STOR 664: FALL 2020
Midterm Exam, September 30, 2020

This is an open-book, remote-learning exam. Time limit: 2 hours. Access to course materials and standard computational tools (in particular, R) is allowed; communication with other students or with anybody via the internet, other than the instructor, is not. The university Honor Code is in effect at all times. Answers may either be typed using Word or Latex, or handwritten and scanned or photographed; if handwritten, it is recommended you use blue or black ink on plain sheets of white paper. They should then be uploaded in sakai. The exam is worth 100 points total (35 for question 1, 65 for question 2); points for each part-question are stated below. Although the questions are intended to be answered in sequence, you may write out your answers in any order and errors in one part-question will not prevent you gaining full credit in other parts of the same question. Attempt all questions.

1. Three objects are to be weighed in a scale, whose true weights are written $\beta_1,\ \beta_2,\ \beta_3$.

   (a) Consider the following weighing scheme.

   i. Each of objects 1, 2, 3 is placed on its own in the scale. The observed weights are $Y_1,\ Y_2,\ Y_3$.

   ii. Each pair of objects (1 and 2, then 1 and 3, then 2 and 3) is placed in the scale together, resulting in observed weights $Y_4,\ Y_5,\ Y_6$ respectively.

   iii. The three objects are weighed together, with observed total weight Y_7.

   Assume each of the weighings has a random error of mean 0 and variance σ^2, and that all the random errors are uncorrelated.

   Without using the computer, find formulas for the least squares estimates $\hat{\beta}_1,\ \hat{\beta}_2,\ \hat{\beta}_3$, as linear combinations of $Y_1, ..., Y_7$. You should state explicitly the coefficients of $Y_1, ..., Y_7$ in these estimators, and show that the common variance of the three estimators is $\frac{3\sigma^2}{8}$. **[18 points]**

   (b) Consider an alternative weighing scheme where object 1 is weighed n_1 times, object 2 is weighed n_2 times, object 3 is weighed n_3 times, where $n_1 + n_2 + n_3 = 7$. Show that, however $n_1,\ n_2,\ n_3$ are chosen, the mean variance of $\hat{\beta}_1,\ \hat{\beta}_2,\ \hat{\beta}_3$ is greater than in (a). **[7 points]**

   (c) Now suppose that the main quantity of interest is not any of $\beta_1,\ \beta_2,\ \beta_3$, but $2\beta_3 - \beta_1 - \beta_2$. (This might be of interest, for instance, if we were substituting objects of types 1 and 2 in a system with objects of type 3, or the other way round, and we wanted to know whether this would result in an increase or a decrease in the total weight of the system.) Find the variance of the estimator $2\hat{\beta}_3 - \hat{\beta}_1 - \hat{\beta}_2$ from part (a), and show that there is at least one estimator derived from the scheme of part (b) that would have smaller variance. **[10 points]**

2. Consider the model

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_i + \epsilon_i, \\ y_{n+i} &= \beta_1 + \beta_3 x_i + \epsilon_{n+i}, \end{aligned}$$

for $i = 1, \ldots, n$. Here, $x_i,\ i = 1, .., n$ is a covariate with $\sum_{i=1}^n x_i = 0$ and $\epsilon_i,\ i = 1, ..., 2n$ are independent $N(0, \sigma^2)$ random variables. Our ultimate interest is in testing the null hypothesis $H_0 :\ \beta_2 = \beta_3$ against the alternative $H_1 :\ \beta_2 \neq \beta_3$.

(a) Assuming H_1, find explicit formulas for the least squares estimators $\hat{\beta}_1,\ \hat{\beta}_2,\ \hat{\beta}_3$ in terms of $y_1, ..., y_{2n}$ and $x_1, ..., x_n$, and calculate their variances. **[10 points]**

(b) Recall that there is a general formula

$$SST \quad = \quad SSR_1 + SSE_1$$

that expresses the total sum of squares ($SST = \sum (y_i - \bar{y})^2$) as the sum of squares due to regression and the sum of squared residuals, which in this case is given by

$$SSE_1 \quad = \quad \sum_{i=1}^{n} \left\{ (y_i - \bar{y} - \hat{\beta}_2 x_i)^2 + (y_{n+i} - \bar{y} - \hat{\beta}_3 x_i)^2 \right\}.$$

The subscript 1 here is to denote that these calculations are made under H_1.
Under the above assumptions, find an explicit formula for SSR_1 in terms of $\hat{\beta}_2,\ \hat{\beta}_3$ and $x_1, ..., x_n$. **[10 points]**

(c) Repeat the calculations of parts (a) and (b) under H_0. In particular, assuming $\beta_3 = \beta_2$, find least squares estimators for β_1 and β_2, their variances, and formulas for SSR_0 and SSE_0 in this case. To distinguish them from the estimators in (a) and (b), write $\tilde{\beta}_1$ and $\tilde{\beta}_2$ for these estimators. **[8 points]**

