**Biost 518: Applied Biostatistics II**
Emerson, Winter 2008

**Homework #2 Key**
February 6, 2008

**A file containing the annotated Stata code used to answer these questions is available on the class web pages.**

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.

**Ans: The following table presents descriptive statistics for measured covariates according to whether or not the patient had relapse of his cancer within 24 months.** *(Note that many would have provided frequencies for the bone scan score and tumor grade instead of the averages I provided. Had I just been presenting the statistics for all patients, I certainly would have taken that tack. In this case, I consider it a toss-up, but there is a chance that some referee would complain.)*

| | Remission > 24 mos | Relapse within 24 mos | All Patients |
|---|---|---|---|
| | Mean (SD; Min – Max) | Mean (SD; Min – Max) | Mean (SD; Min – Max) |
| **Age (years)** | 66.7 (5.84; 58.0 - 81.0) | 68.4 (5.68; 61.0 - 86.0) | 67.4 (5.77; 58.0 - 86.0) |
| **Performance Status** | 83.9 (9.56; 50.0 - 100.0) | 76.5 (11.82; 50.0 - 100.0) | 80.8 (11.08; 50.0 - 100.0) |
| **Bone Scan Score** | 2.32 (0.77; 1.0 - 3.0) | 2.80 (0.41; 2.0 - 3.0) | 2.52 (0.68; 1.0 - 3.0) |
| **Tumor Grade** | 2.08 (0.83; 1.0 - 3.0) | 2.24 (0.75; 1.0 - 3.0) | 2.15 (0.79; 1.0 - 3.0) |
| **Pretreat PSA (ng/ml)** | 617 (1252; 5 - 4377) | 732 (1357; 25 - 4797) | 671 (1288; 5 - 4797) |
| **Nadir PSA (ng/ml)** | 4.1 (17.28; 0.1 - 92.0) | 31.9 (52.50; 0.5 - 183.0) | 16.4 (39.25; 0.1 - 183.0) |

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.

**Ans: The following table provides the desired descriptive and inferential statistics.**

| | | Nadir PSA (ng/ml) | | |
|---|---|---|---|---|
| | n | Mean | SD | 95% CI for Mean |
| **Bone Scan Score < 3** | 18 | 3.5 | 2.11 | (-0.9, 8.0) |

| | | | | |
|---|---|---|---|---|
| **Bone Scan Score = 3** | 30 | 24.9 | 8.88 | (6.7, 43.0) |

b. Using a t test which presumes equal variances across groups, provide inference regarding an association between nadir PSA and bone scan score.

**Ans: The mean nadir PSA among the patients having a bone scan score of 3 was 24.9 ng/ml (95% CI 6.7 to 43.0 ng/ml), while that for the patients having a lower bone score was 3.5 ng/ml (95% CI 0 to 8.0 ng/ml). However, this observed tendency for the nadir PSA to be 21.3 ng/ml higher in the patients with the more involved bone marrow was not beyond that that might be observed due to random chance (P = 0.074; 95% CI 2.1 ng/ml lower to 44.7 ng/ml higher nadir PSA in the group with higher bone scan score).** *(Note that in the text I decided not to report an impossible value for the CI. This is legit: You can always leave out impossible values from your CI and still have just as accurate a statement. Of course, when you get a CI that includes impossible values, you probably ought to worry that your sample size is too small to rely on the normal approximation for your sample mean.)*

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.)*

**Ans: The observed standard deviations were quite different between the two groups, with the group that had the larger sample size also having the larger standard deviation. In such a situation, the t test that presumes equal variances tends to be too conservative in that it would reject the null hypothesis (either weak or null) with probability less than 0.05. Now, if the strong null hypothesis were true, then the variances would have to be equal, and the test would reject the strong null hypothesis with the correct type I error. However, if we were considering the strong null hypothesis, then a situation in which the means were equal but variances were unequal would represent an alternative hypothesis. The fact that the true type I error would be less than 0.05 in such a situation means that the t test that presumes equal variance is a 'biased test'. It is also inconsistent: Even with an infinite sample size, it would not always reject the null strong null hypothesis when the strong null is not true.**

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?

