625.661 - Homework Two

Eric Niblock

February 4, 2022

1. **In a typical multiple linear regression model where x_1 and x_2 are non-random regressors, the expected value of the response variable y given x_1 and x_2 is denoted by $E(y|x_1, x_2)$. Build a multiple linear regression model for $E(y|x_1, x_2)$ such that the value of $E(y|x_1, x_2)$ may change as the value of x_2 changes but the change in the value of $E(y|x_1, x_2)$ may differ in the value of x_1. How can such a potential difference be tested and estimated statistically?**

We need a function for y where $\Delta E(y|x_1, x_2)$ is not necessarily proportional with respect to Δx_1. Here is one such function,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon \tag{1}$$

$$E[y|x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 \tag{2}$$

Now it is clear that the change in $E[y|x_1, x_2]$ differs in the value of x_1 as we shift between two values of x_2,

$$E[y|x_1, x_2^{'}] - E[y|x_1, x_2] = \beta_2(x_2^{'} - x_2) + \beta_{12} x_1 (x_2^{'} - x_2) \tag{3}$$

In other words, in the above example, the change in the expected value is a function of *both* the change in x_2 and the value of x_1.

In order to discern the potential difference in expected value concerning two different values of x_1, we first choose a fixed set of differences $x_2^{'} - x_2$. We then calculate:

$$\gamma_{x_1} = \sum_{i=1}^{n} y_i^{'} - y_i = \sum_{i=1}^{n} \beta_2(x_{2,i}^{'} - x_{2,i}) + \beta_{12} x_1 (x_{2,i}^{'} - x_{2,i}) \tag{4}$$

So, γ_{x_1} would represent an unbiased estimator of the difference in expected value at some fixed x_1 value. We could generate another unbiased estimator of this difference at a different fixed value of x_1, $\gamma_{x_1'}$, and then run a test to see if these values are statistically different. Our null hypothesis would be,

$$H_0 : \gamma_{x_1} - \gamma_{x_1'} = 0 \tag{5}$$

2. For any multiple linear regression model, the total sum of squares can be decomposed into the sum of squares contributed solely by the predictor vector and the sum of squares contributed solely by the residual vector. To assess the importance of a set of the regressors, they can be taken jointly by regressing the response variable on this set of regressors, either including or excluding other regressors.

(a) Discuss, using mathematical proof in vector or matrix expressions, how this set of regressors can be assessed by use of the sums of squares of the residual vectors.

We can assess any individual regressor based on the SSE associated to that specific regressor. In other words, if we have the full model with all k regressors and an intercept, then we have $\hat{y} = X\hat{\beta}$. If we want to find the residual sum of squares associated to an individual regressor, x_j, we set $\hat{\beta}_j = 0$ and create a new model without that coefficient: $\hat{y}_{i \neq j} = X_{i \neq j} \hat{\beta}_{i \neq j}$. This notation implies that $X_{i \neq j}$ lacks the column associated with x_j and $\beta_{i \neq j}$ lacks the j-th entry. Then the difference between the full model and model lacking the j-th coefficient yields the $SS_{Res.}$ caused by the j-th entry:

$$\begin{aligned} SS_{Res.}(\beta_j | \beta_0, ..., \beta_{k-1}) &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T X\beta - (\mathbf{y}^T\mathbf{y} - \mathbf{y}^T X_{i \neq j}\beta_{i \neq j}) \\ &= \mathbf{y}^T X_{i \neq j}\beta_{i \neq j} - \mathbf{y}^T X\beta \end{aligned} \tag{6}$$

It is clear that $SS_{Res.}(\beta_j|\beta_0, ..., \beta_{k-1})$ must be positive since (as shown in Problem 4), the introduction of a regressor must serve to decrease the residual sum of squares, otherwise the coefficient of the regressor would be set to zero.

(b) Will including or excluding regressors change the conclusion in (a)? Why or why not? Provide proof for your assertions.

As shown in Problem 4, including additional regressors serves to decrease (or leave unchanged) the residual sum of squares since,

$$\sum_{i=1}^{n}(y_i-(\hat{\beta}_0+\hat{\beta}_1x_1^{(i)}+...+\hat{\beta}_kx_k^{(i)}))^2 \leq \sum_{i=1}^{n}(y_i-(\hat{\beta}_0+\hat{\beta}_1x_1^{(i)}+...+\hat{\beta}_kx_{k-1}^{(i)}))^2 \tag{7}$$

Given that $SS_T = SS_R + SS_{Res.}$ is always constant provided that the data doesn't change, adding model parameters generally decreases $SS_{Res.}$ while increasing SS_R.

