IEMS 404: 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 $\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}$ *Answer:* $a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2$

   (b) (2 points) Show

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

   Answer:

$$\frac{\partial \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}}{\partial x_1} \\ \frac{\partial \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}}{\partial x_2} \end{pmatrix} = \begin{pmatrix} 2a_{11}x_1 + 2a_{12}x_2 \\ 2a_{12}x_2 + 2a_{22}x_2 \end{pmatrix} = 2\begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 2\mathbf{A}\mathbf{x}$$

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. *Answer:*

$$\begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

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

$$\mathbf{X}^\mathsf{T}\mathbf{X} = \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix}\begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix}$$

   (c) (2 points) Is your matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$ symmetrical? *Answer: Yes,* $\sum_i x_i = \sum_i x_i$

(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 $a_{ij} = a_{ji}$. *Answer: To keep the notation simple, I will write $\mathbf{X}$ as an $n \times p$ matrix, where the first column equals 1 for the intercept, i.e., $x_{i1} = 1$ for $i = 1, \ldots, n$:*

$$\mathbf{A} = \mathbf{X}^\mathsf{T}\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{n1} \\ \vdots & \ddots & \vdots \\ x_{1p} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

We need to show that $a_{ij} = a_{ji}$. Note that a_{ij} is the dot product of row i in $\mathbf{X}^\mathsf{T}$ and column j in $\mathbf{X}$, and a_{ji} is the dot product of the row j in $\mathbf{X}^\mathsf{T}$ and column i in $\mathbf{X}$.

$$a_{ij} = \sum_{k=1}^{n} x_{ki}x_{kj} = \sum_{k=1}^{n} x_{kj}x_{ki} = a_{ji}$$

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? *Answer: The regression function is a horizontal line.*

(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 $a = \bar{y}$. *Answer: Note that RSS $= \sum_{i=1}^{n}(y_i - a)^2$. We want to minimize SSE, so compute $d\mathrm{SSE}/da = -2\sum(y_i - a)$. We set this equal to 0 and solve for $a = \frac{1}{n}\sum y_i = \bar{y}$. Note that the second derivative in $2 > 0$, so we have a minimum.*

(c) (3 points) Prove that the estimate a in the previous part is an unbiased estimator of α. *Answer:* $\mathbb{E}(a) = \mathbb{E}[(1/n)\sum y_i] = (1/n)\sum \mathbb{E}(y_i) = (1/n)\sum \mathbb{E}(\alpha + e_i) = (1/n)n\alpha + 0 = \alpha$.

(d) (3 points) What is the variance of your estimate a? *Answer:* $\mathbb{V}(a) = \mathbb{V}[(1/n)\sum y_i] = (1/n^2)\sum \mathbb{V}(y_i) = (1/n^2)\sum \mathbb{V}(\alpha + e_i) = (1/n^2)n\sigma^2 = \sigma^2/n$

(e) (3 points) Discuss why your estimates are (at least approximately) normally distributed. *Answer: The estimate is a linear combination of random variables e_i. If e_i are normal, then a linear combination is normal. If the errors are not normal, then the central limit theorem tells us that the estimate will become approximately normal as the sample size increases.*

(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? *Answer: Unbiased means that*

$$\alpha \equiv \mathbb{E}(\hat{\alpha}) = \mathbb{E}(\sum_i c_i y_i) = \mathbb{E}[\sum_i c_i(\alpha+e_i)] = \sum_i c_i[\alpha+\mathbb{E}(e_i)] = \sum_i c_i[\alpha+0] = \alpha \sum_i c_i.$$

This implies that the sum of constants c_i must equal 1 for the estimate to be unbiased. This is a constraint on the constants.

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

$$\sum_{i=1}^{n} \frac{d_i}{n} = \frac{1}{n}\sum_{i=1}^{n}\left(c_i - \frac{1}{n}\right) = \frac{1}{n}\left(\sum_{i=1}^{n} c_i - \sum_{i=1}^{n}\frac{1}{n}\right) = \frac{1}{n}(1-1) = 0$$

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

$$\begin{aligned}
\mathbb{V}(\hat{\alpha}) &= \mathbb{V}\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} c_i^2 \mathbb{V}(y_i) = \sum_{i=1}^{n}\left(d_i + \frac{1}{n}\right)^2 \sigma^2 \\
&= \sigma^2\left(\sum_{i=1}^{n} d_i^2 + 2\sum_{i=1}^{n}\frac{d_i}{n} + \sum_{i=1}^{n}\frac{1}{n^2}\right) = \sigma^2\left(\sum_{i=1}^{n} d_i^2 + 2(0) + \frac{1}{n}\right)
\end{aligned}$$

The middle term is 0 because of part ii. Note that the expression is minimized when $d_i^2 \equiv 0$ for all i, in other words when the new estimator equals the OLS estimator.

