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

Homework #3 Key
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.)

Ans: The following table provides descriptive statistics within groups defined by nadir PSA levels (below 2 ng/ml vs 2ng/ml or above). There are 31 patients whose nadir PSA was below 2 ng/ml and 19 patients with nadir PSA greater than 2 ng/ml. Measurements for performance status and bone scan score are missing for two patients, measurements for tumor grade are missing for 9 patients, and pre treatment PSA measurements are missing for 7 patients. Patients with lower nadir PSA have slight trends toward being younger, having higher performance status, having lower bone scan scores, and having lower pre treatment PSA, but having higher tumor grade. None of these trends are very strong. On the other hand, a striking difference is observed between the low and high

nadir groups with respect to relapse: 22.6% of the 31 patients with nadir PSA less than 2 ng/ml relapsed within 24 months compared to 78.0% of the 19 patients with nadir PSA greater than 2 ng/ml.

	Nadir PSA < 2 ng/ml Mean (SD; Min Mdn Max; n) n (%)	**Nadir PSA > 2 ng/ml** Mean (SD; Min Mdn Max; n) n (%)	**All Patients** Mean (SD; Min Mdn Max; n) n (%)
Age (yrs)	66.2 (5.0; 58, 66, 79; n= 31)	69.4 (6.5; 61, 68, 86; n= 19)	67.4 (5.8; 58, 66, 86; n= 50)
Perf Status	82.7 (12.0; 50, 80, 100; n= 30)	77.8 (8.8; 60, 80, 90; n= 18)	80.8 (11.1; 50, 80, 100; n= 48)
Bone Scan Score	2.4 (0.8; 1, 3, 3; n= 30)	2.8 (0.4; 2, 3, 3; n= 18)	2.5 (0.7; 1, 3, 3; n= 48)
BSS=1	5 (17%)	0 (0%)	5 (10%)
BSS=2	9 (30%)	4 (22%)	13 (27%)
BSS=3	16 (53%)	14 (78%)	30 (63%)
Tumor Grade	2.2 (0.8; 1, 2, 3; n= 28)	2.0 (0.8; 1, 2, 3; n= 13)	2.1 (0.8; 1, 2, 3; n= 41)
Grade 1	6 (21%)	4 (31%)	10 (24%)
Grade 2	10 (36%)	5 (38%)	15 (37%)
Grade 3	12 43%)	4 (31%)	16 (39%)
Pre-treat PSA (ng/ml)	565 (1126; 5, 121, 3946; n= 26)	832 (1525; 10, 157, 4797; n= 17)	671 (1288; 5, 127, 4797; n= 43)
Nadir PSA (ng/ml)	0.6 (0.5; 0.1, 0.3, 1.7; n= 31)	42.1 (55.3; 2.2, 14.0, 183.0; n= 19)	16.4 (39.2; 0.1, 1.0, 183.0; n= 50)
Relapse ≤ 24 mos	7 of 31 (22.6%)	15 of 19 (79.0%	22 of 50 (44.0%)

The following table provides descriptive statistics within groups defined by relapse status within 24 months of treatment. There are 28 patients who remained in remission for 24 months and 22 patients who relapsed within 24 months. Measurements for performance status and bone scan score are missing for two patients, measurements for tumor grade are missing for 9 patients, and pre treatment PSA measurements are missing for 7 patients. Patients remaining in remission for 24 months have slight trends toward being younger, having higher performance status, having lower bone scan scores, having lower tumor grade, and having lower pre treatment PSA. None of these trends are very strong. On the other hand, a striking difference is observed between the remission and relapse groups with respect to nadir PSA: a mean nadir PSA of 4.1 ng/ml (median 0.2 ng/ml) was observed for the 28 patients remaining in relapse for 24 months compared to a mean nadir PSA of 31.9 ng/ml (median 10.5 ng/ml) for the 22 patients who relapsed within 24 months.

	Remission at least 24 mos Mean (SD; Min Mdn Max; n) n (%)	**Relapse within 24 mos** Mean (SD; Min Mdn Max; n) n (%)	**All Patients** Mean (SD; Min Mdn Max; n) n (%)
Age (yrs)	66.7 (5.8; 58, 66, 81; n= 28)	68.4 (5.7; 61, 68, 86; n= 22)	67.4 (5.8; 58, 66, 86; n= 50)
Perf Status	83.9 (9.6; 50, 80, 100; n= 28)	76.5 (11.8; 50, 80, 100; n= 20)	80.8 (11.1; 50, 80, 100; n= 48)
Bone Scan Score	2.3 (0.8; 1, 3, 3; n= 28)	2.8 (0.4; 2, 3, 3; n= 20)	2.5 (0.7; 1, 3, 3; n= 48)
BSS=1	5 (18%)	0 (0%)	5 (10%)
BSS=2	9 (32%)	4 (20%)	13 (27%)
BSS=3	14 (50%)	16 (80%)	30 (63%)
Tumor Grade	2.1 (0.8; 1, 2, 3; n= 24)	2.2 (0.8; 1, 2, 3; n= 17)	2.1 (0.8; 1, 2, 3; n= 41)
Grade 1	7 (29%)	3 (18%)	10 (24%)
Grade 2	8 (33%)	7 (41%)	15 (37%)
Grade 3	9 (38%)	7 (41%)	16 (39%)
Pre-treat PSA (ng/ml)	617 (1252; 5, 100, 4377; n= 23)	732 (1357; 25, 174, 4797; n= 20)	671 (1288; 5, 127, 4797; n= 43)
Nadir PSA (ng/ml)	4.1 (17.3; 0.1, 0.2, 92; n= 28)	31.9 (52.5; 0.5, 10.5, 183; n= 22)	16.4 (39.2; 0.1, 1.0, 183; n= 50)

Comments: Either of the above tables could be used to present the descriptive statistics. Either provides general "Materials and Methods" information about the types of patients involved in the study, and either provides some preliminary information about an association between nadir PSA and the probability of remaining in remission for 24 months. Furthermore, either table also provides some information about potential confounding:

