**Problem Set 1** (Version Feb 2, 2020)

**Due**: 5PM February 28, 2020 via Canvas

**Note**: Students may work together with classmates on the problem set but each student must submit their own assignment.

**To be turned in**: Answers to questions, SAS code (not SAS output)

1. Use the frequency distribution data in Table 1 to answer questions 1A through 1D.

|  |  |  |
| --- | --- | --- |
| Table 1. | | |
| **Outcome** | **Exposure=1** | **Exposure=0** |
| Level=0 | 69 | 65 |
| Level=1 | 72 | 66 |
| Level=2 | 77 | 52 |

1A. State the expressions for a nominal logistic model to estimate the crude association between the exposure and the outcome variable. Choose level 0 as the reference group for the outcome.

1B. State the expression for an ordinal logistic model to estimate the crude association between the exposure and outcome variable.

1C. Calculate the crude odds ratios (for the effect of the exposure on the outcome) that would result from part 1a.

1D. Calculate three crude odds ratios for an ordinal logistic model: two odds ratios to evaluate the proportional odds assumptions and a third odds ratio that would result from an ordinal logistic model that assumed the odds were proportional.

1. Consider a binary logistic model with 4 treatment levels and 3 levels of ethnicity, with indicator dummy variables coded as follows:

Treatment T1 T2 T3 Ethnicity E1 E2

1 1 0 0 1 1 0

2 0 1 0 2 0 1

3 0 0 1 3 0 0

4 0 0 0

Now consider two models, one without interaction and the other with interaction:

**Model 1**: logit(D=1) = 𝛽0 + β1T1 + β2T2 + β3T3 + 𝛽4E1 + 𝛽5E2

**Model 2**: logit(D=1) = 𝛽0 + β1T1 + β2T2 + β3T3 + 𝛽4E1 + 𝛽5E2

+ 𝛽6T1×E1 + 𝛽7T1×E2 + 𝛽8T2×E1 + 𝛽9T2\*E2 + 𝛽10T3×E1 + 𝛽11T3×E2

Answer the following:

2A. What is the interpretation of exp(β1) using Model 1?

2B. What is the interpretation of exp(β1) using Model 2?

2C. What is the odds ratio comparing Treatment=3 vs Treatment=4 among those with Ethnicity=1 using Model 1?

2D. What is the odds ratio comparing Treatment=3 vs Treatment=4 among those with Ethnicity=1 using Model 2?

1. Use the permanent SAS dataset named “Nilton” on Canvas (in Problem Set 1 assignment) to answer questions 3A through 3E. The data came from a cross-sectional study of inpatients with Methicillin-resistant staph aureus (MRSA). For simplicity, ignoring missing values in all answers below.

The dataset contains the following variables: METHICSE (dichotomous outcome of interest, coded 1 for MRSA), AGE (continuous), AGECAT (dichotomous coded 1 if age ≥ 55), PREVHOSP (dichotomous coded 1 if hospitalized in the previous 6 months), SEX (dichotomous coded 1 for male), PREANTBU (dichotomous coded 1 for antibiotic use in the previous 3 months). The dichotomous variables are all coded 1 or 0.

3A. Suppose you model AGE and SEX as a predictor of METHISCE in a logistic regression. State the model in terms of the prevalence of MRSA.

3B. Run the model from 3A to estimate the predicted prevalence of MRSA for a 50 year-old male.

3C. Run a binary logistic model to estimate the association between previous hospitalization with prevalent MRSA. In an adjusted model, control of AGECAT, SEX, and PREANTBU. What are the crude and adjusted odds ratios? Interpret the adjusted odds ratio in one sentence.

3D. Run a model with three product terms between the exposure (previous hospitalization) with age category, sex, and antibiotic use. The additional three terms should be the product of PREVHOSP with AGECAT, SEX, and PREANTBU respectively. Include all three product terms in one model. What is the odds ratio for prevalent MRSA in patients with previous hospitalization in the past 6 months compared to those without hospitalization among women less than 55 years old who did not use antibiotics in the past three months?

3E. Run a likelihood ratio test for the addition of three product terms for interaction in part 3D (between previous hospitalization with age category, sex, and antibiotic use). Test the null hypothesis that all three beta coefficients for the interaction terms = 0). What is the -2ln likelihood for the full and reduced models? How many degrees of freedom are there for the likelihood ratio test? What is the p-value for the result? What is your interpretation about interaction?

1. Kaplan Meier Curves with Censoring

A prospective student is considering her course schedule and trying to determine if she should take MSCR534. Use the tables below to determine the two-year survival of two groups, those who recently took MSCR 534 (534=1) and those that did not (534=0).

Table 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Event time  (Months) | Number of survivors | Number of deaths | Number censored | Proportion surviving | 534 status |
| 0 | 24 | 0 | 0 | 1 | 1 |
| 2 | 24 | 1 | 1 |  | 1 |
| 6 |  | 1 | 2 |  | 1 |
| 12 |  | 0 | 2 |  | 1 |
| 24 |  | 2 | 3 |  | 1 |

Table 3

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Event time  (Months) | Number of survivors | Number of deaths | Number censored | Proportion surviving | 534 status |
| 0 | 30 | 0 | 0 | 1 | 0 |
| 4 | 30 | 1 | 2 |  | 0 |
| 12 |  | 2 | 2 |  | 0 |
| 16 |  | 1 | 3 |  | 0 |
| 24 |  | 3 | 4 |  | 0 |

4A. Fill in the missing cells in Table 2 and Table 3 above.

4B. Graph the Kaplan Meier curve for each of the tables (by hand is fine). Graph the survival curves on the same graph. Hint: the Y axis should be from 0 to 1.0 and the X axis should be from 0 to 24 months.

4C. If the data above were true, would you take MSCR 534 again?

1. Write a midterm exam question focused on nominal/ordinal logistic regression, survival analysis, or Cox proportional hazards models. Also include a solution to the question.
2. Use the permanent dataset on Canvas named “Anderson” (in Problem Set 1 assignment) to answer questions 6A through 6D. The dataset consists of remission survival times (SURVT), measured in weeks, on 42 leukemia patients, half of whom get a new therapy (RX=0) and half of whom get a standard therapy (RX=1). Control variables are SEX (1=male, 0=female) and log white blood cell count. The log white blood cell count is in two forms, (LOGWBC - continuous) and a three-level categorical variable (LWBC3). The variable STATUS indicates event (out of remission - coded 1) or censorship (coded 0).

6A. Run PROC LIFETEST three times to perform log rank tests for the effects of 1) treatment, 2) log white blood cell count, and 3) for the effect of gender. State the null hypotheses (there are three) and report the p-values and decision. Examine the survival plots. Why do you think the log rank test for gender was not significant even though the estimated survival curves (for gender) look different? Examine the log (-log) survival plots; does the PH assumption seem violated for any of the predictors?

6B. For part 6B assume the PH assumption is not violated for any of the variables. Run a Cox model with RX, LOGWBC, and SEX in the model. State the model in terms of the hazard function. What are the estimated crude hazard ratio and adjusted hazard ratio for treatment? Interpret the adjusted hazard ratio in one sentence.

6C. Suppose it is decided that the PH assumption is violated just for the gender variable. Run a stratified Cox model for the effect of treatment, adjusted for logwbc, with gender as the stratified variable (assuming no interaction with treatment). What is the estimated hazard ratio for the effect of treatment?

6D. Now include a treatment-gender interaction term in the model. Create this interaction term in a SAS data step. Note that SEX should not be in the model statement, but in the strata statement. What is the Wald test p-value for the product term coefficient? What are the estimated hazard ratios for treatment using the model with the product term?