BTRY/STSCI 4030 - Linear Models with Matrices - Fall 2017
Midterm - Monday, October 15

NAME:

NETID:

Instructions:

It is not necessary to complete numerical calculations (using a calculator) if you clearly show how the answer can be obtained, and if the exact answer is not required in subsequent parts.

A set of formulae and notes is provided with the exam; other outside material is not allowed. You may directly use any result on the notes without proving it.

You may reference any result in the formulae by it's number; e.g. the Eigen-decomposition for a symmetric matrix is in 5.2a.

The questions on this exam are inspired from a consulting meeting that Giles had with a student in Consumer Behavior on October 4 this year. The student was interested in how a student's ecological consciousness affected their preferences for displaying a brand name on a t-shirt. The following description is highly idealized.

- Subjects were given a survey about their ecological attitudes and given a numeric score, x_2, rating their ecological awareness. We will use this as x_2.
- Subject's were also classified as being religious ($x_1 = 1$) or not ($x_1 = 0$).
- Subjects were asked to rate their preference for two t-shirts displaying a brand logo: one large and one small. The difference in their preferences is the response y.

Throughout, we assume the usual framework of a linear regression, that

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \epsilon \sim N(0, \sigma^2 I)$$

for any particular X that we are working with.

We will first only use the categorical variable x_1. For this we assume we have

- n_0 subjects with $x_1 = 0$, with average response $\bar{y}_0$.
- n_1 subjects with $x_1 = 1$ with average response $\bar{y}_1$.
- Totalling $n = n_0 + n_1$ subjects with average response $\bar{y} = (n_0\bar{y}_0 + n_1\bar{y}_1)/n$.

It may be helpful to note that we can write

$$\bar{y}_1 = \frac{\boldsymbol{x}_1^T \boldsymbol{y}}{\boldsymbol{x}_1^T \boldsymbol{x}_1}$$

1. (10 points) Regressing y on x_1, we would use a covariate matrix $X_1 = [\mathbf{1}, \boldsymbol{x}_1]$, express $X_1^T X_1$ and $X_1^T \boldsymbol{y}$ in terms of n_0, n_1, $\bar{y}_0$ and $\bar{y}_1$.

   We have that $\mathbf{1}^T \boldsymbol{x}_1 = n_1$ and $\boldsymbol{x}_1^T \boldsymbol{x}_1 = n_1$ so

$$X_1^T X_1 = \left[\begin{array}{cc} n & n_1 \\ n_1 & n_1 \end{array} \right]$$

   and since $\mathbf{1}^T \boldsymbol{y} = n_0 \bar{y}_0 + n_1 \bar{y}_1$ and $\boldsymbol{x}_1^T \boldsymbol{y} = n_1 \bar{y}_1$

$$X_1^T \boldsymbol{y} = \left[\begin{array}{c} n_0 \bar{y}_0 + n_1 \bar{y}_1 \\ n_1 \bar{y}_1 \end{array} \right].$$

2. (12 points) Hence, express $(X_1^T X_1)^{-1}$ and $\hat{\boldsymbol{\beta}}$ in terms of n_0, n_1, $\bar{y}_0$ and $\bar{y}_1$. It may help to have the following formula

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Give (in words) an interpretation of $\hat{\beta}_0$ and $\hat{\beta}_1$.

Using the formula provided:

$$(X_1^T X_1)^{-1} = \frac{1}{nn_1 - n_1^2} \begin{bmatrix} n_1 & -n_1 \\ -n_1 & n \end{bmatrix} = \frac{1}{n_1 n_0} \begin{bmatrix} n_1 & -n_1 \\ -n_1 & n \end{bmatrix} = \frac{1}{n_0} \begin{bmatrix} 1 & -1 \\ -1 & n/n_1 \end{bmatrix}$$

and

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (X_1^T X_1)^{-1} X_1^T \boldsymbol{y} \\ &= \begin{bmatrix} \frac{1}{n_0}(n_0\bar{y}_0 + n_1\bar{y}_1 - n_1\bar{y}_1) \\ \frac{1}{n_0}(-(n_0\bar{y}_0 + n_1\bar{y}_1) + n\bar{y}_1) \end{bmatrix} \\ &= \begin{bmatrix} \bar{y}_0 \\ \bar{y}_1 - \bar{y}_0 \end{bmatrix}. \end{aligned}$$