iv. Or you can think geometrically. *Answer: For now, consider the $n = 2$ case. The constants come from the $c_1 \times c_2$ plane. The unbiased assumption constrains the points to fall on the line $c_1 + c_2 = 1$. Then $\mathbb{V}(\hat{\alpha}) = \sigma^2(c_1^2 + c_2^2)$ is recognized as a squared distance (multiplied by constant σ^2). Thus, the problem is to find the point on the line $c_1 + c_2 = 1$ that is closest to the origin, which is given by $c_1 = c_2 = 1/2$. In general, you want to find the point where the hyperball $c_1^2 + \cdots + c_p^2$ is tangent to the hyperplane $c_1 + \cdots + c_p = 1$.*

4. (8 points) ACT problem 2.5: show $\overline{y}$ and $\widehat{\beta_1}$ are independent. Hint: it is useful to establish the following lemmas: $\mathsf{C}(aX, bY) = ab\mathsf{C}(X, Y)$ and $\mathsf{C}(X + Y, Z) = \mathsf{C}(X, Z) + \mathsf{C}(Y, Z)$. *Answer: We first establish the lemmas:*

$$\mathsf{C}(aX, bY) = \mathbb{E}[(aX - a\mu_x)(bY - b\mu_y)] = ab\mathbb{E}[(X - \mu_x)(Y - \mu_y)] = ab\mathsf{C}(X, Y)$$

$$
\begin{aligned}
\mathcal{C}(X+Y,Z) &= \mathbb{E}[(X+Y-\mu_x-\mu_y)(Z-\mu_z)] \\
&= \mathbb{E}[(X-\mu_x)(Z-\mu_z)+(Y-\mu_y)(Z-\mu_z)] \\
&= \mathbb{E}[(X-\mu_x)(Z-\mu_z)]+\mathbb{E}[(Y-\mu_y)(Z-\mu_z)] \\
&= \mathcal{C}(X,Z)+\mathcal{C}(Y,Z)
\end{aligned}
$$

Recall that $\hat{\beta}_1 = \sum_i c_i y_i$, *where* $c_i = (x_i - \overline{x})/S_{xx}$. *Using the lemmas,*

$$
\begin{aligned}
\mathcal{C}(\overline{y},\hat{\beta}_1) &= \mathcal{C}(\overline{y},\sum_i c_i y_i) = \sum_i c_i \mathcal{C}(\overline{y},y_i) \\
&= \frac{1}{S_{xx}}\sum_i (x_i-\overline{x})\underbrace{\mathcal{C}(\overline{y},y_i)}_{\sigma^2/n} = \frac{\sigma^2}{nS_{xx}}\underbrace{\sum_i (x_i-\overline{x})}_{0} = 0
\end{aligned}
$$

5. 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. *Answer: (a) Positive associations, roughly linear. (b) Apple has a larger beta (1.2449) than IBM (0.7448). (c) for IBM,* $0.5975(0.0556/0.0446) \approx 0.7448$. *Likewise for Apple. (d) The correlations are similar, but the SD of Apple is nearly twice that of IBM, which means that the Apple stock is almost twice as volatile as the IBM stock. Thus the higher expected return of Apple is accompanied by its higher volatility.*

```
> setwd("/Users/ecm/teach/MLDS401/hw23")
> dat = read.csv("StockBeta.csv")
> plot(dat[,c(2:4)]) # part a
> fit = lm(Apple ~ S_P500, dat); summary(fit) # part b
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.024863   0.008606   2.889  0.00472 **
S_P500      1.244856   0.193007   6.450  3.8e-09 ***
> fit2 = lm(IBM ~ S_P500, dat); summary(fit2) # part b
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.006416   0.004414   1.454    0.149
S_P500      0.744809   0.098977   7.525 2.15e-11 ***

> apply(dat[, c(2:4)], 2, sd) # part c
    S_P500        IBM      Apple
0.04457853 0.05557105 0.10310404
> cor(dat[, c(2:4)])          # part c
          S_P500       IBM     Apple
S_P500 1.0000000 0.5974779 0.5382317
IBM    0.5974779 1.0000000 0.4147253
Apple  0.5382317 0.4147253 1.0000000
```

6. 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.

```
auto = read.table("http://www-bcf.usc.edu/~gareth/ISL/Auto.data",
  header=T, na.strings="?")
auto$origin = factor(auto$origin, 1:3, c("US", "Europe", "Japan"))
> fit = lm(mpg ~ horsepower, auto)  # part a
> summary(fit)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
Residual standard error: 4.906 on 390 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.6059,Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16

> plot(mpg ~ horsepower, auto)
> abline(fit)
> predict(fit, data.frame(horsepower=98), interval="pred")
       fit     lwr      upr
1 24.46708 14.8094 34.12476
> predict(fit, data.frame(horsepower=98), interval="conf", level=.99)
       fit      lwr      upr
1 24.46708 23.81669 25.11747
> confint(fit, level=.90)
                   5 %       95 %
(Intercept) 38.7528707 41.1188513
horsepower  -0.1684719 -0.1472176
> plot(fit)   # part c
```

(a) What is the estimated regression equation? *Answer:* $\hat{mpg} = 39.94 - 0.1578 horsepower$.

