Biost 517: Applied Biostatistics I

Emerson, Fall 2005

Homework #8 Key

December 11, 2005

Written problems: Not to be handed in—just for an example.

> *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.)*

Using the PSA data set, provide at least five distinct alternative analyses that might be used to explore an association between nadir PSA and time in relapse using this dataset. (Note: "Distinct analyses" must differ from each other with respect to the parameter compared across predictor groups, the contrast across treatment groups, or both.) In all analyses, provide descriptive plots (where appropriate) and as complete statistical inference as possible (i.e., provide point estimates, confidence intervals, and p values where possible, along with a statement of your scientific/statististical interpretation and conclusions).

Ans: In answering this homework, I must make decisions on the following issues:

- **The scientific variable (nadir PSA or time to relapse) which will be used as the "response variable" in the statistical analysis. The other scientific variable will be used as the "grouping variable" (predictor of interest).**

- **The summary measure of the distribution of response which will be used to compare the groups. General choices that I typically consider include the following (though I note that not all will be relevant in all cases of the choice of response variable):**
  - **the mean,**
  - **the geometric mean,**
  - **the median,**
  - **the probability of exceeding some scientifically relevant threshold,**
  - **the odds of exceeding some scientifically relevant threshold,**
  - **the (weighted) average of the hazard function, or**
  - **the probability of a randomly selected individual in one group having a measurement of response that exceeds that of a randomly selected individual in the other group.**

- **The basis of comparing the summary measure across groups (a difference or a ratio).**

- **The contrast to be considered across groups defined by the predictor. This is really most often determined by how I model the predictor. General choices that I typically consider include the following (but again I note that not all will be relevant in all cases of the choice of predictor variable):**
  - **dividing the sample into two groups on the basis of some scientifically relevant threshold for the predictor of interest,**
  - **modeling the predictor as an untransformed continuous variable in a regression model to assess a first order trend across multiple ordered groups (in this case the difference between values of the predictor is of greatest interest),**

- **modeling the predictor as a log transformed continuous variable in a regression model to assess a first order trend across multiple ordered groups (in this case the ratio between values of the predictor is of greatest interest),**
- **modeling the predictor according to some scientifically relevant transformation of a continuous variable in a regression model (for instance, we sometimes consider the inverse of creatine in renal disease, because it is related to glomerular filtration rate—but I would need to know about that), or**
- **(more complicated transformations of the predictor which will be covered in Biost 518, because they involve multiple regression).**

- **Technical statistical considerations about exact form of the statistic used for inference. These include**
  - **Choice of probability model (nonparametric, semiparametric, or parametric). In this class I have stressed models that can be regarded as nonparametric. I have done this by urging the use of methods which allow for unequal variances, nonlinearity of effects, and nonproportional hazards.**
  - **Handling of correlated data. (In this homework, all subjects were independent, so I will not discuss this point further.)**
  - **Choice of statistic. There are several variants that we could consider here:**
    - **Exact distributions versus large sample approximations. This comes into play primarily with binary response data. In one sample data, we know the sampling distribution exactly. In two sample data, we can examine the "worst case" exact distribution under the null hypothesis (the adjusted tests that Stata does not do, but something like StatExact does).**
    - **Calculation of standard errors from formulas or by resampling (bootstrapping). We tend to use formulas most times, but in this class we used bootstrapping for the sample median, because the formula was too difficult to use.**
    - **The choice of Wald, score, or likelihood ratio statistics. This comes into play primarily in regression models, though I did mention this issue in the two sample binomial model (where the chi-squared test is a score test, but a likelihood ratio test could also be used) and with two sample censored time to event data (where the logrank test corresponds to the score test in the proportional hazards regression model). Next quarter we will examine the relative behavior of these three asymptotically equivalent tests in more detail.**

Possible analyses that can be considered are given below.

<u>Using Nadir PSA as the Response Variable</u>

I can compare the distribution of nadir PSA across groups defined by time to relapse. Nadir PSA is a continuous variable, and I can consider any of the summary measures given above. Time to relapse was measured subject to censoring, so I am limited in how I can define the groups. Choices I might consider include

- **Dividing subjects into two groups based on whether they relapsed within 24 months or not. This is an option only because the earliest censoring occurred at 24 months. I will call this variable "`relapse24`". It can be computed in Stata using the code**

  ```
  ▪ g relapse24 = 1
  ▪ replace relapse24 = 0 if obstime > 24 | inrem=="yes"
  ```

