MLDS 401: Homework 1
Due: September 29, 17:00
Professor Malthouse

You may work in self-selected groups of at most four. Turn in one copy per group, with all names on it. I encourage you to use Markdown in R.

1. Define the following, where $\mathbf{A}$ is symmetrical:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \text{and} \quad \mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}$$

   (a) (2 points) Find $\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}$

   (b) (2 points) Show

$$\frac{\partial \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$$

2. Suppose that we observe n data pairs: (x_i, y_i), $i = 1, \ldots, n$. Assume that $y_i = \beta_0 + \beta_1 x_i + e_i$, where $e_i \sim \mathcal{N}(0, \sigma^2)$ and the errors (e_i) are independent. This problem asks you to consider the matrix formulation of the regression problem.

   (a) (2 points) Identify $n \times 2$ matrix $\mathbf{X}$ for the model. Hint: the first column is for the intercept and the second for predictor variable x.

   (b) (2 points) Compute $\mathbf{X}^\mathsf{T}\mathbf{X}$. Write the answer in terms such as n and $\sum_i x_i$.

   (c) (2 points) Is your matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$ symmetrical?

   (d) (4 points) Now suppose that you have p predictors instead of 1, so that $\mathbf{X}$ is now $n \times (p + 1)$. Show that $\mathbf{X}^\mathsf{T}\mathbf{X}$ is symmetrical. Hint: if $\mathbf{A} = \mathbf{X}^\mathsf{T}\mathbf{X}$, show that $a_{ij} = a_{ji}$.

3. Consider the regression model $y_i = \alpha + \beta x_i + e_i$, where e_i are independent random variables with $\mathbb{E}(e_i) = 0$ and $\mathbb{V}(e_i) = \sigma^2$ for all i.

   (a) (2 points) What is the implication for the regression function if $\beta = 0$, so that the model is $y_i = \alpha + e_i$? How would the regression function plot on a graph?

   (b) (3 points) Derive the least square estimator a of α for the model above (with $\beta = 0$) and show that it equals the sample mean $a = \bar{y}$.

   (c) (3 points) Prove that the estimate a in the previous part is an unbiased estimator of α.

   (d) (3 points) What is the variance of your estimate a?

   (e) (3 points) Discuss why your estimates are (at least approximately) normally distributed.

(f) The Gauss-Markov theorem states that OLS estimates are best linear unbiased estimates ("BLUE"), i.e., among all linear, unbiased estimates, the OLS estimates have the smallest variance. Show that your esimate from part (b) is BLUE. Hints: Let $\hat{\alpha} = \sum_{i=1}^{n} c_i y_i$ be another linear (it is a linear combination of y_i) unbiased estimate, where c_i are constants. Let $d_i = c_i - 1/n$ be the difference between the constants of the new estimator and those from OLS $(1/n)$. Show that $d_i = 0$ for all i, otherwise the variance will be greater than that of $\bar{y}$ from part (d). When $d_i = 0$ the new estimate is the same as the OLS one.

   i. (2 points) What does the unbiased assumption imply about the sum of c_i?

   ii. (2 points) Show $\sum_i d_i/n = 0$.

   iii. (2 points) Evaluate $\mathbb{V}(\hat{\alpha})$ in terms of d_i and find when it is minimized over the d_i values.

   iv. Or you can think geometrically.

4. ACT Problem 2.9: Beta coefficients for stocks. Note that the data set **IBM*.csv** is on Canvas. The original data cannot be read into R very easily. See `StockBeta.csv` instead.

5. JWHT problem 8a,b on pages 121–2 (Hint: see §2.3.4 on page 48–49.) See `auto.txt` for data. If you use the data from the author's website you will need to read about the `na.strings` option. Note: omit part c for now. Use the `lm` function to regress `mpg` on `horsepower`. Use `summary`, `plot` and `abline` commands to view the results, scatterplot and fitted model. Answer these questions about the output.

   (a) What is the estimated regression equation?

   (b) What does the slope tell you?

   (c) How much uncertainty is associated with the slope estimate?

   (d) What does the residual standard error tell you?

   (e) Using this model, is there a significant relationship between mpg and horsepower?

   (f) What fraction of the variation in mpg is explained by using this linear function of horsepower?

   (g) What is the predicted mpg associated with a horsepower of 98?

   (h) What is the 95% prediction interval for the predicted mpg associated with a horsepower of 98?

   (i) What is the 99% confidence interval for the mean prediction of mpg when horsepower is 98?

   (j) What is a 90% confidence interval for the slope?

   (k) In looking at the scatterplot and fitted model, note any violations of the model assumptions. (You should have done this first!)

6. JWHT problem 9(a)–(c) on page 122. Find the correlation and scatterplot matricies and regress mpg on all other variables except for name. Hint: when finding correlations see the `use="pair"` option. Answer these questions.

```
plot(auto, pch=".") # part a
round(cor(auto[,1:7], use="pair"),4)   # part b
fit = lm(mpg~., auto[,1:8])     # part c
summary(fit)
```

   (a) (3 points) Based on the scatterplots, comment on the relationships between the predictors and mpg.

   (b) (2 points) What is the correlation between mpg and displacement and what does it tell you?

   (c) (2 points) Is there a statistically significant relationship between the predictors and the response?

   (d) (2 points) Which predictors appear to have a statistically significant relationship to the response?

   (e) (2 points) What does the slope coefficient for the year variable suggest?

   (f) (2 points) What does the slope coefficient for the displacement variable suggest?