Assignment 01

Matrix Algebra for Linear Regression

This assignment is worth 20 points.

Unstandardized Regression

1. Write out the design matrix that would be used if we fitted the model `lm(infant ~ 1 + pci + region + pci:region)`. Assume that `Americas` is the reference group in this model.

```
##       [,1] [,2] [,3] [,4]
##  [1,]    1  400    1  400
##  [2,]    1  200    0    0
##  [3,]    1   68    1   68
##  [4,]    1  406    0    0
##  [5,]    1  169    1  169
##  [6,]    1  130    1  130
##  [7,]    1  507    0    0
##  [8,]    1  347    0    0
##  [9,]    1   61    1   61
## [10,]    1  732    0    0
```

2. Write out the elements of the matrix $\mathbf{X}^\intercal\mathbf{X}$, where $\mathbf{X}$ is the design matrix.

```
t(X) %*% X
```

```
##      [,1]    [,2] [,3]   [,4]
## [1,]   10    3020    5    828
## [2,] 3020 1331924  828 213806
## [3,]    5     828    5    828
## [4,]  828  213806  828 213806
```

3. Using matrix algebra, compute the column vector of coefficients from the OLS regression. Report this matrix.

```
b = solve(t(X) %*% X) %*% t(X) %*% Y
b
```

```
##             [,1]
## [1,] 68.16113912
## [2,] -0.05511209
## [3,] 75.28536912
## [4,] -0.11968567
```

4. Using matrix algebra, compute and report the matrix of fitted values for each of the 10 observations.

```
y_hat = X %*% b
y_hat
```

```
##             [,1]
##  [1,]  73.52741
##  [2,]  57.13872
##  [3,] 131.56026
##  [4,]  45.78563
##  [5,] 113.90569
##  [6,] 120.72280
##  [7,]  40.21931
##  [8,]  49.03724
##  [9,] 132.78385
## [10,]  27.81909
```

5. Using matrix algebra, compute and report the matrix of residuals for each of the 10 observations.

```
e = Y - y_hat
e
```

```
##              [,1]
##  [1,]  12.7725938
##  [2,]   3.2612784
##  [3,]  18.4397391
##  [4,]   3.0143684
##  [5,] -58.9056876
##  [6,]  27.5771999
##  [7,]   5.7806892
##  [8,] -10.4372448
##  [9,]   0.1161548
## [10,]  -1.6190911
```

6. Using matrix algebra, compute and report the estimated value for the MSE.

```
mse = (t(e) %*% e) / 6
mse
```

```
##          [,1]
## [1,] 816.3758
```

7. Using matrix algebra, compute and report the variance–covariance matrix of the coefficients.

```
var_cov_b = as.numeric(mse) * solve(t(X) %*% X)
var_cov_b
```

```
##              [,1]         [,2]         [,3]         [,4]
## [1,]  1161.733848 -2.277506127 -1161.733848  2.277506127
## [2,]    -2.277506  0.005195041     2.277506 -0.005195041
## [3,] -1161.733848  2.277506127  1616.937571 -4.040359754
## [4,]     2.277506 -0.005195041    -4.040360  0.015840293
```

8. Based on the variance–covariance matrix you reported in the previous question, find the SE for the coefficient associated with the main-effect of PCI.

```
sqrt(var_cov_b[2, 2])
```

```
## [1] 0.07207664
```

9. Given the assumptions of the OLS model and the MSE estimate you computed in Question 6, compute and report the variance–covariance matrix of the residuals.

```
as.numeric(mse) * diag(10)
```

```
##           [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]     [,8]
##  [1,] 816.3758   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000
##  [2,]   0.0000 816.3758   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000
##  [3,]   0.0000   0.0000 816.3758   0.0000   0.0000   0.0000   0.0000   0.0000
##  [4,]   0.0000   0.0000   0.0000 816.3758   0.0000   0.0000   0.0000   0.0000
##  [5,]   0.0000   0.0000   0.0000   0.0000 816.3758   0.0000   0.0000   0.0000
##  [6,]   0.0000   0.0000   0.0000   0.0000   0.0000 816.3758   0.0000   0.0000
##  [7,]   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000 816.3758   0.0000
##  [8,]   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000 816.3758
##  [9,]   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000
## [10,]   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000
##           [,9]    [,10]
##  [1,]   0.0000   0.0000
##  [2,]   0.0000   0.0000
##  [3,]   0.0000   0.0000
##  [4,]   0.0000   0.0000
##  [5,]   0.0000   0.0000
##  [6,]   0.0000   0.0000
##  [7,]   0.0000   0.0000
##  [8,]   0.0000   0.0000
##  [9,] 816.3758   0.0000
## [10,]   0.0000 816.3758
```

