COMPREHENSIVE WRITTEN EXAMINATION, PAPER III PART 1: FRIDAY AUGUST 12, 2022 9:00 A.M.–11:00 A.M. STOR 664 Theory Question (50 points)

Consider a regression with two predictors $x_{i1},\ x_{i2},\ i=1,\ldots,n$, and assume the model

$$y_i \;=\; \beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i1}x_{i2}+\epsilon_i,\ i=1,\ldots,n \tag{1}$$

where $\beta_0,\ldots,\beta_3$ are unknown parameters and $\epsilon_i \sim N[0,\sigma^2]$ are independent errors with common unknown variance σ^2. Note that the model has *just* the interaction term $\beta_3 x_{i1}x_{i2}$ but no terms in x_{i1}^2 or x_{i2}^2. It is natural to want to test whether $H_0:\ \beta_3=0$.

Defining $S_{jk}=\sum_{i=1}^n x_{i1}^j x_{i2}^k$ let us further assume: $S_{10}=S_{01}=S_{11}=S_{12}=0$ but that $S_{20},\ S_{02},\ S_{21}$ and S_{22} are not 0.

(a) Find explicit expressions for the least squares estimators $\hat\beta_0,\ldots,\hat\beta_3$ in terms of the S_{jk}'s and $\sum y_i,\ \sum y_i x_{i1},\ \sum y_i x_{i2}$ and $\sum y_i x_{i1}x_{i2}$. **[12 points]**

(b) Find expressions for the standard errors of $\hat\beta_0,\ldots,\hat\beta_3$, in terms of $n,\ S_{20},\ S_{02},\ S_{21},\ S_{22}$ and the residual standard deviation s (assuming that s^2 is the standard unbiased estimator of σ^2). **[5 points]**

(c) A t-test of significance level α will reject H_0 if $|\frac{\hat\beta_3}{s}|>C$ for some C which is a combination of $n,\ S_{20},\ S_{02},\ S_{21},\ S_{22}$ and α (alpha). Find C. (You may, if you wish, express your answer as an appropriate R function.) **[5 points]**

(d) What is the power of the test in (c) when $\beta_3\neq 0$? You answer should be expressed in terms of the given parameters and relevant percentage points of the noncentral t or F distributions. (You may choose to express your answer as an R function though alternatives are also acceptable if the derivation behind your answer is clearly explained.) (*Hint:* First find the distribution of $\frac{\hat\beta_3^2}{s^2}$ when $\beta_3\neq 0$.) **[16 points]**

(e) Show that the observation with highest leverage is the index i that maximizes

$$x_{i2}^2(S_{20}S_{22}-S_{21}^2)+x_{i1}^2 S_{02}(S_{22}-2S_{21}x_{i2}+x_{i2}^2 S_{20}).$$

[12 points]

SOLUTIONS

(a) Writing the model as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, we have

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{11}x_{12} \\ 1 & x_{21} & x_{22} & x_{21}x_{22} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n1}x_{n2} \end{pmatrix}, X^TX = \begin{pmatrix} n & 0 & 0 & 0 \\ 0 & S_{20} & 0 & S_{21} \\ 0 & 0 & S_{02} & 0 \\ 0 & S_{21} & 0 & S_{22} \end{pmatrix}, (X^TX)^{-1} = \begin{pmatrix} \frac{1}{n} & 0 & 0 & 0 \\ 0 & \frac{S_{22}}{\Delta} & 0 & -\frac{S_{21}}{\Delta} \\ 0 & 0 & \frac{1}{S_{02}} & 0 \\ 0 & -\frac{S_{21}}{\Delta} & 0 & \frac{S_{20}}{\Delta} \end{pmatrix},$$

where $\Delta = S_{20}S_{22} - S_{21}^2$. Hence

$$\hat\beta_0 = \frac{\sum y_i}{n},\ \hat\beta_1 = \frac{S_{22}\sum y_i x_{i1} - S_{21}\sum y_i x_{i1}x_{i2}}{\Delta}, \hat\beta_2 = \frac{\sum y_i x_{i2}}{S_{02}},\ \hat\beta_3 = \frac{S_{20}\sum y_i x_{i1}x_{i2} - S_{21}\sum y_i x_{i1}}{\Delta}.$$

(b) By the result of (a), the variances of $\hat\beta_j,\ j = 0,1,2,3$ are σ^2 times the diagonal entries of X^TX^{-1}. Therefore, substituting s for σ, the standard errors are $\frac{s}{\sqrt{n}}$, $s\sqrt{\frac{S_{22}}{\Delta}}$, $\frac{s}{\sqrt{S_{02}}}$ and $s\sqrt{\frac{S_{20}}{\Delta}}$.

