Bios 312: Homework #3

February 6, 2013

Written problems due at the beginning of class, Thursday, February 14, 2013.

All questions relate to the question of whether the nadir PSA level following hormonal treatment for prostate cancer is prognostic of time in remission independently of any information from other commonly used covariates. The data is posted on the class web pages (psa.txt). Note that the variable *inrem* is text ("yes" or "no"). You will need to tell Stata that this variable should be stored as a "string" rather than as a number. The following code would do the trick:

```
infile ptid nadir pretx ps bss grade age obstime str8 inrem using psa.txt
```

1. Provide suitable descriptive statistics for this dataset as might be presented in Table 1 of a manuscript appearing in the medical literature.
2. Perform analyses to determine whether there is a difference in mean nadir PSA levels across groups defined by the presence of a bone scan score of 3 at the time of receiving hormonal therapy. (You should create an indicator variable *bss3* that is 1 if the bone scan score is 3 and 0 if the bone scan score is less than 3, and all parts of this problem should consider the dichotomized variable.).)
    a. Provide the sample mean and sample standard deviation in the sample of patients having a bone scan score less than 3 and the sample of patients having a bone scan score of 3. Obtain 95% confidence intervals for the mean nadir PSA in each group.
    b. Using a t test which presumes equal variances across groups, provide inference regarding an association between nadir PSA and bone scan score.
    c. Based on the results of part a, do you think the analysis in part b was appropriate? Consider whether the p value obtained in part b was valid under the strong null hypothesis of equal distributions of nadir PSA in the two groups, as well as under the weak null of equal mean nadir PSA in the two groups. For each of these two null hypotheses, would you expect that the t test provided appropriate inference, anti-conservative inference (i.e., the reported P value would tend to be too low), or conservative inference (i.e., the reported P value would tend to be too high). Explain your reasoning. *(Note: It is highly inappropriate to use the descriptive statistics or any hypothesis test to decide whether you use the analysis in part b or that in part d below. The point here is that you could have anticipated that there might be a problem and used the best analysis to address your scientific question.)*
    d. Using a t test which allows unequal variances across groups, provide inference regarding an association between nadir PSA and bone scan score. How do these results compare with those in part b? Is the difference between the analyses what you would have expected given your answers to part c? What are the implications on how you would report your conclusions scientifically?
    e. Using classical linear regression (so without robust standard errors), provide inference regarding an association between nadir PSA and bone scan score.

f. How do the estimates from your analysis in part e compare to the descriptive statistics you obtained in part a? Explain any similarity or differences.

g. How does the inference about your intercept from part e compare to the inference you obtained in part a? Explain any similarity or differences.

h. How does the inference about your slope from part e compare to the inference you obtained in part b? Explain any similarity or differences. What does this say about the reliability of classical linear regression to detect differences of mean nadir PSA levels across groups defined by whether they had a bone scan score of 3?

i. Using linear regression with the robust standard errors, provide inference regarding an association between nadir PSA and bone scan score.

j. How do the results from your analysis in part i compare to your results in parts a, b, d, and e? Explain any similarity or differences.

3. Perform analyses to determine whether there is a difference in mean nadir PSA levels across groups defined by their bone scan score at the time of receiving hormonal therapy. (You should use the variable *bss* without dichotomization.)

   a. Provide the sample mean and sample standard deviation in each sample of patients defined by their bone scan score. Obtain 95% confidence intervals for the mean nadir PSA in each group.

   b. Using linear regression with the robust standard errors, provide inference regarding an association between nadir PSA and bone scan score. Provide interpretations for the intercept and slope.

   c. Using the regression model, what would be the estimated mean nadir PSA level for groups having a bone scan score of 1, 2, and 3? How do these estimates compare to your results in part a? Explain any similarity or differences.

   d. What are the relative advantages of the analysis in problem 1(i) versus 2(b)? Discuss in terms of both the predicted values for each group as well as your ability to detect associations between nadir PSA and bone scan score.

