Biost 536: Categorical Data Analysis in Epidemiology
Emerson, Fall 2013

Homework #3
November 21, 2013

Written problems: To be submitted as an email attachment in by 5pm on Wednesday, November 27, 2013. See the instructions for peer grading of the homework that are posted on the web pages.

*On this (as all homeworks) unedited Stata output is **TOTALLY** unacceptable. Instead, prepare a table of statistics gleaned from the Stata output. The table should be appropriate for inclusion in a scientific report, with all statistics rounded to a reasonable number of significant digits. (I am interested in how statistics are used to answer the scientific question.)*

Keys to past homeworks from quarters that I taught Biost 517 (e.g. HW #8) or Biost 518 (e.g., HW #3) might be consulted for the presentation of inferential results.

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), with documentation in the file psa.doc. 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
```

Note that all patients were followed for a minimum of 24 months. In all problems we will be considering the probability (or odds) of a patient surviving relapse-free for 24 months following therapy. You can create a variable indicating relapse within 24 months using the following Stata code:

```
g relap24 = 0
replace relap24 = 1 if obstime <= 24 & inrem=="no"
```

1. Provide suitable descriptive statistics for this dataset as might be presented in Table 1 of a manuscript appearing in the medical literature. (Because the primary question is comparing 24 month relapse free survival across groups defined by nadir PSA, you might consider presenting descriptive statistics in groups according to some dichotomization of nadir PSA levels. Alternatively, you could provide descriptive statistics within groups defined by whether the subjects relapse within 24 months or not.)

Table 1: Descriptive Statistics for Covariates by Relapse Status

	Relapse within 24 months			Did not relapse within 24 months			
Variable	**Count**	**Estimate**	**SE**	**Count**	**Estimate**	**SE**	**p-value**
Nadir PSA (ng/ml), mean*	22	31.9	11.2	28	4.12	3.27	0.0113
Pretreatment PSA (ng/ml), mean*	20	732.4	303.5	23	617.2	261.1	0.7738
Performance Status, mean*	20	76.5	2.64	28	83.9	1.81	0.0203
Age, mean*	22	68.4	1.21	28	66.7	1.1	0.3208
Observation time (months), mean*	22	11.1	1.36	28	42.1	2.28	<0.0001
Tumor Grade, proportion**							0.69
Grade 1	3	0.176	0.095	7	0.292	0.095	
Grade 2	7	0.412	0.123	8	0.333	0.095	
Grade 3	7	0.412	0.123	9	0.375	0.101	
Bone Scan Score, proportion**							0.053
1	0	0	0	5	0.179	0.074	
2	4	0.2	0.092	9	0.321	0.090	
3	16	0.8	0.092	14	0.5	0.096	

*Continuous variables evaluated using 2 sample t-tests

**Categorical variables evaluated using Chi-square tests

2. Perform logistic regression analyses to determine whether the distribution of relapse within 24 months differs across groups defined by nadir PSA level after adjustment for bone scan score and performance status. For each of the following models, provide full statistical inference for your measure of association.

   Please note that I am using bss as a continuous variable in all of the following calculations

   a. Perform an adjusted 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.

   The log odds of relapse within 24 months among men post hormonal therapy for prostate cancer is .0333 higher for each unit increase in nadir PSA level measured in ng/ml (95% CI -.058, .125) when controlling for bone scan score and performance status; however this result is not statistically significant which indicates that these results are not beyond what might be expected to occur by chance.

b. Perform an adjusted 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.

The log odds of relapse within 24 months among men post hormonal therapy for prostate cancer is .8597 higher for each unit increase in log nadir PSA level (95% CI .327, 1.48) when controlling for bone scan score and performance status.

c. Perform an adjusted logistic regression comparing the odds of relapse within 24 months across groups defined by the nadir PSA level when modeled as linear splines with knots at 1, 4, and 16 ng/ml.

The log odds of relapse within 24 months among men post hormonal therapy for prostate cancer is 3.39 higher (95% CI 0.306, 6.47) for each unit increase in nadir PSA level for men with a nadir PSA between 0.1 ng/ml and 1 ng/ml when controlling for bone scan score and performance status.

The log odds of relapse within 24 months among men post hormonal therapy for prostate cancer is 0.102 lower (95% CI -1.12, 0.915) for each unit increase in nadir PSA level for men with a nadir PSA between 1 ng/ml and 4 ng/ml when controlling for bone scan score and performance status; however this result is not statistically significant which indicates that these results are not beyond what might be expected to occur by chance.

The log odds of relapse within 24 months among men post hormonal therapy for prostate cancer is 0.322 higher (95% CI -0.058, 0.702) for each unit increase in nadir PSA level for men with a nadir PSA between 4 ng/ml and 16 ng/ml when controlling for bone scan score and performance status; however this result is not statistically significant which indicates that these results are not beyond what might be expected to occur by chance.

