STA 305s14 Regular Assignment Two[1]

This assignment is preparation for Term Test One on Feb. 3d, and for the final exam. Your solutions to these homework problems will not be handed in. Use the formula sheet, which is posted on the course home page. As more material is covered, additional problems will be added at the end of the assignment.

Lecture Unit 4: Introduction to experimental design

1. The point of this exercise is that when we omit or ignore explanatory variables that are related to both the respons variables and to other explanatory variables that are in the model, the result is incorrect estimation of regression parameters. Independently for $i = 1, \ldots, n$, let

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$; the variance σ^2 is an unknown parameter. The $x_{i,j}$ quantities are fixed constants, not random variables. The β_j quantities are unknown parameters. Let us call this the *true model.*

   (a) What is $E(Y_i)$ under the true model?

   (b) What is $E(\overline{Y})$ under the true model? Show your work.

   (c) Unfortunately, the $x_{i,2}$ values are not part of the data set. Suppose we follow the usual practice in applied regression analysis, and just use the data we have. In this case we fit a regression model with just $x_{i,1}$, and estimate the parameter β_1 with the usual least squares estimator

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_{i,1} - \overline{x}_1)(Y_i - \overline{Y})}{\sum_{i=1}^n (x_{i,1} - \overline{x}_1)^2}.$$

   Recall that an estimator T of a parameter θ is said to be *unbiased* if $E(T) = \theta$ for all θ in the parameter space. Is the estimator $\widehat{\beta}_1$ given above an unbiased estimator of β_1 when the true model holds? Answer Yes or No and show your work.

   (d) Compare your expression for $E(\widehat{\beta}_1)$ to a slide from Lecture set 4. Do you see the parallel?

   (e) Now look at the bias term (the difference between $E(\widehat{\beta}_1)$ and β_1) along with the formula for $\widehat{\beta}_1$ above. Do you see a sample regression coefficient in the bias term?

2. The purpose of this exercise is to reinforce the idea of a permutation test. It's based on a small, artificial example so you can do it by hand. In a completely randomized design, two experimental units are randomly assigned to a treatment condition, and two are assigned to a control condition. Values of the response variable are 6, 2 for the treatment condition and 4, 0 for the control condition. The test statistic will be $D = \overline{Y}_1 - \overline{Y}_2$.

   (a) Give the value of D for the observed data.

   (b) Give the complete permutation distribution of D.

   (c) Give the exact two sided permutation $p-$value for a test of no treatment effect. The answer is a number between zero and one.

3. In a(n) _______ study, the allocation of individuals to treatments is not under the control of the investigator.

4. In a(n) _______ study, the system under study (including allocation of individuals to treatments) is mostly under the control of the investigator.

[1] Copyright information is at the end of the last page.

5. The smallest subset of experimental material to which a separate treatment might be applied is called a(n) _______.

6. An omitted variable that is associated with *both* the explanatory and the response variable is sometimes called a(n) _______ variable.

7. In a(n) _______, experimental units are assigned at random to two or more treatment conditions in such a way that all assignments are equally likely.

8. Items are arranged in a random order, so that all orders are equally likely. This is called a(n) _______.

9. Data are allocated to treatments in all possible ways, and the value of the test statistic is calculated for each allocation. Under the null hypothesis of no treatment effects, all such allocations are equally likely. A listing of the values attained by the test statistic, together with their probabilities under the null hypothesis, is called the _______.

10. Using the permutation distribution, we find the probability that the test statistic would equal or exceed the value based on the observed data. This is called the _______ of the _______ test.

11. It is claimed that growing up in a bilingual household causes slower language development in small children (though they eventually catch up).

    (a) What is the experimental unit?

    (b) What is the treatment?

    (c) What is the response?

12. It is claimed that breast-fed babies are more resistant to disease.

    (a) What is the experimental unit?

    (b) What is the treatment?

    (c) What is the response?

13. It is claimed that canned food can have unsafe levels of bacteria if the can is dinted or crushed.

    (a) What is the experimental unit?

    (b) What is the treatment?

    (c) What is the response?

14. Continuing with the canned food example,

    (a) Describe an *observational* study to test the claim that canned food can have unsafe levels of bacteria if the can is dinted or crushed. That is, specify what data you would collect. In the end, you will have a spreadsheet in which the rows are the experimental units. What are the columns?

    (b) Describe an *experimental* study to test the same claim. Give some detail. What are some potential confounding variables that are no longer a problem? Again, in the end you will have a spreadsheet in which the rows are the experimental units. What are the columns?

