Introduction to R - Final Project

An Analysis of Prostate-Specific Antigen's Relation to Advanced Prostate Cancer

Katie Haring

May 11, 2019

Abstract

Prostate-specific antigen (PSA) is a glycoprotein enzyme usually present in small quantities in the blood plasma of men. However, patients with prostate cancer often show elevated levels of PSA (Catalona et al., 1994). PSA levels can be measured to test for prostate cancer, but the medical community is divided on the usefulness of this test, with some believing that the risk of over-diagnosis and false positives outweighs the benefit of early detection (Gomella et al., 2011).

Due to the prevalence of prostate cancer, which had 1.1 million reported new cases in 2014 alone (World Health Organization, 2014), it is important to understand PSA's role throughout the disease, not just in the diagnosis process. This analysis aims to assess if, in a given subset of men with prostate cancer, men with higher Gleason scores, which are used to rate the pathologically determined level of disease, also could be expected to have higher PSA levels. The results support the assessment that Gleason score and PSA level can be linked in this way, but only loosely. More factors would need to be considered to properly model PSA levels.

Introduction

There are 4 stages of prostate cancer, with each progressive stage being more advanced than the last. The Gleason score, which indicates the aggressiveness of the cancer, is one of the factors used in determining the stage. In Stage 1, the Gleason score is less than or equal to 6. In stage 2, the Gleason score can be 7, 8, or higher. In Stages 3 and 4, the cancer has advanced beyond the use of the Gleason score.

A previous study in Western Jamaica found that PSA levels were higher for men with higher Gleason scores (Anderson-Jackson, McGrowder, & Alexander-Lindo, 2012). The primary purpose of this analysis was to test whether PSA levels and Gleason scores are linked, using data from a group of men about to undergo radical prostatectomies. A simple linear statistical model was fitted to the data, and the general linear test with a null hypothesis of β_1=0 was utilized.

Primary Analysis Objectives

1. Determine whether an association exists between Gleason scores and Prostate-Specific Antigen (PSA) levels. The primary objective analysis fit a simple linear statistical model to the data using an F test to hypothesis test:

- H_0: β_1 is not different from 0
- H_1: β_1 is significantly different from 0 when the level of significance (α) is 0.05

Secondary Analysis Objectives

2. Summarize patient demographic and baseline clinical characteristics.
3. Determine if any other measured clinical characteristics are related to PSA levels.

Materials & Methods

Data Sources

The data for this analysis were obtained from Appendix C, Data Set 5 from Kutner, Nachtsheim, and Neter (2014). The complete data set is available here. The data set contains: Identification Number, PSA Level (mg/ml), Cancer Volume (cc), Prostate Weight (gm), Age (years), Amount of Benign Prostatic Hyperplasia (cm^2), Indicator of Seminal Vesicle Invasion, Degree of Capsular Penetration (cm), and Gleason Score for 97 men who were about to undergo radical prostatectomies. The data was collected by a university medical center urology group, and each of the men had advanced prostate cancer.

Statistical Analysis

The data were explored, via graphs. Then, a simple linear regression model was fit to the data, with the Gleason score as the independent variable and the PSA level as the dependent variable. The hypothesis test was ran and model testing was performed. Finally, summary statistics were calculated and additional relationships in the data set were explored.

Model Assumptions

All inferences were conducted using $\alpha = 0.05$ unless stated otherwise. No adjustments for multiplicity were made as this is an exploratory analysis. The error variables (ϵ) were checked for normality and equal variance using the methods described in chapter 5 of *Design and Analysis of Experiments*.

Discrete variables were summarized with proportions and frequencies. Continuous variables were summarized using the following statistics:

- mean
- median
- standard deviation
- quantiles
- minimum
- maximum

Primary Objective Analysis

Preliminary Analysis

To help visual the data, each individual man's Gleason score was plotted against his PSA level (Figure 1):

