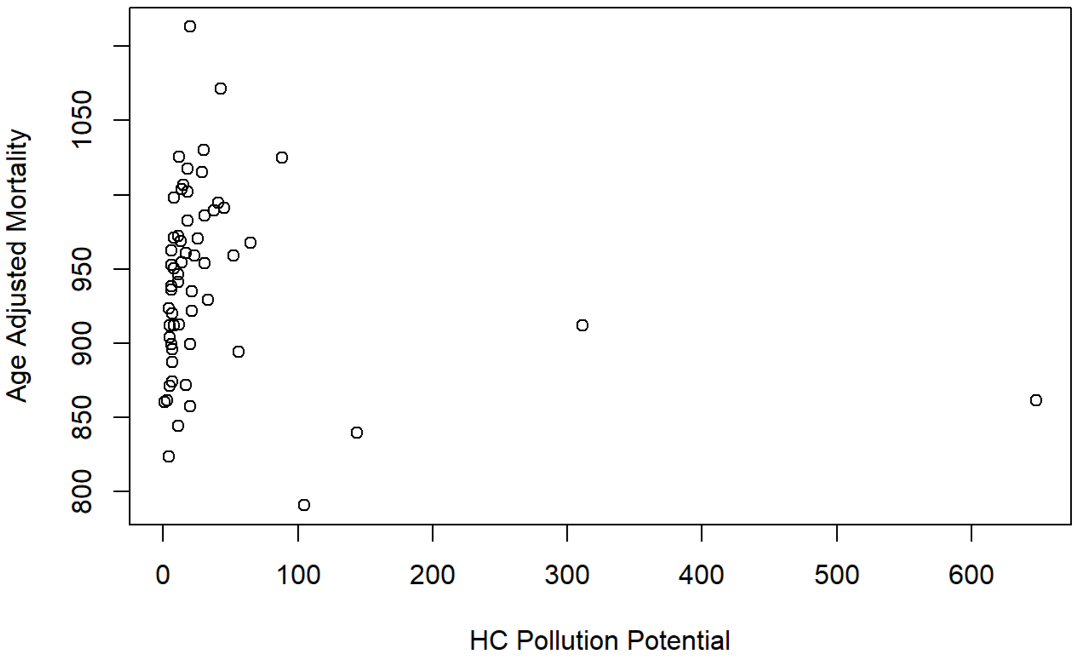
```
poldata = read.delim("pollute_data.csv", header=TRUE, sep=",")
poldata = poldata[complete.cases(poldata),]   # remove rows with any n/a values
model1 = lm(Mortality~HCPot, data=poldata)
summary(model1)
```

```
## 
## Call:
## lm(formula = Mortality ~ HCPot, data = poldata)
## 
## 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
```

No, there does not seem to be a statistically significant relationship between HCPot and Mortality. The p-value is high at 0.161, meaning that unless we adopted a very large alpha, we don’t have significant evidence to reject the null hypothesis that the coefficient on HCPot is not zero.

**1b) Create a scatterplot**

```
attach(poldata)
plot(HCPot, Mortality, xlab="HC Pollution Potential", ylab="Age Adjusted Mortality")
```