- ***The first table provides information about potential associations in the sample between the nadir PSA and the other predictors.***
    - ***Variables that are not associated with nadir PSA in the sample cannot be a confounder, whether they are associated with time in relapse or not.***
    - ***Any variables that are associated with nadir PSA in the sample <u>might</u> be a confounder, providing we believe that the variable is causally associated with time in relapse independent of any association between nadir PSA and time in relapse. We generally would use our prior understanding of the scientific setting to decide whether we thought such a causal relationship existed. To the extent that we might want to use this data to demonstrate such presumed associations, the first table cannot be used to examine this: I did not provide any descriptive statistics directed toward associations between time in relapse for any variables except nadir PSA. And even if I had, I would have wanted to examine those associations with time in relapse while controlling for nadir PSA (i.e., within strata having nadir PSA (nearly) constant) .***
- ***The second table provides information about potential associations in the sample between the other predictors and time in relapse.***
    - ***To be a confounder, there must be a causal relationship in the population between the third variable and the time in relapse. Furthermore, this causal association must be independent of any association of interest between nadir PSA and the time in relapse. To the extent that population based sampling was used in the study design, we might expect that any causal associations that exist in the population would also appear as an association in the sample. So it may not be completely irrelevant to want to examine the associations between all the predictors and the time in relapse. But there are many links in this chain of evidence that are left unexamined: I can not evaluate whether the asoociation is causal using descriptive statistics with observational data, and I did not examine the associations within strata that hold nadir PSA constant.***
    - ***To be a confounder, there must also be a relationship between the additional predictors and the predictor of interest, which in this case is the nadir PSA. The second table presented no information on this aspect of confounding.***

In light of the above, I have an overall preference in this study for the first table as opposed to the second table. We have a clear predictor of interest (nadir PSA), and the primary question relates to trying to quantify whether any association between the response variable (relapse) and the POI (nadir PSA) is merely a reflection of confounding by other variables already known to be associated with poor outcomes in prostate cancer. Hence, I would rather be able to get some information about the associations in the sample between my POI and other variables.

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.

General comments common to all regressions in problems 2 and 3: In all problems, I asked you to adjust for performance status and bone scan score. I did not specify how you would do such adjustment, and there would be some variability in how this was done. Here I discuss the options you could have considered.

For performance status:

- ***Issues related to type of measurement:***
  - ***Performance status is conceptually a continuous random variable that is measured on a somewhat arbitrary 100 point scale.***
    - ***Most often we expect continuous random variables to similarly have smoothly continuous relationships with the response (as opposed to "step functions" or "jump discontinuities" that might be modeled with dummy variables).***
  - ***However, it is generally recorded on in multiples of 10, and given that we most often restrict attention to patients who are not in the severest categories, performance status usually has a discrete distribution in the sample.***
    - ***Mere "round off" error in recording data is not sufficient reason to avoid continuous models of a covariate. Even when we have a covariate that is measured quite coarsely (e.g., recording age only to the decade), if the true relationship between a finely measured version of the variable (e.g., age to the nearest nanosecond) and the response variable were linear, we do far, far, <u>far</u> better to model the rounded off version as a linear continuous predictor. This holds true even if the widths of the intervals vary across categories (e.g., age categorized as 10-20, 20-40, and 40-90), though it would be important to code each interval by the mean age within that group (best), the median age within that group (second best choice), or midpoint of the range (this would likely be my choice if I had no other information, but other approaches could be considered based on the relative sizes of the categories).***
    - ***If there is a markedly nonlinear relationship between the covariate and the response and the covariate is measured coarsely, then the relative advantages of using dummy variables or splines versus a continuously modeled covariate will depend upon how nonlinear is the relationship and how coarsely recorded is the measurement.***
  - ***In this sample, there were only 6 distinct values observed (50 through 100).***
    - ***We only have 48 observations with performance status measurements available.***
    - ***That means on average there will be only 8 subjects in each category. Most times there is wide variation in the counts within each category, with the middle categories having a lot of observations and the lowest categories having only a few.***
      - ***In such a setting, we may have very little data to estimate the regression parameter for some dummy variables or some splines.***

        - ***The problem is made worse when using a binary or time to event response variable. We have to have enough events and enough nonevents in each group to be able to estimate the regression parameter for the dummy variable***
            - ***If we do have a parameter that is estimated from only a single observation, the resulting analysis is equivalent to omitting that observation.***
            - ***Trying to fit too many parameters unnecessarily also tends to decrease our statistical precision overall.***
- ***Issues related to the statistical role of variable:***
    - ***We are primarily interested in adjusting for this variable because it might be a confounder.***
    - ***Failure to model true relationships between the confounder and the response variable will mean that there might exist some residual confounding. Hence, we generally want to have more flexible models than might be used for our POI or for precision variables.***
- ***After considering all of those issues, I chose to model performance status as a continuous linear predictor.***

For bone scan score (I will not repeat discussion of general issues):

- ***Bone scan score is again a continuous value recorded as one of three broad categories.***
- ***We have relatively few observations with a bone scan score of 1, thus making it more difficult to fit dummy variables in any type of regression. With a binary response variable, the fact that there are no relapses in that group means that we could not estimate an odds in logistic regression (we would be trying to estimate negative infinity for that group).***
- ***While I could have decided to dichotomize bone scan score as 1 or 2 versus 3, I chose to model it continuously.***

Modeling interactions between performance status and bone scan score:

- ***When asked to adjust for confounders, it is rare that data analysts explicitly consider adjusting for interactions. Instead they adjust for interactions if the analysis model demands it:***
    - ***In regression models, the most common approach is to only model the "main effects".***
    - ***In stratified analyses (e.g., Mantel-Haenszel), we have to model the interaction as well as the main effects.***
- ***In the absence of knowledge that the interaction is confounding (i.e., the interaction is causally associated with the response independent of the POI and the interaction is associated with the POI in the sample), I, too, tend not to adjust for the interaction of confounders in regression models.***

Inference

- ***In all models, I choose to use classical logistic regression. I note that when modeling a linear continuous covariate, there can be a bit more difference between the classical SE and the robust SE, because the robust SE will be better at addressing the possibility of "model***

misspecification" (i.e., lack of linearity in the modeled predictors). However, the robust SE do not behave really well until we have a larger sample size. I consider it somewhat of a wash for this data set.

