BTRY/STSCI 4030/5030 - Linear Models with Matrices - Fall 2018

Midterm - Tuesday, October 16

NAME: Giles Hooker

NETID: gjh27

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.

For the first part of the exam, we will use the following table as motivation. This was derived from a data set collected to assess the speed of putting greens using three different types of grass: C1, C2, and C3. They were also grown in two regions, R1 and R2. We treat C1 and R1 as references. The design is balanced, meaning that 8 putting greens were planted for each type of grass, 4 in each region for a total of 24 observations.

In the results below, C2, C3 and R2 are indicator functions for grass types 2 and 3, and region 2 repsectively. We report sequential ANOVA decompositions for a model where the effects are entered in this order, and in the reverse.

```
> mod1 = lm(Speed ~ R2 + C2 + C3,data=data)
> anova(mod1)
          Df  Sum Sq Mean Sq F value    Pr(>F)
R2         1  0.5612  0.5612  2.0998    0.1628
C2         1  0.0200  0.0200  0.0749    0.7872
C3         1 12.9600 12.9600 48.4913 9.271e-07 ***
Residuals 20  5.3453  0.2673

> mod2 = lm(Speed ~ C3 + C2 + R2,data=data)
> anova(mod2)
          Df  Sum Sq Mean Sq F value    Pr(>F)
C3         1 10.1660 10.1660 38.0373 5.025e-06 ***
C2         1  2.8140  2.8140 10.5289  0.004058 **
R2         1  0.5612  0.5612  2.0998  0.162816
Residuals 20  5.3453  0.2673
```

Both of these models also include an intercept.

If it helps, the data are given by

```
> data
   Speed R2 C2 C3
1   7.56  0  0  0
2   8.88  0  1  0
3   8.20  0  0  1
4   7.41  0  0  0
5   8.20  0  1  0
6   9.15  0  0  1
7   7.64  0  0  0
8   7.20  0  1  0
9   9.24  0  0  1
10  6.81  0  0  0
11  7.12  0  1  0
12  8.31  0  0  1
13  6.86  1  0  0
14  8.16  1  1  0
15  9.42  1  0  1
16  6.86  1  0  0
17  8.68  1  1  0
18  9.26  1  0  1
19  7.22  1  0  0
20  8.25  1  1  0
21  8.93  1  0  1
22  7.64  1  0  0
23  8.22  1  1  0
24  9.89  1  0  1
```

The questions below will (mostly) not refer to the specific numbers in this table, but you may use them if it helps you to think about them.

1. We observe that the sum of squares for R2 does not change between the models. This happens for orthogonal covariates. Is R2 orthogonal to any of the other columns in the regression matrix X?

   No, we see that $R2^T\mathbf{1} = 12$, $R2^T C2 = 4$ *and* $R2^T C3 = 4$.

2. To resolve the apparent paradox from the previous question, find a transformation of R2, *using only itself and the intercept*, so that the result is orthogonal to all the other columns of X. (Hint: you will need to use the fact that the design is balanced) This means that the hat matrices are all the same as if you used this new orthogonal version.

   Here we see that $\tilde{R} = -2R2 + \mathbf{1}$ *is +1 for Region 1 and -1 for Region 2. Since the design is balanced, we have that* $\tilde{R}^T\mathbf{1} = 0$, $\tilde{R}^T C2 = 4 - 4 = 0$ *and* $\tilde{R}^T C3 = 4 - 4 = 0$

3. The first ANOVA table isn't quite what the researchers wanted. They would like to measure the total effect of grass type. In terms of sums of squares matrices, that means they are interested in the sum of squares associated with the difference in hat matrices $H_3 - H_1$. Is there any way to obtain this from the ANOVA table above? Justify your answer; carry out the calculation if you are able to.

   The second two terms in the ANOVA are $H_3 - H_2$ *and* $H_2 - H_1$, *so summing them gives* $H_3 - H_2 + H_2 - H_1 = H_3 - H_1$. *In this table 0.02+12.96 = 12.98.*

4. Given the interest in $H_3 - H_1$, what are its degrees of freedom?

   The degrees of freedom are $tr(H_3 - H_1) = tr(H_3) - tr(H_1) = 4 - 2 = 2$.

5. We also notice that the sums of squares for C2 and C3 *do* change. What are they not orthogonal to? Can you transform these so that C3 is orthogonal to everything?