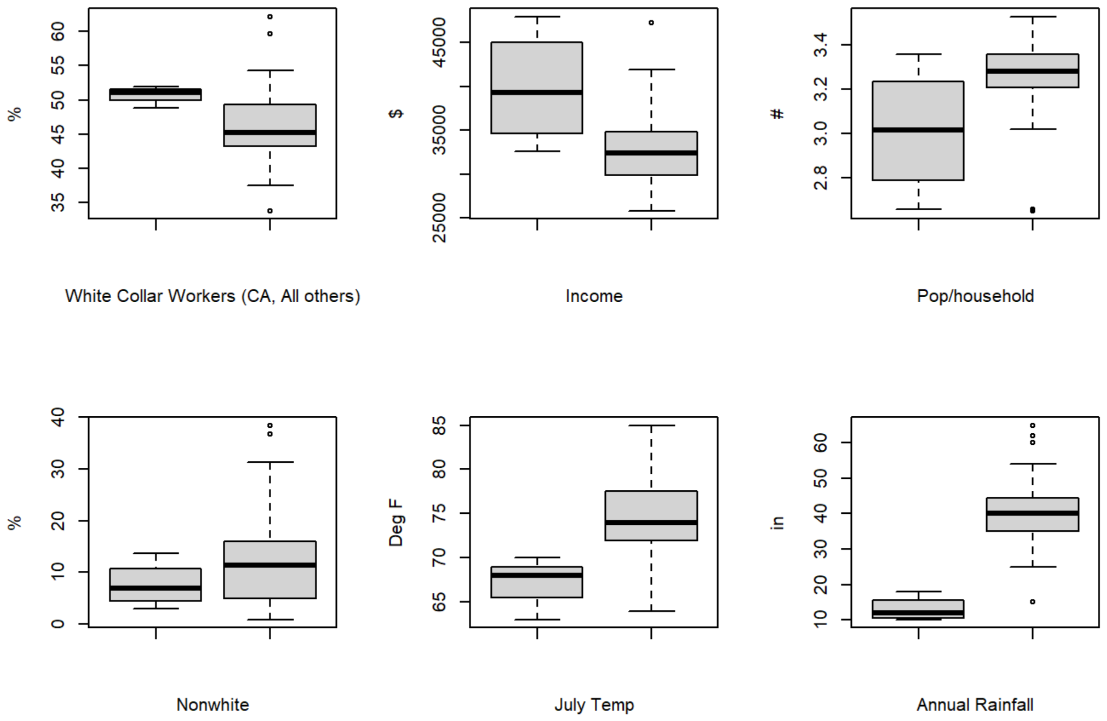
The scatterplot shows at least 3 cities have abnormally high levels of HC pollution potential compared to the rest of the sample. These 3 cities are LA, SF and SD in California. Their HC values are so far off the mean that they will have a very large influence on the calculation of the estimated slope parameter.

**1c) Compare the CA cities to all others**

```
cavec = c(grep(', CA',city))  # identify index for CA cities
frame2ca = data.frame(poldata[cavec,c("X.WC","income","pop.house","X.NonWhite","JulyTemp","Rain")])
frame2c = data.frame(poldata[-cavec,c("X.WC","income","pop.house","X.NonWhite","JulyTemp","Rain")])
# create table of means
frame2cfinal=data.frame(cbind(round(apply(poldata[cavec,c("X.WC","income","pop.house","X.NonWhite","JulyTemp","Rai
n")],2,mean),digits=2),
                   round(apply(poldata[-cavec,c("X.WC","income","pop.house","X.NonWhite","JulyTemp","Rain")],2,mea
n),digits=2)))
colnames(frame2cfinal) = c("CA (mean of)","All Others (mean of)")
rownames(frame2cfinal) = c("Percent of white-collar workers","Median income","Population per household","Percent n
on-white residents","Mean July temperature","Annual rainfall")
frame2cfinal
```

```
##                                 CA (mean of) All Others (mean of)
## Percent of white-collar workers        50.75                46.07
## Median income                       39792.50             32770.60
## Population per household                3.01                 3.26
## Percent non-white residents             7.60                12.19
## Mean July temperature                  67.25                74.93
## Annual rainfall                        13.00                40.36
```

```
# also wants boxplots
par(mfrow=c(2,3))
boxplot(frame2ca[,1],frame2c[,1],ylab="%", xlab="White Collar Workers (CA, All others)")
boxplot(frame2ca[,2],frame2c[,2],ylab="$", xlab="Income")
boxplot(frame2ca[,3],frame2c[,3],ylab="#", xlab="Pop/household")
boxplot(frame2ca[,4],frame2c[,4],ylab="%", xlab="Nonwhite")
boxplot(frame2ca[,5],frame2c[,5],ylab="Deg F", xlab="July Temp")
boxplot(frame2ca[,6],frame2c[,6],ylab="in", xlab="Annual Rainfall")
```



From the boxplots, it appears Californian cities have higher median incomes, lower July temperatures, lower people per household (slightly) and lower annual rainfall than the other cities.