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.

Ans: From a logistic regression model of the odds of 24 month relapse on nadir PSA and adjusting for performance status and bone scan score at the time of treatment, we estimate that the odds of relapse within 24 months is 1.03 times higher for every 1 ng/ml difference in nadir PSA between populations having similar performance status and bone scan score. Based on a 95% confidence interval, we find that the precision of our study was such that these results would not be unusual if the true odds ratio per 1 unit difference in nadir PSA were anywhere between 0.987 and 1.08. A two-sided p value of 0.156 suggests that we can not with high confidence reject the hypothesis that there was no true association between the nadir PSA and probability of relapsing within 24 months.

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.

Ans: From a logistic regression model of the odds of 24 month relapse on logarithmically transformed nadir PSA and adjusting for performance status and bone scan score at the time of treatment, we estimate that the odds of relapse within 24 months is 1.81 times higher for every doubling of nadir PSA between populations having similar performance status and bone scan score. Based on a 95% confidence interval, we find that the precision of our study was such that these results would not be unusual if the true odds ratio per doubling of nadir PSA were anywhere between 1.27 and 2.58. A two-sided p value of 0.001 suggests that we can with high confidence reject the hypothesis that there was no true association between the nadir PSA and probability of relapsing within 24 months in favor of there being a significant tendency toward higher odds of relapse for populations having higher nadir PSA.

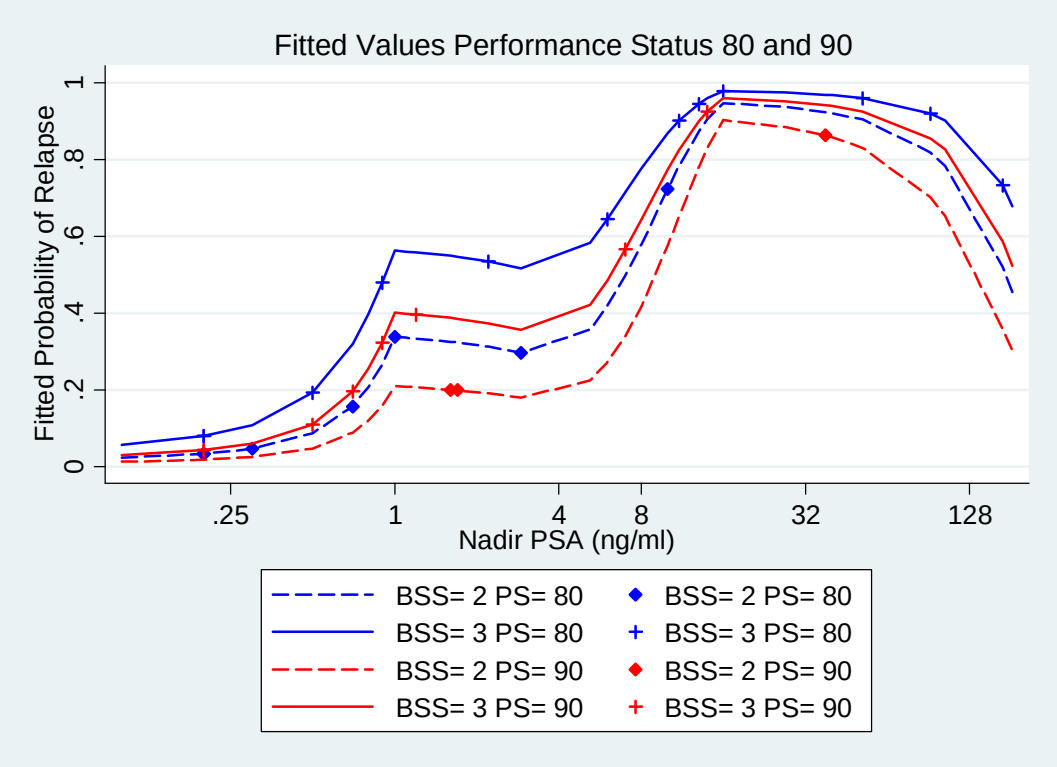
Comments: I found it useful to make comparisons based on a doubling of the nadir PSA. To do this, I had to raise the reported OR per 1 unit difference in log nadir to the log(2).

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.

Ans: From a logistic regression model of the odds of 24 month relapse on nadir PSA using linear splines and adjusting for performance status and bone scan score at the time of treatment, we find a statistically significant association between odds of relapse and nadir PSA (two-sided P= 0.0260). Figure 1 displays the estimated probability of relapse from the linear spline analysis. Fitted curves over the entire range of observed nadir PSA are displayed for strata having performance status 80 or 90 and bone scan score 2 or 3. Points are displayed for the actual values of nadir PSA observed within each stratum. Apparent is an estimated general trend toward higher probability of relapse with higher nadir PSA in each stratum. It should be noted that the fitted curves were linear in nadir PSA on the log odds scale between the knots at 1, 4, and 16 ng/ml. Lack of smoothness and lack of monotonicity are possibly due to the sparseness of the data. Vertical separation of the fitted curves displays the estimated effect associated with a 1 unit difference in bone scan score (solid for BSS= 3

versus dashed for BSS= 2) and a 10 unit difference in performance status (blue for PS= 80 versus red for PS= 90).

Comments: Our multiple partial test of all the linear spline coefficients is testing whether the true relationship could be adequately described by a horizontal (flat) line corresponding to a constant relapse probability for all values of the nadir PSA. It does not tell us anything about upward or downward trends. Those latter questions would be difficult to dissect out of the linear splines.



