**Homework #3**

**November 27, 2013**

1. Provide suitable descriptive statistics for this dataset as might be presented in Table 1 of a manuscript appearing in the medical literature

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1. Descriptive statistics of prostate cancer patient level variables by relapse status.** | | | | | | | | |
|  |  | Relapse within 24 months | | | No Relapse within 24 months | | | P-value2 |
| Variable | | Count | Estimate1 | SE | Count | Estimate1 | SE |
| Lowest PSA (ng/ml) post-therapy, mean | | 22 | 31.9 | 11.192 | 28 | 4.1 | 3.265 | 0.011 |
| PSA (ng/ml) prior to therapy, mean | | 20 | 732.4 | 303.511 | 23 | 617.2 | 261.077 | 0.774 |
| Performance (0=worst, 100=best), mean | | 20 | 76.5 | 2.643 | 28 | 83.9 | 1.807 | 0.020 |
| Bone scan score, percent | |  |  |  |  |  |  | 0.053 |
|  | Least disease spread | 0 | 0.0% | 0.000 | 5 | 17.9% | 0.074 |  |
|  | Moderate disease spread | 4 | 20.0% | 0.092 | 9 | 32.1% | 0.090 |  |
|  | Most disease spread | 16 | 80.0% | 0.092 | 14 | 50.0% | 0.096 |  |
| Tumor grade, percent | |  |  |  |  |  |  | 0.690 |
|  | Least aggressive | 3 | 17.7% | 0.095 | 7 | 29.2% | 0.095 |  |
|  | Moderately aggressive | 7 | 41.2% | 0.123 | 8 | 33.3% | 0.098 |  |
|  | Most aggressive | 7 | 41.2% | 0.123 | 9 | 37.5% | 0.101 |  |
| Patient’s age (years), mean | | 22 | 68.4 | 1.211 | 28 | 66.7 | 1.104 | 0.321 |
| Time in remission (months), mean | | 22 | 11.1 | 1.365 | 28 | 42.1 | 2.278 | < 0.001 |

1 The estimates provided are mean values for continuous variables and percent values for categorical variables

2 P-values were derived using two-sample t-tests for continuous variables and chi-square test for categorical variables.

1. 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.

In the following logistic regression models, I chose to dichotomize bone scan score (*bss*). As noted in Table 1, the bss variable has no observations when among those with the least disease spread in the bones who have relapsed within 24 months. One option is to treat bss as a continuous variable, but having it as a categorical variables produces a more saturated model. Leaving bss as three categories will cause the logistic regression to kick out the entire category because of the empty cell and thus eliminating 5 observations from the entire regression. Therefore, the first and second categories were collapsed into one forming a binary version of bss. The chi-square p-value for each logistic regression model tells us that the model as a whole is a better fit than a model with no predictors. When comparing this p-value for the logistic regressions model using the 3-category or binary form of bss, the binary form had a better p-value (0.3588 vs. 0.1478, respectively). You loose some information when collapsing categories 1 and 2, but you have more observations and a better model than the alternative approaches. The following regressions were produced using the logit command on STATA. The logit command was chosen over logistic command in STATA to be able to interpret the intercept for part D (Logistic produces odds ratios, but does not provide the intercept for the analysis).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 2. Log odds of relapse within 24 months by post therapy PSA (Nadir PSA) controlling for bone scan score and performance score using Logit in STATA.** | | | | | | |
| Predictor of Interest | | Log Odds | | Standard Error | 95% Confidence Interval  Min Max | |
| *Model 1: Regress Relapse on Continuous, Untransformed PSA Post Therapy* | | |  | | | |
|  | Nadir PSA (ng/ml) | 0.03 | | 0.05 | -0.06 | 0.132 |
| *Model 2: Regress Relapse on Continuous, Log Transformed PSA Post Therapy* | | |  | | | |
|  | Log Nadir PSA | 0.88 | | 0.32 | 0.26 | 1.495 |
| *Model 3: Regress Relapse on Linear Splines of Nadir PSA* | | |  | | | |
|  | Nadir PSA 0.1-1ng/ml | 3.49 | | 1.53 | 0.48 | 6.495 |
|  | Nadir PSA >1-4ng/ml | -0.11 | | 0.52 | -1.12 | 0.906 |
|  | Nadir PSA >4-16ng/ml | 0.32 | | 0.19 | -0.06 | 0.704 |
|  | Nadir PSA >16ng/ml | -0.02 | | 0.01 | -0.04 | -0.001 |

Note: bone scan score and performance score were not associated with relapse status at the α=0.05 level for all three models.

