CPH 931 — Fall 2008 — Dr. Charnigo

Final Examination

Complete two exercises of your choice. This Final Examination is due on Tuesday 16 December at 2:30 p.m. and is a strictly individual activity.

[50] 1. Acquire the Channing House data from {http://www.mcw.edu/biostatistics/Faculty/Faculty/JohnPKleinPhD/SurvivalAnalysisBook/DataSetsBothEditions.htm}. The time-to-event response variable is the "Difference between the above two ages", hereafter denoted T. The censoring variable is "Death status". The explanatory variables are "Gender" and "Age of entry into retirement home", hereafter denoted Z_1 and Z_2.

[10] a. Consider the Weibull model

$$\log T_i = \beta_0 + \beta_1 z_{1,i} + \sigma W_i.$$

Which of the following would indicate that women tend to live longer than men once entering retirement homes: $\beta_1 < 0$ or $\beta_1 > 0$? Please give a brief explanation for your answer.

[10] b. Fit the Weibull model from part a. Do women tend to live longer than men once entering retirement homes, do men tend to live longer than women, or do the data fail to resolve the question? [Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ and, if H_0 is rejected, note whether $\widehat{\beta}_1$ is positive or negative.]

[10] c. Create a plot of estimated survival functions (or estimated cumulative distribution functions) for males and for females based on the Weibull model.

[10] d. Create a plot of estimated survival functions (or estimated cumulative distribution functions) for males and for females using the Kaplan-Meier method. Comment on how this plot is similar to or different from that based on the Weibull model. [Some template SAS code for Kaplan-Meier estimation may be found at {http://www.richardcharnigo.net/STA580F08/SASsurvival.txt}.]

[10] e. Now consider the expanded Weibull model

$$\log T_i = \beta_0 + \beta_1 z_{1,i} + \beta_2 z_{2,i} + \sigma W_i.$$

When you control for the age of entry into the retirement home, do women tend to live longer than men once entering retirement homes, do men tend to live longer than women, or do the data fail to resolve the question?

[50] 2. The Excel file {Cushing931F08.xls} contains a modified version of the Cushing's syndrome data set referenced in *Pattern Recognition and Neural Networks* by B. D. Ripley (1996). "Chemical1" is a laboratory measurement of tetrahydrocortisone, expressed on a logarithmic scale. "Chemical2" is a laboratory measurement of pregananetriol, also expressed on a logarithmic scale. "Type" identifies the specific variety of Cushing's syndrome with which a patient has been diagnosed (0 = adenoma, 1 = bilateral hyperplasia, 2 = carcinoma). Note that "Type" is a nominal but not an ordinal variable.

[10] a. Propose a statistical model by which a medical professional could predict the variety of Cushing's syndrome with which a patient would be diagnosed, given that patient's laboratory measurements of (log) tetrahydrocortisone and (log) pregananetriol. Specifically, state the name of the statistical model (linear regression, logistic regression, proportional hazards regression, ... ?) and write out formulas to express the relationships among variables; be sure to define all symbols used in your formulas.

[10] b. State symbolically and then test the null hypothesis that your statistical model can be simplified through the removal of tetrahydrocortisone. [Regardless of whether you accept or reject the null hypothesis, please keep tetrahydrocortisone in the model for parts c through e.]

[10] c. Use your statistical model to estimate the probabilities of type 0, type 1, and type 2 Cushing's syndrome for a patient with Chemical1 = 2.00 and Chemical2 = 1.00.

[10] d. Mimic your solution to part c to find formulas for the estimated probabilities of type 0, type 1, and type 2 Cushing's syndrome for a patient with Chemical1 = x_1 and Chemical2 = x_2, where x_1 and x_2 are generic numbers. [Your formulas will be functions of x_1 and x_2. You can check your work by setting $x_1 = 2.00$ and $x_2 = 1.00$ and seeing whether you recover your answers to part c.]

[10] e. Use Excel or SAS to implement your answers to part d for all of the patients in {Cushing931F08.xls}. If a medical professional had made predictions based on your statistical model, how many of the patients in {Cushing931F08.xls} would have been diagnosed correctly? [Assume that the diagnosis would have been the type of Cushing's syndrome with the largest estimated probability.]

[50] 3. The *American Journal of Epidemiology* paper by Pesch et al (2002) investigated whether environmental arsenic exposure is related to nonmelanoma skin cancer. Among 129 males with skin cancer, 99 lived within 10 km of a coal-burning power plant (and hence were deemed to have a high level of environmental arsenic exposure) and 30 did not. Among 142 male controls, 104 lived within 10 km of the power plant and 38 did not.

[10] a. Obtain a point estimate of the odds ratio (odds of nonmelanoma skin cancer for males living within 10 km of a coal-burning power plant divided by odds for males not living within 10 km of the power plant). Then obtain a 95% confidence interval for the odds ratio.

[10] b. Assuming that a 15% increase in the risk of nonmelanoma skin cancer is important, comment on the matter of statistical significance versus clinical significance for your point estimate of the odds ratio. [You may assume that the prevalence is low enough for the odds ratio to approximate the relative risk.]

[10] c. An unsophisticated reader states, "The estimated risk of skin cancer among males living within 10 km of a coal-burning power plant is $99/203 = 0.488$, the estimated risk of skin cancer among males not living within 10 km of the power plant is $30/68 = 0.441$, and so the estimated relative risk is $0.488/0.441 = 1.11$." Explain how you would convince the reader of his error. [Assume that the reader, being unsophisticated, will not understand Bayes' Theorem.]

[10] d. Apply Bayes' Theorem to determine an estimate of the relative risk as a function of an assumed value for the prevalence. What is the estimated relative risk with an assumed value of 0.005 for the prevalence? With what assumed value for the prevalence is the estimated relative risk equal to 1.11?

[10] e. Recalling part b, compare the two estimated relative risks in part d (1.11 and one other that you calculated) regarding their implications for clinical significance. Explain why the unsophisticated reader's mistake is fairly serious in this instance.