**Ans: The mean nadir PSA among the patients having a bone scan score of 3 was 24.9 ng/ml (95% CI 6.7 to 43.0 ng/ml), while that for the patients having a lower bone score was 3.5 ng/ml (95% CI 0 to 8.0 ng/ml). This observed tendency for the nadir PSA to be 21.3 ng/ml higher in the patients with the more involved bone marrow was beyond that that might be observed due**

**to random chance (P = 0.026; 95% CI 2.71 to 39.9 ng/ml higher nadir PSA in the group with higher bone scan score). The use of the t test that allows unequal variances would be expected to provide greater precision when the variances are not equal and the group with the higher variance has a higher sample size. The fact that this test (asymptotically) has the correct type I error when the means are equal means that I can definitely claim with 95% confidence that the two groups have different means. In the most general case, when using the t test that presumes equal variances can only be interpreted as rejecting the strong null (the distributions are different) with high confidence, as opposed to rejecting the weak null (the means are different). However, as the sample size was larger in the group with the larger sample variance, we probably do not have to worry about our rejection of the null hypothesis merely indicating a difference in variances.**

e. Using classical linear regression (so without robust standard errors), provide inference regarding an association between nadir PSA and bone scan score.

**Ans: From linear regression we estimate a tendency for the nadir PSA to be 21.3 ng/ml higher in the patients with the more involved bone marrow, however that was not beyond that that might be observed due to random chance (P = 0.074; 95% CI 2.1 ng/ml lower to 44.7 ng/ml higher nadir PSA in the group with higher 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.

**Ans: The estimated intercept is exactly the sample mean for the low bone scan score group. The estimated slope is exactly the difference in sample means when subtracting the low bone scan score sample mean from the high bone scan score sample mean.**

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.

**Ans: Even though the interpretation of the intercept is that of the true mean for the low bone scan score and the estimated intercept is the sample mean, the 95% CI computed from regression does not at all agree with that computed based on a single sample. This is because when constructing the CI for the intercept, linear regression uses a estimate of the within group variance that is pooled from both bone scan score groups.**

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?

**Ans: The inference about the slope from classical linear regression is exactly the same as that from the t test that presumes equal variances. Because of this, a rejection of the null hypothesis must be interpreted as rejection of the strong null, rather than the weak null.**

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

**Ans: From linear regression (with robust standard errors) we estimate a tendency for the nadir PSA to be 21.3 ng/ml higher in the patients with the more involved bone marrow, and that beyond that that might be observed due to random chance (P = 0.025; 95% CI 2.86 lower to 39.8 ng/ml higher nadir PSA in the group with higher 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.

**Ans: The relationship between the regression parameter estimates and the sample means is unchanged: Use of robust SE does not change the intercept or slope estimates. The robust SE approach does not use a pooled variance, so the inference about the slope is very nearly the same as that from the t test which allows for the possibility of unequal variances. For this same reason, the 95% CI for the intercept is in fairly close agreement with the 95% CI for the low bone score group mean when the latter was calculated using one sample methods. I note, however, that there are some very slight differences between the way the standard errors are computed when based on a single sample and when using linear regression with robust SE. Mostly this has to do with the number of degrees of freedom used for the critical value in the CI, but we also handle differently the question of n vs n-p in estimating the within group variance.**

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.

**Ans: The following table provides the desired descriptive and inferential statistics.** *(There are certainly ample signs that we probably have too small a sample size to trust our inference in a couple of these groups. Note that there is nothing in our data to suggest that the group with a bone scan score of 1 does not have a normal distribution with a very small variance. I am afraid we just have to recognize that 5 is a pretty small number.)*

| | | Nadir PSA (ng/ml) | | |
|---|---|---|---|---|
| | **n** | **Mean** | **SD** | **95% CI for Mean** |
| **Bone Scan Score =1** | 5 | 0.2 | 0.00 | (0.2, 0.2) |
| **Bone Scan Score = 2** | 13 | 4.8 | 10.34 | (-1.4, 11.1) |
| **Bone Scan Score = 3** | 30 | 24.9 | 8.88 | (6.7, 43.0) |

   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.

