Stat 218 Linear Regression Exercise

Rommel Bartolome

January 30, 2019

In this exercise, we will try to find out if pollution kills people. The “pollution.csv” dataset used was from a study of 5 Standard Metropolitan Statistical Areas (SMSAs) in the United States from 1959-1961. In this problem, let MORT be the dependent variable and PRECIP, EDUC, and NONWHITE are the non-pollutant independent variables while logNOX and logSO2 are the pollutant independent variables.

1. Is there sufficient evidence that the full model in the study can explain mortality in cities?

We will assume that the “full model” here is the one with the five independent variables previously mentioned. We load the data using read.csv and create a linear model for such. Checking the summary of the model, we see the following:

```r
pol.data <- read.csv('pollution.csv')
mort.fullmodel <- lm(MORT ~ PRECIP + EDUC + NONWHITE +
                   logNOX + logSOX, data = pol.data)
summary(mort.fullmodel)
```

```
##
## Call:
## lm(formula = MORT ~ PRECIP + EDUC + NONWHITE + logNOX + logSOX,
##     data = pol.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -102.222  -19.547    0.239   20.084   95.386
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 940.6584    94.0551  10.001 6.81e-14 ***
## PRECIP        1.9467     0.7007   2.778   0.0075 **
## EDUC        -14.6645     6.9379  -2.114   0.0392 *
## NONWHITE      3.0289     0.6685   4.531 3.29e-05 ***
## logNOX        6.7164     7.3990   0.908   0.3680
## logSOX       11.3578     5.2955   2.145   0.0365 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.3 on 54 degrees of freedom
## Multiple R-squared:  0.6883, Adjusted R-squared:  0.6594
## F-statistic: 23.85 on 5 and 54 DF,  p-value: 1.418e-12
```

We are interested in knowing if the full model can explain the mortality so we will check the last line where the p-value is seen. Since the p-value is less than our level of significance 0.05, we can say that there is sufficient evidence that the full model can explain the mortality in the cities.

2. How much of the variability in mortality among cities is explained by the 5 variables?

We will check the value of *Adjusted* R^2 for this one. I think that the *Adjusted* R^2 will be the better metric since it penalizes the number of terms used in our model. This value has already been seen above. Calling it:

```
summary(mort.fullmodel)$adj.r.squared
```

```
## [1] 0.6594116
```

We know that around 66% of variability can be explained by the 5 variables.

3. Write the estimated regression line of the problem.

To get the regression line, we will need the coefficients in our model. Fortunately, this has already been seen above. Calling it again:

```
coef(mort.fullmodel)
```

```
## (Intercept)      PRECIP        EDUC    NONWHITE      logNOX      logSOX
##  940.658446    1.946748  -14.664523    3.028928    6.716432   11.357829
```

Our line will be of the form `MORT = 1.946748(PRECIP) - 14.664523(EDUC) + 3.028928(NONWHITE) + 6.716432(logNOX) + 11.357829(logSOX) + 940.658446`.

4. Analyse the coefficients of the models, telling also how the independent variables might affect mortality. Which variables are significant in explaining mortality?

For this question, we check again p-values of the each of the coefficients. Again, this has already been seen above. Calling it again:

```
summary(mort.fullmodel)$coefficients[,4]
```

```
##  (Intercept)       PRECIP         EDUC     NONWHITE       logNOX 
## 6.811116e-14 7.500865e-03 3.917799e-02 3.290529e-05 3.680457e-01 
##       logSOX 
## 3.648581e-02
```

We see that all values except `logNOX` are deemed significant (< 0.05) in explaining mortality.

We also check the coefficients of the model. We can see that the highest p-value among all coefficients is `PRECIP`. Although should still be closely examined, it is quite logical to see rainfall in the region to be an indicator of mortality with rain typically being a by-product of natural disasters such as cyclones and typhoons. In addition, except for `EDUC` everything is positive. This is logical for the pollutant variables as we would expect that higher concentrates of pollutants will result to higher mortality. As for the `NONWHITE` variable, this is also logical. It should be recalled that this data was from 1959-1961 where non-whites have limited access to health care due to segregation and racism. As for the `EDUC` variable, it is also logical to see that as people become more educated, the less likely they will die as they would have more access to health care. This is can be related to their capacity to pay for these services.

5. Test whether the pollutant variables jointly explain the variability of mortality given that nonpollutant variables have been considered.