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

Ans: In the logistic regression model from problem 2a, we estimate that subjects with a nadir PSA of 0, performance status of 0, and a bone scan score of 0 would have a log odds of relapse within 24 months of 0.729, which corresponds to an odds of relapse of $e^{0.729}$ = 2.07 or a probability of relapse of 2.07 / (1 + 2.07) = 0.675. We note that a nadir PSA of 0 is likely non-physiologic (if reported, it is probably just below the limit of detection), a performance status of 0 corresponds to someone who is dead, and a bone scan score of 0 was not sampled in this study (it would correspond to someone who had no bone metastases, and such a patient would not be likely to undergo hormonal treatment for PSA). Hence, the intercept in this model is not of any scientific value. *(Note, however, that none of the variables in this model were statistically significant. Hence this intercept will likely tend to be influenced heavily by the overall relapse rate among all subjects.)*

In the logistic regression model from problem 2b, we estimate that subjects with a log nadir PSA of 0 (so a nadir PSA of 1), performance status of 0, and a bone scan score of 0 would have a log odds of relapse within 24 months of 1.12, which corresponds to an odds of relapse of $e^{1.12}$ = 3.06 or a probability of relapse of 3.06 / (1 + 3.06) = 0.753. We note that while a nadir PSA of 1 ng/ml is very much within the range sampled in our data (it was close to the median value), a performance status of 0 corresponds to someone who is dead, and a bone scan score of 0 was not sampled in this study

(it would correspond to someone who had no bone metastases, and such a patient would not be likely to undergo hormonal treatment for PSA). Hence, the intercept in this model is not of any scientific value.

In the logistic regression model from problem 2c, we estimate that subjects with a nadir PSA of 0, performance status of 0, and a bone scan score of 0 would have a log odds of relapse within 24 months of -0.679, which corresponds to an odds of relapse of $e^{-0.679} = 0.507$ or a probability of relapse of 0.507 / (1 + 0.5077) = 0.336. We note that a nadir PSA of 0 is likely non-physiologic (if reported, it is probably just below the limit of detection), a performance status of 0 corresponds to someone who is dead, and a bone scan score of 0 was not sampled in this study (it would correspond to someone who had no bone metastases, and such a patient would not be likely to undergo hormonal treatment for PSA). Hence, the intercept in this model is not of any scientific value. ***(Note, that in this model, the linear splines as a group were statistically significant, hence the estimated intercept differs from that observed in problem 2a.)***

Some additional comments about interpretations of the linear splines:

- ***We used knots at 1, 4, and 16 ng/ml.***
- ***The values of ndr1 were equal to the values of nadir, when nadir < 1. When nadir > 1, the values of ndr1 were 1.***
- ***The values of ndr2 were equal to 0 when nadir < 1. When 1 < nadir < 4, the values of ndr2 were equal to nadir – 1, When nadir > 4, the values of ndr2 were 4.***
- ***The values of ndr3 were equal to 0 when nadir < 4. When 4 < nadir < 16, the values of ndr3 were equal to nadir – 4, When nadir > 16, the values of ndr3 were 16.***
- ***The values of ndr4 were equal to 0 when nadir < 16. When 16 < nadir, the values of ndr4 were equal to nadir – 16.***
- ***Hence, all of the linear spline variables (ndr1, ndr2, ndr3, and ndr4) were equal to 0 only when nadir=0.***
- ***Note also that under this coding, if all the coefficient parameters are equal, then this corresponds to a straight line relationship. I will discuss this in the answer to problem 4.***

3. 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.
    a. 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.

Ans: From a linear regression model of the mean nadir PSA on the indicator of relapse within 24 months and adjusting for performance status and bone scan score at the time of treatment, we estimate that the mean nadir PSA is 23.5 ng/ml higher in a population that relapses within 24 months compared to a population having similar performance status and bone scan score, but remaining in remission for at least 24 months. Based on a 95% confidence interval computed using the robust standard error computed using the Huber-White sandwich estimator (and thereby

accounting for possible heteroscedasticity), we find that the precision of our study was such that these results would not be unusual if the true difference in mean nadir PSA were anywhere between 0.476 and 46.6 ng/ml. A two-sided p value of 0.046 suggests that we can with 95% confidence reject the hypothesis that there was no true association between the nadir PSA and probability of relapsing within 24 months in favor of there being a statistically significant tendency toward higher mean nadir PSA in populations that relapse in 24 months.

b. 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.)

Ans: From a linear regression model of the mean log nadir PSA on the indicator of relapse within 24 months and adjusting for performance status and bone scan score at the time of treatment, we estimate that the geometric mean nadir PSA is 13.7 times higher in a population that relapses within 24 months compared to a population having similar performance status and bone scan score, but remaining in remission for at least 24 months. Based on a 95% confidence interval computed using the robust standard error computed using the Huber-White sandwich estimator (and thereby accounting for possible heteroscedasticity), we find that the precision of our study was such that these results would not be unusual if the true ratio of geometric means for nadir PSA were anywhere between 4.13 and 45.2. A two-sided p value less than 0.0005 suggests that we can with high confidence reject the hypothesis that there was no true association between the nadir PSA and probability of relapsing within 24 months in favor of there being a statistically significant tendency toward higher geometric mean nadir PSA in populations that relapse in 24 months.

Comment: Note that I almost always back transform my estimates in order to make inference about the geometric means, rather than discussing log geometric means. Note also that I prefer to use the Stata regress command, rather than the glm command. Even though the glm command will back transform the estimates for you, it fits the robust SE slightly differently. I think the more standard linear regression approach is preferable. (You can compare the results given above to that that was obtained with glm.)

4. 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?

Ans: Logistic regression has the advantage of answering the "forward" question: Given a nadir PSA value, what is the probability (or odds) of relapse. Logistic regression suffers more when sample sizes are so small that we do not have enough events and nonevents in relevant groups. For instance, I would not trust trying to adjust a logistic regression for the interaction between performance status and bone scan score in this relatively small data set. On the other hand, linear regression would probably handle that fairly well.