Here $\hat{\beta}_0$ is the average for those subjects with $x_1 = 0$ and $\hat{\beta}_1$ is the difference between the averages at $x_1 = 1$ and $x_1 = 0$.

3. (12 points) Write the prediction for a new subject with $x_1 = 1$ in terms of $\hat{\beta}_0$ and $\hat{\beta}_1$. Show that it's variance is σ^2/n_1.

When $x_1 = 1$, from the previous answer

$$\hat{\beta}_0 + \hat{\beta}_1 x_1 = \bar{y}_0 + \bar{y}_1 - \bar{y}_0 = \bar{y}_1$$

and the variance of $\bar{y}_1$ is σ^2/n_1.

Alternatively,

$$\begin{aligned}\text{var}\left([1,1]\hat{\boldsymbol{\beta}}\right) = \sigma^2[1,1](X_1^T X_1)^{-1}\left[\begin{array}{c}1\\1\end{array}\right] \\ &= \frac{\sigma^2}{n_0}(1 - 1 - 1 + n/n_1) \\ &= \frac{\sigma^2}{n_0}(n/n_1 - 1) \\ &= \sigma^2/n_1\end{aligned}$$

where we have made use of the fact that $n - n_1 = n_0$.

We will now also consider x_2. Using both categorical (x_1) and continuous (x_2) covariates often referred to as the *Analysis of Covariance (ANCOVA)*, even if Giles thinks it's all just part of linear regression.

For this, we will write the average value of x_2 among subjects with $x_1 = 0$ to be $\bar{x}_{2,0}$ and among subjects with $x_1 = 1$ to be $\bar{x}_{2,1}$ and write $\tilde{x}_2$ to be x_2 with the group mean subtracted:

$$\tilde{x}_{i2} = \begin{cases} x_{i2} - \bar{x}_{2,0} & \text{if } x_{i1} = 0 \\ x_{i2} - \bar{x}_{2,1} & \text{if } x_{i1} = 1 \end{cases} = (I - H_1)\boldsymbol{x}_2$$

and we will set $X_2 = [\mathbf{1}, \boldsymbol{x}_1, \tilde{\boldsymbol{x}}_2]$.

4. (10 points) Show that $\tilde{\boldsymbol{x}}_2$ can be written as $\boldsymbol{x}_2 - \alpha_1 \mathbf{1} - \alpha_2 \boldsymbol{x}_1$. What are α_1 and α_2? You may find earlier questions useful.

   Setting $\alpha_1 = \bar{x}_{2,0}$ and $\alpha_2 = \bar{x}_{2,1} - \bar{x}_{2,0}$ we have $\tilde{\boldsymbol{x}}_2 = \boldsymbol{x}_2 - \alpha_1 \mathbf{1} - \alpha_2 \boldsymbol{x}_2$.

5. (12 points) Write out $X_2^T X_2$ for this new model. Show that your estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are unchanged from Question 2.

   If we are interested in β_1, was there any point to adding x_2?

   First we observe that $X_1^T \tilde{\boldsymbol{x}}_2 = X_1^T (I - H_1)\boldsymbol{x}_2 = 0$ so

$$X_2^T X_2 = \left[\begin{array}{cc} X_1^T X_1 & \mathbf{0} \\ \mathbf{0} & \tilde{\boldsymbol{x}}_2^T \tilde{\boldsymbol{x}}_2 \end{array} \right]$$

   and

$$(X_2^T X_2)^{-1} = \left[\begin{array}{cc} (X_1^T X_1)^{-1} & \mathbf{0} \\ \mathbf{0} & 1/(\tilde{\boldsymbol{x}}_2^T \tilde{\boldsymbol{x}}_2) \end{array} \right]$$

   from which

$$\left[\begin{array}{c} \hat{\beta}_0 \\ \hat{\beta}_1 \end{array} \right] = (X_1^T X_1)^{-1} X_1^T \boldsymbol{y}$$

   as we obtained without using $\tilde{x}_2$.