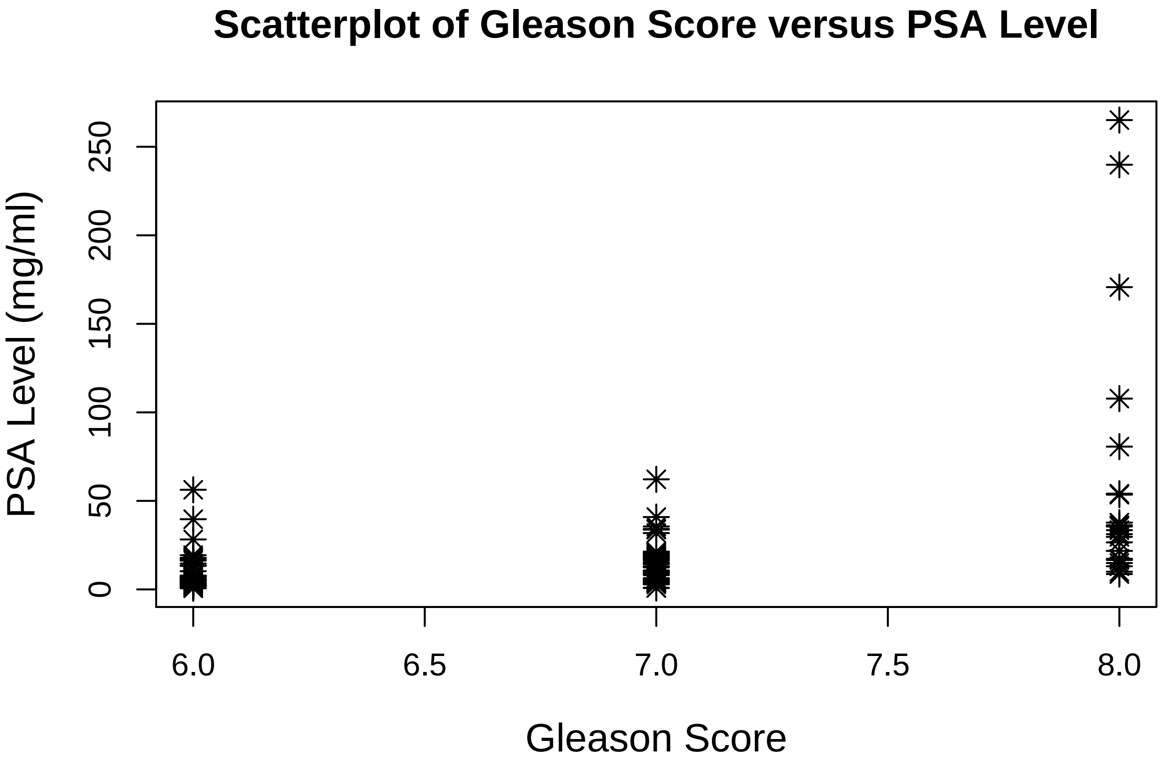

Figure 1: Scatterplot of Gleason Score vs PSA Level Data

Box plots for PSA Level and Gleason Score were also produced (Figure 2):

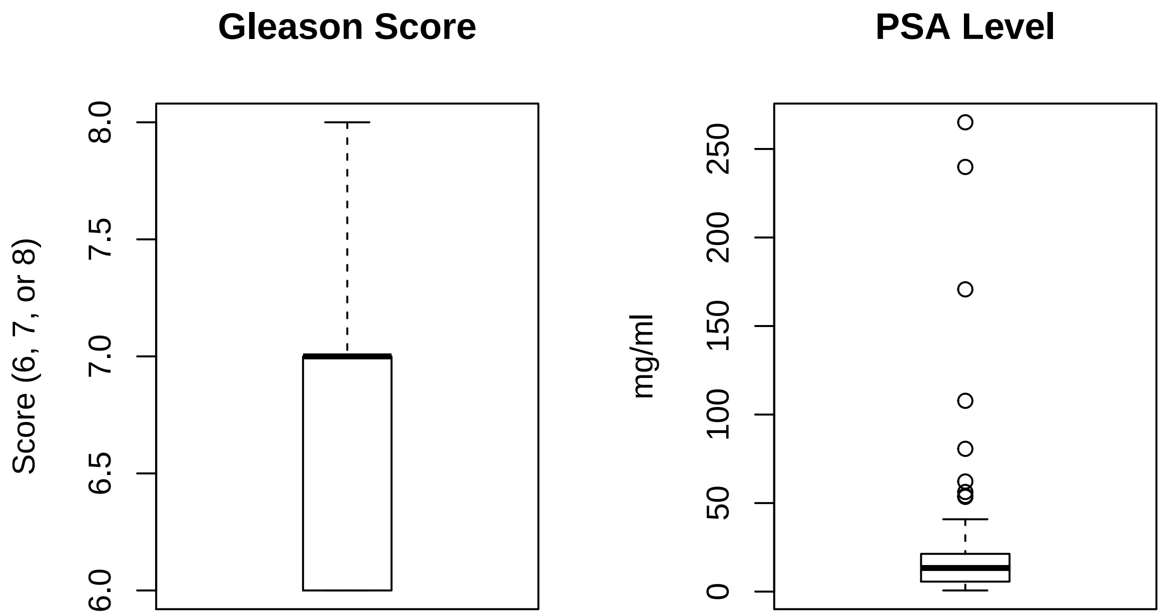

Figure 2: Box Plots of Gleason Score and PSA Level Data

It was noted that PSA level did not appear to be normally distributed. This issue was analyzed further in the model testing step.

Fitting the Model

A simple linear regression model was fitted to the data, of the form:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \text{ i=1,2,...,97}$$

where:
Y_i = the PSA level of the i^{th} randomly selected man
X_i = the Gleason score of the i^{th} randomly selected man
$\epsilon_i \sim idd\text{N}(0,\sigma^2)$
and β_0, β_1, σ^2 were the unknown parameters of interest.