- **Dividing subjects into groups according to the number of months they were in remission during the first 24 months. Again, this an option only because the earliest censoring occurred at 24 months. I will call this variable "`remtime24`". It can be computed in Stata using the code**

  ```
  1. g remtime24= obstime
  2. replace remtime24= 24 if obstime > 24
  ```

So now the analyses I can consider are

Inference on Mean Nadir PSA

1. **Comparing mean nadir PSA across groups defined by whether they relapsed within 24 months or not. I would do this using a t test which allows unequal variances. The relevant estimate would be the estimated difference in mean nadir PSA between the two groups. Stata code: Either of the following would be acceptable, because they are roughly equivalent in large samples. But the t test would be a little more "standard":**

   ```
   ▪ ttest nadir, by(relapse24) unequal
   ▪ regress nadir relapse24, robust
   ```

2. **Examining the first order trend in mean nadir PSA across groups defined by the number of months spent in remission during the first 24 months post treatment using a contrast comparing the difference in times in remission. I would do this using a linear regression of the nadir PSA on the untransformed variable measuring time in remission during the first 24 months. The slope parameter would be the estimated difference in mean nadir PSA between two groups which differ by 1 month in their time spent in remission during the first 24 months. I would use robust standard error estimates. Stata code:**

   ```
   ▪ regress nadir remtime24, robust
   ```

3. **Examining the first order trend in mean nadir PSA across groups defined by the number of months spent in remission during the first 24 months post treatment using a contrast comparing the ratio of times in remission. I would do this using a linear regression of the nadir PSA on the log transformed variable measuring time in remission during the first 24 months. The slope parameter multiplied by log(2) would be the estimated difference in mean nadir PSA between two groups in which one group spent twice as long in remission during the first 24 months than the other group. I would use robust standard error estimates. Stata code:**

   ```
   ▪ g logrmtm24= log(remtime24)
   ▪ regress nadir logrmtm24, robust
   ```

Inference on Geometric Mean Nadir PSA

I will need to perform statistical analyses on log transformed nadir PSA using the following code

```
▪ g lognadir= log(nadir)
```

1. **Comparing geometric mean nadir PSA across groups defined by whether they relapsed within 24 months or not. I would do this using a t test which allows unequal variances on the log transformed nadir PSA. The exponentiated estimated difference in mean log nadir PSA would be the estimated ratio of geometric mean nadir PSA between the two groups. Stata code: Either**

of the following would be acceptable, because they are roughly equivalent in large samples. But the t test would be a little more "standard":

- `ttest lognadir, by(relapse24) unequal`
- `regress lognadir relapse24, robust`

2. **Examining the first order trend in geometric mean nadir PSA across groups defined by the number of months spent in remission during the first 24 months post treatment using a contrast comparing the difference in times in remission. I would do this using a linear regression of the log nadir PSA on the untransformed variable measuring time in remission during the first 24 months. The exponentiated slope parameter would be the estimated ratio of geometric mean nadir PSA between two groups which differ by 1 month in their time spent in remission during the first 24 months. I would use robust standard error estimates. Stata code:**
   - `regress lognadir remtime24, robust`
3. **Examining the first order trend in geometric mean nadir PSA across groups defined by the number of months spent in remission during the first 24 months post treatment using a contrast comparing the ratio of times in remission. I would do this using a linear regression of the nadir PSA on the log transformed variable measuring time in remission during the first 24 months. The 2 raised to the power of the slope parameter would be the estimated ratio of geometric mean nadir PSA between two groups in which one group spent twice as long in remission during the first 24 months than the other group. I would use robust standard error estimates. Stata code:**
   - `g logrmtm24= log(remtime24)`
   - `regress lognadir logrmtm24, robust`

Inference on Median Nadir PSA

I will need to use bootstrapping to get the standard errors. Also, we have not covered any methods that could be used to do this in a regression setting, so I will only consider the two sample problem.

1. **Comparing median nadir PSA across groups defined by whether they relapsed within 24 months or not. I would do this using the sample median computed in each group separately, and bootstrapped estimates of SEs also computed for each group. The analysis would proceed as was done in Homework # 6, problem 3. The difference in sample medians would estimate the difference in population medians.**
2. **(I could use the ratio of medians, using the formulas for the SE of a ratio as given in class.)**

Inference on Proportion with Nadir PSA Exceeding 4 ng/ml

I will need to perform statistical analyses on dichotomized nadir PSA using the following code