(d) What are the relationships between $\tilde{\beta}_1$ and $\tilde{\beta}_2$ from part (c) and $\hat{\beta}_1,\ \hat{\beta}_2,\ \hat{\beta}_3$ in (a)? **[4 points]**

(e) Prove the formula

$$SSE_0 - SSE_1 \quad = \quad SSR_1 - SSR_0 \ = \ \frac{(\hat{\beta}_2 - \hat{\beta}_3)^2}{2} \sum_{1}^{n} x_i^2.$$

[*Note:* If you didn't succeed in proving this formula, nevertheless you should assume it is correct for the remaining parts of the question.] **[6 points]**

(f) Show how to formally test H_0 against H_1 at significance level 0.05. Specifically, you should define a relevant test statistic (which may be expressed in terms of $SSE_0,\ SSE_1$, or any quantities developed in previous parts of the question), and explain how to define the rejection region so that the test has the desired significance level. **[8 points]**

(g) Explain how you would calculate the power of this test for given values of $\beta_2,\ \beta_3,\ \sigma$ and $x_1, ..., x_n$. To illustrate your answer, calculate the power of the test when $\left| \frac{\beta_2 - \beta_3}{\sigma} \right| = \frac{1}{2}$, $n = 10$ and $\sum_1^n x_i^2$ is any of (i) 20, (ii) 40, (iii) 60, (iv) 80. **[11 points]**

(h) In practice, the main "design of the experiment" issue may well be the value of $\sum x_i^2$, which the experimenter may be able to adjust by choosing suitable values of the x_i's. Assuming the other parameters are as in part (g), what value of $\sum x_i^2$ would be needed to achieve power 0.8? **[8 points]**

[*Note:* The last two parts are the only places on the exam where you are expected to use the computer; if you use R, you should indicate clearly which functions are being used, and how.]

Solutions

1. (a) Writing $Y_1 = \beta_1 + \epsilon_1$, etc., the matrix formulation of the model is

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_7 \end{pmatrix} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},\ X = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix},\ X^TX = \begin{pmatrix} 4 & 2 & 2 \\ 2 & 4 & 2 \\ 2 & 2 & 4 \end{pmatrix}.$$

Here, $X^TX = 2I_3 + 2J_3$, and using the general formula for $(aI_n + bJ_n)^{-1} = \frac{1}{a}I_n - \frac{b}{a(a+nb)}J_n$, we deduce $(X^TX)^{-1} = \frac{1}{2}I_3 - \frac{1}{8}J_3 = \frac{1}{8}\begin{pmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{pmatrix}$. Hence

$$\begin{aligned} \begin{pmatrix} \hat\beta_1 \\ \hat\beta_2 \\ \hat\beta_3 \end{pmatrix} &= \frac{1}{8}\begin{pmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{pmatrix}\begin{pmatrix} Y_1+Y_4+Y_5+Y_7 \\ Y_2+Y_4+Y_6+Y_7 \\ Y_3+Y_5+Y_6+Y_7 \end{pmatrix} \\ &= \frac{1}{8}\begin{pmatrix} 3Y_1 - Y_2 - Y_3 + 2Y_4 + 2Y_5 - 2Y_6 + Y_7 \\ -Y_1 + 3Y_2 - Y_3 + 2Y_4 - 2Y_5 + 2Y_6 + Y_7 \\ -Y_1 - Y_2 + 3Y_3 - Y_4 + 2Y_5 + 2Y_6 + Y_7 \end{pmatrix}. \end{aligned}$$

So the coefficients of $\hat\beta_1$ are $\frac{3}{8},\ -\frac{1}{8},\ -\frac{1}{8},\ \frac{1}{4},\ \frac{1}{4},\ -\frac{1}{4},\ \frac{1}{8}$, and similarly for $\hat\beta_2$ and $\hat\beta_3$. Since the diagonal entries of $(X^TX)^{-1}$ are all $\frac{3}{8}$, it follows that the common variance of the three estimators is $\frac{3}{8}\sigma^2$.

(b) The most even allocation would be to make one of $n_1,\ n_2,\ n_3$ equal to 3 and the other two equal to 2, but then the average variance of the three estimated weights is $\frac{1}{3}\left(\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} + \frac{\sigma^2}{n_3}\right) = \frac{4\sigma^2}{9} > \frac{3\sigma^2}{8}$.