(c) Using R notation, the two-sided test will reject H_0 at significance level α (alpha) if

$$\left|\frac{\hat\beta_3}{s}\right| \ > \ \sqrt{\frac{S_{20}}{\Delta}}\texttt{qt}(\texttt{1} - \texttt{alpha/2}, \texttt{n} - \texttt{4}) \ = \ C.$$

(d) We have that $\frac{\hat\beta_3}{\sigma}\sqrt{\frac{\Delta}{S_{20}}}$ is distributed $\mathcal{N}[\frac{\beta_3}{\sigma}\sqrt{\frac{\Delta}{S_{20}}}, 1]$ where $\mathcal{N}$ is the normal distribution. Hence $\frac{\hat\beta_3^2\Delta}{\sigma^2 S_{20}}$ has the distribution $\chi'^2_{1,\delta}$ where $\delta = \frac{\beta_3}{\sigma}\sqrt{\frac{\Delta}{S_{20}}}$. Defining $T_1 = \frac{\hat\beta_3^2\Delta}{\sigma^2 S_{20}}$, $T_2 = \frac{(n-4)s^2}{\sigma^2}$ with respective distributions $\chi'^2_{1,\delta}$ and χ^2_{n-4}, we have that $\frac{T_1(n-4)}{T_2} = \frac{\hat\beta_3^2\Delta}{s^2 S_{20}}$ has the distribution $F'_{1,n-4,\delta}$ and hence $\frac{\hat\beta_3}{s}\sqrt{\frac{\Delta}{S_{20}}}$ has the distribution $t'_{n-4,\delta}$. The notation here follows p. 134 of the course text; however, in class we also defined $\lambda = \delta^2$ and in this case λ is called the noncentrality parameter. The power of the test is $\Pr\left\{\frac{\hat\beta_3}{s} > C\right\} + \Pr\left\{\frac{\hat\beta_3}{s} < -C\right\}$ which can also be written $\Pr\left\{\frac{\hat\beta_3}{s}\sqrt{\frac{\Delta}{S_{20}}} > x\right\} + \Pr\left\{\frac{\hat\beta_3}{s}\sqrt{\frac{\Delta}{S_{20}}} < -x\right\}$ where $x = \frac{C}{s}\sqrt{\frac{\Delta}{S_{20}}} =$ $\texttt{qt}(\texttt{1} - \texttt{alpha/2}, \texttt{n} - \texttt{4})$. Hence the final answer, in R notation and writing `lambda` in place of λ, is `1-pt(x,4,lambda)+pt(-x,4,lambda)` or equivalently `1-pf(x*x,1,4,lambda)`. (Minor discrepancies in notation will not be penalized so long as the basic method is correct.)

(e) The ith diagonal element of $H = X(X^TX)^{-1}X^T$ is

$$\begin{aligned} h_i \ &= \ \begin{pmatrix} 1 & x_{i1} & x_{i2} & x_{i1}x_{i2} \end{pmatrix} \begin{pmatrix} \frac{1}{n} & 0 & 0 & 0 \\ 0 & \frac{S_{22}}{\Delta} & 0 & -\frac{S_{21}}{\Delta} \\ 0 & 0 & \frac{1}{S_{02}} & 0 \\ 0 & -\frac{S_{21}}{\Delta} & 0 & \frac{S_{20}}{\Delta} \end{pmatrix} \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ x_{i1}x_{i2} \end{pmatrix} \\ &= \ \begin{pmatrix} 1 & x_{i1} & x_{i2} & x_{i1}x_{i2} \end{pmatrix} \begin{pmatrix} \frac{1}{n} \\ \frac{S_{22}x_{i1} - S_{21}x_{i1}x_{i2}}{\Delta} \\ \frac{x_{i2}}{S_{02}} \\ \frac{S_{20}x_{i1}x_{i2} - S_{21}x_{i1}}{\Delta} \end{pmatrix} \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{n} + \frac{x_{i1}^2(S_{22} - S_{21}x_{i2})}{\Delta} + \frac{x_{i2}^2}{S_{02}} + \frac{x_{i1}^2 x_{i2}(S_{20}x_{i2} - S_{21})}{\Delta} \\
&= \frac{1}{n} + \frac{x_{i2}^2}{S_{02}} + \frac{x_{i1}^2 S_{22} - 2S_{21}x_{i1}^2 x_{i2} + x_{i1}^2 x_{i2}^2 S_{20}}{S_{20}S_{22} - S_{21}^2} \\
&= \frac{1}{n} + \frac{x_{i2}^2(S_{20}S_{22} - S_{21}^2) + x_{i1}^2 S_{02}(S_{22} - 2S_{21}x_{i2} + x_{i2}^2 S_{20})}{S_{02}(S_{20}S_{22} - S_{21}^2)}
\end{aligned}$$

Since the values of x_{i1} and x_{i2} enter into only the numerator of the second term of this expression, the term of highest leverage is the one that maximizes that.

