Exam 1 Review

A medical center urology group is studying the association between prostate-specific antigen (PSA) levels and several clinical measurements on 97 men with cancer about to undergo prostatectomies as treatment. We focus here on the association of PSA levels (this is a noninvasive 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 in some form as a file beginning “prostate.xxx” 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 (cm2)

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 (observational or experimental)? Explain your reasoning.

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:

i. population

The population consists of men with advanced prostate cancer; we do not know how the sample was collected from the population, making inferences beyond these 97 men questionable.

ii. sample

The sample consists of the 97 men with cancer about to undergo treatment.

iii. 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 (or ordinal). Since the Gleason scores are ordinal (6, 7, or 8), it should be the explanatory variable.

iv. response variable

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

v. confounding variables

Some of the confounding variables are measured, like age, weight, and other cancer-related measurements. Others you could mention are smoking status, alcohol intake, diet, etc. Basically, anything that might influence PSA levels as well as Gleason scores would be acceptable as an answer, so long as you realize that the researchers attempted 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.

**Normality**: The data appear to be heavily right-skewed and nonnormal according to the histograms and q-q plots. With the smallest sample size of 21, the Central Limit Theorem may ensure that the sample means are normally distributed, but it may not. It could be a borderline case.

**Equal Standard deviations**: The axes of the histograms and the scatter plot all suggest that the standard deviations are not equal. The widest range occurring in the group with the smallest sample size (Gleason = 8 group) is further evidence against equal standard deviations. Note that we could have included box plots as well.

**Independence**: We do not know much about how this data was selected, but we will assume independence going forward.

\*To address ANOVA assumptions on original data with histograms and q-q plots;

proc univariate data = prostatedata;

by gleason;

var psa;

histogram psa;

qqplot psa;

run;

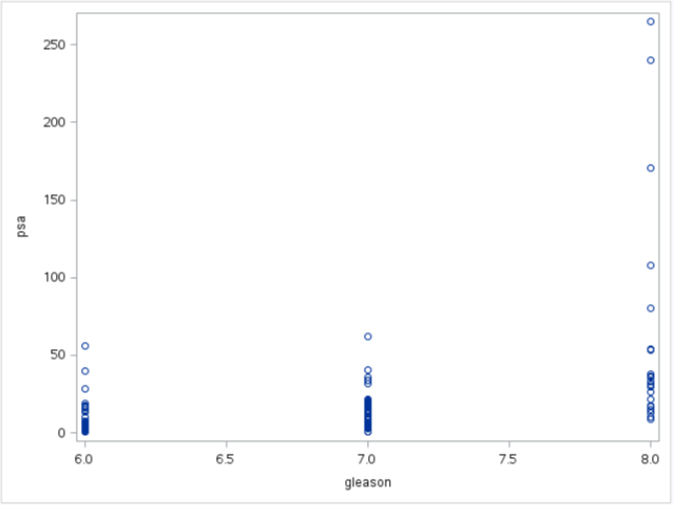
| Gleason = 6 | Gleason = 7 | Gleason = 8 |
| --- | --- | --- |
|  |  |  |
|  |  |  |

\*To address ANOVA assumptions on original data with a scatter plot;

proc sgplot data = prostatedata;

scatter x= gleason y = psa;

run;



Due to both normality and equal variances assumptions being violated, we could try a transformation or try nonparametric methods of analysis.

(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 something similar to LOGY=LOG(Y) and SQRTY=SQRT(Y) in the DATA step of your SAS program (or equivalent in R).

i. Do the transformed data satisfy the model assumptions?

Log-transformed data:

**Normality**: The histograms and q-q plots do not suggest any obvious departures from normality. 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.

**Equal Standard deviations**: The axes of the histograms and the scatter plot do not suggest an obvious violation of the equal standard deviations assumption.