All of that having been said, I would likely use logistic regression.

Now, in terms of how to model the nadir PSA in the regression models, I would have *a priori* guessed that a logarithmic transformation would be best. This is often the case in disease: multiplicative changes are more important. (And if I had chosen the "backwards" analyses of

problem 3, I would for the same reason prefer inference on the geometric mean to inference on the mean.)

Analyses using the linear splines are too difficult to interpret. (If you do not believe that, reread the vagueness with which I answered problem 2c.)

Comments: For what it is worth, we could use the linear splines to investigate the adequacy of a model fitting the untransformed nadir PSA versus a model fitting the log transformed PSA.

- ***A straight line model is a special case of the linear splines. It corresponds to the case when all the regression parameters are equal.***
    - ***We could test this in Stata using test ndr1=ndr2=ndr3=ndr4. That test returns a p value of 0.025.***
    - ***Alternatively, we could have fit a model including nadir and all the linear splines. Stata would have dropped one of the spline variables (it chose ndr4). Then we test that the other splines add nothing using testparm ndr*. This gives the exact same result as the preceding test.***
    - ***If linearity of effect had been my question, I would reject that hypothesis based on this analysis.***
- ***I could similarly have used the spline variables to test for departures from a straight line association between log odds of relapse and log nadir.***
    - ***In this case, the log nadir model is not a special case of the linear spline model. So I could fit a model including lnadir and all the spline variables.***
    - ***I would then test for model fit of lnadir using the Stata command testparm ndr*.***
    - ***This analysis finds a p value of 0.54.***
    - ***Hence, if my question had been: Does a model with log nadir fit the data well?, I would not have been able to reject the hypothesis that there is a straight line relationship between log odds of relapse and the log nadir PSA (after adjusting for performance status and bone scan score).***

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.)

Ans: Both the nadir PSA and the relapse status are determined some time after treatment. This means that the nadir PSA could just be a marker for when the relapse occurs and not have any prognostic value in advance. For instance, if we imagine that after treatment, nadir PSA decreases by 20% every month until the patient dies, and, furthermore, that after relapsing, the patient dies immediately, then the nadir PSA would occur exactly at the time of relapse and thus be perfectly predictive <u>at the time of relapse</u>. It would provide no useful prediction into the future.

To try to address this, I looked at the ability of the nadir PSA to predict time in relapse after the first 18 months, because all patients achieved their nadir in 18 months. There was still a significant trend, but this did not solve all the problems.

STATA Code

In this key, I include the Stata code used to answer the questions so you can see where I obtained the numbers. This would be unacceptable on your homework.

Question 1:

```
. tabstat age ps bss grade pretx nadir, by(nadirge2) stat(n mean sd min q max) col(stat) long

nadirge2 variable |    N      mean        sd    min    p25     p50       p75       max
------------------+------------------------------------------------------------------
0             age |   31  66.22581  5.018032     58     63      66        69        79
               ps |   30  82.66667  12.01532     50     80      80        90       100
              bss |   30  2.366667  .7648905      1      2       3         3         3
            grade |   28  2.214286  .7867958      1      2       2         3         3
            pretx |   26  565.3192  1126.252    4.8     45     121       387      3946
            nadir |   31  .5709677  .4838688     .1     .2      .3        .9       1.7
------------------+------------------------------------------------------------------
1             age |   19  69.42105  6.483447     61     64      68        71        86
               ps |   18  77.77778  8.782038     60     70      80        80        90
              bss |   18  2.777778  .4277926      2      3       3         3         3
            grade |   13         2  .8164966      1      1       2         3         3
            pretx |   17       832   1524.68     10     65     157       536      4797
            nadir |   19  42.12105  55.34177    2.2      7      14        52       183
------------------+------------------------------------------------------------------
Total         age |   50     67.44  5.771711     58     63      66        70        86
               ps |   48  80.83333  11.07678     50     80      80        90       100
              bss |   48  2.520833  .6838434      1      2       3         3         3
            grade |   41  2.146341  .7924953      1      2       2         3         3
            pretx |   43  670.7512  1287.638    4.8     46     127       429      4797
            nadir |   50     16.36   39.2462     .1     .2     .95        10       183
-------------------------------------------------------------------------------------

. tabulate nadirge2 bss, row

+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

 RECODE of |               bss
     nadir |         1          2          3 |     Total
-----------+---------------------------------+----------
         0 |         5          9         16 |        30
           |     16.67      30.00      53.33 |    100.00
-----------+---------------------------------+----------
         1 |         0          4         14 |        18
           |      0.00      22.22      77.78 |    100.00
-----------+---------------------------------+----------
     Total |         5         13         30 |        48
           |     10.42      27.08      62.50 |    100.00

. tabulate nadirge2 grade, row

+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
```

```
+----------------+

 RECODE of |               grade
     nadir |         1          2          3 |     Total
-----------+---------------------------------+----------
         0 |         6         10         12 |        28
           |     21.43      35.71      42.86 |    100.00
-----------+---------------------------------+----------
         1 |         4          5          4 |        13
           |     30.77      38.46      30.77 |    100.00
-----------+---------------------------------+----------
     Total |        10         15         16 |        41
           |     24.39      36.59      39.02 |    100.00

. tabulate nadirge2 relap24, row

+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

 RECODE of |        relap24
     nadir |         0          1 |     Total
-----------+----------------------+----------
         0 |        24          7 |        31
           |     77.42      22.58 |    100.00
-----------+----------------------+----------
         1 |         4         15 |        19
           |     21.05      78.95 |    100.00
-----------+----------------------+----------
     Total |        28         22 |        50
           |     56.00      44.00 |    100.00

. tabstat age ps bss grade pretx nadir, by(relap24) stat(n mean sd min q max) col(stat) long

relap24  variable |    N      mean        sd    min    p25     p50       p75       max
------------------+--------------------------------------------------------------------
0             age |   28  66.71429  5.842736     58     63    65.5      69.5        81
               ps |   28  83.92857  9.560445     50     80      80        90       100
              bss |   28  2.321429  .7723735      1      2     2.5         3         3
            grade |   24  2.083333  .8297022      1      1       2         3         3
            pretx |   23   617.187   1252.08    4.8     45     100       387      4377
            nadir |   28  4.117857  17.27921     .1     .2      .2       .95        92
------------------+--------------------------------------------------------------------
1             age |   22  68.36364  5.678241     61     64      68        71        86
               ps |   20      76.5  11.82103     50     70      80        80       100
              bss |   20       2.8  .4103913      2      3       3         3         3
            grade |   17  2.235294   .752447      1      2       2         3         3
            pretx |   20    732.35  1357.341     25   69.5     174       530      4797
            nadir |   22  31.94091  52.49686     .5    1.2    10.5        38       183
------------------+--------------------------------------------------------------------
Total         age |   50     67.44  5.771711     58     63      66        70        86
               ps |   48  80.83333  11.07678     50     80      80        90       100
              bss |   48  2.520833  .6838434      1      2       3         3         3
            grade |   41  2.146341  .7924953      1      2       2         3         3
            pretx |   43  670.7512  1287.638    4.8     46     127       429      4797
            nadir |   50     16.36   39.2462     .1     .2     .95        10       183
-------------------------------------------------------------------------------------

. tabulate relap24 bss, row
```

```
+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

           |                bss
   relap24 |         1          2          3 |     Total
-----------+---------------------------------+----------
         0 |         5          9         14 |        28
           |     17.86      32.14      50.00 |    100.00
-----------+---------------------------------+----------
         1 |         0          4         16 |        20
           |      0.00      20.00      80.00 |    100.00
-----------+---------------------------------+----------
     Total |         5         13         30 |        48
           |     10.42      27.08      62.50 |    100.00


. tabulate relap24 grade, row

+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

           |               grade
   relap24 |         1          2          3 |     Total
-----------+---------------------------------+----------
         0 |         7          8          9 |        24
           |     29.17      33.33      37.50 |    100.00
-----------+---------------------------------+----------
         1 |         3          7          7 |        17
           |     17.65      41.18      41.18 |    100.00
-----------+---------------------------------+----------
     Total |        10         15         16 |        41
           |     24.39      36.59      39.02 |    100.00
```