10. Compute the hat-matrix and show how you would use the values in the hat-matrix to find $\hat{y}_1$ (the predicted value for Algeria).

```
h = X %*% solve(t(X) %*% X) %*% t(X)
sum(h[1, ] * Y)
```

```
## [1] 73.52741
```

Standardized Regression

11. Write out the design matrix that would be used to fit the model. Again, assume that `Americas` is the reference group in this model.

```
##       [,1]       [,2] [,3]       [,4]
##  [1,]    1  0.4537145    1  0.4537145
##  [2,]    1 -0.4722335    0  0.0000000
##  [3,]    1 -1.0833591    1 -1.0833591
##  [4,]    1  0.4814929    0  0.0000000
##  [5,]    1 -0.6157554    1 -0.6157554
##  [6,]    1 -0.7963153    1 -0.7963153
##  [7,]    1  0.9490967    0  0.0000000
##  [8,]    1  0.2083383    0  0.0000000
##  [9,]    1 -1.1157673    1 -1.1157673
## [10,]    1  1.9907882    0  0.0000000
```

Use for Questions 12–13

Using the standardized infant mortality `z_infant` and per-capita income `z_pci` variables. Fit the model: `lm(z_infant ~ z_pci + americas + z_pci:americas - 1)`.

12. How is the design matrix for this model different than the design matrix for the model fitted in Question 11? What effect does this have on the vector of coefficient values?

The formula suppresses the intercept; thus the design matrix will not have a column of ones. As a result, the vector of coefficients will be 3×1 (one less coefficient).

13. Using matrix algebra, compute and report the estimates for each of the coefficients, the standard errors of the coefficients, and the RMSE. (3pts)

```
# Compute coefficients
b = solve(t(X) %*% X) %*% t(X) %*% Y
b
```

```
##            [,1]
## [1,] -0.5967810
## [2,]  0.2411228
## [3,] -0.2012573
```

```r
y_hat = X %*% b
e = Y - y_hat
mse = (t(e) %*% e) / 7

# Compute SEs
sqrt(diag(as.numeric(mse) * solve(t(X) %*% X)))
```

```
## [1] 0.2950487 0.4545426 0.6091257
```

```r
# Compute RMSE
sqrt(as.numeric(mse))
```

```
## [1] 0.6832315
```

ANOVA Model via Regression

14. Write out the design matrix that would be used to fit the model.

```
##       [,1] [,2]
##  [1,]    1    1
##  [2,]    1   -1
##  [3,]    1    1
##  [4,]    1   -1
##  [5,]    1    1
##  [6,]    1    1
##  [7,]    1   -1
##  [8,]    1   -1
##  [9,]    1    1
## [10,]    1   -1
```

15. Using matrix algebra, compute and report the column vector of coefficients from the OLS regression.

```r
b = solve(t(X) %*% X) %*% t(X) %*% Y
b
```

```
##              [,1]
## [1,] 4.163336e-17
## [2,] 7.450812e-01
```

16. Using matrix algebra, compute and report the variance–covariance matrix for the coefficients.

```
y_hat = X %*% b
e = Y - y_hat
mse = (t(e) %*% e) / 7

# Compute variance-covariance matrix
as.numeric(mse) * solve(t(X) %*% X)
```

```
##             [,1]       [,2]
## [1,] 0.04926487 0.00000000
## [2,] 0.00000000 0.04926487
```

17. Explain why the sampling variances for the coefficients are the same and why the sampling covariance is zero by referring to computations produced in the matrix algebra. (2pts)

The reason is the computation of the inverse of $\mathbf{X}^\intercal\mathbf{X}$. This produces a 2×2 diagonal matrix that has 0.1 on the main diagonal and zeros on the off-diagonal. Since the diagonal elements are the same, the sampling variances will be equal; since the off-diagonal elements are zero, the covariance will also be zero.

```
solve(t(X) %*% X)
```

```
##      [,1] [,2]
## [1,]  0.1  0.0
## [2,]  0.0  0.1
```