The ANOVA table for the model and overall measures of model fit can be found in Tables 1 and 2, respectively.

Table 1: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(Gleason.score)	2	39887	19943	15.65	1.359e-06
Residuals	94	119785	1274	NA	NA

Table 2: Measures of Model Fit

r.squared	adj.r.squared	sigma	statistic	p.value
0.25	0.234	35.697	15.65	0

As shown in table 1, the F test for the Gleason Score was significant. The null hypothesis was rejected; β_1 is significantly different from 0 at the 0.05 level of significance used for the test. It significantly different all the way down to the 0 level of significance.
According to the R^2, Gleason score explains 25% of the variation in PSA level.

Model Testing

The normality of error terms was tested using a QQ-Plot (Figure 3) and a Shapiro-Wilk normality test:

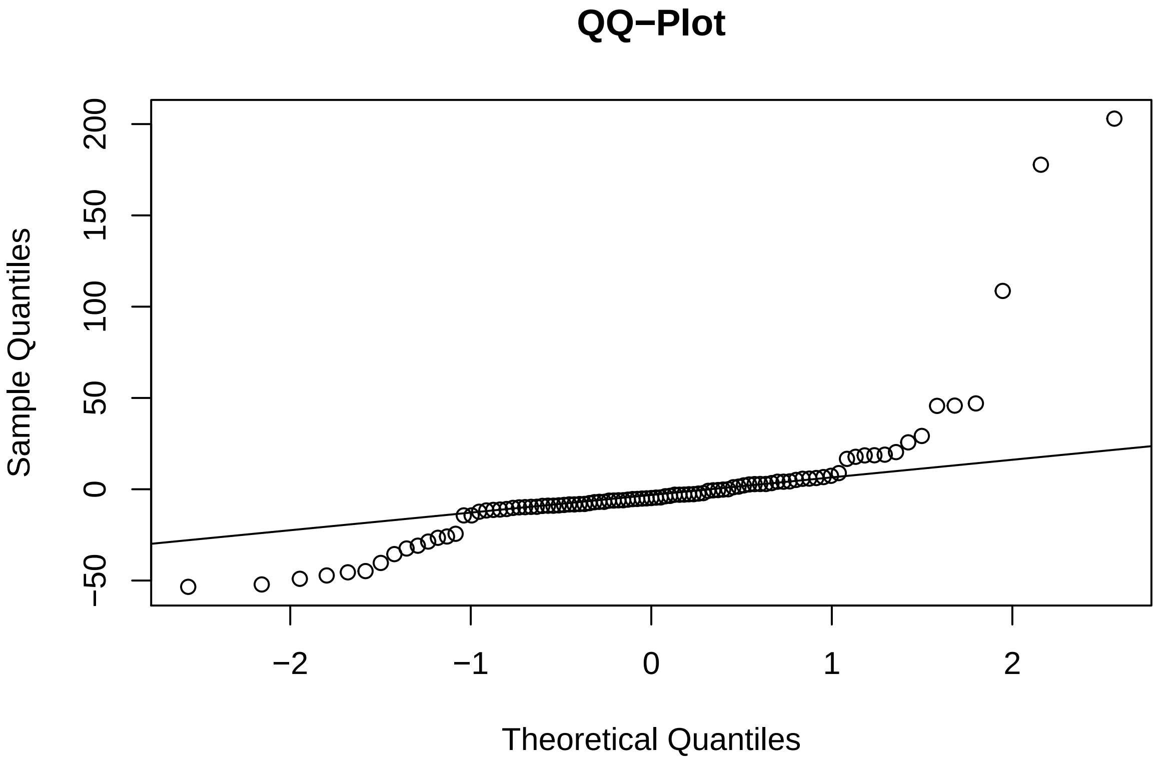

Figure 3: Q-Q Plot

```
	Shapiro-Wilk normality test

data:  residuals(mymodel)
W = 0.63624, p-value = 3.751e-14
```

The assumption of normality was rejected.

Residuals were plotted against the fitted values produced by the model in Figure 4:

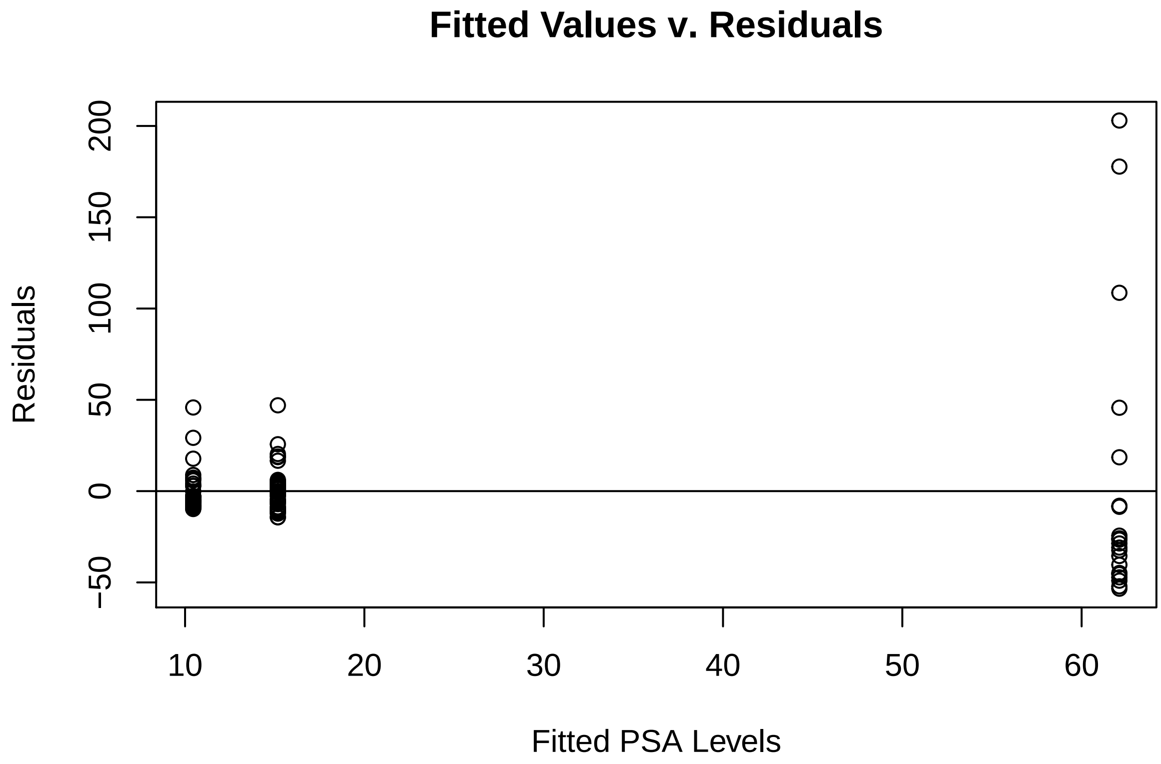

Figure 4: Residual Plot

The plot suggested that the model did not have homoscedasticity, so a non-constant variance test was conducted:

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 136.559, Df = 1, p = < 2.22e-16
```

The test confirmed that variance was not constant.

Refitting the Model

Since the original model didn't meet the assumption requirements, a transformation was needed. Given that the Preliminary Analysis showed that PSA Level was left-skewed, a new model was created that transformed it using the logarithm function. The new model took the following form:

$$\log(Y_i) = \beta_0 + \beta_1 X_i + \epsilon_i, \text{ i=1,2,...,97}$$

The ANOVA table for the new model and overall measures of model fit are shown Tables 3 and 4, respectively.

Table 3: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(Gleason.score)	2	7.578	3.789	21.56	1.967e-08
Residuals	94	16.52	0.1758	NA	NA

Table 4: Measures of Model Fit

r.squared	adj.r.squared	sigma	statistic	p.value
0.314	0.3	0.419	21.558	0

As seen in Table 3, the F test for the Gleason Score was still highly significant. In the new model, Gleason score explained 31.4% of the variation in log(PSA level). How to potentially build a better model will be addressed in the second half of the secondary analysis.

The Shapiro-Wilk normality and non-constant variance test results are shown below:

```

	Shapiro-Wilk normality test

data:  residuals(model2)
W = 0.98826, p-value = 0.5502
```

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.03257535, Df = 1, p = 0.85677
```

The new model passed both assumption tests and still came to the conclusion that PSA Level was significantly related to Gleason score.

Below is a summary of the final model coefficients (Table 5) where the default Gleason score is assumed to be 6.

Table 5: Model Coefficients Summary

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.812	0.073	11.129	0.000
factor(Gleason.score)7	0.224	0.097	2.310	0.023
factor(Gleason.score)8	0.762	0.117	6.512	0.000

Finally, a Tukey's Highly Significant Differences (HSD) test was conducted, and the following results were produced (Table 6):

Table 6: Tukey's HSD Test Result and Pairwise Confidence Intervals

	Difference	95% Lower	95% Upper	P-value
7-6	0.224	-0.007	0.455	0.059
8-6	0.762	0.483	1.041	0.000
8-7	0.538	0.272	0.804	0.000

The differences between 8-6 and 8-7 were significant, whereas the difference between 7-6 was not.

Secondary Objective Analyses

Data Summarization

Summary statistics were produced for baseline characteristics (see Tables 7, 8, and 9, below).