The log odds of relapse within 24 months among men post hormonal therapy for prostate cancer is 0.018 lower (95% CI -0.036, -0.001) for each unit increase in nadir PSA level for men with a nadir PSA above 16 ng/ml when controlling for bone scan score and performance status.

d. For each of the above regression models, provide an interpretation of the intercept.

Model a: the log odds of relapse within 24 months among men post hormonal therapy for prostate cancer is .7287 (95% CI -5.500, 6.958) for a nadir PSA level of 0ng/ml, and all other covariates set to 0. In this case, setting all covariates to zero is not a useful measure as it is outside the range of our data. Furthermore, this result is not statistically significant which indicates that these results are not beyond what might be expected to occur by chance.

Model b: the log odds of relapse within 24 months among men post hormonal therapy for prostate cancer is 1.1186 (95% CI -6.1817, 8.4190) for a log nadir PSA level of 1ng/ml, and all other covariates set to 0. In this case, setting all

covariates to zero is not a useful measure as it is outside the range of our data. Furthermore, this result is not statistically significant which indicates that these results are not beyond what might be expected to occur by chance.

Model c: the log odds of relapse within 24 months among men post hormonal therapy for prostate cancer is -.67916 (95% CI -8.6104, 7.2521) for a nadir PSA level of 0ng/ml, and all other covariates set to 0. In this case, setting all covariates to zero is not a useful measure as it is outside the range of our data. Furthermore, this result is not statistically significant which indicates that these results are not beyond what might be expected to occur by chance.

In this longitudinal study, we could instead have considered the "reverse" analyses in which nadir PSA is used as the response and the predictor is the indicator of relapse within 24 months.

e. Perform linear regression analyses to determine whether there is an association between mean nadir PSA level and relapse within 24 months after adjustment for bone scan score and performance status. Make clear the statistical analysis you perform. Provide full statistical inference for your measure of association.

For men post hormonal therapy for prostate cancer who relapse within 24 months, the mean nadir PSA level is approximately 23.5 ng/ml higher (95% CI 0.476, 46.56) compared to men post hormonal therapy for prostate cancer who do not relapse, after controlling for bone scan score and performance status.

f. Perform linear regression analyses to determine whether there is an association between geometric mean nadir PSA level and relapse within 24 months after adjustment for bone scan score and performance status. Make clear the statistical analysis you perform. Provide full statistical inference for your measure of association. (Recall that inference on the geometric mean is obtained by performing linear regression on log transformed response variables.)

For men post hormonal therapy for prostate cancer who relapse within 24 months, the geometric mean nadir PSA level is approximately 2.61 ng/ml higher (95% CI 1.42, 3.81) compared to men post hormonal therapy for prostate cancer who do not relapse, after controlling for bone scan score and performance status.

3. Consider the analyses performed in problems 2 and 3 above.

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

Logistic regression with continuous untransformed predictor variable: easily interpretable, efficient, doesn't assume normality/linearity/homoscedasticity, robust

Logistic regression with continuous log transformed predictor variable: depending on scientific question may model variable better, other benefits same as above

Logistic regression with linear splines: flexible (can fit non-linear distributions), allow more informative comparison of lower values to higher values, other benefits same as above

Linear regression with untransformed response variable: easily interpretable, efficient

Linear regression with log transformed response variable: depending on scientific question may model variable better, addresses non-normality/non-linearity, still interpretable on the multiplicative scale

I prefer the logistic regression model with the non-transformed predictor variable due to the efficiency, flexibility and interpretability. Splines have arbitrary cut points and can be difficult to interpret. Linear regression requires us to assume that the relationship between the predictor and outcome is linear, and can be problematic with small sample sizes in regards to non-normality and heteroscadasticity.

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.)

One problem with the study design is accurate measurement of the outcome. We are not taking into account individual person time so we are not making the same comparison for each individual. A better method would be to consider a survival type of analysis in which time to censoring/outcome is taken into account for each individual.