HOMEWORK 4
MATH 463 - SPRING 2017

ALEX THIES

INTRODUCTION. I collaborated with Joel Bazzle, Torin Brown, Ashley Ordway and Seth Temple on this assignment.

PROBLEM 1

Let $x_1 = (1,1,1,1,1,1)'$, $x_2 = (3,1,4,6,3,3)'$, $x_3 = (7,3,2,0,3,3)'$, $x_4 = (8,4,9,5,4,4)'$, $Y = (4,36,44,12,16,8)'$, and $V = \mathscr{L}(x_1,x_2,x_3,x_4)$. Suppose we wish to test $H_0 : \beta_4 = 0, \beta_2 = \beta_3$.

(a) Find two matrix A such that H_0 is equivalent to $A\beta = 0$.
(b) Find $\hat{\beta}$, $\hat{Y} = X\hat{\beta}$, and $Z = A\hat{\beta}$ for one of your choices of A.
(c) Define V_0 so that $\theta := \mathbb{E}[Y|X] \in V_0$ if and only if $A\beta = 0$. Find $\hat{Y_0} = \Pi_{V_0} Y$, $Y - \hat{Y}$ and $\hat{Y_1} = \hat{Y} - \hat{Y_0}$.
(d) Determine $\mathrm{SS_{Res}} = ||\mathrm{Y} - \hat{\mathrm{Y}}||^2$, $\mathrm{SS_{Res}(V_0)} = ||\mathrm{Y} - \hat{\mathrm{Y}_0}||^2$, and the F-statistic.
(e) Verify that $||\hat{Y} - \hat{Y_0}||^2 = Z'[A(X'X)^{-1}A']^{-1}Z$.

Solution. (a) Given that $H_0 : \beta_4 = 0, \beta_2 = \beta_3$, and upon brushing up on linear algebra, we find the following matrices. Note that A_i must have two rows because there are two linear constrains in H_0. We will use A_0 as our choice for A_i for the rest of this problem.

$$A_0 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & -1 & 0 \end{pmatrix}$$
$$A_1 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \end{pmatrix}$$

(b) We compute $\hat{\beta}$, $\hat{Y} = X\hat{\beta}$ and $Z = A\hat{\beta}$ in `R`.

$$\hat{\beta} = (41, -3, -8, 3)',$$
$$\hat{Y} = (-1.07 \times 10^{-14}, 32, 40, 8, 20, 20)',$$
$$Z = (3, 5)'.$$

(c) As before, most of the computational work is done in `R`. Note that in order to define V_0 as requested, we combine β_2 and β_3 thus, $w = \beta_2 + \beta_3$. We define $V_0 = x_1 + w$. Computing $\hat{Y_0}$, $Y - \hat{Y}$ and $\hat{Y_1}$ yields the following,

$$\hat{Y_0} = (4, 36, 20, 20, 20, 20)',$$
$$Y - \hat{Y} = (4, 4, 4, 4, -4, -12)', \text{and}$$
$$\hat{Y_1} = (-4, -4, 20, -12, -3.55 \times 10^{-14}, -3.55 \times 10^{-14})'.$$

(d) Again, we compute the requested Residual Sums of Squares and F-statistic in `R`.

$$\begin{aligned} \mathrm{SS}_{\mathrm{Res}} &= 224, \\ \mathrm{SS}_{\mathrm{Res}}(V_0) &= 800, \\ F &= \frac{800 - 224/2}{224/2}, \\ &= 2.57. \end{aligned}$$

(e) We compute that $||\hat{Y} - \hat{Y}_0||^2 = 576$ and $Z'[A(X'X)^{-1}A']^{-1}Z = 576$. Observe that $576 = 576$.

□

Problem 2

Consider the data in the dataset `teengamb` in the package faraway:

```
install.packages("faraway")
library(faraway)
data(teengamb)
```

The last line should bring up a description of the variables. Is there a difference between males and females as relates to gambling behavior? Fit any appropriate model(s) and carry out any appropriate test(s).