   We notice that although $C2^T C3 = 0$ *we have* $C2^T \mathbf{1} = C3^T \mathbf{1} = 8$*. If we replace C3 with* $\tilde{C}_3 = 2C3 - C2 - \mathbf{1}$*, the new column will be 1 in region C3, 0 in C2 and -1 in C1. Hence* $\tilde{C}_3^T \mathbf{1} = 0$ *and* $\tilde{C}_3^T R2 = 0$ *as well as* $\tilde{C}_3^T C2 = 0$*.*

Here we will turn to examining predictions of a linear regression. That is, having fit a model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, resulting in $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$. X will be treated completely generally and we will make the usual normality assumptions $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I)$.

We now wish to use this model to predict y^*, the response for a new data point: $y^* = \mathbf{z}^T \beta + \epsilon^*$. Here $\mathbf{z}$ is a vector of covariate values for the new data point (treated as a column by default) and ϵ^* is a new error term, independent of anything else.

We don't get to see y^*, of course, so our prediction will use the estimated coefficients: $\hat{y}^* = \mathbf{z}^T \hat{\boldsymbol{\beta}}$.

6. Show that $\hat{y}^*$ has expectation $\mathbf{z}^T \beta$, the true mean of y^*. Give an expression for $\mathrm{var}(\hat{y}^*)$.

   We have that

   $$E\hat{y}^* = E\mathbf{z}^T \hat{\boldsymbol{\beta}} = \mathbf{z}^T E\hat{\boldsymbol{\beta}} = \mathbf{z}^T \boldsymbol{\beta}$$

   and

   $$\mathrm{var}(\hat{y}^*) = \mathbf{z}^T \mathrm{var}\hat{\boldsymbol{\beta}}\mathbf{z} = \sigma^2 \mathbf{z}^T (X^T X)^{-1} \mathbf{z}.$$

7. By subtracting it's expectation, dividing by its standard deviation, and substituting MSE for σ^2, obtain a t-statistic for $\hat{y}^*$. Hence provide a confidence interval for $\mathbf{z}^T \boldsymbol{\beta}$.

   Observing that $\hat{\beta} \sim N(0, \sigma^2 (X^T X)^{-1})$ *and hence* $\hat{y}^* \sim N(\mathbf{z}\boldsymbol{\beta}, \sigma^2 \mathbf{z}^T (X^T X)^{-1} \mathbf{z})$. *Therefore*

   $$\frac{\hat{y}^* - \mathbf{z}^T \boldsymbol{\beta}}{\sqrt{\sigma^2 \mathbf{z}^T (X^T X)^{-1} \mathbf{z}}} \sim N(0,1)$$

   so that setting $\hat{\sigma}^2 = MSE = SSE/(n-p-1)$

   $$t = \frac{\hat{y}^* - \mathbf{z}^T \boldsymbol{\beta}}{\sqrt{\hat{\sigma}^2 \mathbf{z}^T (X^T X)^{-1} \mathbf{z}}} = \frac{\hat{y}^* - \mathbf{z}^T \boldsymbol{\beta}}{\sqrt{\sigma^2 \mathbf{z}^T (X^T X)^{-1} \mathbf{z}} \sqrt{\hat{\sigma}^2 / \sigma^2}} \sim t_{n-p-1}$$

because $(n-p-1)MSE/\sigma^2 \sim \chi^2_{n-p-1}$.

This results in a confidence interval of

$$(\mathbf{z}^T\hat{\boldsymbol{\beta}} - t_{n-p-1}^{1-\alpha/2}\sqrt{\hat{\sigma}^2\mathbf{z}^T(X^TX)^{-1}\mathbf{z}},\ \mathbf{z}^T\hat{\boldsymbol{\beta}} + t_{n-p-1}^{1-\alpha/2}\sqrt{\hat{\sigma}^2\mathbf{z}^T(X^TX)^{-1}\mathbf{z}}).$$

8. What is the joint covariance (2×2 matrix) of y^* and $\mathbf{z}^T\hat{\boldsymbol{\beta}}$?

   Since ϵ^* *is independent* $\hat{\beta}$ *their covariance is zero and their covariance matrix is*

$$\begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2\mathbf{z}^T(X^TX)^{-1}\mathbf{z} \end{pmatrix}$$

9. Obtain the mean and variance of the error $y^* - \mathbf{z}^T\hat{\boldsymbol{\beta}}$. (Note that this is only a theoretical quantity until you have observed y^*.)