3. Consider the gasoline mileage data in Table B.3. Complete the following using 20 randomly selected data points.

Note that this problem was completed by hand before realizing that statistical software use was required. The end of the attached PDF shows the results provided by use of statistical software.

(a) Fit a multiple linear regression model relating gasoline mileage y (miles per gallon) to engine displacement x_1 and the number of carburetor barrels x_6.

The attached PDF reveals the following fitted model:

$$\hat{y} = 33.92 - 0.057x_1 + 1.060x_6 \tag{8}$$

(b) Construct the analysis-of-variance table and test for significance of regression.

The attached PDF provides the table and resulting p-value. The regression is significant given the p-value is approximately 0.00001.

(c) Calculate R^2 and R^2_{Adj} for this model. Compare this to the R^2 and the R^2_{Adj} for the simple linear regression model relating mileage to engine displacement in Problem 2.4.

The attached PDF provides the three requested values. The R^2 of our multiple linear regression is higher than that of our simple linear regression, which must be the case. Furthermore R^2_{Adj} is lower, than R^2 for our multiple linear regression, though still higher than our R^2 for simple linear regression, suggesting that the addition of x_6 may produce a slightly improved predictive model.

(d) Find a 95% CI for β_1.

The attached PDF provides the 95% confidence interval of $(-0.0741, -0.0407)$ for β_1.

(e) Compute the t statistics for testing H_0: $\beta_1 = 0$ and H_0: $\beta_6 = 0$. What conclusions can you draw?

The attached PDF reveals the t statistics and p-values, revealing that x_1 is statistically significant, though x_6 is not.

(f) Find a 95% CI on the mean gasoline mileage when $x_1 = 275$ in.^{3} and $x_6 = 2$ barrels.

The attached PDF reveals a 95% CI on the mean gasoline mileage when $x_1 = 275$ in.^{3} and $x_6 = 2$ barrels of $(18.428, 22.081)$

(g) Find a 95% prediction interval for a new observation on gasoline mileage when $x_1 = 275$ in.^{3} and $x_6 = 2$ barrels.

The attached PDF reveals a 95% PI on the mean gasoline mileage when $x_1 = 275$ in.^{3} and $x_6 = 2$ barrels of $(13.131, 27.378)$

4. Will R^2 value decrease by excluding regressor(s) from any linear regression model? Will R^2_{adj} value decrease by excluding regressor(s) from any linear regression model? Provide mathematical proof for your answers to the two questions.

Recall the formula for R^2 is given by,

$$\begin{aligned} R^2 &= 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \\ &= 1 - \frac{\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)} + ... + \hat{\beta}_k x_k^{(i)}))^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \end{aligned} \tag{9}$$

It is clear that only the numerator depends on the regressors. Furthermore we have that,

$$\hat{\beta}_0, ..., \hat{\beta}_k = \min_{\beta_0,...,\beta_k} \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_1^{(i)} + ... + \beta_k x_k^{(i)}))^2 \tag{10}$$

And therefore,

$$\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)} + ... + \hat{\beta}_k x_k^{(i)}))^2 \leq \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)} + ... + \hat{\beta}_k x_{k-1}^{(i)}))^2 \tag{11}$$

It is clear that if the introduction of $\hat{\beta}_k$ did not serve to reduce the function further, it would be set to zero within the minimization problem. Therefore the exclusion of specific regressors can only decrease or leave R^2 unchanged.

It is not true that R^2_{adj} will decrease if we exclude a regressor from any linear regression model. We have that,

$$R^2_{adj,k} = 1 - \frac{(1 - R_k^2)(n-1)}{n-k-1} \tag{12}$$

So in order for R^2_{adj} to decrease when removing a regressor we must have that,

$$1 - \frac{(1 - R^2_{k-1})(n-1)}{n-(k-1)-1} \leq 1 - \frac{(1-R^2_k)(n-1)}{n-k-1} \tag{13}$$

$$\frac{1 - R^2_{k-1}}{1 - R^2_k} \geq \frac{n-k}{n-k-1} \tag{14}$$

The right side of the inequality is strictly greater than one, however, as explained above, there could be instances where $R^2_{k-1} = R^2_k$. This is an obvious example of when the inequality does not hold, and therefore, it is not always true that removing a regressor causes R^2_{adj} to decrease.

5. In a typical multiple linear regression model that includes k regressors and an intercept, derive the expectation of the sum of squares of the n predicted y values.