   However, in this case, if $\beta_2 \neq 0$ then without accounting for $\tilde{\boldsymbol{x}}_2 \beta_2$, we would absorb this term into the error, inflating our estimate of σ^2 and widening our confidence intervals.

6. (10 points) Give an expression for the variance inflation factor for $\hat{\beta}_2$ in terms of $\tilde{\boldsymbol{x}}_2$ and $\boldsymbol{x}_2$.

$$\begin{aligned} VIF &= \frac{\boldsymbol{x}_2^T C \boldsymbol{x}_2}{\boldsymbol{x}_2 (I - H_1) \boldsymbol{x}_2} \\ &= 1 + \frac{\boldsymbol{x}_1^T H_1 C H_1 \boldsymbol{x}_1}{\tilde{\boldsymbol{x}}_2^T \tilde{\boldsymbol{x}}_2} \\ &= 1 + \frac{\frac{n_1 n_2}{n} (\bar{x}_{2,1} - \bar{x}_{2,0})^2}{\tilde{\boldsymbol{x}}_2^T \tilde{\boldsymbol{x}}_2} \end{aligned}$$

   Although only the first line was needed for a correct solution.

7. (14 points) By writing out the prediction equation $\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 \tilde{x}_2$ in terms of x_2, find $\hat{\beta}_1^*$, the estimate of $\hat{\beta}_1$ in a model where we used $X_2^* = [\mathbf{1}, \boldsymbol{x}_1, \boldsymbol{x}_2]$ instead of X.

   Why has $\hat{\beta}_2$ not changed? What is the variance of $\hat{\beta}_1^*$?

   Using the previous questions

$$\begin{aligned}\hat{\beta}_0 + \hat{\beta}_1 \boldsymbol{x}_1 + \hat{\beta}_2 \tilde{\boldsymbol{x}}_2 &= \hat{\beta}_0 + \hat{\beta}_1 \boldsymbol{x}_1 + \hat{\beta}_2(\boldsymbol{x}_2 - \alpha_1 \mathbf{1} - \alpha_2 \boldsymbol{x}_1) \\ &= (\bar{y}_0 - \hat{\beta}_2 \bar{x}_{2,0}) + (\bar{y}_1 - \bar{y}_0 - \hat{\beta}_2(\bar{x}_{2,1} - \bar{x}_{2,0}))\boldsymbol{x}_1 + \hat{\beta}_2 \boldsymbol{x}_2\end{aligned}$$

   Here the fitted values must uniquely determine the values of $\hat{\boldsymbol{\beta}}$ and we see from this equation that $\hat{\beta}_2$ hasn't changed.

   Here we have

$$\begin{aligned}\mathrm{var}(\hat{\beta}_1^*) &= \mathrm{var}(\hat{\beta}_1) + (\bar{x}_{2,1} - \bar{x}_{2,0})^2 \mathrm{var}(\hat{\beta}_2) \\ &= \frac{n\sigma^2}{n_1 n_0} + \frac{\sigma^2(\bar{x}_{2,1} - \bar{x}_{2,0})^2}{\tilde{\boldsymbol{x}}_2^T \tilde{\boldsymbol{x}}_2} \\ &= \frac{n\sigma^2}{n_1 n_0}\left(1 + \frac{\sigma^2 \frac{n_0 n_1}{n}(\bar{x}_{2,1} - \bar{x}_{2,0})^2}{\tilde{\boldsymbol{x}}_2^T \tilde{\boldsymbol{x}}_2}\right) \qquad = \mathrm{var}(\hat{\beta}_1) VIF\end{aligned}$$

8. (10 points) There is a concern that the slope on x_2 (awareness) might be different between the $x_1 = 1$ group and the $x_1 = 0$ group. For this reason, the researcher considers adding an interaction term to produce a design matrix $X = [\mathbf{1}, \boldsymbol{x}_1, \tilde{\boldsymbol{x}}_2, \boldsymbol{x}_1\tilde{\boldsymbol{x}}_2]$ where the last column is the *element-wise* product of x_1 and $\tilde{x}_2$.