   We have

$$E(y^* - \mathbf{z}^T\hat{\boldsymbol{\beta}}) = \mathbf{z}^T\boldsymbol{\beta} - \mathbf{z}^T\boldsymbol{\beta} = 0$$

   and

$$\begin{aligned} \mathrm{var}(y^* - \mathbf{z}^T\hat{\boldsymbol{\beta}}) &= \mathrm{var}(y^*) + \mathrm{var}(\mathbf{z}^T\hat{\boldsymbol{\beta}}) \\ &= \sigma^2 + \sigma^2\mathbf{z}^T(X^TX)^{-1}\mathbf{z} \\ &= \sigma^2\left(1 + \mathbf{z}^T(X^TX)^{-1}\mathbf{z}\right) \end{aligned}$$

Here we will turn this into a means of measuring test set error. We will consider a scenario where we obtain a second *vector* of measurements at *exactly the same covariates* as the first. That is $\mathbf{y}^* = X\boldsymbol{\beta} + \boldsymbol{\epsilon}^*$. (Yes this is a little contrived, but it will make the math easier).

Our fitted values will still be obtained using only the first vector of measurements $\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}}$ so that these are independent of $\mathbf{y}^*$.

10. What is the joint $(2n \times 2n)$ covariance of $\mathbf{y}^*$ and $X\hat{\boldsymbol{\beta}}$?

    Observing that $X\hat{\boldsymbol{\beta}} \sim N(X\boldsymbol{\beta}, \sigma^2 X(X^T X)^{-1} X^T) = N(X\boldsymbol{\beta}, \sigma^2 H)$ *we have that* $\mathbf{y}^*$ *is independent of* $\hat{\boldsymbol{\beta}}$ *hence the covariance we want is*

    $$\begin{pmatrix} \sigma^2 I & 0 \\ 0 & \sigma^2 H \end{pmatrix}$$

11. Hence give the expected value of the *test set* error $\sum (y_i^* - \mathbf{x}_i \hat{\boldsymbol{\beta}})^2$ in terms of σ^2 and a degrees of freedom.

    We observe that $E(y_i^* - \mathbf{x}_i \hat{\boldsymbol{\beta}}) = 0$ *and* $var(y_i^* - \mathbf{x}_i \hat{\boldsymbol{\beta}}) = \sigma^2(1 + h_{ii})$ *so that*

    $$\begin{aligned} E \sum (y_i^* - \mathbf{x}_i \hat{\boldsymbol{\beta}})^2 &= \sum \sigma^2 (1 + h_{ii}) \\ &= \sigma^2 (n + \mathrm{tr}(H)) \\ &= \sigma^2 (n + p + 1) \end{aligned}$$

    this can also be written as $\sigma^2 tr(I + H)$ *with degrees of freedom* $n + p + 1 = tr(I + H)$.

12. If given only the test set errors, how would you estimate σ^2? Can this be distributed as a χ^2?

    An unbiased estimate would be

    $$\hat{\sigma}^2 = \frac{1}{\mathrm{tr}(I + H)} \sum (y_i^* - \mathbf{x}_i \hat{\boldsymbol{\beta}})^2$$

We can write this as

$$\mathbf{y}^* - X\hat{\boldsymbol{\beta}} = \boldsymbol{y}^* - X\boldsymbol{\beta} + X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) = \boldsymbol{\epsilon}^* - H\boldsymbol{\epsilon}$$

which we can turn into a sum of square errors

$$\sum (y_i^* - \mathbf{x}_i\hat{\boldsymbol{\beta}})^2 = \left[\boldsymbol{\epsilon}^{*T},\ \boldsymbol{\epsilon}^T\right] \begin{bmatrix} I & -H \\ -H & H \end{bmatrix} \begin{bmatrix} \boldsymbol{\epsilon}^* \\ \boldsymbol{\epsilon} \end{bmatrix}$$

where we observe the matrix in the middle is not idempotent hence this sum of square is not necessarely χ^2*.*

Bonus Using your results in Part 9, write down a *prediction interval* around $\mathbf{z}^T\hat{\boldsymbol{\beta}}$ that captures y^* with 95% probability.

Using the variance found in Part 9, we can produce

$$(\mathbf{z}^T\hat{\boldsymbol{\beta}} - t_{n-p-1}^{1-\alpha/2}\sqrt{\hat{\sigma}^2(1 + \mathbf{z}^T(X^TX)^{-1}\mathbf{z})},\ \mathbf{z}^T\hat{\boldsymbol{\beta}} + t_{n-p-1}^{1-\alpha/2}\sqrt{\hat{\sigma}^2(1 + \mathbf{z}^T(X^TX)^{-1}\mathbf{z})})$$

Bonus Part 7 and the bonus above both suggest we could produce a confidence interval at *every possible* value of $\mathbf{z}$. For one covariate, that would mean putting a band around the linear regression line. Why might this be problematic?

This would not account for multiple testing. That is, the interval would be valid at each individual $\mathbf{z}$*, but not if we wanted the band to contain all values of the regression at once.*