Question 2a:

```
. logistic relap24 nadir ps bss

Logistic regression                               Number of obs   =         48
                                                  LR chi2(3)      =      15.10
                                                  Prob > chi2     =     0.0017
Log likelihood = -25.052835                       Pseudo R2       =     0.2315

------------------------------------------------------------------------------
     relap24 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       nadir |   1.033877   .0242646     1.42   0.156     .9873965    1.082545
          ps |   .9522044   .0314029    -1.49   0.138      .892603    1.015786
         bss |   2.624249    1.65688     1.53   0.126     .7613409    9.045463
------------------------------------------------------------------------------
```

Question 2b:

```
. g lnadir= log(nadir)
```

```
. logistic relap24 lnadir ps bss

Logistic regression                                     Number of obs   =         48
                                                        LR chi2(3)      =      28.17
                                                        Prob > chi2     =     0.0000
Log likelihood = -18.518315                             Pseudo R2       =     0.4320

------------------------------------------------------------------------------
     relap24 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lnadir |   2.362528   .6142753     3.31   0.001     1.419247    3.932745
          ps |   .9490084   .0347574    -1.43   0.153     .8832728    1.019636
         bss |   2.345473    1.92422     1.04   0.299     .4697933    11.70992
------------------------------------------------------------------------------
```

Question 2c:

```
. mkspline ndr1 1 ndr2 4 ndr3 16 ndr4= nadir

. logistic relap24 ndr1 ndr2 ndr3 ndr4 ps bss

Logistic regression                                     Number of obs   =         48
                                                        LR chi2(6)      =      31.17
                                                        Prob > chi2     =     0.0000
Log likelihood =  -17.01874                             Pseudo R2       =     0.4780

------------------------------------------------------------------------------
     relap24 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ndr1 |   29.61699   56.15361     1.79   0.074     .7205578    1217.343
        ndr2 |   .9034492   .5289488    -0.17   0.862     .2867773    2.846182
        ndr3 |   1.379926   .3083346     1.44   0.150      .890555    2.138211
        ndr4 |   .9818103   .0175701    -1.03   0.305     .9479706    1.016858
          ps |   .9367181   .0385458    -1.59   0.112      .864136    1.015397
         bss |    2.52226   2.294262     1.02   0.309      .424162     14.9985
------------------------------------------------------------------------------

. testparm ndr*

 ( 1)  [relap24]ndr1 = 0
 ( 2)  [relap24]ndr2 = 0
 ( 3)  [relap24]ndr3 = 0
 ( 4)  [relap24]ndr4 = 0

           chi2(  4) =   11.05
         Prob > chi2 =    0.0260

. g realps=ps
(2 missing values generated)

. g realbss=bss
(2 missing values generated)

. replace ps=80
(29 real changes made)

. replace bss=2
(37 real changes made)

. predict fit280
(option pr assumed; Pr(relap24))

. replace bss=3
(50 real changes made)
```

```
. predict fit380
(option pr assumed; Pr(relap24))

. replace ps=90
(50 real changes made)

. predict fit390
(option pr assumed; Pr(relap24))

. replace bss=2
(50 real changes made)

. predict fit290
(option pr assumed; Pr(relap24))

. replace ps= realps
(36 real changes made, 2 to missing)

. replace bss=realbss
(37 real changes made, 2 to missing)

twoway (scatter fit280 nadir, connect(l) sort(nadir) msymb(i) lp(dash) color(blue) )
       (scatter fit280 nadir if bss==2 & ps==80, msymb(d) color(blue)) ///
       (scatter fit380 nadir, connect(l) sort(nadir) msymb(i) lp(solid) color(blue)) ///
       (scatter fit380 nadir if bss==3 & ps==80, msymb(+) color(blue)) ///
       (scatter fit290 nadir, connect(l) sort(nadir) msymb(i) lp(dash) color(red)) ///
       (scatter fit290 nadir if bss==2 & ps==90, msymb(d) color(red))  ///
       (scatter fit390 nadir, connect(l) sort(nadir) msymb(i) lp(solid) color(red)) ///
       (scatter fit390 nadir if bss==3 & ps==90, msymb(+) color(red)), ///
       t1("Fitted Values Performance Status 80 and 90") ///
       ytitle("Fitted Probability of Relapse") ///
       xtitle("Nadir PSA (ng/ml)") xscale(log) xlabel(0.25 1 4 8 32 128) ///
       legend(label(1 BSS= 2 PS= 80) label(2 BSS= 2 PS= 80) ///
       label(3 BSS= 3 PS= 80) label(4 BSS= 3 PS= 80) ///
       label(5 BSS= 2 PS= 90) label(6 BSS= 2 PS= 90) ///
       label(7 BSS= 3 PS= 90) label(8 BSS= 3 PS= 90))
```