(c) The question is equivalent to asking the variance of $\mathbf{c}^T\hat{\boldsymbol{\beta}}$ where $\mathbf{c} = \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix}$ and the general formula for that is $\mathbf{c}^T(X^TX)^{-1}\mathbf{c}\sigma^2$. However, you can calculate directly that

$$\frac{1}{8}\begin{pmatrix} -1 & -1 & 2 \end{pmatrix}\begin{pmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{pmatrix}\begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix} = 3,$$

so the variance of the estimator following the scheme in (a) is $3\sigma^2$. However, if you followed scheme (b) with $n_1 = 2,\ n_2 = 2,\ n_3 = 3$, the variance of $-\hat\beta_1 - \hat\beta_2 + 2\hat\beta_3$ is $\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2} + \frac{4}{n_2}\right) = \frac{7}{3}\sigma^2$ which is better.

2. (a) We calculate

$$X = \begin{pmatrix} 1 & x_1 & 0 \\ 1 & x_2 & 0 \\ 1 & \vdots & 0 \\ 1 & x_n & 0 \\ 1 & 0 & x_1 \\ 1 & 0 & x_2 \\ 1 & 0 & \vdots \\ 1 & 0 & x_n \end{pmatrix},\ X^TX = \begin{pmatrix} 2n & 0 & 0 \\ 0 & \sum x_i^2 & 0 \\ 0 & 0 & \sum x_i^2 \end{pmatrix},\ (X^TX)^{-1} = \begin{pmatrix} \frac{1}{2n} & 0 & 0 \\ 0 & \frac{1}{\sum x_i^2} & 0 \\ 0 & 0 & \frac{1}{\sum x_i^2} \end{pmatrix},$$

Also $X^TY = \begin{pmatrix} \sum_1^{2n} y_i, \\ \sum_1^n x_i y_i, \\ \sum_1^n x_i y_{n+i}, \end{pmatrix}$, so $\hat\beta_1 = \frac{1}{2n}\sum_1^{2n} y_i = \bar y,\ \hat\beta_2 = \sum_1^n x_i y_i / \sum_1^n x_i^2,\ \hat\beta_3 = \sum_1^n x_i y_{n+i} / \sum_1^n x_i^2$, with variances $\sigma^2/(2n),\ \sigma^2/\sum_1^n x_i^2,\ \sigma^2/\sum_1^n x_i^2$.

(b) $\sum_{i=1}^{2n}(y_i - \bar y)^2$ may be written as

$$\sum_{i=1}^{n}\left\{(y_i - \bar y - \hat\beta_2 x_i + \hat\beta_2 x_i)^2 + (y_{n+i} - \bar y - \hat\beta_3 x_i + \hat\beta_3 x_i)^2\right\} = \left\{(y_i - \bar y - \hat\beta_2 x_i)^2 + \hat\beta_2^2 x_i^2 + (y_{n+i} - \bar y - \hat\beta_3 x_i)^2 + \hat\beta_3^2 x_i^2\right\}$$

where for the first term we have used

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \bar y - \hat\beta_2 x_i + \hat\beta_2 x_i)^2 &= \sum_{i=1}^{n}(y_i - \bar y - \hat\beta_2 x_i)^2 + \hat\beta_2^2\sum_{i=1}^{n} x_i^2 + 2\hat\beta_2\sum_{i=1}^{n}(y_i - \bar y - \hat\beta_2 x_i)x_i \\ &= \sum_{i=1}^{n}(y_i - \bar y - \hat\beta_2 x_i)^2 + \hat\beta_2^2\sum_{i=1}^{n} x_i^2 + 2\hat\beta_2\left\{\hat\beta_2\sum_{i=1}^{n} x_i^2 - \hat\beta_2\sum_{i=1}^{n} x_i^2\right\} \end{aligned}$$

and similarly for the second term.

Hence if $SST = \sum_1^{2n}(y_i - \bar y)^2,\ SSE_1 = \sum_1^n\left\{(y_i - \bar y - \hat\beta_2 x_i)^2 + (y_{n+i} - \bar y - \hat\beta_3 x_i)^2\right\}$, we deduce

$$SST = SSE_1 + (\hat\beta_2^2 + \hat\beta_3^2)\sum_1^n x_i^2.$$

Therefore, the explicit formula for SSR_1 is $(\hat\beta_2^2 + \hat\beta_3^2)\sum_1^n x_i^2$.

Alternatively: Just use the fact that $SSR = \sum_{i=1}^{2n}(\hat y_i - \bar y)^2$ where $\hat y_i - \bar y = \hat\beta_2 x_i$ and $\hat y_{n+i} - \bar y = \hat\beta_3 x_i$ for $1 \le i \le n$.

(c) In this case, $X^TX = \begin{pmatrix} 2n & 0 \\ 0 & 2\sum_1^n x_i^2 \end{pmatrix},\ X^TY = \begin{pmatrix} \sum_1^{2n} y_i \\ \sum_1^n (y_i + y_{n+i})x_i \end{pmatrix}$, so $\tilde\beta_1 = \bar y,\ \tilde\beta_2 = \left(\sum_1^n (y_i + y_{n+i})x_i\right) / \left(2\sum_1^n x_i^2\right)$ and their respective variances are $\frac{\sigma^2}{2n},\ \frac{\sigma^2}{2\sum_1^n x_i^2}$.