### 1d) Write the linear regression model

$$Mor_i = \beta_0 + \beta_1 * log(HCPot_i) + \beta_2 * X.WC_i + \beta_3 * inc_i + \beta_4 * pop.house_i + \beta_5 * X.NonW_i + \beta_6 * JulyT_i + \beta 7 * Rain_i + \epsilon_i$$

### 1e) Interpret coefficients from model 1d

```
model1d = lm(Mortality~log(HCPot)+X.WC+income+pop.house+X.NonWhite+JulyTemp+Rain, data=poldata, x=TRUE) #add x=TRU
E to capture design matrix
summary(model1d)
```

```
##
## Call:
## lm(formula = Mortality ~ log(HCPot) + X.WC + income + pop.house +
##     X.NonWhite + JulyTemp + Rain, data = poldata, x = TRUE)
##
## 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
```

#### 1e) Interpret coefficients

For every 1% increase in the HC pollution potential, the model expects the age adjusted mortality (“aam”) to increase by 16.99, on average and controlling for the other variables.
For every 1 percentage point increase in white collar workers, the model expects the aam to decrease by 2.39, on average and controlling for the other variables. For every $1 increase in median income, the model expects the aam to decrease by 0.0012 on average and controlling for the other variables. For every additional person in the average household, the model expects the aam to increase by 40.65 on average and controlling for the other variables. For every 1 percentage point increase in nonwhite residents, the model expects the aam to increase by 3.17 on average and controlling for the other variables. For every 1 degree Fahrenheit increase in temperature, the model expects the aam to decrease by 0.86 on average and controlling for the other variables. For every additional 1 inch in annual rainfall, the model expects the aam to increase by 2.13 on average and controlling for the other variables. The log(pollution potential), annual rainfall and percent non white residents are all highly significant at the .01 level.

#### 1f) Likelihood ratio test all betas=0 except for HCPot

```
model1fres = lm(Mortality~log(HCPot), data = poldata) # fit the restricted model
diff01 = as.numeric(logLik(model1fres))-as.numeric(logLik(model1d))  #loglik rest-unrest
teststat1 = -2*diff01
df1 = df.residual(model1fres)-df.residual(model1d)
pval1 = 1-pchisq(teststat1, df1)
print(pval1)
```

```
## [1] 5.312966e-10
```

The p-value is very small, much smaller than alpha=.05, so we have sufficient evidence to reject the null hypothesis that the coefficients for all the other explanatory variables are 0.

#### 1g) Now test the null hypothesis that the coefficients for percent white collar and percent non-white sum to zero.

```
model1g = lm(Mortality~log(HCPot)+I(X.WC-X.NonWhite)+income+pop.house+JulyTemp+Rain, data=poldata, x=TRUE) #add x=
TRUE to
summary(model1g)
```

```
##
## Call:
## lm(formula = Mortality ~ log(HCPot) + I(X.WC - X.NonWhite) +
##     income + pop.house + JulyTemp + Rain, data = poldata, x = TRUE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.507 -22.512  -4.641  23.159 115.594
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          830.452900 212.409284   3.910 0.000269 ***
## log(HCPot)            18.711540   6.601621   2.834 0.006523 **
## I(X.WC - X.NonWhite)  -2.844799   0.758663  -3.750 0.000446 ***
## income                -0.001087   0.001368  -0.795 0.430430
## pop.house             41.562634  35.645035   1.166 0.248929
## JulyTemp              -0.359349   1.650850  -0.218 0.828534
## Rain                   2.192965   0.600934   3.649 0.000610 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.23 on 52 degrees of freedom
## Multiple R-squared:  0.6089, Adjusted R-squared:  0.5637
## F-statistic: 13.49 on 6 and 52 DF,  p-value: 3.696e-09
```

```
diff02 = as.numeric(logLik(model1g))-as.numeric(logLik(model1d))  #loglik rest-unrest
teststat2 = -2*diff02
df2 = df.residual(model1g)-df.residual(model1d)
pval2 = 1-pchisq(teststat2, df2)
print(pval2)
```

```
## [1] 0.6111189
```

```
#linearHypothesis(model1d,"X.WC  =-1*X.NonWhite", test=c("Chisq"))  # Verify with Linear Hypothesis Test (asymptot
ic, similar result)
```

The p-value is very high at 0.611 and more than alpha. We do not have sufficient evidence to reject the null hypothesis that the coefficients sum to zero.