**Independence**: We do not know much about how this data was selected, but we will assume independence going forward.

\*To perform a log transformation;

data prostatedata;

set prostatedata;

logpsa = log(psa);

run;

\*To address ANOVA assumptions on logged data with histograms and q-q plots;

proc univariate data = prostatedata;

by gleason;

var logpsa;

histogram logpsa;

qqplot logpsa;

run;

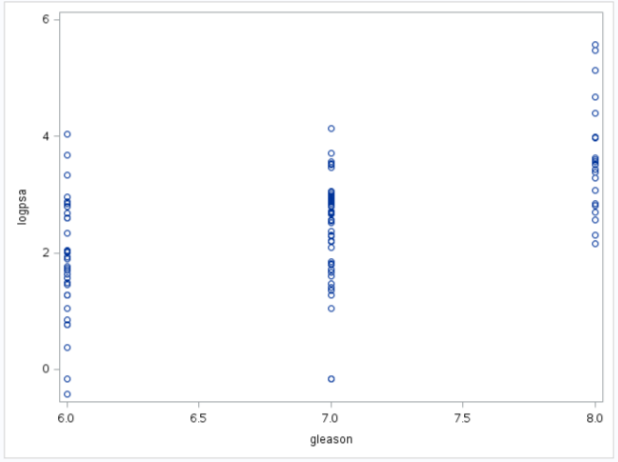
| Gleason = 6 | Gleason = 7 | Gleason = 8 |
| --- | --- | --- |
|  |  |  |
|  |  |  |

\*To address ANOVA assumptions on logged data with a scatter plot;

proc sgplot data = prostatedata;

scatter x= gleason y = logpsa;

run;



The log-transformed data appear to meet the normality, equal standard deviations, and independence assumptions of parametric methods of analysis, such as ANOVA.

Square root-transformed data:

**Normality**: The histograms show a slight skew, and the q-q plots show some curvature, indicating deviations from normality. This transformation, while an improvement in the area of normality, still could be better. The central limit theorem is likely enough to overcome this normality deviation.

**Equal Standard deviations**: The axes of the histograms and the scatter plot suggest that the third group (Gleason = 8) has a standard deviation greater than the other two groups.

**Independence**: We do not know much about how this data was selected, but we will assume independence going forward.

\*To perform a square root transformation;

data prostatedata;

set prostatedata;

sqrtpsa = sqrt(psa);

run;

\*To address ANOVA assumptions on square-rooted data with histograms and q-q plots;

proc univariate data = prostatedata;

by gleason;

var sqrtpsa;

histogram sqrtpsa;

qqplot sqrtpsa;

run;

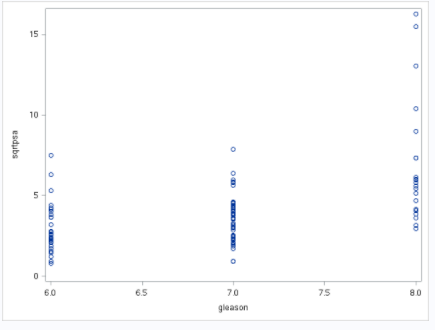
| Gleason = 6 | Gleason = 7 | Gleason = 8 |
| --- | --- | --- |
|  |  |  |
|  |  |  |

\*To address ANOVA assumptions on square-rooted data with a scatter plot;

proc sgplot data = prostatedata;

scatter x= gleason y = sqrtpsa;

run;



While the square root transformed data still do not meet the equal standard deviations assumption, the improvement in distribution shape could make normality assumable, in conjunction with the central limit theorem. We would need to employ methods that allow for this situation, like Welch’s ANOVA.

ii. Which transformation works best?

The logged data appears to meet the normality and equal standard deviations assumptions; whereas the square rooted data does not meet both. If we opt to analyze transformed data, we should opt for 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? You do not need to perform a complete analysis, but information you would put in the concluding paragraph of a complete analysis should suffice.

Code and output are included for completeness but are not necessary for the answer. We will perform pure ANOVA on logged data.