**Ans: Interpreting the slope: From linear regression (with robust standard errors) we estimate a tendency for the nadir PSA to be on average 14.7 ng/ml higher for every one unit difference in bone scan scores, with the patients with the more involved bone marrow having the higher mean nadir PSA, and that is beyond that that might be observed due to random chance (P = 0.018; 95% CI 2.7 lower to 26.7 ng/ml higher nadir PSA per 1 unit difference in bone scan score between groups).**

**Interpreting the intercept: The estimated mean nadir PSA in a group having a bone scan score of 0 is -20.2 ng/ml. Because there is no such possible bone scan score, it is not entirely surprising that the estimate is not a possible measurement of PSA. In this case, the intercept is just a mathematical construct to describe the best fitting straight line. It is certainly of no scientific interest in and of itself.**

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.

**Ans: From the linear regression model, we estimate an average nadir PSA level of -5.49, 9.21, or 23.9 ng/ml, respectively, in groups having bone scan scores of 1, 2, or 3. Clearly these do not agree with the sample means found in part a, and the fitted value for the bone scan score of 1 is impossible. This happens because the sample means do not appear to lie on a straight line.**

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.

**Ans: If our goal is merely to define an association using the first order trend in means across bone scan score groups, we will have more precision to detect such a trend quite often fitting a straight line to the data using a predictor modeled continuously, even when the observed trend appears to be nonlinear. This will tend to be true unless there is a quite large "step" in the pattern of the sample means AND we are good at dichotomizing the data exactly at that step (without looking at the data).** *(We will discuss this tradeoff in modeling of the predictor more later in the course.)*

**If our goal is estimating the group means, then nonlinearity causes the linear regression model to provide poor fits, because borrowing data across groups in a linear fashion is not reasonable. Of course, the dichotomized predictor *bss3* would not allow fitting separate values to the bone scan score 1 and bone scan score 2 groups. We will later in the course discuss the use of "dummy variables" (separate indicator variables for each group) in order to estimate each group separately.**

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?

**Ans: Patients were followed for variable amounts of time. So the *inrem* variable is indicating whether a patient relapsed over variable amounts of time.**

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

**Ans: The earliest censoring observation is at 24 months. So we have complete information for every patient regarding his remission status at 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.

**Ans: The following table provides the desired descriptive and inferential statistics.** *(I provided CI computed four ways: Using the exact binomial distribution, using the asymptotic distribution for the estimated proportion while ignoring the mean-variance relationship, using the score statistic (which is equivalent to using the asymptotic distribution for the estimated proportion while*

*accounting for the mean-variance relationship), and using the asymptotic distribution for the estimated log odds while ignoring the asymptotic distribution. The exact method is generally regarded as the best. Just as an aside: There are truly an infinite number of valid 95% CI, and we usually try to find the one that is shortest. In this case, however, there is the distinct possibility that the sample size is too small for the asymptotic CI to have the correct coverage probability for the observed sample proportion. This is especially true for the low bone scan score group, which has a proportion further from 0.5 and lower sample size, and this probably is the reason behind the greater disagreement between the asymptotic methods and the exact methods.)*

| | n | Prob | 95% CI for Prob | Odds | 95% CI for Odds |
|---|---|---|---|---|---|
| **Bone Scan Score < 3** | 18 | 0.222 | Exact (0.064, 0.476)<br>Asym – p (0.030, 0.414)<br>Score (0.074, 0.481)<br>Asym – lo (0.086, 0.465) | 0.286 | Exact (0.068, 0.910)<br>Asym – p (0.031, 0.707)<br>Score (0.080, 0.926)<br>Asym – lo (0.094, 0.868) |
| **Bone Scan Score = 3** | 30 | 0.533 | Exact (0.343, 0.717)<br>Asym – p (0.355, 0.712)<br>Score (0.346, 0.712)<br>Asym – lo (0.358, 0.701) | 1.14 | Exact (0.523, 2.53)<br>Asym – p (0.550, 2.47)<br>Score (0.530, 2.47)<br>Asym – lo (0.558, 2.34) |

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

**Ans: The proportion of patients relapsing within 24 months among the patients having a bone scan score of 3 was 0.533 (95% CI 0.343 to 0.717), while that for the patients having a lower bone score was 0.222 (95% CI 0.064 to 0.476). This observed absolute difference of relapse rates of 0.311 was beyond that that would be expected due to random chance in the absence of a true difference in probability of relapse (P = 0.0343 from the chi square test; 95% CI 0.049 to 0.573 absolute difference in relapse rates with the higher relapse probability in the group with higher bone scan score).** *(Note that I chose to view the chi squared test as a test of the difference in proportions, but it is just as correct to view it as a test of the odds ratio. I also chose to report the exact CI for each of the individual relapse probabilities. Note that there is substantial overlap between the two CI (no matter how they were calculated), but that the chi squared test is statistically significant. The CI reported for the difference in relapse rates is computed using an asymptotic distribution.)*

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.