Table 7: Summary Statistics for Continuous Variables

Metric	**PSA Level (mg/ml)**	**Cancer Volume (cc)**	**Prostate Weight (gm)**	**Age**	**Benign Prostatic Hyperplasia (cm^2)**	**Capsular Penetration (cm)**
Minimum	0.65	0.26	10.70	41.00	0.00	0.00
1st Quartile	5.64	1.67	29.37	60.00	0.00	0.00
Median	13.33	4.26	37.34	65.00	1.35	0.45
Mean	23.73	7.00	45.49	63.87	2.53	2.25
3rd Quartile	21.33	8.41	48.42	68.00	4.76	3.25
Max	265.07	45.60	450.34	79.00	10.28	18.17
Std Dev	40.78	7.88	45.71	7.45	3.03	3.78

Table 8: Seminal Vesicle Invasion Indicator Proportion Table

Seminal.vesicle.invasion	n	Proportion
0	76	0.784
1	21	0.216

Table 9: Gleason Score Proportion Table

Gleason.score	n	Proportion
6	33	0.340
7	43	0.443
8	21	0.216

Other Data's Relation to PSA Level

The scatterplot matrix was produced for all the variables in the data (including Gleason Score) in Figures 5 and 6, and the correlation matrix was produced in Table 10. ID number was included to check for any potential bias from experimental design:

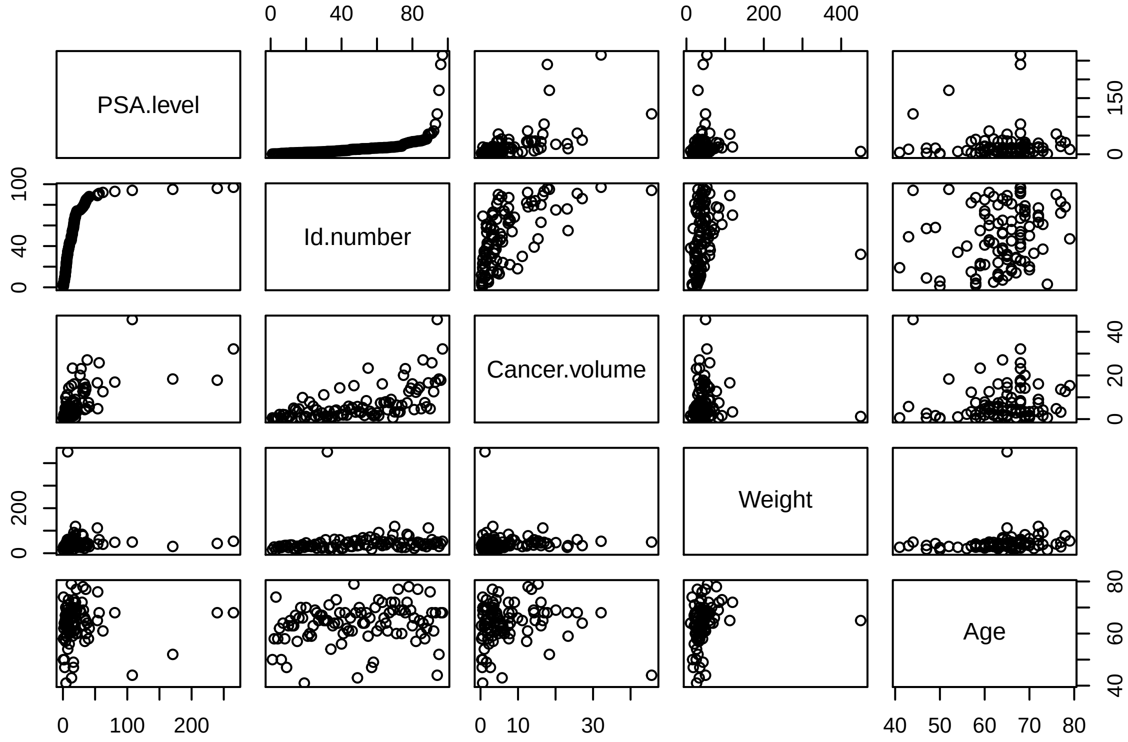

Figure 5: Scatterplot Matrix

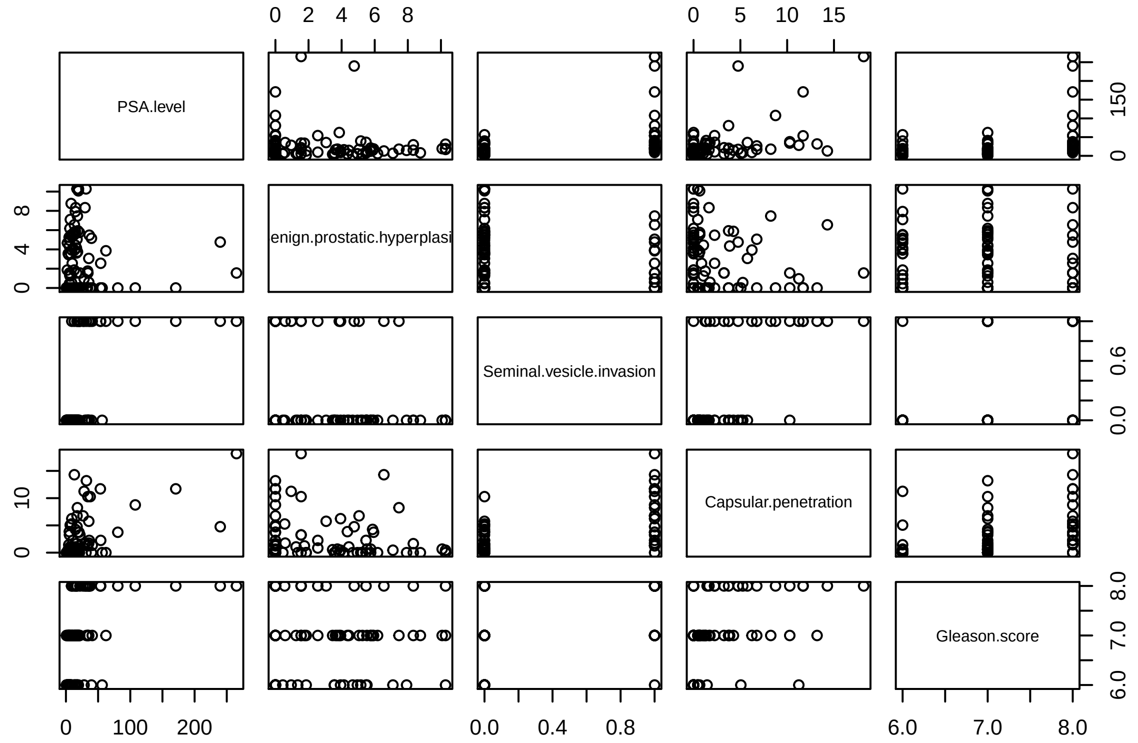

Figure 6: Scatterplot Matrix

Table 10: Correlations

	ID #	**PSA Level**	**Cancer Volume**	**Weight**	**Age**	**Ben. Prost. Hyp.**	**Sem. Ves. Inv.**	**Caps. Pen.**	**Gleason Score**
ID #	1.000	0.603	0.621	0.114	0.197	0.165	0.567	0.477	0.538
PSA Level	0.603	1.000	0.624	0.026	0.017	-0.016	0.529	0.551	0.430
Cancer Volume	0.621	0.624	1.000	0.005	0.039	-0.133	0.582	0.693	0.481
Weight	0.114	0.026	0.005	1.000	0.164	0.322	-0.002	0.002	-0.024
Age	0.197	0.017	0.039	0.164	1.000	0.366	0.118	0.100	0.226
Ben. Prost. Hyp.	0.165	-0.016	-0.133	0.322	0.366	1.000	-0.120	-0.083	0.027
Sem. Ves. Inv.	0.567	0.529	0.582	-0.002	0.118	-0.120	1.000	0.680	0.429

Caps. Pen.	0.477	0.551	0.693	0.002	0.100	-0.083	0.680	1.000	0.462
Gleason Score	0.538	0.430	0.481	-0.024	0.226	0.027	0.429	0.462	1.000

The factors most correlated with PSA Level were: Cancer Volume (0.624), ID Number (0.603), Capsular Penetration (0.551), Seminal Vesicle Invasion (0.529), and Gleason Score (0.430). The fact that ID number was highly correlated with PSA level was concerning. If ID Numbers were assigned randomly, it would be highly unlikely to see any correlation with PSA Level. This paper recommends that the process of numbering the patients be assessed for any systematic flaws (e.g., measurement equipment error leading to greater PSA Levels over time, etc.).

Based on these results, a more appropriate model than our initial simple linear model might use Cancer Volume , Capsular Penetration, Seminal Vesicle Invasion, and Gleason Score as the independent variables. It is possible that when the other variables are included, the Gleason score is not as useful of a predictor comparatively and may need to be dropped from the model. However, many of these independent variables also share a high correlation with each other, so the potential for multicollinearity would be a concern. It is outside the scope of this paper to determine the best model for PSA level.

Conclusion & Discussion

The general linear F-test was utilized to determine that a model using Gleason Score is better than the null model at predicting PSA levels in a sample of 97 men who were about to undergo radical prostatectomies. This is in-line with the findings of Anderson-Jackson, McGrowder, & Alexander-Lindo (2012). However, the simple linear regression model for PSA Levels using Gleason score alone did not satisfy the linear regression model requirements of error term normality and constant variance. A slightly modified model with a log-transformed dependent variable was also determined to be better than the null model using the F test and met both the assumption requirements, further proving that the two variables were linearly linked and could be modeled as such.

Other factors included in the data with relatively high correlations to PSA Level include: Cancer Volume, ID Number, Capsular Penetration, Seminal Vesicle Invasion, and Gleason Score. Many of these factors are also highly correlated with each other, suggesting possible multicollinearity in a model incorporating all 5 of them. The correlation between ID number and PSA Level should be explored further, as no correlation would be expected to exist between them in a properly randomized design.

