BIOS 665: Problem Set 4

Assigned: October 26, 2017

Due: November 9, 2017

Reminder: For all hypothesis tests, please state the method, the null hypothesis, the test statistic, the degrees of freedom, the p-value, and the interpretation of the test using a two-sided significance level of 5%, unless otherwise stated.

Helpful hints: For estimates and tests, simply copying and pasting SAS output without any commentary will not earn full credit, especially on exams. Highlighting is not considered commentary. However, commentary can be as simple as: The 95% CI for the odds ratio is (\_\_, \_\_).

I have followed the Honor Code. Signed: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. A social scientist is interested in how region and education level are associated with people’s opinions regarding physical activity. She conducts a survey on randomly selected individuals, asking them their geographic region (West Coast, Midwest, and East Coast) and education level (college graduate, high school graduate, less than high school). She also asks how strongly they agree with the following statement: “U.S. employers should offer greater incentives to have physically active employees.” Participants could choose one of these options: “Disagree,” “Neutral,” or “Agree.” Where applicable, let “less than high school” education level and “West Coast” region be the reference groups.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Education Level | Region | Level of agreement with statement regarding physical activity | | |
| Disagree | Neutral | Agree |
| College Graduate | West Coast | 18 | 15 | 48 |
|  | Midwest | 13 | 19 | 21 |
|  | East Coast | 28 | 28 | 52 |
| High School Graduate | West Coast | 46 | 23 | 24 |
|  | Midwest | 22 | 20 | 21 |
|  | East Coast | 48 | 18 | 23 |
| Less than high school | West Coast | 13 | 15 | 28 |
|  | Midwest | 15 | 16 | 17 |
|  | East Coast | 15 | 17 | 24 |

* 1. Mathematically specify a proportional odds regression model for level of agreement (ordered from more agreement to less agreement) with main effects for region and education level (using all 3 levels). State assumptions and interpret all model parameters. Then, assess goodness of fit of the proportional odds model, being sure to justify these methods.
  2. Conduct a test to assess whether proportional odds is a reasonable assumption for these data. State your null hypothesis, the p-value of your test, the criteria for making the decision of your test, and write a sentence explaining the results of your test. Then,

1. test whether education level has an effect on agreement at the two-sided 0.05 significance level. State your null hypothesis, the p-value of your test, the criterion for making the decision of your test, and write a sentence explaining the results of your test.
2. provide an estimate and 95% confidence interval for the odds ratio of agree vs. (neutral or disagree) comparing college graduates with those having less than high school education. What do you conclude about the statistical significance of this effect from the confidence interval? Briefly discuss how this estimate compares to a comparable estimate for (agree or neutral) vs. disagree.
3. Consider the data from Question 1 above.
   1. Mathematically specify and fit a generalized logits regression model for level of agreement, treating the level of agreement as nominal. Include main effects for education level and region. Let “neutral” be your reference level for the outcome variable. State assumptions, and interpret all model parameters.
   2. Use the model from Part 2.a. to test whether education level has an effect on agreement at the two-sided 0.05 significance level. State your null hypothesis, the p-value of your test, the criterion for making the decision of your test, and write a sentence explaining the results of your test.
   3. Using this model, provide an estimate and 95% confidence interval for the odds ratio of agree vs. neutral comparing college graduates with those having less than high school education. Repeat for disagree vs. neutral, as well as for agree vs. disagree. What do you conclude about the statistical significance of each effect from these confidence intervals? *Hint: You might invoke a separate PROC LOGISTIC by changing the reference level to obtain the odds ratio for agree vs disagree.*
4. The following data are from a study to compare two treatments with respect to the relationship between dose and dichotomous response concerning pain relief.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Treatment | Dose | Favorable | Unfavorable | Sample Size |
| A | 1 mg | 21 | 39 | 60 |
| A | 10 mg | 24 | 36 | 60 |
| A | 100 mg | 42 | 18 | 60 |
| B | 2 mg | 23 | 37 | 60 |
| B | 20 mg | 31 | 29 | 60 |
| B | 200 mg | 50 | 10 | 60 |

1. For Treatment A only, use logistic regression to describe the relationship between favorable response and log10(dose). Note: log10(.) denotes the log base 10 transformation. *Hint: recall Problem Set #2, Problem 3.*
2. State assumptions, and mathematically specify the model.
3. Provide estimates and 95% confidence limits (or fiducial limits) for the doses at which each of 25%, 50%, and 75% favorable response are predicted by the model (i.e. ED25, ED50, and ED75, respectively).
4. Use a probit analysis to calculate all the estimates and confidence intervals requested in Part 3.a.ii. How might your assumptions change when using a probit model vs a logistic model?
5. Briefly compare and contrast your results from Part 3.a.ii. and Part 3.a.iii.
6. For the data from both Treatment A and Treatment B, use logistic regression to describe the relationship between favorable response and log10(dose) for the data, allowing for separate effects for each treatment group, as illustrated in class.
7. State assumptions, and mathematically specify the model.
8. Evaluate goodness of fit of the model.
9. Provide a point estimate and its 95% confidence interval for the relative potency of Treatment B relative to Treatment A.
10. The table shown below displays the cross-classification of maternal age groups (in years) and the number of births with a particular disorder in a specific geographic area during a specific time period, as well as the corresponding numbers of all births. When necessary, use ‘20-24 years’ as the reference group for maternal age group and ‘1’ as the reference group for birth order.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Birth Order | 1 | 2 | 3 |  | 1 | 2 | 3 |
| Maternal Age | Number of Births with Disorder | | |  | Total Number of Births | | |
| 20-24 | 128 | 152 | 71 |  | 329,462 | 326,735 | 175,682 |
| 25-29 | 54 | 112 | 101 |  | 114,987 | 208,692 | 207,060 |
| 30-34 | 41 | 79 | 109 |  | 39,473 | 83,224 | 117,312 |
| 35-39 | 38 | 89 | 99 |  | 14,202 | 28,478 | 45,015 |
| 40+ | 22 | 44 | 83 |  | 3,046 | 5,381 | 8,654 |

1. Specify the mathematical structure of a statistical model to describe the variation in the rates of the disorder per 100,000 live births with respect to maternal age group and birth order.
2. Interpret the estimated parameters of this model, and provide appropriate two-sided 95% confidence intervals for those pertaining to birth order.
3. Use the model from Part 4.a. to obtain predicted values for the rates of the birth disorder for the respective birth order subpopulations corresponding to ‘30-34 years’ for maternal age group.