The sum of squares of the predicted values is given by,

$$\sum_{i=1}^{n} \hat{y}_i^2 = \hat{\mathbf{y}}^T \hat{\mathbf{y}} = (\mathbf{X}\hat{\beta})^T \mathbf{X}\hat{\beta} = \hat{\beta}^T \mathbf{X}^T \mathbf{X} \hat{\beta} \tag{15}$$

And furthermore, we know that $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, yielding,

$$\begin{aligned} \hat{\beta}^T \mathbf{X}^T \mathbf{X} \hat{\beta} &= ((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y})^T \mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= ((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y})^T \mathbf{X}^T\mathbf{y} \\ &= \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \end{aligned} \tag{16}$$

Since this result is a real number, we can take the expected value of the trace. For simplicity, we introduce the hat matrix $H = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$,

$$
\begin{aligned}
E[tr(\mathbf{y}^T\mathbf{H}\mathbf{y})] &= E[tr(\mathbf{y}\mathbf{y}^T\mathbf{H})] \\
&= tr(\mathbf{H}E[\mathbf{y}\mathbf{y}^T]) \\
&= tr(\mathbf{H}(Var(\mathbf{y}) + E[\mathbf{y}]E[\mathbf{y}^T]) \\
&= \frac{\sigma^2}{n}tr(\mathbf{H}) + tr(\mathbf{H}E[\mathbf{y}]E[\mathbf{y}^T]) \\
&= \frac{\sigma^2}{n}tr(\mathbf{H}) + E[\mathbf{y}^T]\mathbf{H}E[\mathbf{y}] \qquad (17) \\
&= \frac{\sigma^2(k+1)}{n} + E[\mathbf{y}^T]\mathbf{H}E[\mathbf{y}] \\
&= \frac{\sigma^2(k+1)}{n} + \beta^T\mathbf{X}^T\mathbf{X}\beta \\
&= \frac{\sigma^2(k+1)}{n} + \mathbf{y}^T\mathbf{y}
\end{aligned}
$$

6. Companies considering the purchase of a computer must first assess their future needs in order to determine the proper equipment. A computer scientist collected data from seven similar company sites so that a forecast equation of computer-hardware requirements for inventory management could be developed. In the regression fit, calculation of the confidence interval and prediction interval, what are the assumptions I made?

The following assumptions are required for the regression fit:

(a) Z_1 and Z_2 are not highly correlated.

(b) The residuals have a mean of zero and a constant σ^2.

(c) The errors are independent.

A confidence interval requires the following assumptions:

(a) Z_1 and Z_2 are not highly correlated.

(b) The residuals have a mean of zero and a constant σ^2.

(c) The errors are independent.

(d) The residuals follow a normal distribution.

A prediction interval requires the same assumptions as the confidence interval, though it is more sensitive regarding the normality assumption. All three require a random sample of the data.

Data Loading and Selection

In [37]:
```
# Load Data

import pandas as pd
import numpy as np
from scipy import stats
df = pd.read_excel(r'data-table-B3.XLS')
```

```
WARNING *** file size (10909) not 512 + multiple of sector size (512)
WARNING *** OLE2 inconsistency: SSCS size is 0 but SSAT size is non-zero
```

In [38]:
```
## Relevant vectors
use = df[['y','x1','x6']]

n = 20
k=2

sample = use.sample(n)
# sample = use.iloc[[28,4,21,9,23,6,26,30,7,15,13,10,12,20,5,0,19,8,11,16]] ## Used for

ones = np.ones(len(sample))
x1 = np.array(sample['x1'])
x6 = np.array(sample['x6'])
Y = np.array(sample['y'])
sample
```

Out[38]:

	y	x1	x6
28	13.90	351.0	2
4	20.07	225.0	1
21	21.47	360.0	2
9	30.40	96.9	2
23	31.90	96.9	2
6	22.12	231.0	2
26	23.90	133.6	2
30	13.77	360.0	4
7	21.47	262.0	2
15	17.80	302.0	2
13	19.70	258.0	1
10	16.50	350.0	4
12	21.50	171.0	2
20	23.54	231.0	2
5	11.20	440.0	4

	y	x1	x6
0	18.90	350.0	4
19	16.41	318.0	2
8	34.70	89.7	2
11	36.50	85.3	2
16	14.39	500.0	4

Part (a)

In [39]:
```
## Matrix X

X = np.zeros((len(ones),3))
X[:,0] = ones
X[:,1] = x1
X[:,2] = x6
```

In [40]:
```
## Calculation of Model Parameters

betas = np.linalg.inv(X.T@X)@X.T@Y
print('beta0:  ', betas[0])
print('beta1:  ', betas[1])
print('beta2:  ', betas[2])
```

```
beta0:   33.92155846550538
beta1:   -0.05740406277911794
beta2:   1.0596742386872364
```

Part (b)