References

Catalona, W.J., Richie, J.P., Ahmann, F.R., Hudson, M.A., Scardino, P.T., Flanigan, R.C., … Dalkin, B.L. (May 1994). Comparison of digital rectal examination and serum prostate specific antigen in the early detection of prostate cancer: results of a multicenter clinical trial of 6,630 men. *The Journal of Urology, 151 (5)*: 1283–90.

Gomella, L.G., Liu, X.S., Trabulsi, E.J., Kelly, W.K., Myers, R., Showalter, T., … Wender, R. (Oct 2011). Screening for prostate cancer: the current evidence and guidelines controversy. *The Canadian Journal of Urology, 18 (5)*: 5875–83.

Stewart, B.W. & Wild, C.P. (Eds.). (2014). *World Cancer Report 2014*. Lyon, France: International Agency for Research on Cancer.

Anderson-Jackson, L., McGrowder, D.A., & Alexander-Lindo, R. (2012). Prostate specific antigen and Gleason score in men with prostate cancer at a private diagnostic radiology centre in Western Jamaica. *Asian Pac J Cancer Prev. ;13(4)*: 1453-6.

Kutner, M.H., Nachtsheim, C.J., & Neter, J. (2014). *Applied Linear Regression Models* (4th ed.). New York, NY; McGraw-Hill/Irwin.

Dean, A., Voss, D., & Draguljic, D. (2017). *Design and Analysis of Experiments* (2nd ed.). New York, NY; Springer International Publishing AG.

Document Information

All of the statistical analyses in this document will be performed using R version 3.5.2 (2018-12-20).

```
sessionInfo()
```

```
## R version 3.5.2 (2018-12-20)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 17134)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
```

```
## [1] kableExtra_1.0.1 dplyr_0.8.0.1    knitr_1.22
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.0        rstudioapi_0.9.0  xml2_1.2.0
##  [4] magrittr_1.5      hms_0.4.2         munsell_0.5.0
##  [7] rvest_0.3.2       tidyselect_0.2.5  viridisLite_0.3.0
## [10] colorspace_1.4-0  R6_2.3.0          rlang_0.3.1
## [13] httr_1.4.0        stringr_1.3.1     highr_0.7
## [16] tools_3.5.2       webshot_0.5.1     xfun_0.4
## [19] htmltools_0.3.6   yaml_2.2.0        assertthat_0.2.0
## [22] digest_0.6.18     tibble_2.0.1      crayon_1.3.4
## [25] formatR_1.6       purrr_0.3.0       readr_1.3.1
## [28] glue_1.3.0        evaluate_0.12     rmarkdown_1.11
## [31] stringi_1.3.1     compiler_3.5.2    pillar_1.3.1
## [34] scales_1.0.0      pkgconfig_2.0.2
```

