Assignment 01

Matrix Algebra for Linear Regression

This goal of this assignment is to give you experience using matrix algebra for regression. Turn in a printed version of your responses to each of the questions on this assignment.

In questions that ask you to "use matrix algebra" to solve the problem, you can either show your syntax and output from carrying out the matrix operations, or you can use Equation Editor to input the matrices involved in your calculations.

In addition, please adhere to the following guidelines for further formatting your assignment:

- All graphics should be set to an appropriate aspect ratio and sized so that they do not take up more room than necessary. They should also have an appropriate caption.
- Any typed mathematics (equations, matrices, vectors, etc.) should be appropriately typeset within the document.
- Syntax or computer output should not be included in your assignment unless it is specifically asked for.

This assignment is worth 20 points. (Each question is worth 1 point unless otherwise noted.)

Data Set

The data set you will use to answer the questions in this assignment contains measurements for 10 countries on: infant mortality rate per 1000 live births (`infant`), the per-capita income (`pci`) and world region (`region`) of the country.

country	infant	pci	region
Algeria	86.3	400	Africa
Bolivia	60.4	200	Americas
Burundi	150.0	68	Africa
Dominican Republic	48.8	406	Americas
Kenya	55.0	169	Africa
Malawi	148.3	130	Africa
Nicaragua	46.0	507	Americas
Paraguay	38.6	347	Americas
Rwanda	132.9	61	Africa
Trinidad & Tobago	26.2	732	Americas

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.
2. Write out the elements of the matrix $\mathbf{X}^\top\mathbf{X}$, where $\mathbf{X}$ is the design matrix.
3. Using matrix algebra, compute and report the column vector of coefficients from the OLS regression.
4. Using matrix algebra, compute and report the matrix of fitted values for each of the 10 observations.
5. Using matrix algebra, compute and report the matrix of residuals for each of the 10 observations.
6. Using matrix algebra, compute and report the estimated value for the MSE.
7. Using matrix algebra, compute and report the variance–covariance matrix of the coefficients.
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.
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.
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).

Standardized Regression: Part I

Standardize the infant mortality `z_infant` and per-capita income `z_pci` variables. Refit the model using the standardized variables: `lm(z_infant ~ 1 + z_pci + region + z_pci:region)`.

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.

Standardized Regression: Part II

Using the standardized infant mortality `z_infant` and per-capita income `z_pci` variables. Fit the model: `lm(0 + z_infant ~ z_pci + region + z_pci:region)`.

12. How is the design matrix for this model different than the design matrix for the model fitted in Question 12? What effect does this have on the vector of coefficient values?
13. Using matrix algebra, compute and report the estimates for each of the coefficients, the standard errors of the coefficients, and the RMSE. (**3pts**)

ANOVA Model via Regression

Now consider fitting the Analysis-of-Variance (ANOVA) model to the data to examine whether there is an effect of region on infant mortality. This model is:

$$\text{Infant Mortality}_i = \mu + \alpha_{\text{Region}} + \epsilon_i$$

To fit this model, rather than dummy coding to code the region predictor, we use effects-coding which has the following constraint: $\sum \alpha_{\text{Region}} = 0$. (See Fox pp. 156–159 for more information).

14. Write out the design matrix that would be used to fit the model.
15. Using matrix algebra, compute and report the column vector of coefficients from the OLS regression.
16. Using matrix algebra, compute and report the variance–covariance matrix for the coefficients.
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**)