**Ans: The proportion of patients relapsing within 24 months among the patients having a bone scan score of 3 was 0.533 (95% CI 0.344 to 0.723), while that for the patients having a lower bone score was 0.222 (95% CI 0.009 to 0.435). This observed absolute difference of relapse rates of 0.311 was beyond that that would be expected due to random chance in the absence of a true difference in probability of relapse (P = 0.0347 from the chi square test; 95% CI 0.023 to 0.599 absolute difference in relapse rates with the higher relapse probability in the group with higher bone scan score).** *(Note that because I was asking you to take the rather unorthodox approach of using the t test here, I chose to report the CI for each of the individual relapse probabilities based on the t statistic as well. Those CI would use a critical value based on the t distribution rather than the standard normal, as well as using a slightly different standard error estimate.)*

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.

**Ans: The proportion of patients relapsing within 24 months among the patients having a bone scan score of 3 was 0.533 (95% CI 0.344 to 0.723), while that for the patients having a lower bone score was 0.222 (95% CI 0.009 to 0.435). This observed absolute difference of relapse rates of 0.311 was beyond that that would be expected due to random chance in the absence of a true difference in probability of relapse (P = 0.0284 from the chi square test; 95% CI 0.035 to 0.588 absolute difference in relapse rates with the higher relapse probability in the group with higher bone scan score).**

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.

**Ans: The chi squared test is the most standard of these three analyses, though in sufficiently large samples, there would be little difference between them. With respect to the p values obtained, the t test that presumes equal variances would tend to be in closest agreement with the chi square test, because under the null hypothesis of no differences in proportions, the variance would be the same in both groups. This is because with a comparison of independent binary data across two groups, there is no difference between the weak and strong null hypotheses. However, with respect to the CI, the t test that allows for unequal variances would tend to be in closest agreement with the asymptotic CI for the difference in proportions, because under alternative hypotheses, the variances would be unequal.**

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.)

**Ans: From the logistic regression model we estimate that the odds of patients relapsing within 24 months among the patients having a bone scan score of 3 was 1.14, which corresponds to a proportion 0.533 relapsing within 24 months. In patients having a lower bone score, the estimated odds of patients relapsing within 24 months was 0.286 (95% CI 0.094 to 0.868), which corresponds to a proportion 0.222 relapsing within 24 months. From logistic regression, we estimate an odds ratio of 4.00, and this observed odds ratio is beyond that that would be expected due to random chance in the absence of a true difference in odds of relapse (P = 0.040; 95% CI 1.07 to 15.0).**

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.

**Ans: The proportions and odds estimated for each group using logistic regression are exactly equal to the sample proportions: The intercept from the logistic regression corresponds exactly to the log of the sample odds for the group with low bone scan scores, and the slope from the logistic regression corresponds exactly to the log of the ratio of the odds for the high bone scan score group to the odds for the low bone scan score group as calculated from descriptive statistics derived independently for each group.**

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.

**Ans: The 95% CI for the intercept in the logistic regression agrees exactly with the CI calculated for the low bone scan score group using the asymptotic distribution of the log odds. This is to be expected, because there is the linearity of the log odds must hold exactly in a two sample comparison (two points make a line), and logistic regression uses model based estimates of variability (i.e., the estimated variability comes from the mean-variance relationship for binary data when using the predicted proportion derived from the logistic regression model).**

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.

**Ans: The p value for the slope agrees approximately with the p value from the chi squared test. The chi squared test is equivalent to the score test from logistic regression with a binary predictor, while the p value reported in part f represents a Wald test computed from the slope estimate and its estimated standard error. Asymptotically, these tests are equivalent under the null hypothesis, but in small samples they will differ somewhat. As noted above, the t tests are also asymptotically equivalent to the chi squared test under the null hypothesis.** *(The logistic regression output also gave a p value for the likelihood ratio test, which was 0.030. Asymptotically, this is equivalent to all the others under the null hypothesis.)*

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

**Ans: From logistic regression with robust standard errors, we estimate an odds ratio of 4.00, and this observed odds ratio is beyond that that would be expected due to random chance in the absence of a true difference in odds of relapse (P = 0.042; 95% CI 1.05 to 15.2).**

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.