Question 2d:

```
. logit relap24 nadir ps bss

Iteration 0:   log likelihood = -32.601277
Iteration 1:   log likelihood = -25.316792
Iteration 2:   log likelihood = -25.060302
Iteration 3:   log likelihood = -25.052843
Iteration 4:   log likelihood = -25.052835
Iteration 5:   log likelihood = -25.052835

Logistic regression                               Number of obs   =         48
                                                  LR chi2(3)      =      15.10
                                                  Prob > chi2     =     0.0017
Log likelihood = -25.052835                       Pseudo R2       =     0.2315

------------------------------------------------------------------------------
     relap24 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       nadir |   .0333158   .0234695     1.42   0.156    -.0126836    .0793151
          ps |  -.0489755   .0329791    -1.49   0.138    -.1136134    .0156623
         bss |   .9647946   .6313732     1.53   0.126    -.2726741    2.202263
       _cons |   .7286543   3.178092     0.23   0.819    -5.500292    6.957601
------------------------------------------------------------------------------
```

```
. logit relap24 lnadir ps bss

Iteration 0:   log likelihood = -32.601277
Iteration 1:   log likelihood = -18.811778
Iteration 2:   log likelihood = -18.527518
Iteration 3:   log likelihood = -18.518323
Iteration 4:   log likelihood = -18.518315
Iteration 5:   log likelihood = -18.518315

Logistic regression                                     Number of obs   =         48
                                                        LR chi2(3)      =      28.17
                                                        Prob > chi2     =     0.0000
Log likelihood = -18.518315                             Pseudo R2       =     0.4320

------------------------------------------------------------------------------
     relap24 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lnadir |    .859732   .2600077     3.31   0.001     .3501263    1.369338
          ps |  -.0523377   .0366249    -1.43   0.153    -.1241212    .0194458
         bss |   .8524872   .8203975     1.04   0.299    -.7554624    2.460437
       _cons |   1.118636   3.724742     0.30   0.764    -6.181725    8.418997
------------------------------------------------------------------------------

. logit relap24 ndr1 ndr2 ndr3 ndr4 ps bss

Iteration 0:   log likelihood = -32.601277
Iteration 1:   log likelihood = -17.679096
Iteration 2:   log likelihood = -17.039807
Iteration 3:   log likelihood = -17.018764
Iteration 4:   log likelihood =  -17.01874
Iteration 5:   log likelihood =  -17.01874

Logistic regression                                     Number of obs   =         48
                                                        LR chi2(6)      =      31.17
                                                        Prob > chi2     =     0.0000
Log likelihood =  -17.01874                             Pseudo R2       =     0.4780

------------------------------------------------------------------------------
     relap24 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ndr1 |   3.388348   1.895993     1.79   0.074    -.3277297    7.104426
        ndr2 |  -.1015354   .5854771    -0.17   0.862    -1.249049    1.045979
        ndr3 |   .3220296   .2234429     1.44   0.150    -.1159104    .7599696
        ndr4 |  -.0183571   .0178956    -1.03   0.305    -.0534318    .0167175
          ps |  -.0653729   .0411499    -1.59   0.112    -.1460251    .0152794
         bss |   .9251551    .909606     1.02   0.309    -.8576399     2.70795
       _cons |  -.6791581    4.04664    -0.17   0.867    -8.610427    7.252111
------------------------------------------------------------------------------
```

Question 3a:

```
. regress nadir relap24 ps bss, robust

Linear regression                                       Number of obs =      48
                                                        F(  3,    44) =    2.47
                                                        Prob > F      =  0.0741
                                                        R-squared     =  0.1786
                                                        Root MSE      =  37.454

------------------------------------------------------------------------------
             |               Robust
```

```
       nadir |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     relap24 |   23.51765   11.43273     2.06   0.046     .4764958    46.55881
          ps |  -.5099777   .6183742    -0.82   0.414    -1.756229    .7362735
         bss |    6.84555   4.688718     1.46   0.151    -2.603939    16.29504
       _cons |    31.0281   53.12237     0.58   0.562      -76.033    138.0892
------------------------------------------------------------------------------
```