- `g nadirgt4= 0`
- `replace nadirgt4= 1 if nadir > 4`

1. **Comparing proportion with high nadir PSA (defined here as exceeding 4 ng/ml) across groups defined by whether they relapsed within 24 months or not. I would do this using a chi squared test (alternatively a Fisher's exact test). The estimated difference in proportions will estimate the corresponding quantity in the population. Stata code: Either of the following would be acceptable, because they are roughly equivalent in large samples. The most popular choice would be to use the chi square test because the event rate is sufficiently high to meet the rule of thumb criterion. Some people would use Fisher's exact test in small samples, but as discussed in class, it is better to use an adjusted version of this test.**
   - `cs nadirgt4 relapse24`

- `cs nadirgt4 relapse24, exact`

Inference on Odds of Having Nadir PSA Exceeding 4 ng/ml

I will need to perform statistical analyses on dichotomized nadir PSA using the following code

- `g nadirgt4= 0`
- `replace nadirgt4= 1 if nadir > 4`

1. **Comparing odds of having high nadir PSA (defined here as exceeding 4 ng/ml) across groups defined by whether they relapsed within 24 months or not. I would do this using logistic regression on the binary predictor. There is no real need (nor a real disadvantage) in using the robust SE with a binary predictor. The odds ratio will compare the odds of high nadir for the group relapsing early to the odds of high nadir for the group remaining in remission for 24 months. Stata code: Using the logistic regression command that does not demand back transformation**
   - `logistic nadirgt4 relapse24`
2. **Comparing odds of having high nadir PSA (defined here as exceeding 4 ng/ml) across groups defined by defined by the number of months spent in remission during the first 24 months post treatment using a contrast comparing the difference in times in remission. I would do this using a logistic regression on the untransformed variable measuring time in remission during the first 24 months. I would use the robust SE to protect against nonlinearity causing inaccurate inference about a first order trend. The odds ratio will compare the odds of high nadir between two groups which differ by 1 month in their time spent in remission during the first 24 months. Stata code:**
   - `logistic nadirgt4 remtime24, robust`
3. **(I suppose I could do logistic regression on the log transformed remtime24 variable, as well.)**

Inference on Hazard Function for Distribution of Nadir PSA

These analyses would be extremely unusual, but my point is they could be done. Providing an intuitive interpretation for the hazard is very difficult when the variable is not measuring time. (What does it mean to be "at risk for a nadir PSA of 7" merely because your PSA was not less than 7?) For these analyses, I have to create a variable that indicates that none of the nadir measurements were censored.

- `g nocensor= 1`

1. **Comparing hazard functions for nadir PSA across groups defined by whether they relapsed within 24 months or not. I would do this using proportional regression on the binary predictor. There is no real need (nor a real disadvantage) in using the robust SE with a binary predictor. The hazard ratio will compare the instantaneous "risk" of high nadir (whatever that might mean) for the group relapsing early to that of the group remaining in remission for 24 months. I note that a hazard ratio less than 1 will indicate that the group relapsing early will tend to have higher nadir PSAs. Stata code:**
   - `stset nadir nocensor`
   - `stcox relapse24`
2. **Comparing hazard functions for nadir PSA across groups defined by defined by the number of months spent in remission during the first 24 months post treatment using a contrast comparing the difference in times in remission. I would do this using proportional hazards regression on the untransformed variable measuring time in remission during the first 24**

months. I would use the robust SE to protect against nonlinearity causing inaccurate inference about a first order trend. The hazard ratio will compare the instantaneous "risk" of high nadir (whatever that might mean) for the group relapsing early to that of the group remaining in remission for 24 months. I note that a hazard ratio less than 1 will indicate that the group relapsing early will tend to have higher nadir PSAs. Stata code:

- `stset nadir nocensor`
- `stcox remtime24, robust`

3. **(I suppose I could do proportional hazards regression on the log transformed remtime24 variable, as well.)**

Inference on Probability of Nadir PSA in Relapse Group Exceeding that in Remission Group

1. **Comparing to 0.5 the probability that a randomly chosen patient who relapsed within 24 months would have a higher nadir PSA than a randomly chosen patient in the group that remained in remission for 24 months. I would do this using the Wilcoxon rank sum test. There are no corresponding estimates of scientific interest. Stat command:**
   - **ranksum nadir, by(relapse24)**

<u>Using Time to Relapse as the Response Variable</u>

Some version of the following would actually be my preferred analyses, because they are using the present to predict the future. That is, when we measure the nadir PSA, we want to know about the future. (In the previous section, we were looking at groups defined by whether a patient had relapsed, and then looked back in time to see what the nadir PSA was.) The most standard of the following analyses would be the proportional hazards regression. Furthermore, knowing the behavior of PSA from other studies, I would prefer the log transformation of the predictor.