**Ans: When using a binary predictor, there is very little difference between using classical logistic regression or using logistic regression with robust standard errors. This is because the model based variance estimates have to be correct. So with a binary predictor, there is no advantage in using the robust standard errors.** *(With more than two groups, the robust standard errors can address possible anti-conservative behavior of logistic regression in the presence of nonlinearity of the log odds across groups. But as two points make a line, there is no problem with a binary predictor.)*

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.

**Ans: The following table provides the desired descriptive and inferential statistics.** *(There are certainly ample signs that we probably have too small a sample size to trust our inference in a couple of these groups. Note that I only provide exact CI in this table, having illustrated in problem 4 the different CI we can obtain when different methods are used.)*

| | n | Prob | 95% CI for Prob | Odds | 95% CI for Odds |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| **Bone Scan Score = 1** | 5 | 0.000 | (0.000, 0.522) | 0.000 | (0.000, 1.09) |
| **Bone Scan Score = 2** | 13 | 0.308 | (0.091, 0.614) | 0.444 | (0.100, 1.59) |
| **Bone Scan Score = 3** | 30 | 0.533 | (0.342, 0.717) | 1.14 | (0.522, 2.53) |

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.

**Ans: Interpreting the slope: From classical logistic regression we estimate a tendency for the odds of relapse within 24 months to be 3.71 times higher for every one unit difference in bone scan scores between two groups, with the patients with the more involved bone marrow having the higher odds of relapse, and that is beyond that that might be observed due to random chance (P = 0.025; 95% CI 1.18 to 11.7 times higher odds of relapse per 1 unit difference in bone scan score between groups).**

**Interpreting the intercept: By exponentiating the intercept from logistic regression, we estimate that the odds of relapse within 24 months would be 0.0237 in a group having a bone scan score of 0. Because there is no such possible bone scan score, the intercept is just a mathematical construct to describe the best fitting straight line. It is certainly of no scientific interest in and of itself.**

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.

**Ans: From the logistic regression model, we estimate that the odds of relapse within 24 months would be 0.088, 0.326, and 1.21 in groups having bone scan scores of 1, 2, and 3, respectively. These estimates would correspond to estimated probability of relapse of 0.081, 0.246, and 0.547, respectively. Any lack of agreement between these estimates and those presented in part a can be attributed to nonlinearity of the log odds across the groups.**

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.

**Ans: From classical logistic regression with robust standard errors, we estimate a tendency for the odds of relapse within 24 months to be 3.71 times higher for every one unit difference in bone scan scores between two groups, with the patients with the more involved bone marrow having the higher odds of relapse, and that is beyond that that might be observed due to random chance (P = 0.010; 95% CI 1.37 to 10.0 times higher odds of relapse per 1 unit difference in bone scan score between groups).**

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.

**Ans: When performing logistic regression across more than two groups (i.e., with something other than a single binary predictor), the model based estimates of within group variability can be incorrect due to nonlinearity in the log odds across predictor groups. In this case, the robust standard errors led to a smaller p value. While we must worry that our sample sizes are not**

**large enough to allow the robust standard error estimates to be reliable, this may also be a reflection of a more accurate handling of any nonlinearity in the log odds.**

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.

**Ans: The mean nadir PSA among the patients having remained in remission for 24 months was 4.18 ng/ml (SD 3.27 ng/ml, range 0.1 to 92 ng/ml; 95% CI for mean 0.0 to 10.8 ng/ml), while that for the patients relapsing within 24 months was 31.9 ng/ml (SD 11.2 ng/ml, range 0.5 to 183 ng/ml; 95% CI for mean 8.67 to 55.2 ng/ml). Based on the t test that allows for unequal variances, this observed tendency for the nadir PSA to be 27.8 ng/ml higher in the patients having relapsed within 24 months was beyond that that might be observed due to random chance in the absence of a difference between the groups (P = 0.025; 95% CI 3.79 to 51.9 ng/ml higher nadir PSA in the group having relapsed).**

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.