Appendix

R code

```
knitr::opts_chunk$set(fig.pos = 'H')
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)

# load packages
library(knitr)

# determine output format dynamically
out_type <- knitr::opts_knit$get("rmarkdown.pandoc.to")

# load data sets
#install.packages("data.table")
library(data.table)
data<-
  fread('http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/
  textdatasets/KutnerData/Appendix%20C%20Data%20Sets/APPENC05.txt')


colnames(data) <- c("Id.number",     "PSA.level",     "Cancer.volume",
"Weight",    "Age",   "Benign.prostatic.hyperplasia",
"Seminal.vesicle.invasion",
"Capsular.penetration", "Gleason.score")

attach(data)
plot(PSA.level~Gleason.score,
     main="Scatterplot of Gleason Score versus PSA Level",
     xlab="Gleason Score",
     ylab="PSA Level (mg/ml)",
     pch=8, cex.main=1.25, cex.lab=1.25, cex=1.25 )

par(mfrow=c(1,2))
boxplot(data$Gleason.score,
         boxwex = 0.5, main="Gleason Score", ylab="Score (6, 7, or 8)")
boxplot(data$PSA.level,
         boxwex = 0.5, main="PSA Level", ylab="mg/ml")

library(kableExtra)
library(broom)
library(pander)

mymodel<-lm(PSA.level~factor(Gleason.score), data=data)
```

```
pander(anova(mymodel))
glance(mymodel)[c(1,2,3,4,5)] %>% kable(caption = "Table 2",
  longtable = TRUE, digits = 3)

qqnorm(residuals(mymodel),main="QQ-Plot")
qqline(residuals(mymodel))

shapiro.test(residuals(mymodel))

#Plot fitted values v residuals
plot(fitted(mymodel),residuals(mymodel), main="Fitted Values v. Residuals",
  ylab="Residuals",xlab="Fitted PSA Levels")
abline(h=0)

#non-constant variance test
require(car);ncvTest(mymodel)

#fit second model
model2<-lm(log10(PSA.level)~factor(Gleason.score),data=data)

pander(anova(model2))
glance(model2)[c(1,2,3,4,5)] %>%
kable(caption = "Measures of Model Fit", longtable = TRUE, digits = 3)

shapiro.test(residuals(model2))

require(car);ncvTest(model2)

library(knitr)
library(kableExtra)

text_tbl <- data.frame("Metric" = c("Minimum", "1st Quartile",
   "Median","Mean","3rd Quartile","Max","Std Dev"),
  "PSA Level (mg/ml)" =
    c(round(min(PSA.level),2),
      round(quantile(PSA.level,.25),2),
      round(quantile(PSA.level,.5),2),
      round(mean(PSA.level),2),
      round(quantile(PSA.level,.75),2),
      round(max(PSA.level),2),
      round(sqrt(var(PSA.level)),2)),
  "Cancer Volume" =
    c(round(min(Cancer.volume),2),
      round(quantile(Cancer.volume,.25),2),
      round(quantile(Cancer.volume,.5),2),
      round(mean(Cancer.volume),2),
      round(quantile(Cancer.volume,.75),2),
      round(max(Cancer.volume),2),
      round(sqrt(var(Cancer.volume)),2)),
  "Prostate Weight (gm)" =
    c(round(min(Weight),2),
      round(quantile(Weight,.25),2),
      round(quantile(Weight,.5),2),
      round(mean(Weight),2),
      round(quantile(Weight,.75),2),
      round(max(Weight),2),
      round(sqrt(var(Weight)),2)),
  "Age" =
    c(round(min(Age),2),
      round(quantile(Age,.25),2),
      round(quantile(Age,.5),2),
      round(mean(Age),2),
      round(quantile(Age,.75),2),
      round(max(Age),2),
      round(sqrt(var(Age)),2)),
  "Benign Prostatic Hyperplasia (cm^2)"=
    c(round(min(Benign.prostatic.hyperplasia),2),
      round(quantile(Benign.prostatic.hyperplasia,.25),2),
      round(quantile(Benign.prostatic.hyperplasia,.5),2),
      round(mean(Benign.prostatic.hyperplasia),2),
      round(quantile(Benign.prostatic.hyperplasia,.75),2),
      round(max(Benign.prostatic.hyperplasia),2),
      round(sqrt(var(Benign.prostatic.hyperplasia)),2)),
  "Capsular Penetration (cm)" =
    c(round(min(Capsular.penetration),2),
      round(quantile(Capsular.penetration,.25),2),
      round(quantile(Capsular.penetration,.5),2),
      round(mean(Capsular.penetration),2),
      round(quantile(Capsular.penetration,.75),2),
      round(max(Capsular.penetration),2),
      round(sqrt(var(Capsular.penetration)),2))
)

kable(text_tbl, col.names=
        c("Metric", "PSA Level (mg/ml)",
          "Cancer Volume (cc)","Prostate Weight (gm)",
          "Age","Benign Prostatic Hyperplasia (cm^2)",
          "Capsular Penetration (cm)"),
      caption = "Summary Statistics for Continuous Variables",
      longtable = TRUE
      ) %>%
  kable_styling(bootstrap_options = c("striped")) %>%
        column_spec(1, width = "2cm") %>%
        column_spec(2, width = "1.75cm") %>%
        column_spec(3:4, width = "1.75cm")   %>%
        column_spec(5, width = "1cm")   %>%
        column_spec(6:7, width = "2.5cm") %>%
        row_spec(0,bold=TRUE)

library(dplyr)
library(knitr)
data%>%
  group_by(Seminal.vesicle.invasion)%>%
  summarize(n=n())%>%
  mutate(Proportion=round(n/sum(n),3))%>%
  kable(caption="Seminal Vesicle Invasion Indicator Proportion Table",
      longtable = TRUE)

data%>%
  group_by(Gleason.score)%>%
  summarize(n=n())%>%
  mutate(Proportion=round(n/sum(n),3))%>%
  kable(caption="Gleason Score Proportion Table",
      longtable = TRUE)

pairs(PSA.level~Id.number+Cancer.volume+Weight+Age, data=data)
pairs(PSA.level~Benign.prostatic.hyperplasia+Seminal.vesicle.invasion
    +Capsular.penetration+Gleason.score, data=data)


corrdata<-round(cor(data),3)
rownames(corrdata) <- c("ID #","PSA Level","Cancer Volume", "Weight","Age",
  "Ben. Prost. Hyp.","Sem. Ves. Inv.","Caps. Pen.","Gleason Score")

library(kableExtra)
kable(corrdata, col.names=
        c("ID #","PSA Level","Cancer Volume", "Weight","Age",
        "Ben. Prost. Hyp.",
        "Sem. Ves. Inv.","Caps. Pen.","Gleason Score"),
      caption="Correlations",
      longtable = TRUE) %>%
  kable_styling(bootstrap_options = c("striped")) %>%
        column_spec(c(4,10), width = "1.5cm") %>%
          column_spec(c(1,2,3,5,6,7,8,9), width = "1.25cm") %>%
row_spec(0, bold=TRUE)
```