Solution. We compare the following three models, and perform a F test between each using an `ANOVA` table. The first model considers all variables except sex, the second introduces sex as a variable, and the third model introduces sex as a dummy variable. We can see from the table that upon introducing sex as a descriptive variable we have significance, i.e., sex does have an impact on gambling behavior. Upon comparing sex generally with sex as a category among all other variables we again determine that this is significant, however less so than just considering sex alone.

We should note however that there are several descriptive variables which haven't been measured/observed which are being absorbed by our error terms that are not independent from the variables which are controlling. These could be things such as parent's educational attainment, whether or not gambling is legal in the jurisdiction(s) from which we are performing observations, etc.

```
tg.lm0 <- lm(gamble~verbal+status+income, data = teengamb)
tg.lm1 <- lm(gamble~verbal+status+income+sex, data = teengamb)
tg.lm2 <- lm(gamble~verbal+status+income+sex+verbal:sex
             +status:sex+income:sex, data = teengamb)
xtable(anova(tg.lm0, tg.lm1, tg.lm2))
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	43	25359.56				
2	42	21623.77	1	3735.79	8.49	0.0059
3	39	17163.54	3	4460.23	3.38	0.0277

□

Problem 3

Suppose that 11 plots of land are plotted with three varieties of corn. The following lists the yields for the three varieties:

I	II	III
52	64	53
56	57	55
60	62	58
56		50

(a) Test the hypothesis that the three varieties all have the same expected yield.
(b) Suppose that for the corn yield the true means were 70, 75, 95 and that $\sigma = 20$. Find the power of the $\alpha = 0.05$ level test for equal means.
(c) How large should n_0, the number of observations per treatment (number of plots per treatment) be in order to have power at least 0.90 for the parameters in (a)?

Solution. (a) In order to test $H_0 : \mu_1 = \mu_2 = \mu_3$ we perform an F test, in this case we refer to RSS_{Full} as the within group variability (SSW), and $\text{RSS}_{\text{Small}}$ as the between group variability (SSB). We find that $\mu_1 = 56$, $\mu_2 = 61$, and $\mu_3 = 54$, thus the grand mean is $\mu_{\text{G}} = n^{-1} \sum_{i=1}^{3} \mu_i = 57$. With these quantities we can compute the SSB as follows,[1]

$$\begin{aligned} \text{SSB} &= \sum_{i=1}^{3} n_i \left(\bar{Y}_{i,\cdot} - \bar{Y}_{\cdot,\cdot} \right)^2, \\ &= 88. \end{aligned}$$

For SSW we sum over the variances of each column, weighted by their size,

$$\begin{aligned} \text{SSW} &= \sum_{i=1}^{3} \sum_{j=1}^{n_i} \left(Y_{i,j} - \bar{Y}_{i,\cdot} \right)^2, \\ &= 127. \end{aligned}$$

We compute the F-statistic as follows,

$$\begin{aligned} F &= \frac{SSB/df_B}{SSW/df_W}, \\ &= \frac{88/2}{127/8}, \\ &= 2.77. \end{aligned}$$

We find that the p-value is 0.12, given that this is greater than most reasonable level α's, we fail to reject H_0

(b) In order to compute the power for $H_1 : \mu_1 = 70, \mu_2 = 75$ and $\mu_3 = 95$, with $\sigma^2 = 400$ we must find the grand mean for these supposed μi's, and with that compute the Corrected Sum of Squares (CSS). We do these computations in `R` and find that the new grand mean $\mu_{G_1} = \sum_{i=1}^{3} \frac{n_i}{11} \bar{Y}_{i,\cdot} = 80.45$. With this new grand mean we find that CSS = 1372.73. When computing the power we must compute our non-centrality parameter δ, for one-way analysis of variance this is simply $\delta = \text{CSS}/\sigma^2 = 3.43$. If we suppose $\alpha = 0.05$, we compute that

[1] More detailed computations are shown in the R chunks.

the power is 0.26, which is too low for us to say that our conclusion in (a) is well-founded.

(c) Here I use Seth's idea to use a for loop to try new sample sizes starting at $n_i = 20$ down to $n_i = 4$ (our largest sample), with the loop terminating when the power dips below our desired value of 0.9. Given this method we determine that the least sample size for which we have a power of at least 0.9 is $n_i = 16$. □

E-mail address: athies@uoregon.edu