Note added after the exam

Some students tried to evaluate δ or λ in part (d) by following the substitution rule which was talked about in class. You *can* do it this way, but that method is a lot more complicated than the above (another way is to use formula (3.42) of the course text, which comes out fairly easily in this case). To use the substitution rule, you first have to write

$$\begin{aligned}
&\sum (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i1}x_{i2})^2 \\
= &\sum (y_i - \beta_0)^2 + \beta_1^2 S_{20} + \beta_2^2 S_{02} + \beta_3^2 S_{22} - 2\beta_1 \sum y_i x_{i1} - 2\beta_2 \sum y_i x_{i2} - 2\beta_3 \sum y_i x_{i1}x_{i2} + 2\beta_1\beta_3 S_{21}.
\end{aligned}$$

Assume that the parameter estimates under H_1 are $\hat\beta_0, \hat\beta_1, \hat\beta_2, \hat\beta_3$ (as previously derived) and that the corresponding estimates under $H_0 : \ \beta_3 = 0$ are $\tilde\beta_0, \tilde\beta_1, \tilde\beta_2, \tilde\beta_3$. You can quickly check that $\tilde\beta_0 = \hat\beta_0 = \bar y, \ \tilde\beta_1 = \frac{\sum y_i x_{i1}}{S_{20}}, \ \tilde\beta_2 = \hat\beta_2, \ \tilde\beta_3 = 0$. So

$$\begin{aligned}
&SSE_0 - SSE_1 \\
= &\sum (y_i - \tilde\beta_0)^2 + \tilde\beta_1^2 S_{20} + \tilde\beta_2^2 S_{02} + \tilde\beta_3^2 S_{22} - 2\tilde\beta_1 \sum y_i x_{i1} - 2\tilde\beta_2 \sum y_i x_{i2} - 2\tilde\beta_3 \sum y_i x_{i1}x_{i2} + 2\tilde\beta_1\tilde\beta_3 S_{21} \\
&- \sum (y_i - \hat\beta_0)^2 - \hat\beta_1^2 S_{20} - \hat\beta_2^2 S_{02} - \hat\beta_3^2 S_{22} + 2\hat\beta_1 \sum y_i x_{i1} + 2\hat\beta_2 \sum y_i x_{i2} + 2\hat\beta_3 \sum y_i x_{i1}x_{i2} - 2\hat\beta_1\hat\beta_3 S_{21} \\
= &(\tilde\beta_1^2 - \hat\beta_1^2)S_{20} - \hat\beta_3^2 S_{22} - 2(\tilde\beta_1 - \hat\beta_1)\sum y_i x_{i1} + 2\hat\beta_3 \sum y_i x_{i1}x_{i2} - 2\hat\beta_1\hat\beta_3 S_{21}
\end{aligned}$$

Following the substitution rule, we evaluate the above expression replacing each y_i by its expected value under H_1. So $\sum y_i x_{i1}$ is replaced by $\sum(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}x_{i2})x_{i1} = \beta_1 S_{20} + \beta_3 S_{21}$ and $\sum y_i x_{i1}x_{i2}$ is replaced by $\sum(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}x_{i2})x_{i1}x_{i2} = \beta_1 S_{21} + \beta_3 S_{22}$. Also, each $\hat\beta_j$ is replaced by β_j (since these estimates are unbiased under H_1) while $\tilde\beta_1$ is replaced by $\frac{\sum y_i x_{i1}}{S_{20}}$ which is replaced by $\beta_1 + \beta_3 \frac{S_{21}}{S_{20}}$. Putting this all together, $SSE_0 - SSE_1$ is replaced by

$$2\beta_1\beta_3 S_{21} + \beta_3^2 \frac{S_{21}^2}{S_{20}} - \beta_3^2 S_{22} - 2\beta_3 \frac{S_{21}}{S_{20}}(\beta_1 S_{20} + \beta_3 S_{21}) + 2\beta_3(\beta_1 S_{21} + \beta_3 S_{22}) - 2\beta_1\beta_3 S_{21}$$

which, on further manipulation, reduces to $\sigma^2\delta^2 = \beta_3^2\left(S_{22} - \frac{S_{21}^2}{S_{20}}\right)$ which leads to the same formula for δ is in part (d) above. So this method does work, but it obviously is not the simplest solution in this case!