Question 3b:

```
. regress lnadir relap24 ps bss, robust

Linear regression                                      Number of obs =      48
                                                       F(  3,    44) =   14.67
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4611
                                                       Root MSE      =  1.6545

------------------------------------------------------------------------------
             |               Robust
      lnadir |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     relap24 |   2.614223   .5934664     4.41   0.000      1.41817    3.810275
          ps |  -.0072296   .0276556    -0.26   0.795    -.0629659    .0485066
         bss |   .4817541   .2977627     1.62   0.113    -.1183472    1.081855
       _cons |  -1.166384   2.496694    -0.47   0.643    -6.198141    3.865372
------------------------------------------------------------------------------

. glm lnadir relap24 ps bss, robust eform

Iteration 0:   log pseudolikelihood = -90.187567

Generalized linear models                         No. of obs      =        48
Optimization     : ML                             Residual df     =        44
                                                  Scale parameter =  2.737239
Deviance         =  120.4384976                   (1/df) Deviance =  2.737239
Pearson          =  120.4384976                   (1/df) Pearson  =  2.737239

Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =  3.924482
Log pseudolikelihood = -90.18756683               BIC             = -49.89435

------------------------------------------------------------------------------
             |               Robust
      lnadir |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     relap24 |   13.65659   7.841804     4.55   0.000     4.431705    42.08371
          ps |   .9927964   .0265657    -0.27   0.787     .9420704    1.046254
         bss |   1.618912   .4664133     1.67   0.094     .9204272    2.847455
------------------------------------------------------------------------------
```

Question 4a:

```
. logit relap24 ndr1 ndr2 ndr3 ndr4 ps bss

Iteration 0:   log likelihood = -32.601277
Iteration 1:   log likelihood = -17.679096
Iteration 2:   log likelihood = -17.039807
Iteration 3:   log likelihood = -17.018764
Iteration 4:   log likelihood =  -17.01874
```

```
Iteration 5:   log likelihood =  -17.01874

Logistic regression                                     Number of obs   =         48
                                                        LR chi2(6)      =      31.17
                                                        Prob > chi2     =     0.0000
Log likelihood =  -17.01874                             Pseudo R2       =     0.4780

------------------------------------------------------------------------------
     relap24 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ndr1 |   3.388348   1.895993     1.79   0.074    -.3277297    7.104426
        ndr2 |  -.1015354   .5854771    -0.17   0.862    -1.249049    1.045979
        ndr3 |   .3220296   .2234429     1.44   0.150    -.1159104    .7599696
        ndr4 |  -.0183571   .0178956    -1.03   0.305    -.0534318    .0167175
          ps |  -.0653729   .0411499    -1.59   0.112    -.1460251    .0152794
         bss |   .9251551    .909606     1.02   0.309    -.8576399     2.70795
       _cons |  -.6791581    4.04664    -0.17   0.867    -8.610427    7.252111
------------------------------------------------------------------------------

. test ndr1=ndr2=ndr3=ndr4

 ( 1)  [relap24]ndr1 - [relap24]ndr2 = 0
 ( 2)  [relap24]ndr1 - [relap24]ndr3 = 0
 ( 3)  [relap24]ndr1 - [relap24]ndr4 = 0

           chi2(  3) =    9.35
         Prob > chi2 =    0.0250

. logit relap24 nadir ndr1 ndr2 ndr3 ndr4 ps bss

note: ndr4 omitted because of collinearity
Iteration 0:   log likelihood = -32.601277
Iteration 1:   log likelihood = -17.679096
Iteration 2:   log likelihood = -17.039807
Iteration 3:   log likelihood = -17.018764
Iteration 4:   log likelihood =  -17.01874
Iteration 5:   log likelihood =  -17.01874

Logistic regression                                     Number of obs   =         48
                                                        LR chi2(6)      =      31.17
                                                        Prob > chi2     =     0.0000
Log likelihood =  -17.01874                             Pseudo R2       =     0.4780

------------------------------------------------------------------------------
     relap24 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       nadir |  -.0183571   .0178956    -1.03   0.305    -.0534318    .0167175
        ndr1 |   3.406705   1.898212     1.79   0.073     -.313722    7.127133
        ndr2 |  -.0831782   .5805131    -0.14   0.886    -1.220963    1.054607
        ndr3 |   .3403868   .2361048     1.44   0.149    -.1223701    .8031436
        ndr4 |  (omitted)
          ps |  -.0653729   .0411499    -1.59   0.112    -.1460251    .0152794
         bss |   .9251551    .909606     1.02   0.309    -.8576399     2.70795
       _cons |  -.6791581    4.04664    -0.17   0.867    -8.610427    7.252111
------------------------------------------------------------------------------

. testparm ndr*

 ( 1)  [relap24]ndr1 = 0
 ( 2)  [relap24]ndr2 = 0
 ( 3)  [relap24]ndr3 = 0

           chi2(  3) =    9.35
```

```
         Prob > chi2 =    0.0250

. logit relap24 lnadir ndr1 ndr2 ndr3 ndr4 ps bss

Iteration 0:   log likelihood = -32.601277
Iteration 1:   log likelihood = -17.669655
Iteration 2:   log likelihood = -17.046486
Iteration 3:   log likelihood = -17.018782
Iteration 4:   log likelihood = -17.018708
Iteration 5:   log likelihood = -17.018708

Logistic regression                               Number of obs   =         48
                                                  LR chi2(7)      =      31.17
                                                  Prob > chi2     =     0.0001
Log likelihood = -17.018708                       Pseudo R2       =     0.4780

------------------------------------------------------------------------------
     relap24 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lnadir |   .0292092   3.681405     0.01   0.994    -7.186212    7.244631
        ndr1 |   3.326554   8.014468     0.42   0.678    -12.38151    19.03462
        ndr2 |  -.1147405   1.764621    -0.07   0.948    -3.573334    3.343853
        ndr3 |   .3179469    .560666     0.57   0.571    -.7809383    1.416832
        ndr4 |  -.0187395   .0514125    -0.36   0.715    -.1195062    .0820273
          ps |  -.0654293   .0417651    -1.57   0.117    -.1472875    .0164289
         bss |   .9253011   .9098915     1.02   0.309    -.8580533    2.708656
       _cons |  -.6179445   8.712338    -0.07   0.943    -17.69381    16.45792
------------------------------------------------------------------------------

. testparm ndr*

 ( 1)  [relap24]ndr1 = 0
 ( 2)  [relap24]ndr2 = 0
 ( 3)  [relap24]ndr3 = 0
 ( 4)  [relap24]ndr4 = 0

           chi2(  4) =    3.11
         Prob > chi2 =    0.5399
```