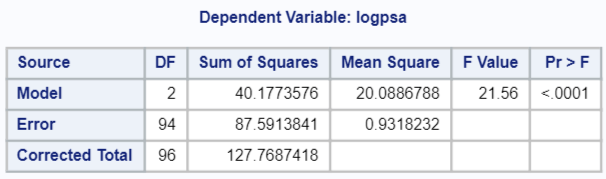
\*To test if any medians are different (ANOVA on log trasnformed data);

proc glm data = prostatedata;

class gleason;

model logpsa = gleason;

run;



Performing an ANOVA on the logged PSA scores suggests that there are differences in the mean logged PSA scores among the three groups (and therefore the median PSA level). This is clear from the F-test of that the three groups (Gleason = 6,7,8) do NOT have the same mean log PSA score (same median on original scale). This null hypothesis of equal medians is clearly rejected (F2,94 = 21: 56; p-value < : 0001).

(f) Identify the pairs of groups which differ significantly using the Bonferroni approach. You do not need to perform a complete analysis, but you should provide evidence to support your conclusion.

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 (p-value = 0.0693).

\*To perform bonferroni adjusted lsmeans t-tests;

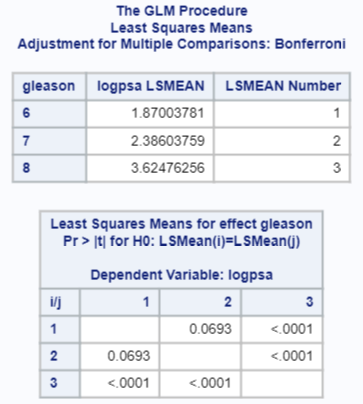
proc glm data = prostatedata;

class gleason;

model logpsa = gleason;

lsmeans gleason/ adjust = bon pdiff;

run;



(g) Define and analyze contrasts to address the following:

i. linear trend of PSA vs. Gleason scores.

For the linear trend, use c = (- 1, 0, 1), although this is not the only choice. This yields a highly significant result (F = 42. 41; p-value < 0.0001). This means we reject the null hypothesis and find in favor of a linear trend.

\*To examine a linear trend;

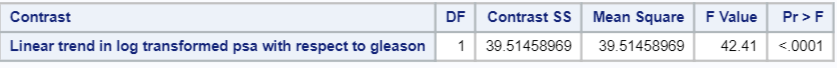
proc glm data = prostatedata order = data;

class gleason;

model logpsa = gleason;

contrast "Linear trend in log transformed PSA with respect to Gleason" gleason -1 0 1;

run;



ii. quadratic trend of PSA levels vs. Gleason scores.

For the quadratic trend, use c = (1, -2, 1), although this is not the only choice, which yields a marginally significant result (F = 3.28; p-value = 0.0734). This makes us suspect some non-linearity (a quadratic trend), but it is not very strong evidence.

\*To examine a quadratic trend;

proc glm data = prostatedata order = data;

class gleason;

model logpsa = gleason;

contrast "quadratic trend in log transformed PSA with respect to Gleason" gleason 1 -2 1;

run;



iii. What does it mean if both contrasts are significant?

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.

\*Note that the quadratic contrast vector (-1, 2, -1) used above is known as the pure quadratic because it is orthogonal to the linear contrast. This means that the two models can be examined separately, and the presence of a linear trend will not confound the quadratic contrast (or vice versa). Secondly, to understand what is happening, examine the sample means. For logpsa, these are 1.87 when Gleason = 6, 2.39 when Gleason = 7, and 3.62 when Gleason = 8. There is an increasing trend, 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.

Try to connect to the science rather than just report on statistical methods. Here's one angle:

An observational study of 97 men with prostate cancer is used to analyze the association between Gleason scores (cancer severity) and PSA levels (blood test results). Preliminary analysis of the data suggests that an examination of logged 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 an 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.

SAS hints: Use SAS PROC UNIVARIATE and a normal probability plot for each group defined by the Gleason score to assess the assumptions in (a); a scatterplot of PSA vs. Gleason score may also be useful. PROC GLM can produce residual plots for ANOVA and regression, but these may not be drawn by group.