15. The point of this question is that the statistical analysis will be better if the statistician tries to see the bigger picture. Recall the experiment on scab disease, which also appear in Computer Assignment Two. Scab disease is a fungal infection that affects potatoes. The fungus does not grow well in acidic soil, so investigators designed a study to see whether adding sulphur to the soil would reduce the scab disease. In a completely randomized design, eight plots of land were assigned to a control condition, and four plots of land each were assigned to several levels of sulphur that was spread on the land in the Fall. The amounts of sulphur were either 300 pounds per acre, 600 pounds per acre or 1200 pounds per acre. The potatoes were harvested at the end of the growing season. One hundred potatoes were randomly selected from each plot of land. The potatoes were washed, and then a lab assistant estimated the percent of each potato's surface that was infected with scab disease. The response variable is, for each plot of land, the mean percent of the potato's surface covered with scab disease. The explanatory variable is pounds of sulphur per acre, in hundreds of pounds; the control is zero.

    (a) Is this an experimental study, an observational study, or both?

    (b) What is the experimental unit in this study?

    (c) What is the treatment?

    (d) What is the response?

    (e) Why would it be a good idea to harvest all the potatoes at about the same time?

    (f) Give another example of an explanatory variable that was probably held constant in this study, though it is not mentioned.

    (g) It seems to me that scab disease must have been a problem in all the plots of land that were used in this study, though again this is not mentioned. Using this as a hint, give an example of a potentially powerful omitted variable that was likely *not* held constant.

    (h) Do you think this omitted variable is systematically related to the treatment? Why or why not?

    (i) Do you think the omitted variable in question might be systematically related to the treatment if this were a purely observational study? Why or why not?

    (j) Give an example of at least one additional omitted variable that could have a strong effect on the average amount of scab disease.

    (k) There are lots of hypotheses that could be tested here. How would you do a randomization test to compare the control to just the 600 pound per acre treatment? What's a reasonable test statistic for the randomization test?

    (l) Do you think there is a case for one-sided tests here? Give one advantage of a one-sided test, say for the comparison of 600 pounds per acre to the control. Can you think of a disadvantage?

    (m) Suppose you wanted to test for any differences among the three sulphur amounts, with a single randomization test. Suggest a test statistic that you might want to use. Would you want to randomize *all* the data?

    (n) If you were involved in this study before the data were collected, you would *not* agree to assess the effect of the treatment on just this one response. Give at least two more good response variables, and briefly indicate why each one is desirable.

Lecture unit 5: Estimating a treatment effect

16. The experimental treatment adds the same constant Δ to the response of each unit receiving the treatment. This is called the assumption of _______.

17. Make up an example of a study for which the assumption if unit-treatment additivity is *not* justified. Try to make it different from the examples given in class.

18. we have N experimental units, and we randomly select n without replacement to receive the experimental treatment. $Z_i = 1$ if unit i is chosen, zero otherwise. If all experimental units were in the control condition, their response variable values would have been $y_1, \ldots, y_N$.

    (a) Following standard notation from Sample Surveys, let $\overline{y}_u = \frac{1}{N}\sum_{i=1}^N y_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^N Z_i y_i$. Show $E(\overline{y}) = \overline{y}_u$.

    (b) Under the assumption of unit-treatment additivity, the sample mean response of the units receiving the treatment is $\overline{y}_1 = \frac{1}{n}\sum_{i=1}^N Z_i(y_i + \Delta)$, while the sample mean response of the units in the control condition is $\overline{y}_2 = \frac{1}{N-n}\sum_{i=1}^N (1-Z_i)y_i$. Is $\widehat{\Delta} = \overline{y}_1 - \overline{y}_2$ an unbiased estimator of Δ? Answer Yes or No and carry out the calculation. Show *all* your work.

    (c) It can be shown that $Var(\widehat{\Delta}) = \frac{S^2}{n\left(1-\frac{n}{N}\right)}$ For fixed N, what choice of n minimizes this variance (thus making the estimate as precise as possible)? Show your work.

19. Under the *Random Sampling Model*, a total of n experimental units (the notation was N in the preceding question) are a simple random sample from some very large population with expected value μ and variance σ^2. Randomly dividing them into a treatment group and a control group yields two independent random samples, a fact you do not need to prove. The sample sizes are n_1 for the treatment group and n_2 for the control group, with $n_1 + n_2 = n$. Let $Y_{i,1}$ denote the value of the response variable for unit i in the treatment condition, *if the treatment had not been applied*. These values are not observable. Let $Y_{i,2}$ denote the value of the response variable for unit i in the control condition. These values *are* observable. Let

$$\overline{Y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}(Y_{i,1} + \Delta) \quad \text{and} \quad \overline{Y}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_{i,2}.$$

    (a) Is $\widehat{\Delta} = \overline{Y}_1 - \overline{Y}_2$ an unbiased estimator of the treatment effect Δ? Answer Yes or No and show your work.

    (b) Calculate $Var(\widehat{\Delta})$. Show your work.

    (c) For fixed n, what choice of n_1 minimizes this variance (thus making the estimate as precise as possible)? Show your work.

This assignment was prepared by Jerry Brunner, Department of Statistical Sciences, University of Toronto. The LaTeX source code is available from the course website: http://www.utstat.toronto.edu/~brunner/oldclass/305s14