For SSR_0 and SSE_0, we use the same decomposition of SST as in part (b), but with $\tilde\beta_2$ in place of both $\hat\beta_2$ and $\hat\beta_3$. Therefore,

$$\begin{aligned} SSE_0 &= \sum_{i=1}^{n}\left\{(y_i - \bar y - \tilde\beta_2 x_i)^2 + (y_{n+i} - \bar y - \tilde\beta_2 x_i)^2\right\}, \\ SSR_0 &= 2\tilde\beta_2^2\sum_1^n x_i^2. \end{aligned}$$

(d) By direct comparison of the algebraic formulas, we deduce $\tilde{\beta}_1 = \hat{\beta}_1$, $\tilde{\beta}_2 = (\hat{\beta}_2 + \hat{\beta}_3)/2$.

(e) We have

$$\begin{aligned} SSE_0 - SSE_1 &= SSR_1 - SSR_0 \\ &= (\hat{\beta}_2^2 + \hat{\beta}_3^2)\sum_1^n x_i^2 - 2\left(\frac{\hat{\beta}_2 + \hat{\beta}_3}{2}\right)^2 \sum_1^n x_i^2 \\ &= \frac{(\hat{\beta}_2 - \hat{\beta}_3)^2}{2}\sum_1^n x_i^2. \end{aligned}$$

(f) The test statistic is

$$T = \frac{SSE_0 - SSE_1}{1} \Big/ \frac{SSE_1}{2n-3}.$$

The degrees of freedom are 1 in the numerator because there is only one free parameter (β_3) being tested, and $2n-3$ in the denominator because there are $2n$ observations minus 3 estimated parameters under H_1. When H_0 is true, the distribution of T is $F_{1,2n-3}$, therefore, the test rejects H_0 when $T > c$, where $c = F_{1,2n-3;0.95}$ (the 0.95 quantile of the $F_{1,2n-3}$ distribution).

(g) The distribution of T under H_1 is $F'_{1,2n-3;\lambda}$, where the noncentrality parameter λ is defined by $\lambda\sigma^2 = \frac{(\beta_2-\beta_3)^2}{2}\sum_1^n x_i^2$ (substitution rule). Therefore, the power of the test is $\Pr\{T > c\}$ when $T \sim F'_{1,2n-3;\lambda}$ and c is the critical value found in part (f).

When $\left|\frac{\beta_2-\beta_3}{\sigma}\right| = \frac{1}{2}$, this formula reduces to $\lambda = \sum_1^n x_i^2/8$. Under the four given values for $\sum_1^n x_i^2$, we therefore want to evaluate the power for $\lambda = 2.5,\ 5,\ 7.5,\ 10$. Based on the $F_{1,17}$ and $F'_{1,17;\lambda}$ distributions, we deduce $c = 4.451$ and the four values for power are 0.32, 0.56, 0.73, 0.85 (sample R code: `c=qf(0.95,1,17), pow=1-pf(c,1,17,ncp=2.5)` yields the value 0.3202709).

(h) The problem is to find λ so that $\Pr\{F'_{1,17;\lambda} > c\} = 0.8$, where c is the critical value found in part (g). By trial and error, find $\lambda = 8.838$ (R code: `1-pf(c,1,17,ncp=8.838)` yields the answer 0.7999904). Since $\sum_1^n x_i^2 = 8\lambda$, this means we need $\sum_1^n x_i^2$ to be about 70.7.

Alternate solution to (g) and (h). The alternative way to calculate λ is with formula (3.42) of the course notes. The null hypothesis is written in the form $C\beta = h$ if we define $C = \begin{pmatrix} 0 & 1 & -1 \end{pmatrix}$, $h = 0$. In this case $C\beta = \beta_2 - \beta_3$. The alternative then has $h' = \pm\sigma/2$,

$$C(X^TX)^{-1}C^T = \begin{pmatrix} 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \frac{1}{2n} & 0 & 0 \\ 0 & \frac{1}{\sum_1^n x_i^2} & 0 \\ 0 & 0 & \frac{1}{\sum_1^n x_i^2} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} = \frac{2}{\sum_1^n x_i^2},$$

$$\sigma^2\lambda = (h-h')\{C(X^TX)^{-1}C^T\}^{-1}(h-h') = \frac{\sigma}{2}\cdot\frac{\sum_1^n x_i^2}{2}\cdot\frac{\sigma}{2} = \sigma^2\frac{\sum_1^n x_i^2}{8}$$

so $\lambda = \sum_1^n x_i^2/8$ and hence as in the first solution.