PBHS 32410 / STAT 22401 Problem Set 3 Winter 2025

Possible points 80.

1. [22 pts total] Modified from Exercises 5.7

Three types of fertilizer are to be tested to see which one yields more corn crop. Forty similar plots of land were available for testing purposes. The 40 plots are divided at random into four groups, 10 plots in each group. Fertilizer 1 was applied to each of the 10 corn plots in Group 1. Similarly, Fertilizers 2 and 3 were applied to the plots in Groups 2 and 3, respectively. The corn plants in Group 4 were not given any fertilizer; it will serve as the control group. Table 5.17 gives the corn yield y_{ij} for each of the 40 plots. Data are located on the website in the file

fert.txt

(a) [4 pt] Create three indicator variables, F_1, F_2, F_3, one for each of the three fertilizer groups (to compare against control)

(b) [4 pt] Fit the model $y_{ij} = \mu_0 + \mu_1 F_{i1} + \mu_2 F_{i2} + \mu_3 F_{i3} + \epsilon_{ij}$.

(c) [4 pts] Test the hypothesis that none of the three types of fertilizer has an effect on corn crops. Specify the hypothesis to be tested, the test used, and your conclusions at the 5% significance level.

(d) [5 pts] Test the hypothesis that the three types of fertilizer have equal effects on corn crop. Specify the hypothesis to be tested, the test used and your conclusions at the 5% significance level.

(e) [5 pts] Irrespective of the results in (d), test whether there is a common effect of fertilizer of any kind relative to control. (hint: this involves creating a new predictor variable)

2. [38 pts total] Modified from C&H

Analysis of the Education Expenditures data in Ch. 5 using the ideas presented in recent lectures. We will examine the relationship between a per capita expenditure variable Y and measures for income, per capita school age children, and urbanicity (X_1, X_2, X_3), and a year interval categorical variable (*yearint*). The data (already in STATA) are located on the website

`educ_expend_12.dta`

You can read into R using

```
> library(foreign)
> expend <- read.dta("educ_expend_12.dta")
```

(a) [2 pts] Check relation among outcome and predictors X_1, X_2, X_3 via plots and correlation

(b) [4 pts] test the effects of X_1, X_2, X_3 on Y. Start w/overall test and proceed to evaluating effect of each variable (note: this is straightforward regression testing as we have done, test and comment on results)

(c) [4 pts] There is a variable called *yearint*, representing the relevant year interval for the values. Create appropriate indicator variables for year interval, and evaluate the effect of year interval on Y in the multivariable model. Use interval 1 as the baseline.

(d) [4 pts] Change the baseline level for *yearint* to level 2. Why does the coefficient and significance level for level 3 change so much?

(e) [4 pts] Create the appropriate interaction term(s) to test whether the effects of X_2 remain unchanged over year intervals. For this model, set the baseline year back to interval 1.

(f) [4 pts] Based on findings in (e), you should find that separate regressions by year interval need to be reported. Report coefficients for X_2 variables separately by year interval (ie., run stratified analysis by year) and interpret

(g) [4 pts] Based on the stratified models in (f), is there a suggestion that other predictors may vary over year interval?

(h) [6 pts] Based on the models with interaction terms in (e), Write out the slopes for the three year intervals in terms of the estimated β values from the model. How do these compare to the strata-specific slopes?

(i) [6 pts] Finally, use the model in (e) to compute the predicted expenditure (that is, the $\hat{y}$) for each of the three year intervals. For the calculation, set the X_1 (income) and X_3 (urbanicity) values at the mean of those covariates (which you can calculate from the data).

3. [20 pts total] Interaction Model Concepts

(this problem is just conceptual and has no data). Evaluating treatment effect on a continuous outcome measure. We are evaluating circulating tumor cell DNA (ctDNA) in relation to randomly assigned treatment groups (coded 0 and 1) and also taking into account a categorical tumor feature (EGFR receptor, coded 0 (negative) and 1(positive)). The full model fit is:

$$E[Y|Trt, EGFR] = \beta_0 + \beta_1 Trt + \beta_2 EGFR + \beta_3 Trt \times EGFR$$

(a) [5 pts] In the above model, what is the treatment effect (in terms of β parameters) in EGFR-negative patients? What is the treatment effect in EGFR-positive patients?

(b) [5 pts] If in the above model, the interaction term is found to be not needed (that is, β_3 not significantly different from zero), then what is the treatment effect in EGFR-

negative patients? In EGFR-positive patients?

(c) [5 pts] For the model with the interaction term, what is the mean ctDNA in treatment 0 and in treatment 1 when EGFR = 1

(d) [5 pts] For the model without the interaction term ($E[Y|Trt, EGFR] = \beta_0 + \beta_1 Trt + \beta_2 EGFR$), what is the mean ctDNA in treatment 0 and treatment 1 when EGFR = 1?