(b) What does the slope tell you? *Answer: Every unit increase in horsepower is associated with a .1578 decrease in mpg on the average.*

(c) How much uncertainty is associated with the slope estimate? *Answer: Standard error is 0.006446.*

(d) What does the residual standard error tell you? *Answer: Typical size of residuals.*

(e) Using this model, is there a significant relationship between mpg and horsepower? *Answer: Yes,* $P < 2 \times 10^{-16} < .05$.

(f) What fraction of the variation in mpg is explained by using this linear function of horsepower? *Answer:* $R^2 = .6059$

(g) What is the predicted mpg associated with a horsepower of 98? *Answer: 24.47 mpg.*

(h) What is the 95% prediction interval for the predicted mpg associated with a horsepower of 98? *Answer: 14.8094 to 34.12476.*

(i) What is the 99% confidence interval for the mean prediction of mpg when horsepower is 98? *Answer: 23.81669 to 25.11747.*

(j) What is a 90% confidence interval for the slope? *Answer:* $[-0.1684719, -0.1472176]$.

(k) In looking at the scatterplot and fitted model, note any violations of the model assumptions. (You should have done this first!) *Answer: The scatterplot shows that the relationship is not linear and the error variance is not constant.*

7. 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)
```

Answer: The regression estimates depend on how you treat origin, which is a nominal variable and should be treated as a `factor` in R. I did not tell you to do this and we have not covered dummies yet, so I will give full credit for either answer. I have shown how to redefine it as a factor in the instructions of the previous problem.

```
                 mpg cylinders displace horse-  weight  acceler    year
                                   ment   power           ation
mpg           1.0000   -0.7763  -0.8044 -0.7784 -0.8317  0.4223  0.5815
cylinders    -0.7763    1.0000   0.9509  0.8430  0.8970 -0.5041 -0.3467
displacement -0.8044    0.9509   1.0000  0.8973  0.9331 -0.5442 -0.3698
horsepower   -0.7784    0.8430   0.8973  1.0000  0.8645 -0.6892 -0.4164
weight       -0.8317    0.8970   0.9331  0.8645  1.0000 -0.4195 -0.3079
acceleration  0.4223   -0.5041  -0.5442 -0.6892 -0.4195  1.0000  0.2829
year          0.5815   -0.3467  -0.3698 -0.4164 -0.3079  0.2829  1.0000

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  -1.795e+01  4.677e+00  -3.839 0.000145 ***
cylinders    -4.897e-01  3.212e-01  -1.524 0.128215
displacement  2.398e-02  7.653e-03   3.133 0.001863 **
horsepower   -1.818e-02  1.371e-02  -1.326 0.185488
weight       -6.710e-03  6.551e-04 -10.243  < 2e-16 ***
acceleration  7.910e-02  9.822e-02   0.805 0.421101
year          7.770e-01  5.178e-02  15.005  < 2e-16 ***
originEurope  2.630e+00  5.664e-01   4.643 4.72e-06 ***
originJapan   2.853e+00  5.527e-01   5.162 3.93e-07 ***

Residual standard error: 3.307 on 383 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.8242,Adjusted R-squared:  0.8205
F-statistic: 224.5 on 8 and 383 DF,  p-value: < 2.2e-16
```

(a) (3 points) Based on the scatterplots, comment on the relationships between the predictors and mpg. *Answer: There are nonlinear relationships between mpg and displacement, horsepower, and weight; in all three cases, as x increases mpg decreases, but at a decreasing rate, suggesting the need for a log for square root transform. Cylinders have a negative linear association with mpg. Acceleration and year have positive linear associations. There is heteroscedasticity in most of the plots, with the variance increasing with the mean.*

(b) (2 points) What is the correlation between mpg and displacement and what does it tell you? *Answer: $r = -.8044$, so larger displacement is associated with smaller mpg.*

(c) (2 points) Is there a statistically significant relationship between the predictors and the response? *Answer: Yes, $H_0 : \beta_1 = \ldots = \beta_6 = 0$ versus H_1 : at least one $\beta_j \neq 0$ for $j = 1, \ldots 6$. $P < 2.2 \times 10^{-16}$, so reject H_0.*

(d) (2 points) Which predictors appear to have a statistically significant relationship to the response? *Answer: Displacement, weight, year and origin. Note that the tests are questionnable because we have a misspecified model.*

(e) (2 points) What does the slope coefficient for the year variable suggest? *Answer: $b = .777$ suggests that gas milage improves over time.*

(f) (2 points) What does the slope coefficient for the displacement variable suggest? *Answer: $b = 0.02398$ suggests that larger displacement is associated with higher gas milage. This contradicts part b because of multicollinearity. It is an example of a sign flip.*