* 1. 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.

In Table 2, for every 1 unit increase in PSA post-therapy, there is a 0.0349 increase in log-odds of having a relapse within 24 months when holding bone scan score and performance status constant. However, this is not statistically significant at the α=0.05 level.

* 1. 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.

In Table 2, for every 1 unit increase in log PSA post-therapy, there is a 0.8757 increase in log-odds of having a relapse within 24 months when holding bone scan score and performance status constant. This is statistically significant at the α=0.05 level.

* 1. 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.

Bone scan score and performance score are held constant for the following logistic regression results using linear splines, which is depicted in the following line graph:



Shown in Table 2, among those with post therapy PSAs <= 1ng/ml, for every unit increase in PSA there is a 3.489 increase in log-odds of having a relapse within 24 months, which is statistically significant at the α=0.05 level. Among those with PSAs >1ng/ml and <=4ng/ml, for every unit increase in PSA there is a 0.109 decrease in log-odds of having a relapse within 24 months, which is not significant. Among those with PSAs >4ng/ml and <=16ng/ml, for every unit increase in PSA there is a 0.3238 decrease in log-odds of having a relapse within 24 months, which is not significant. Among those with PSAs >16ng/ml, for every unit increase in PSA there is a 0.3238 decrease in log-odds of having a relapse within 24 months, which is statistically significant.

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

For each logistic regression, the intercept is the expected log odds of relapsing within 24 months when all covariates in the model are equal to zero. So for regression 2a, the log odds of relapsing within 24 months is 2.889 when post-therapy PSA, bone scan score, and performance score equal zero. For regression 2b, the log odds of relapsing within 24 months is 2.921 when the log PSA post-therapy, bone scan score, and performance score equal zero…

Even though interpretation of the intercepts were given, these are not meaningful interpretation. Some of the values for each covariate do not and cannot have a value equal to zero. For instance, the lowest value for PSA post-therapy is .1 and it is not possible for someone to have a PSA level of zero. Therefore, the intercept is extrapolating information outside the range of the data.

1. 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.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 3. Log odds of relapse within 24 months by controlling for bone scan score and performance score using Regress in STATA.** | | | | | | |
| Predictor of Interest | | Risk Difference | | Standard Error | 95% Confidence Interval | | |
| *Model 1: Regress Continuous, Untransformed of PSA Post Therapy on Relapse* | | |  | | | |
|  | Relapse Status | 22.81 | | 11.26 | 0.11 | 45.502 | |
| *Model 2: Regress Continuous, Log Transformed PSA Post Therapy on Relapse* | | |  | | | |
|  | Relapse Status | 2.68 | | 0.59 | 1.49 | 3.878 | |

* 1. 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.

The mean post-therapy PSA level for those who relapsed within 24 months was 66.03 ng/ml and 43.32 ng/ml for those who did not relapse within 24 months. Those who relapsed within 24 months had a post-therapy PSA level of 22.81 ng/ml higher than those who did not relapse within 24 months. This difference is statistically significant at the α=0.05 level.

* 1. 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.)

The log mean PSA level post-therapy for those who relapsed within 24 months was 2.65 and -0.04 for those who did not relapse within 24 months. Those who relapsed within 24 months had a 2.68 higher log PSA level post-therapy than those who did not relapse within 24 months. This difference is statistically significant at the α=0.05 level.

1. Consider the analyses performed in problems 2 and 3 above.
   1. What are the relative merits of the five analyses. Which might you prefer *a priori*? Why?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 4. AIC and BIC results for each analysis** | | | | |
| Model | Obs | df | AIC | BIC |
| Analysis 1 | 48 | 4 | 59.07 | 66.551 |
| Analysis 2 | 48 | 4 | 45.25 | 52.737 |
| Analysis 3 | 48 | 7 | 48.15 | 61.246 |
| Analysis 4 | 48 | 4 | 487.31 | 494.791 |
| Analysis 5 | 48 | 4 | 189.26 | 196.740 |

The first analysis used a binary outcome (Relapse Status) and a continuous, untransformed predictor of interest (Nadir PSA). With this analysis, we did not find a statistically significant association between relapsing within 24 months and post therapy PSA. It is not as easy to interpret log odds to the general population, but if you exponentiate the coefficient to get odds ratios (OR=1.03), you have a measure that is more commonly used among the public health and medical community for binary outcomes.

