Exam 1 Review Solutions

A medical center urology group is studying the association between prostate-specific antigen (PSA) levels and a number of clinical measurements on 97 men with cancer about to undergo prostectomies as treatment. We focus here on the association of PSA levels (this is a non-invasive blood test result) and Gleason scores (cancer severity score ranging from 6 to 8 with higher scores indicating worse prognosis). Note that PSA levels and Gleason scores are higher for this group than they would be for a random sample of men because all of these men have advanced cancers. The data are available as prostate.txt, prostate.xls, prostate.csv on the course website. There are 9 variables for each man:

id = identification number
PSA = prostate-specific antigen level (mg/ml)
cancvol = estimate of prostate cancer volume (cc)
weight = prostate weight (gm)
age = age of patient (yrs)
bph = amount of benign prostatic hyperplasia (cm**2**)
sem = presence/absence of seminal vesicle invasion
capspen = degree of capsular penetration (cm) gleason = grade of disease (6, 7, or 8 for these men)

(a) What type of study is this, and explain how you know.

This is an observational study. The subjects (men with prostate cancer) were not randomized to a particular group. In other words, the researcher is not active in assigning treatments to the subjects.

(b) Identify the following as specifically as possible: population, sample, explanatory variable, response variable, confounding variables

Population: men with advanced prostate cancer

Sample: 97 men with cancer about to undergo treatment

Explanatory variable: We are looking at an association between PSA levels and Gleason scores. Either of these could be the explanatory variable for a regression analysis. In this case, since we are doing an ANOVA, the explanatory variable should be categorical. Since the Gleason score are categorical (6, 7, or 8), it should be the explanatory variable.

Response variable: Using the same reasoning as above, the response is the PSA score, which is quantitative.

Confounding variables: Some of the confounding variables are measured, like age, weight, and other cancer-related measurements. Other ones you could mention are smoking status, alcohol intake, diet, etc. Basically, anything that might influence PSA levels as well as Gleason score would be acceptable as an answer, as long as you realize that the researchers made an attempt to measure some of these variables, as given in the list above.

(c) Before examining the results of an ANOVA we should check our model assumptions. This can be done in two ways: by examining the data in each group for normality and constant variance or by running an ANOVA and examining the residuals. Here it is

sufficient to take the first approach. Comment on the assumptions for ANOVA.

Examining the results of PROC UNIVARIATE or PROC MEANS or plotting the PSA scores against Gleason score demonstrates non-constant variance. (The latter two approaches are used in the SAS program below.) In particular, the Gleason = 8 group has much higher variance than the others. Normal probability plots show that none of the groups appear to follow a normal distribution.

(d) A logarithmic or square root transformation often helps with continuous positive measurements such as the PSA measurements. Reconsider the assumptions using log transformed data and square-root transformed data (create new variables by including LOGY=LOG(Y); and SQRTY=SQRT(Y); in the DATA step of your SAS program). Do the transformed data satisfy the model assumptions? Which transformation works best?

The square root transformation helps with both of the identified problems but doesn't solve them. The logarithm transformation seems to produce data for which constant variance is plausible. The log data looks fairly normal as well. If you did formal tests of normality (which are NOT recommended!!), you would see that the log transformation didn't work for all Gleason groups. Clearly though the log data is much closer to what we need for ANOVA than the untransformed data so it makes sense to proceed with the log transformed data.

(e) For the remainder of the question use the log transformed data. Is there evidence that median PSA levels vary across groups? Explain.

Doing an ANOVA on the log PSA scores suggests that there are differences in the mean log PSA scores among the three groups (and therefore the median PSA level). This is clear from the F-test of the hypothesis that the three groups (Gleason = 6,7,8) have the same mean log PSA score. This hypothesis is clearly rejected ($F_{2;94} = 21{:}56; p < {:}0001$).

(f) Identify the pairs of groups which differ significantly using the Bonferroni approach.

Using the multiple comparisons approach (even though there are only three paired comparisons) with the Bonferroni correction shows that the group with Gleason score 8 has significantly higher mean log PSA score than the other two groups. The other two groups are not significantly different at the 0.05 level but the difference there is nearly significant. Note that because the groups have unequal sample size the pairwise comparison results are given line-by-line separately for each pairwise comparison. (When the sample sizes are the same then the SE for the difference in means is sqrt(MSE(2/n)) which is the same for each pairwise comparison. This is what allows for just listing the means and identifying by letter which ones are different. When the sample sizes are different in the different groups, then the SE for the difference in means varies across pairwise comparisons and you get the kind of output you see here.)

(g) Define and analyze contrasts to see if there is a:

a. linear trend of PSA vs Gleason scores
b. quadratic trend of PSA levels vs Gleason scores.

What does it mean if both contrasts are significant?

For the linear trend use c = (- 1, 0, 1) which yields a highly significant result (F = 42. 41; p <

0.0001). This means we reject the null hypothesis and find in favor of a linear trend. For the quadratic trend, use c = (-1, 2, 1), which yields a marginally significant result (F = 3.28; p = 0.0734). This makes us suspect non-linearity (a quadratic trend but it is not very strong evidence). If both were deemed significant then it would suggest the means differ in ways that are consistent with both the presence of linear trend and the presence of a quadratic trend. A couple of important comments here: The quadratic I used above is known as the pure quadratic because it is orthogonal to the linear contrast. Second, if you want to understand what is happening look at the sample means. For log PSA these are 1.87 when Gleason = 6, 2.39 when Gleason = 7 and 3.62 when Gleason = 8. These increase but there seems to be some curvature (the change from 6 to 7 is less than the change from 7 to 8).

(h) Describe your findings on these data in a paragraph.

The key is to try and connect to the science rather than just report on statistical methods. Here's mine: An observational study of 97 men with prostrate cancer is used to analyze the association between Gleason scores (measuring cancer severity) and PSA levels (blood test results). Preliminary analysis of the data suggests that analysis of log PSA levels was more appropriate because of the wide range of such measurements. Our analysis shows significant differences among the mean log PSA scores in the three groups defined by Gleason scores ($p < 0.0001$ using ANOVA F-test). Higher log PSA scores are associated with higher Gleason scores (mean log PSA = 3.62 when Gleason = 8, 2.39 when Gleason = 7 and 1.87 when Gleason = 6). These data suggest a strong relationship but we note that this sample covers only a very narrow range of Gleason scores.

Sample SAS Program (if yours isn't like this but accomplishes the task, that's OK!)

```
data psa;
infile 'c:\appropriate pathname\prostate.txt' firstobs=2;
input id psa cancvol weight age bph sem capspen gleason;
logpsa = log(psa);
sqrtpsa = sqrt(psa); run;
proc sort;
by gleason; run;
proc rank normal=blom out=norm;
var psa;
by gleason;
ranks nrm; run;
proc gplot data=norm;
plot psa*nrm sqrtpsa*nrm logpsa*nrm;
by gleason; run;
proc gplot;
plot psa*gleason sqrtpsa*gleason logpsa*gleason; run;
proc means mean stddev min max n;
var psa sqrtpsa logpsa;
by gleason; run;
proc glm order=data;
class gleason;
model logpsa = gleason;
means gleason / lsd bon tukey scheffe;
contrast 'lin' gleason -1 0 1;
contrast 'quad' gleason -1 2 -1;
   run;
```