**Ans: The geometric mean nadir PSA among the patients having remained in remission for 24 months was 0.520 ng/ml (range 0.1 to 92 ng/ml; 95% CI for geometric mean 0.294 to 0.918 ng/ml), while that for the patients relapsing within 24 months was 8.31 ng/ml (range 0.5 to 183 ng/ml; 95% CI for geometric mean 3.62 to 19.1 ng/ml). Based on the t test that allows for unequal variances applied to log transformed nadir PSA measurements, this observed tendency for the nadir PSA to be 16.0 times higher in the patients having relapsed within 24 months was beyond that that might be observed due to random chance in the absence of a difference between the groups (P < 0.0005; 95% CI 5.99 to 42.8 times higher nadir PSA in the group having relapsed).**

    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?

**Ans: To the extent that we are aware of a tendency for the distribution of PSA to be quite skewed, we might consider whether we believe the increased risk of relapse likely to be linear in the nadir PSA. A "normal" PSA is thought to be less than 4 ng/ml. A PSA that is as high as 10 ng/ml had been found to be associated with a markedly increased risk of prostate cancer. As values in the hundreds or thousands are sometimes seen, it is highly likely that the risk behaves more multiplicatively than additively. Hence, inference based on the geometric mean might be expected to be more relevant scientifically and more precise statistically. I know of no compelling scientific reason that would drive me to use the mean in this setting.**

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

**Ans: The following table provides descriptive and inferential statistics within strata defined by nadir PSA level.** *(I did not ask you to do this, but I provide it as an example of what I would have done for this analysis. Note that I used intervals based on a multiplicative scale—a four-fold difference across an interval.)*

| Nadir PSA | n | Prob | 95% CI for Prob | Odds | 95% CI for Odds |
|---|---|---|---|---|---|
| **≤ 0.5 ng/ml** | 19 | 0.105 | (0.013, 0.331) | 0.118 | (0.013, 0.496) |
| **0.5 – 2.0 ng/ml** | 12 | 0.417 | (0.152, 0.723) | 0.714 | (0.179, 2.61) |
| **2.0 – 8.0 ng/ml** | 6 | 0.500 | (0.118, 0.881) | 1.00 | (0.134, 7.47) |
| **8.0 – 32.0 ng/ml** | 6 | 1.000 | (0.541, 1.000) | ∞ | (1.18, ∞) |
| **> 32.0 ng/ml** | 7 | 0.857 | (0.421, 0.996) | 6.00 | (0.728, 276) |

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.

**Ans: Interpreting the slope: From logistic regression using robust standard errors, we estimate a tendency for the odds of relapse within 24 months to be 4.15% higher for every one unit difference in nadir PSA between two groups, with the patients with the higher nadir PSA having the higher odds of relapse. This difference is not beyond that that might be observed due to random chance when there is no true difference (P = 0.391; 95% CI odds ratio is 5.11% lower to 14.3% higher for every 1 unit difference in nadir PSA).**

**Interpreting the intercept: By exponentiating the intercept from logistic regression, we estimate that the odds of relapse within 24 months would be 0.508 in a group having a nadir PSA of 0 ng/ml. Such a group is marginally outside the range of values measured in our sample. Furthermore, such an estimate would only be valid if the log odds of relapse were linear across groups defined by the nadir PSA.**

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.

**Ans: Interpreting the slope: From logistic regression using robust standard errors, we estimate a tendency for the odds of relapse within 24 months to be 1.85 times higher for every doubling of the nadir PSA between two groups, with the patients with the higher nadir PSA having the higher odds of relapse. This difference is beyond that that might be observed due to random chance when there is no true difference (P = 0.001; 95% CI odds ratio 1.28 to 2.69 times higher for every doubling of nadir PSA).** *(See the annotated Stata log file for how I computed these values from the regression parameters.)*

**Interpreting the intercept: By exponentiating the intercept from logistic regression, we estimate that the odds of relapse within 24 months would be 0.491 in a group having a nadir PSA of 1 ng/ml. Such a group is within the range of values measured in our sample, however, such an estimate would only be valid if the log odds of relapse were linear across groups defined by the nadir PSA.**

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)?