The second analysis used a binary outcome (Relapse Status) and a continuous, log transformed predictor of interest (Log Nadir PSA). With this analysis, we did find that the relapse status and the log nadir PSA are statistically associated. However, it is difficult to interpret what one unit increase on the log scale means in practice. This model, does have the lowest BIC, which means this is the best fitting model. However, the BIC does penalize complexity more and chooses the simpler models.

The third analysis used a binary outcome (Relapse Status) and linear splines of the predictor of interest (Nadir PSA with knots 1, 4, & 16). With this analysis we see that the first and fourth spline are statistically associated with the outcome. Using splines gives us more flexibility in determining the relationship between the predictor of interest and outcome. It also has the lowest AIC value, which tells us that this model is the best fit as well. One should also note the degrees of freedom are larger so you your estimates are not as precise as the other models.

The Fourth analysis used a continuous, untransformed outcome (Nadir PSA) and a binary predictor of interest (Relapse Status). The analysis does find that the outcome and predictor of interest are associated, and it can relay the information using a measure (Risk Differences) that is easily understandable to the general audience. However the AIC and BIC are quite high.

The Fifth analysis used a continuous, log transformed outcome (Log Nadir PSA) and a binary predictor of interest (Relapse Status). This analysis also finds that the outcome and predictor of interest are associated. However it is difficult to interpret the measure on the log-scale. This model also has very high AIC and BIC values.

The original research question was: “To determine if the PSA nadir (the lowest value observed post therapy) is highly associated with time of remission, and whether any association between the PSA nadir and time of remission is independent of an effect due to performance status or tumor mass.” Each of these analysis can tell you whether or not there is an association between PSA nadir and relapse status, and that the association is independent of effects due to the covariates. However, the spline was one of the best fit models, and it gives you the most information as to which PSA’s levels are associated with Nadir PSA and what direction that association is (for instance, the first spline has a positive association while the 4th spline has a negative association). So I would choose analysis 3.

* 1. 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.)

The problem that this study suffers from is not collecting person-time to account for any attrition due to patients missing to follow-up and death in the study. Death due to prostate cancer is possible within the 24-month time-frame of the study. I’m assuming that those who were lost at some point due to follow-up or death were excluded from the study if they did not reach the 24-month follow-up. Therefore, collecting person-time might also help their current problem of having a small sample size, which yields very small numbers in each cell and can lead to not having enough power in the study (probability of having a type II error increases).

**Stata Do File**

quietly: infile ptid nadirpsa pretxpsa ps bss grade age obstime str8 inrem using psa.txt

drop if ptid==.

gen inrem\_num = .

replace inrem\_num=1 if inrem=="yes"

replace inrem\_num=0 if inrem=="no"

tab inrem\_num inrem, m

g relap24 = 0

replace relap24 = 1 if obstime <= 24 & inrem\_num==0

sum ptid nadirpsa pretxpsa ps bss grade age obstime inrem\_num relap24

gen ndrpsa\_ln = ln(nadirpsa)

sum ndrpsa\_ln nadirpsa

gen bss\_bi = bss==3

replace bss\_bi =. if bss==.

tab bss bss\_bi, m

\* Problem 1

\* Means, Proportions & Significance tests

ttest nadirpsa, by(relap24)

ttest pretxpsa, by(relap24)

ttest ps, by(relap24)

ttest age, by(relap24)

ttest obstime, by(relap24)

proportion bss, over (relap24)

tab bss relap24, chi col

proportion grade, over (relap24)

tab grade relap24, chi col

\* Problem 2

logit relap24 nadirpsa bss\_bi ps, robust nolog

estat ic

logit relap24 ndrpsa\_ln bss\_bi ps, robust nolog

estat ic

mkspline snadir1 1 snadir2 4 snadir3 16 snadir4 = nadir, displayknots

sum snadir\*

preserve

logit relap24 snadir1-snadir4 i.bss\_bi ps, nolog robust

estat ic

predict p, p

replace bss\_bi=1

replace ps=80.8

predict p1, p

sort nadirpsa

line p1 nadirpsa

restore

\* Problem 3

regress nadirpsa relap24 bss\_bi ps, robust

estat ic

regress ndrpsa\_ln relap24 bss\_bi ps, robust

estat ic