I can compare the distribution of time to relapse across groups defined by nadir PSA. Time to relapse was measured subject to censoring, and that will affect the degree to which I can consider any of the summary measures given above. Because the earliest censoring was at 24 months, I can consider dichotomizing the data at any time prior to 24 months without problem (I will use `relapse24`, because I tend to think longer times are more clinically relevant). I can also consider the uncensored variable measuring the number of months in remission during the first 24 months following therapy (as created in the variable `remtime24` above). The mean of this variable would be called a "24 month resticted mean" for the time in remission. Other analyses will use the censored measurements as recorded in `obstime` and `relapse`.

Nadir PSA is a continuous measurement, and thus I can consider dichotomizations of the random variable (I will consider `nadirgt4` defined above) and log transformations (as created in `lognadir` above) in addition to modeling of the untransformed random variable.oring occurred at 24 months. I will call this variable "`relapse24`". It can be computed in Stata using the code

So now the analyses I can consider are

Inference on Restricted Mean Time to Relapse

1. **Comparing 24 month restricted mean time to relapse across groups defined by whether nadir PSA was greater than 4 ng/ml or not. I would do this using a t test which allows unequal**

variances. The relevant estimate would be the estimated difference in 24 month restricted mean time to relapse between the two groups. Stata code: Either of the following would be acceptable, because they are roughly equivalent in large samples. But the t test would be a little more "standard":

- `ttest remtime24, by(nadirgt4) unequal`
- `regress remtime24 nadirgt4, robust`

2. **Examining the first order trend in 24 month restricted mean time to relapse across groups defined by nadir PSA using a contrast comparing the difference in nadir PSA as a continuous variable. I would do this using a linear regression on the untransformed variable measuring nadir PSA. The slope parameter would be the estimated difference in restricted mean time to relapse between two groups which differ by 1 ng/ml in their nadir PSA. I would use robust standard error estimates. Stata code:**
   - `regress remtime24 nadir, robust`
3. **Examining the first order trend in24 month restricted mean time to relapse across groups defined by nadir PSA using a contrast comparing the ratio of nadir PSA. I would do this using a linear regression of the variable measuring months in remission during the first 24 months on the log transformed variable measuring nadir PSA. The slope parameter multiplied by log(2) would be the estimated difference in restricted mean time to relapse between two groups in which one grouphad a nadir PSA twice as high as the other group. I would use robust standard error estimates. Stata code:**
   - `regress remtime24 lognadir, robust`
4. **(There are methods to do analysis on restricted means defined up to time points that might include some censored observations. These methods would use the area under the Kaplan-Meier curve. Some statistical packages provide such estimates, but I don't know of any giving inference.)**

Inference on Restricted Geometric Mean Time to Relapse

Analyses analogous to the above can also be done on log transformed variables.

- `g logrmtm24= log(remtime24)`

Inference on Median (Reseticted) Time in Remission

Analyses using the median can be done on the variable measured only over the first 24 months, in which case the analyses will proceed exactly as in Homework #6, problem 3. On the other hand, providing the censoring distribution allows an estimate of the unrestricted time in remission, Kaplan-Meier estimates of the median can be used. Again, bootstrapping will be the best bet for computing the SEs.

Inference on Proportion with (or Odds of) Time in Remission Exceeding Some Threshold

If the time chosen is prior to any censoring, e.g 24 months in this sample, the analysis can proceed exactly as above (i.e., chi squared test if nadir PSA was dichotomized, logistic regression if not). On the other hand, if the time threshold is later than the earliest censoring, the Kaplan-Meier estimates will need to be used, and the analysis would have to proceed as in Homework #7, problem 1 for a dichotomized predictor variable.

1. **In particular, I would consider comparing the odds of remaining in remission for 24 months across groups defined by nadir PSA as a dichotomized variable, an untransformed variable, or a log transformed variable. Stata code:**

- **logistic relapse24 nadirgt4**
- **logistic relapse24 nadir, robust**
- **logistic relapse24 lognadir, robust**

Inference on Hazard of Relapse

1. **Comparing hazard of relapse across groups defined by nadir PSA. I would do this using proportional regression on the binary predictor, the untransformed predictor, or the log transformed predictor. Stata code:**
    - **`stset obstime relapse`**
    - **`stcox nadirgt4`**
    - **`stcox nadir, robust`**
    - **`stcox lognadir`**