4. Perform analyses to determine whether there is a difference in the distribution of relapse. across groups defined by the presence of a bone scan score of 3 at the time of receiving hormonal therapy. (All parts of this problem should consider the dichotomized variable.)

   a. Why is it scientifically not of interest (and thus not addressing the question posed by this problem) to compare groups merely according to whether they have relapsed (so *inrem* = “no”) or not (so *inrem*=”yes”) while under observation?

   b. Why is it acceptable to compare groups according to whether the patients have relapsed within 24 months?

   c. Provide the probability and odds of a patient having a relapse within 24 months in the sample of patients having a bone scan score less than 3 and the sample of patients having a bone scan score of 3. Obtain 95% confidence intervals for the probability and odds of relapse in each group.

   d. Using the chi squared test, provide inference regarding an association between relapse and a bone scan score of 3.

e. Using a t test which presumes equal variances across groups, provide inference regarding the probability of relapse within 24 months across groups defined according to whether they have a bone scan score of 3.

f. Using a t test which allows unequal variances across groups, provide inference regarding the probability of relapse within 24 months across groups defined according to whether they have a bone scan score of 3.

g. Compare the results obtained in parts d, e, and f. Which would be the most accepted method of analysis? Under what situations are the others acceptable? Discuss with respect to the strong and weak null hypotheses.

h. Using classical logistic regression (so without robust standard errors), provide inference regarding an association between relapse within 24 months and bone scan score. Provide estimates of the probability and odds of relapse with 24 months as derived from the regression model. (You will want to consider both the logit command (which provides estimates on the log odds scale) and the logistic command (which provides estimates on the odds ratio scale) to answer this problem.)

i. How do the estimates from your analysis in part h compare to the descriptive statistics you obtained in part c? Explain any similarity or differences.

j. How does the inference about your intercept from part h compare to the inference you obtained in part c? Explain any similarity or differences.

k. How does the inference about your slope from part h compare to the inference you obtained in part d, e, and f? Explain any similarity or differences.

l. Using logistic regression with the robust standard errors, provide inference regarding an association between relapse within 24 months and bone scan score.

m. How do the results from your analysis in part l compare to your results in parts c, d, e, f, and h? Explain any similarity or differences.

5. Perform analyses to determine whether there is a difference in the distribution of relapse within 24 months across groups defined by their bone scan score at the time of receiving hormonal therapy. (You should use the variable *bss* without dichotomization.)

a. Provide the sample probability and odds of relapse within 24 months for each sample of patients defined by their bone scan score. Obtain 95% confidence intervals for the probability and odds of relapse within 24 months for each group.

b. Using classical logistic regression (so without the robust standard errors), provide inference regarding an association between relapse within 24 months and bone scan score. Provide interpretations for the intercept and slope.

c. Using the regression model, what would be the estimated probability and odds of relapse within 24 months for groups having a bone scan score of 1, 2, and 3? How do these estimates compare to your results in part a? Explain any similarity or differences.

d. Using logistic regression with the robust standard errors, provide inference regarding an association between relapse within 24 months and bone scan score. How does this analysis differ from that in part c? Explain any similarity or differences.

e. What are the relative advantages of the analyses in problem 4 versus problem 5? Discuss in terms of both the predicted values for each group as well as your ability to detect associations between relapse and bone scan score.

6. Perform analyses to determine whether there is an association between mean nadir PSA level and relapse.

    a. Provide suitable descriptive statistics and inference comparing mean nadir PSA levels across groups defined by whether they have relapsed within 24 months. Make clear the statistical analysis you performed.

7. Perform analyses to determine whether there is an association between geometric mean nadir PSA level and relapse.

    a. Provide suitable descriptive statistics and inference comparing geometric mean nadir PSA levels across groups defined by whether they have relapsed within 24 months. Make clear the statistical analysis you performed.

    b. Why might you *a priori* prefer inference based on the geometric mean to that based on the mean? What considerations would make you prefer inference based on the mean?

8. Perform analyses to determine whether the distribution of relapse differs across groups defined by nadir PSA level.

    a. Perform a logistic regression comparing the odds of relapse within 24 months across groups defined by the nadir PSA level when modeled as a continuous, untransformed variable. Provide interpretation for the intercept and slope.

    b. Perform a logistic regression comparing the odds of relapse within 24 months across groups defined by the nadir PSA level when modeled as a continuous, log transformed variable. Provide interpretation for the intercept and slope.

    c. Why might you *a priori* prefer inference based on the log transformed nadir PSA value? What considerations would make you prefer inference based on the untransformed variable instead? What consideration might make you prefer a dichotomization of nadir PSA (which analysis I did not make you perform, but you can if you want)?

9. Consider the analyses performed in problems 6 through 8 above.

    a. What are the relative merits of the four analyses. Which might you prefer *a priori*? Why?

    b. All of these analyses suffer from a serious definitional problem inherent in this study. Can you deduce this problem? (Hint: There is no analysis that you can do to address this problem. It is a problem with the study design.)