Assignment 2: Multivariate_Regression

Serena Choi

2/1/2022

Matrices and Multivariate Regression

Part I. Matrix Operation

First, we will create matrices, X and y. I will assign each vectors (as column) of X as x1, x2 and x3, then combine them as columns to create the matrix X.

```
x1 <- rep(1, times = 9)
x2 <- c(3, 0, 3, 4, 1, 1, 2, 5, 0)
x3 <- c(4, 1, 9, 3, 2, 0, 6, 0, 7)
X <- cbind(x1, x2, x3)
is.matrix(X)
```

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

Similarly, we will create the matrix y, which has only one column.

```
y1 <- c(3, 4, 5, 0, 1, 2, 2, 4, 1)
y <- matrix(y1, ncol = 1)
is.matrix(y)
```

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

1. Matrix calculations We will calculate the following matrices: X', X'X, X'y

X' looks like below:

```
t.X <- t(X)
t.X
```

```
##    [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
## x1    1    1    1    1    1    1    1    1    1
## x2    3    0    3    4    1    1    2    5    0
## x3    4    1    9    3    2    0    6    0    7
```

The calculation for X'X looks like below:

```
t.X.X <- t(X) %*% X
t.X.X
```

```
##    x1 x2  x3
## x1  9 19  32
## x2 19 65  65
## x3 32 65 196
```

Similarly, the calculation for X'y is as follows:

```
t.X.y <- t(X) %*% y
t.X.y
```

```
##    [,1]
## x1   22
## x2   51
## x3   82
```

2. Matrix operations Given that X is a design matrix and y is an column vector for the response variable, we will find the least squares estimates by using the formula that OLS is (X'X)ˆ-1 * X'y). The least squares estimates, also known as beta coefficients, are calculated below:

```
inv.Xt <- solve (t.X.X)
OLS <- inv.Xt %*% t.X.y
OLS
```

```
##          [,1]
## x1 1.86262868
## x2 0.18835449
## x3 0.05180021
```

We will check the estimates above by inputting X and y as a data set and using lm regression function.

```
#Setting X and y as a data set
Xdata <- as.data.frame(X)
ydata <- as.data.frame(y)

#lm model
model1 <- lm (y ~ x2 + x3, data=Xdata)
summary(model1)
```

```
##
## Call:
## lm(formula = y ~ x2 + x3, data = Xdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.77145 -1.15458 -0.05098  1.19560  2.10611
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.8626     1.2792   1.456    0.196
## x2            0.1883     0.3771   0.499    0.635
## x3            0.0518     0.2075   0.250    0.811
##
## Residual standard error: 1.878 on 6 degrees of freedom
## Multiple R-squared:  0.04742,    Adjusted R-squared:  -0.2701
## F-statistic: 0.1493 on 2 and 6 DF,  p-value: 0.8644
```

We can see that the beta coefficients estimates from the model summary statistics and those from the matrix formula are the same.

Part II. Multivariate Regression

Using the OECD data set, we will perform a multivariate linear regression analysis. The *research questions* are:

1) How does the coefficient for one variable change when the other independent variables are included in the model, and why; and

2) What are the effects of economic development, left politics, and age demographics on welfare state development.

```
oecd <- read.csv(file="oecd.csv")
```

In order to answer the **first research question**, we will look at three models with decommodification as the response variable.

Model 1 is the bivariate model that we created in the Assignment 1, with left politics as an independent variable. **Model 2** will add economic development as another independent variable, and **Model 3** will consider all three as independent variables.

According to the model 1 below, the beta coefficient of left politics is 1.6140, indicating the *1.6140 unit increase* in welfare state development for a one unit increase in left politics. This is statistically significant at p<=0.05.

```
model1 <- lm(decom ~ left, data=oecd)
summary(model1)
```

```
##
## Call:
## lm(formula = decom ~ left, data = oecd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.4314  -3.8619   0.3309   5.6139   7.6142
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.6877     2.3496   9.656 4.46e-08 ***
## left          1.6140     0.6249   2.583     0.02 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.686 on 16 degrees of freedom
## Multiple R-squared:  0.2943, Adjusted R-squared:  0.2502
## F-statistic: 6.672 on 1 and 16 DF,  p-value: 0.02002
```

In the model 2, this beta coefficient *decreases to 1.1512*, indicating left politics has less impact on decommodification when holding economic development constant. Another change is that p-value has increased to 0.08, and therefore, it is marginally significant at 0.05 level.

```
model2 <- lm(decom ~ left + gdp, data = oecd)
summary(model2)
```

```
##
## Call:
## lm(formula = decom ~ left + gdp, data = oecd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.966  -3.970   1.226   5.331   5.885
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.1835     6.0442   1.850   0.0841 .
## left          1.1512     0.6148   1.873   0.0808 .
## gdp           1.1674     0.5733   2.036   0.0598 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.112 on 15 degrees of freedom
## Multiple R-squared:  0.4471, Adjusted R-squared:  0.3734
## F-statistic: 6.065 on 2 and 15 DF,  p-value: 0.01174
```

Similary, in the model 3, the coefficient *further decreases to 0.3313*, indicating only 0.3313 unit increases in mean of decommodification for one unit increase in left politics, net of economic development and age demographics. The p-value is 0.714, and therefore, it is marginally significant at 0.05 level.

```
model3 <- lm(decom ~ left + gdp + gt65, data = oecd)
summary(model3)
```

```
##
## Call:
## lm(formula = decom ~ left + gdp + gt65, data = oecd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.788  -2.612   0.542   3.969   7.656
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.001412  10.669271   0.000    1.000
## left         0.331306   0.886923   0.374    0.714
## gdp          0.785102   0.638967   1.229    0.239
## gt65         1.386225   1.099634   1.261    0.228
##
## Residual standard error: 5.996 on 14 degrees of freedom
## Multiple R-squared:  0.5035, Adjusted R-squared:  0.3971
## F-statistic: 4.732 on 3 and 14 DF,  p-value: 0.01756
```

In conclusion, the coefficient estimate for left politics continues to decrease as more independent variables are included in the model. It is expected since adding another variable helps us to isolate the effect of the variable of interest, left politics. The rule of thumb is that the more variables are considered, the better the model explains.

The **second research question** requires us to examine the model 3 more closely. In this multivariate regression, the *null hypotheses* are the beta coefficients for all three independent variables are 0.

Let's look at the summary statistics of the model 3 again:

```
summary(model3)
```

```
##
## Call:
## lm(formula = decom ~ left + gdp + gt65, data = oecd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.788  -2.612   0.542   3.969   7.656
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.001412  10.669271   0.000    1.000
## left         0.331306   0.886923   0.374    0.714
```

```
## gdp          0.785102   0.638967   1.229    0.239
## gt65         1.386225   1.099634   1.261    0.228
##
## Residual standard error: 5.996 on 14 degrees of freedom
## Multiple R-squared:  0.5035, Adjusted R-squared:  0.3971
## F-statistic: 4.732 on 3 and 14 DF,  p-value: 0.01756
```

The key analyses from the model above is as follows:

1. The *beta estimates* for the left politics, gdp, gt65 are 0.0014, 0.3313 and 0.7851, respectively. That is, all three variables have positive relationship with the response variable, decommodification, when controlling one another constant. As we learned in the lecture, we cannot compare the magnitudes of these coefficients as they are in different metric units.
2. Looking at p-values, while left politics is only marginally significant at 0.05 level, the other two variables are very significant at 0.05. Overall, we can *reject the null hypotheses*, and therefore, conclude that *all three variables have impacts on welfare state development.*
3. The *model fit is good* looking at R-squared and F-statistics join p-value.