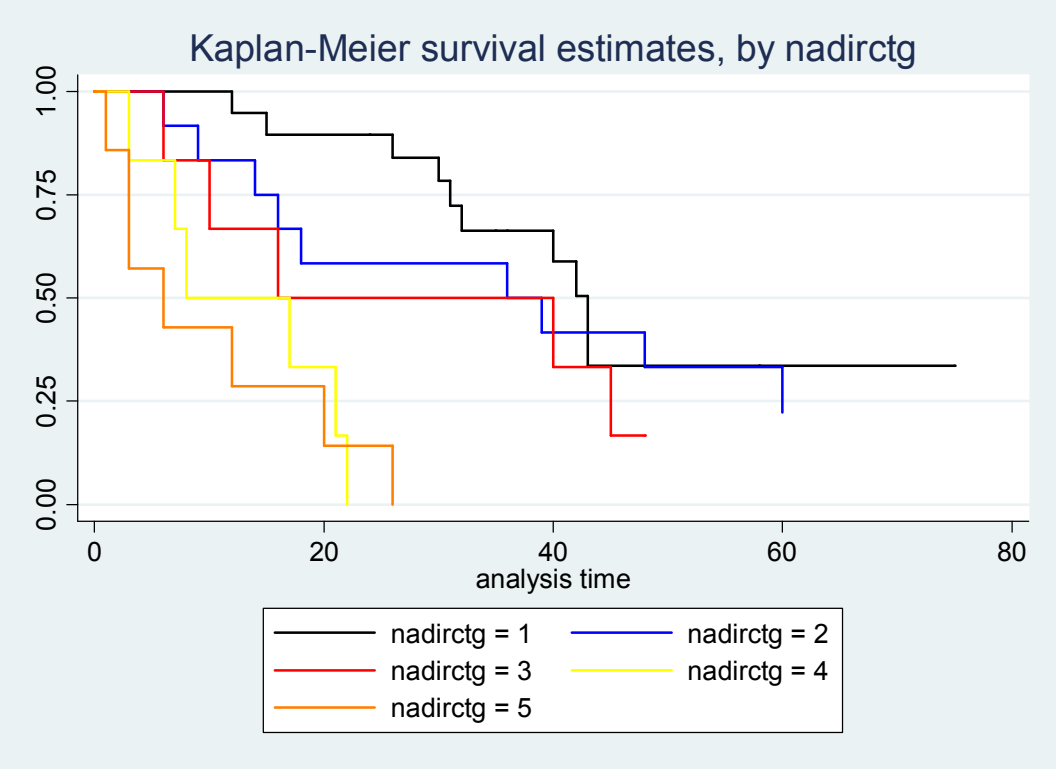
**Ans: To the extent that we are aware of a tendency for the distribution of PSA to be quite skewed, we might consider whether we believe the increased risk of relapse likely to be linear in the nadir PSA. A "normal" PSA is thought to be less than 4 ng/ml. A PSA that is as high as 10 ng/ml had been found to be associated with a markedly increased risk of prostate cancer. As values in the hundreds or thousands are sometimes seen, it is highly likely that the risk behaves more multiplicatively than additively. Hence, inference based on the log transformed nadir PSA might be expected to be more relevant scientifically and more precise statistically. If, however, we thought that the risk did increase linearly in nadir PSA, I would not log transform the predictor, even though its distribution was skewed in our sample. I would tend to use the dichotomized data only if 1) I knew it behaved like a threshold in terms of the risk of relapse (highly unlikely), or 2) I was going to use a particular threshold as a diagnostic cutoff and I wanted to be able to provide the prognostic value of that threshold (but then I would need to know that I had a representative sample of measurements both above and below the threshold).**

9. Perform analyses to determine whether the distribution of time to relapse differs across groups defined by nadir PSA level. (You can perform the proportional hazards regressions required for this problem using the Stata commands stset and stcox.)

    a. Provide suitable descriptive statistics regarding the distribution of time to relapse according to nadir PSA level.