**1h) Report a 95% confidence interval for the sum of the coefficients for percent white collar and percent non-white. Interpret the result, and also test the hypothesis of (g) using this CI.**

```
xtx.inv <- solve(t(model1d$x)%*%model1d$x)   # Calculate (X'X)-1
semodel <- sigma(model1d)   #Get the estimate of sigma
A <- cbind(0,0,1,0,0,1,0,0) # calculate the se of the sum, set up matrix A and A'
tA <- t(A)
AXTX.invA <- A%*%xtx.inv%*%tA
var1h <- AXTX.invA*(semodel)^2  # calculate the variance of B6-B7
se1h <- sqrt(var1h)  # calc std error for B6+B7
coefB3 <- summary(model1d)$coefficients[3,1] #Pull B3, B6 coefficients from regression output
coefB6 <- summary(model1d)$coefficients[6,1]
(coefB3+coefB6)-qt(0.975,51)*se1h # calc coeff,   df=n-#parameters = 59-8=51
```

```
##           [,1]
## [1,] -2.518867
```

```
(coefB3+coefB6)+qt(0.975,51)*se1h  # 0 is in the interval, so do not reject H0
```

```
##          [,1]
## [1,] 4.072779
```

The CI is (-2.52,4.07). As 0 is in the CI, this indicates that we cannot reject the null hypothesis in 1g that the sum of the coefficients does not equal zero.

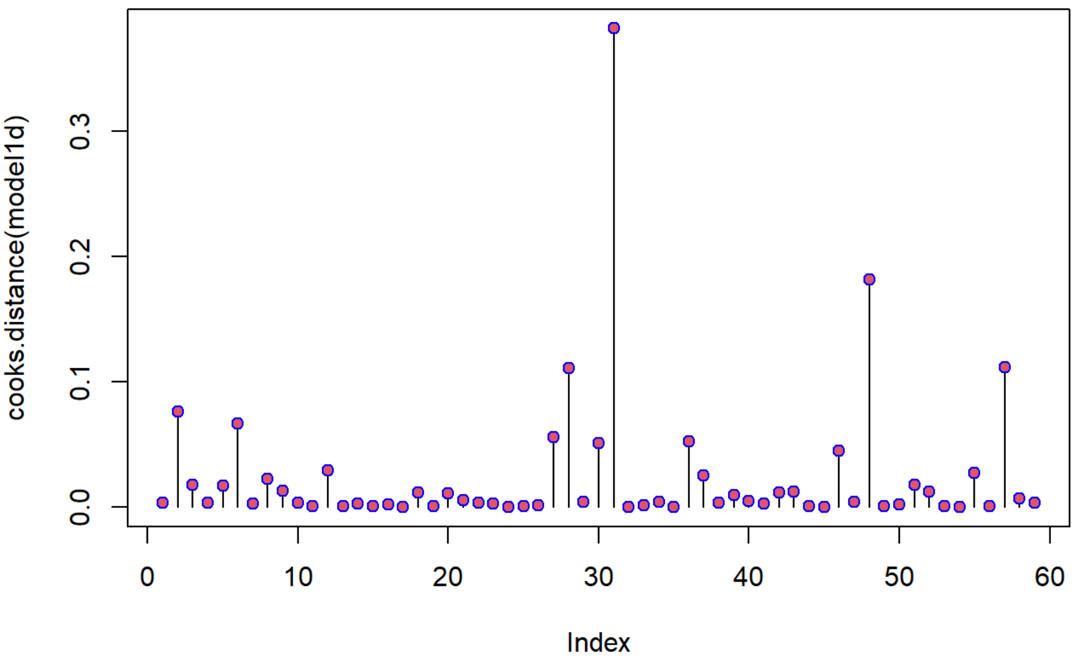
**1i) For model1d, identify any potential leverage or influential points from this data.**

```
threshold = 2*8/59  # calculate threshold for leverage points
poldata[hatvalues(model1d)>threshold,1]  # identify which cities are leverage points
```

```
## [1] "Bridgeport-Milford, CT"      "Dallas, TX"
## [3] "Los Angeles, Long Beach, CA" "Miami-Hialeah, FL"
## [5] "San Francisco, CA"           "York, PA"
```

The above cities are likely leverage points.

```
plot(cooks.distance(model1d), type="h")
points(cooks.distance(model1d), pch=21, col="blue", bg=2)
```



```
poldata[cooks.distance(model1d)>.15,1]  # identify which cities are influential points, those >.15 have much greater Cooks D.
```

```
## [1] "Miami-Hialeah, FL" "San Jose, CA"
```

Miami-Hialeah and San Jose are influential points.