We now create a model using the pollutant variables `logNOX` and `logSOX`, then ANOVA test it with the full model:

```
mort.polmodel <- lm(MORT ~ logNOX + logSOX, data = pol.data)
anova(mort.polmodel, mort.fullmodel)
```

```
## Analysis of Variance Table
## 
## Model 1: MORT ~ logNOX + logSOX
## Model 2: MORT ~ PRECIP + EDUC + NONWHITE + logNOX + logSOX
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
## 1     57 191172
## 2     54  71159  3    120013 30.358 1.228e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again, we can see than the model's p-value `Pr(>F)` is less than our significance level of 0.05 and as such, we can say that the pollutant variables `logNOX` and `logSOX` can jointly explain the variability.

6. Suppose that a city has an annual mean precipitation of 20 inches, 33% of the population being non-white, an average of 12 years in education, logNOX at 0, and logSOX at 0. With 95% confidence, what is the expected level of mortality? Generate the 99% prediction interval for mortality in the city as well.

We create a data frame containin the following predictions first:

```
newdata <- data.frame(PRECIP=20, EDUC=12, NONWHITE=33,
                      logNOX = 0, logSOX = 0)
```

Now, we fit this dataset to our model, with a 95% confidence interval:

```
predict(mort.fullmodel, newdata, interval='confidence')
```

```
##        fit      lwr      upr
## 1 903.5738 845.4334 961.7142
```

Here we can see that the expected level of mortality to be around 903.6 with its lower (`lwr`) and upper(`upr`) bounds, respectively.

This will also be the case for a 99% prediction interval, with the upper and lower bounds to be:

```
predict(mort.fullmodel, newdata, interval='prediction', level = 0.99)
```

```
##        fit      lwr      upr
## 1 903.5738 779.5206 1027.627
```

7. Test whether the demographic variables (EDUC, NONWHITE) can jointly explain mortality in the city, given than PRECIP, logNOX and logSO2 are considered.

Similar to 5, we will use linear regression using `EDUC` and `NONWHITE` as the independent variables and compare it with the full model:

```
mort.demogmodel <- lm(MORT ~ EDUC + NONWHITE, data = pol.data)
anova(mort.demogmodel, mort.fullmodel)
```

```
## Analysis of Variance Table
##
## Model 1: MORT ~ EDUC + NONWHITE
## Model 2: MORT ~ PRECIP + EDUC + NONWHITE + logNOX + logSOX
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1     57 99830
## 2     54 71159  3     28671 7.2525 0.000356 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value `Pr(>F)` is less than 0.05. As such, we can say that demographic variables can jointly explain the mortality.

8. Suppose now that we include the "Region" variable in the regression by creating the necessary dummy variables, with "West" as baseline. Is there sufficient evidence that the regions can explain mortality jointly, given that we have our previous variables?

Since we want "West" as a baseline for our "Region" variable, we would need to relevel first our data. We do this using the relevel function:

```
repol.data <- within(pol.data, Region <- relevel(pol.data$Region,"West"))
```

We shall use this data to create a full model with Regions as an additional coefficient:

```
mort.fullmodelwReg <- lm(MORT ~ PRECIP + EDUC + NONWHITE +
                           logNOX + logSOX + Region, data = repol.data)
```

We now compare this with a Region only model:

```
mort.reggmodel <- lm(MORT ~ Region, data = repol.data)
anova(mort.reggmodel, mort.fullmodelwReg)
```

```
## Analysis of Variance Table
##
## Model 1: MORT ~ Region
## Model 2: MORT ~ PRECIP + EDUC + NONWHITE + logNOX + logSOX + Region
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1     56 174324
## 2     51  64566  5    109758 17.34 5.379e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value is less than our level of significance of 0.05, we can say that the regions can explain the mortality.

9. Test whether the pollutant variables jointly explain the variability of mortality given that nonpollutant and region variables have been considered.

Now, we will use the model we created in 8 (full model with regions) for ANOVA testing with our model in 5 (model with pollutant variables):

```
anova(mort.polmodel, mort.fullmodelwReg)
```

```
## Analysis of Variance Table
##
## Model 1: MORT ~ logNOX + logSOX
## Model 2: MORT ~ PRECIP + EDUC + NONWHITE + logNOX + logSOX + Region
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     57 191172
## 2     51  64566  6    126607 16.668 1.581e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, we can see that the pollutant variables jointly explain the variability of mortality given that the non-pollutant and region variables have been considered since the p-value `Pr(>F)` is less than 0.05.

10. Suppose that a city has an annual mean precipitation of 10 inches, 10% of the population being non-white, an average of 11 years in education, logNOX at 1, and logSOX at 1, and is located at the West region. With 95% confidence, what is the expected level of mortality? Generate the 99% prediction interval for mortality in the city as well.

We create our new dataset with the said specifications:

```r
newdatawReg <- data.frame(PRECIP=10, EDUC=11, NONWHITE=10,
                       logNOX = 1, logSOX = 1, Region="West")
```

We predict the expected level of mortality with 95% confidence interval, with the following lower (`lwr`) and upper (`upr`) bounds:

```r
predict(mort.fullmodelwReg, newdatawReg, interval='confidence')
```

```
##        fit      lwr      upr
## 1 833.3435 774.7103 891.9768
```

We also generate the 99% prediction interval:

```r
predict(mort.fullmodelwReg, newdatawReg, interval='prediction', level = .99)
```

```
##        fit      lwr      upr
## 1 833.3435 710.1739 956.5132
```