**Ans: The following table and figure provide descriptive statistics within strata defined by nadir PSA level. The numbering of categories in the figure correspond to the strata in the table, with the lowest category corresponding to the lowest level of nadir PSA**

| | | Kaplan-Meier Estimates of Relapse Free Survival Probabilities | | | |
|---|---|---|---|---|---|
| **Nadir PSA** | **n** | **1 yr** | **2 yr** | **3 yr** | **4 yr** |
| **≤ 0.5 ng/ml** | 19 | 0.95 | 0.89 | 0.66 | 0.34 |
| **0.5 – 2.0 ng/ml** | 12 | 0.83 | 0.58 | 0.50 | 0.33 |
| **2.0 – 8.0 ng/ml** | 6 | 0.67 | 0.50 | 0.50 | 0.17 |
| **8.0 – 32.0 ng/ml** | 6 | 0.50 | 0.00 | 0.00 | 0.00 |
| **> 32.0 ng/ml** | 7 | 0.29 | 0.14 | 0.00 | 0.00 |




b. Perform a proportional hazards regression comparing the instantaneous risk of relapse across groups defined by the nadir PSA level when modeled as a continuous, untransformed variable. Provide interpretation for the slope.

**Ans: Interpreting the slope: From proportional hazards regression using robust standard errors, we estimate a tendency for the instantaneous risk of relapse to be 1.55% higher for every one unit difference in nadir PSA between two groups, with the patients with the higher nadir PSA having the higher risk of relapse. This difference is beyond that that might be observed due to random chance when there is no true difference (P < 0.0005; 95% CI risk of relapse is 0.81% to 2.30% higher for every 1 unit difference in nadir PSA).**

c. Perform a proportional hazards regression comparing the instantaneous risk of relapse across groups defined by the nadir PSA level when modeled as a continuous, log transformed variable. Provide interpretation for the slope.

**Ans: Interpreting the slope: From proportional hazards regression using robust standard errors, we estimate a tendency for the instantaneous risk of relapse to be 34.6% higher for every doubling of nadir PSA between two groups, with the patients with the higher nadir PSA having the higher risk of relapse. This difference is beyond that that might be observed due to random chance when there is no true difference (P < 0.0005; 95% CI risk of relapse is 19.6% to 51.6% higher for every doubling of nadir PSA).**

d. 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)?

**Ans: To the extent that we are aware of a tendency for the distribution of PSA to be quite skewed, we might consider whether we believe the increased risk of relapse likely to be linear in the nadir PSA. A "normal" PSA is thought to be less than 4 ng/ml. A PSA that is as high as 10 ng/ml had been found to be associated with a markedly increased risk of prostate cancer. As values in the hundreds or thousands are sometimes seen, it is highly likely that the risk behaves more multiplicatively than additively. Hence, inference based on the log transformed nadir PSA might be expected to be more relevant scientifically and more precise statistically. If, however, we thought that the risk did increase linearly in nadir PSA, I would not log transform the predictor, even though its distribution was skewed in our sample. I would tend to use the dichotomized data only if 1) I knew it behaved like a threshold in terms of the risk of relapse (highly unlikely), or 2) I was going to use a particular threshold as a diagnostic cutoff and I wanted to be able to provide the prognostic value of that threshold (but then I would need to know that I had a representative sample of measurements both above and below the threshold).**

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

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

**Ans: The analyses in problems 6, 7, and 8 all required the dichotomization of the time to relapse. This will tend to have lower statistical precision than would an analysis that used the continuous (but censored) time to relapse. I would therefore prefer the proportional hazards analysis, and because of the behavior of the PSA values seeming to behave more multiplicatively (which should be known or determined from another data set), I would prefer the analyses based on the log transformed PSA.**

    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: Nadir PSA is determined post therapy. Patients were followed with monthly measurements of PSA until they relapsed. This means that necessarily patients with longer times in remission had a higher number of PSA measurements. If the nadir PSA were defined merely as the lowest available measurement prior to relapse, then we could be seeing just an artifact of sample size: We expect to see a lower minimum in a larger sample, even if the populations all have the exact same distribution.**

**Even if the nadir was defined as the last value before the patient's measurements increased, a tendency toward similar steady declines across all patients would still mean that we would see a lower nadir PSA in patients who had longer remission times just because we followed them longer.**

**These points were a major concern of mine when I was first presented with the data. Descriptively, we could see that the typical patient had a U-shape in PSA over time, and that clinical relapse tended to be diagnosed 7-9 months post observation of the nadir. In all cases, the nadir was observed in the first year or so. I therefore did an analysis restricted to the data that was at least 18 months after therapy (and after all patients' nadirs had been observed). Such an analysis obviously excluded any patient who had relapsed within 18 months. But even in this restricted group, we found that the nadir value was highly associated with time to relapse.**