   Define a sum of squares to measure the total contribution of $\tilde{x}_2$ to the model in this case.

   Setting $\boldsymbol{x}_3 = \boldsymbol{x}_1\boldsymbol{x}_2$ then we can compare a model with only $\boldsymbol{x}_1$ (ie, that doesn't use x_2 at all) to one with using $[\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3]$, then the sum of squared changes in fitted values is $\boldsymbol{y}^T(H_3 - H_1)\boldsymbol{y}$.

9 (10 points) In the general regression model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, when describing VIFs, we have described $\sigma^2/(\boldsymbol{x}_1^T C \boldsymbol{x}_1)$ as the "minimum possible variance" that could be achieved for β_1.

To see this, write $X = [\boldsymbol{x}_1, X_{-1}]$ to separate $\boldsymbol{x}_1$ from the other covariates, and assume $\boldsymbol{x}_1$ is centered.

We'll consider $\tilde{X}_{-1} = X_{-1} - \boldsymbol{x}_1\boldsymbol{\alpha}$ where $\boldsymbol{\alpha}$ is a $p-1$-dimensional row vector and use a new design matrix $\tilde{X} = [\boldsymbol{x}_1, \tilde{X}_{-1}]$.

Show that the variance of β_1 is minimized when $\boldsymbol{\alpha}$ is chosen so that $\tilde{X}_{-1}^T\boldsymbol{x}_1 = \mathbf{0}$.

The following formula may be helpful

$$(\tilde{X}^T\tilde{X})^{-1} = \left[\begin{array}{cc} \frac{1}{r} & -\frac{1}{r}(\tilde{X}_{-1}^T\tilde{X}_{-1})^{-1}\tilde{X}_{-1}^T\boldsymbol{x}_1 \\ -\frac{1}{r}\boldsymbol{x}_1^T\tilde{X}_{-1}(\tilde{X}_{-1}^T\tilde{X}_{-1})^{-1} & \left(\tilde{X}_{-1}^T\tilde{X}_{-1} - \frac{\tilde{X}_{-1}^T\boldsymbol{x}_1\boldsymbol{x}_1^T\tilde{X}_{-1}}{\boldsymbol{x}_1^T\boldsymbol{x}_1}\right)^{-1} \end{array}\right]$$

with $r = \boldsymbol{x}_1^T\boldsymbol{x}_1 - \boldsymbol{x}_1^T\tilde{X}_{-1}(\tilde{X}_{-1}^T\tilde{X}_{-1})^{-1}\tilde{X}_{-1}^T\boldsymbol{x}_1$.

Here the variance of $\hat{\beta}_1$ is σ^2/r. We observe that

$$\boldsymbol{x}_1^T\tilde{X}_{-1}(\tilde{X}_{-1}^T\tilde{X}_{-1})^{-1}\tilde{X}_{-1}^T\boldsymbol{x}_1 \geq 0$$

because $(\tilde{X}_{-1}^T\tilde{X}_{-1})^{-1}$ is positive definite. The larger this term, the smaller is r and the larger the variance.

This variance is therefore minimized if $\tilde{X}_{-1}^T\boldsymbol{x}_1 = 0$, in which case

$$\boldsymbol{x}_1^T\tilde{X}_{-1}(\tilde{X}_{-1}^T\tilde{X}_{-1})^{-1}\tilde{X}_{-1}^T\boldsymbol{x}_1 = 0.$$