R Code:

##Read in the data

prostate <- read.csv('C:/Users/Charles/Documents/SMU/Online Teaching/MSDS 6371 - Statistical Foundations for Data Science/UNIT 7 Midterm/Spring 2018/prostate.csv')

##########################

###Raw Data Assumptions###

##########################

##Boxplots

boxplot(psa ~ gleason, data=prostate)

##This is an efficient way to plot everything at once, but is not the only way

##You could use individual subset statements and make 3 separate calls to hist/qqline/qqnorm

par(mfrow=c(2,2))

lapply(unique(prostate$gleason), function(x){

hist(subset(prostate, gleason=x)$psa, xlab='PSA', main=paste('Gleason =',x))})

par(mfrow=c(2,2))

lapply(unique(prostate$gleason), function(x){

qqnorm(subset(prostate, gleason=x)$psa, main=paste('Gleason =',x))

qqline(subset(prostate, gleason=x)$psa, main=paste('Gleason =',x))

})

####################################

###Log and Square Root Assumptions##

####################################

##Create the transformed data

prostate$log.psa <- log(prostate$psa)

prostate$sqrt.psa <- sqrt(prostate$psa)

##Boxplots

boxplot(log.psa ~ gleason, data=prostate)

boxplot(sqrt.psa ~ gleason, data=prostate)

##Histogram and QQ Plots - Log##

par(mfrow=c(2,2))

lapply(unique(prostate$gleason), function(x){

hist(subset(prostate, gleason=x)$log.psa, xlab='Log of PSA', main=paste('Gleason =',x))})

par(mfrow=c(2,2))

lapply(unique(prostate$gleason), function(x){

qqnorm(subset(prostate, gleason=x)$log.psa, main=paste('Gleason =',x))

qqline(subset(prostate, gleason=x)$log.psa, main=paste('Gleason =',x))

})

##Histogram and QQ Plots - Square Root

par(mfrow=c(2,2))

lapply(unique(prostate$gleason), function(x){

hist(subset(prostate, gleason=x)$sqrt.psa, xlab='Square Root of PSA', main=paste('Gleason =',x))})

par(mfrow=c(2,2))

lapply(unique(prostate$gleason), function(x){

qqnorm(subset(prostate, gleason=x)$sqrt.psa, main=paste('Gleason =',x))

qqline(subset(prostate, gleason=x)$sqrt.psa, main=paste('Gleason =',x))

})

########################

##Analysis of Variance##

########################

##Note the use of as.factor, otherwise R treats gleason as a continuous variable

prostate$gleason <- as.factor(prostate$gleason)

prostate.aov <- aov(log.psa ~ gleason, data=prostate)

summary(prostate.aov)

##Bonferroni corrected group differences

library(agricolae)

prostate.bonf <- LSD.test(prostate.aov, 'gleason', p.adj='bonferroni')

#############

##Contrasts##

#############

mean.gleason <- with(prostate, tapply(log.psa, gleason, mean))

var.gleason <- with(prostate, tapply(log.psa, gleason, var))

n <- c(33, 43, 21)

##Extract RMSE from AOV table

s.pooled <- sqrt(summary(prostate.aov)[[1]][[3]][[2]])

################

##Linear Trend##

################

##Define contrast on all groups, then calculate

contr <- c(-1, 0, 1)

g <- sum(contr \* mean.gleason)

se.g <- s.pooled \* sqrt(sum(contr^2/n))

t.94 <- qt(0.975, 97-3)

##Calculate t

t.g <- g/se.g

##Calculate p-value

2\*pt(t.g, 94, lower.tail=F)

##Calculation of CI

g + c(-1,1) \* t.94 \* se.g

###################

##Quadratic Trend##

###################

##Define contrast on all groups, then calculate

contr2 <- c(1, -2, 1)

g2 <- sum(contr2 \* mean.gleason)

se.g2 <- s.pooled \* sqrt(sum(contr2^2/n))

t.94 <- qt(0.975, 97-3)

##Calculate t

t.g2 <- g2/se.g2

##Calculate p-value

2\*pt(t.g2, 94, lower.tail=F)

##Calculation of CI

g2 + c(-1,1) \* t.94 \* se.g2