Assignment 2

Linear Models in Matrix Form

This goal of this assignment is to give you experience in using matrix algebra applied to linear models. Turn in a printed document that includes your responses to each of the questions on this assignment. Please adhere to the following guidelines for further formatting your assignment:

- All graphics should be resized so that they do not take up more room than necessary and should have an appropriate caption.
- Any typed mathematics (equations, matrices, vectors, etc.) should be appropriately typeset within the document using Markdown's display equations.
- No syntax should be included unless specifically asked for.

Part I

Use the following data to answer the questions in this section.

wage	age	sex
12.00	32	M
8.00	33	F
16.26	32	M
13.65	33	M
8.50	26	M

Consider the regression model that includes an intercept, and the main-effects of `age` and `sex` to predict `wage`. The `lm()` formula for this model would be `wage ~ 1 + age + sex`. For consistency, use females as the reference group.

1. Write out the design matrix (i.e., the X-matrix) for the model.

2. What are the dimensions of the design matrix?

3. Using matrix algebra, compute and report the **b** vector (i.e., the vector of the regression coefficients). Show any relevant work.

4. Using matrix algebra, compute and report the standard errors for each of the regression coefficients in the model. Show any relevant work.

5. Using the values from Questions 3 and 4, compute and report the t-statistic for each of the regression coefficients. Show any relevant work. You may need to refresh your memory about what how a t-value is computed. One place to start may be your introductory statistics textbook, or any of a number of websites online.

6. Use the `pt()` function to compute the p-value (two-sided) for each of the regression coefficients. Show any relevant work. Again, you may need to refresh your memory about what a p-value is, and how they are computed.

Part II

Now consider the regression model that includes an intercept, the main-effects of `age` and `sex`, and the interaction effect between `age` and `sex` to predict `wage`. The `lm()` formula for this model would be `wage ~ 1 + age + sex + age:sex`. For consistency, use females as the reference group.

7. Write out the design matrix for the model.
8. Try to compute the **b** vector. You get an error message saying: `Error in solve.default(t(X) %*% X) : system is computationally singular`. Explain, using the language of matrix algebra, what this error message means.
9. We could have predicted whether $\mathbf{X}'\mathbf{X}$ is computationally singular by determining whether it was rank-deficient. Compute the rank of $\mathbf{X}'\mathbf{X}$ and explain why it is rank-deficient.

Part III

Consider the one-factor analysis of variance model to predict variation in wages based on sex,

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij},$$

for $j \in \big[\text{Male},\ \text{Female}\big]$.

10. Write out the matrix form of the overparametrized ANOVA model.
11. Explain why this model is overparameterized.
12. Explain why adding the constaint $\sum \alpha_j = 0$ leads to a full-rank matrix.