In [41]:
```
Ydiff = Y - (np.ones(len(Y))*np.mean(Y))
TSS = Ydiff.T@Ydiff
```

In [42]:
```
ypred_diff = X@betas - (np.ones(len(Y))*np.mean(Y))
RSS = ypred_diff.T@ypred_diff
```

In [43]:
```
ESS = (Y.T@Y - betas.T@X.T@Y)
```

In [44]:
```
## Analysis of variance

types = ['Regression', 'Residual', 'Total']
SS = [RSS, ESS, TSS]
degf = [k, n-k-1, n-1]
MS = [RSS/degf[0], ESS/degf[1], '-']
F0 = [MS[0]/MS[1], '-', '-']
p = [0.00, '-', '-']
```

```
pd.DataFrame({'Type':types, 'Sum Squares': SS, 'DoF': degf, 'Mean Square': MS, 'F_0':F0
```

Out[44]:

	Type	Sum Squares	DoF	Mean Square	F_0	p-Value
0	Regression	766.67352	2	383.337	35.9958	0
1	Residual	181.04130	17	10.6495	-	-
2	Total	947.71482	19	-	-	-

Part (c)

In [45]:

```
## R^2 and R_adj^2

r2 = 1 - (ESS/TSS)
r2adj = 1 - ((1-r2)*(n-1)/(n-k-1))
```

In [46]:

```
print('R^2 value: ', r2)
print('R_adg^2 value: ', r2adj)
print('Previous R^2 value: ', 0.7475956798276959)
```

```
R^2 value:  0.8089706982435291
R_adg^2 value:  0.7864966627427677
Previous R^2 value:  0.7475956798276959
```

Part (d)

In [47]:

```
C = MS[1]*np.linalg.inv(X.T@X)

se = C[1,1]**0.5
tval = stats.t.ppf(1-0.025, n-(k+1))

print('95% CI for beta1:    (',betas[1]-tval*se,', ',betas[1]+tval*se,')')
```

```
95% CI for beta1:    ( -0.07410292415287431 ,  -0.04070520140536157 )
```

Part (e)

In [48]:

```
T1 = betas[1]/se
pval_T1 = 2*(1 - stats.t.cdf(abs(T1), n-(k+1)))

T6 = betas[2]/(C[2,2]**0.5)
pval_T6 = 2*(1 - stats.t.cdf(abs(T6), n-(k+1)))

print('beta1 t-value: ', T1)
print('beta6 t-value: ', T6)
print('beta1 p-value: ', pval_T1)
print('beta6 p-value: ', pval_T6)
```

```
beta1 t-value:  -7.2527092219972715
```

```
beta6 t-value:  1.0987458018325706
beta1 p-value:  1.3506989091638388e-06
beta6 p-value:  0.28719694300786536
```

Part (f)

In [49]:

```
x_i = np.array([1, 275, 2]).T

ypred = betas@x_i
pm = tval*(x_i.T@C@x_i)**0.5

print('95% CI for x1, x2 values:    (',ypred-pm,', ',ypred+pm,')')
```

```
95% CI for x1, x2 values:    ( 18.428314756433632 ,  22.0812646008112 )
```

Part (g)

In [50]:

```
x_i = np.array([1, 275, 2]).T

ypred = betas@x_i
pm = tval*(x_i.T@C@x_i + MS[1])**0.5

print('95% PI for x1, x2 values:    (',ypred-pm,', ',ypred+pm,')')
```

```
95% PI for x1, x2 values:    ( 13.13156642121319 ,  27.37801293603164 )
```

Results Using Statistical Software

In [52]:

```
import statsmodels.api as sm

X = np.array(sample[['x1','x6']])
X = sm.add_constant(X)
y = Y.reshape((20, 1))
mod = sm.OLS(y, X)
results = mod.fit()
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.809
Model:                            OLS   Adj. R-squared:                  0.786
Method:                 Least Squares   F-statistic:                     36.00
Date:                Mon, 14 Feb 2022   Prob (F-statistic):           7.75e-07
Time:                        20:43:35   Log-Likelihood:                -50.409
No. Observations:                  20   AIC:                             106.8
Df Residuals:                      17   BIC:                             109.8
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         33.9216      2.043     16.604      0.000      29.611      38.232
x1            -0.0574      0.008     -7.253      0.000      -0.074      -0.041
x2             1.0597      0.964      1.099      0.287      -0.975       3.094
==============================================================================
```

```
Omnibus:                        1.308   Durbin-Watson:                   1.287
Prob(Omnibus):                  0.520   Jarque-Bera (JB):                1.089
Skew:                           0.522   Prob(JB):                        0.580
Kurtosis:                       2.533   Cond. No.                         827.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specifi
ed.
```