# STAT 6340 (Statistical and Machine Learning), Spring 2019

Mini Project 2 (Solution)

March 18, 2019

Consider the prostate cancer dataset available on eLearning as `prostate cancer.csv`. It consists of data on 97 men with advanced prostate cancer. We would like to understand how `PSA` level is related to the other predictors in the dataset. Note that `vesinv` is a qualitative variable. You can treat `gleason` as a quantitative variable.

(a) **Perform an exploratory analysis of data.**

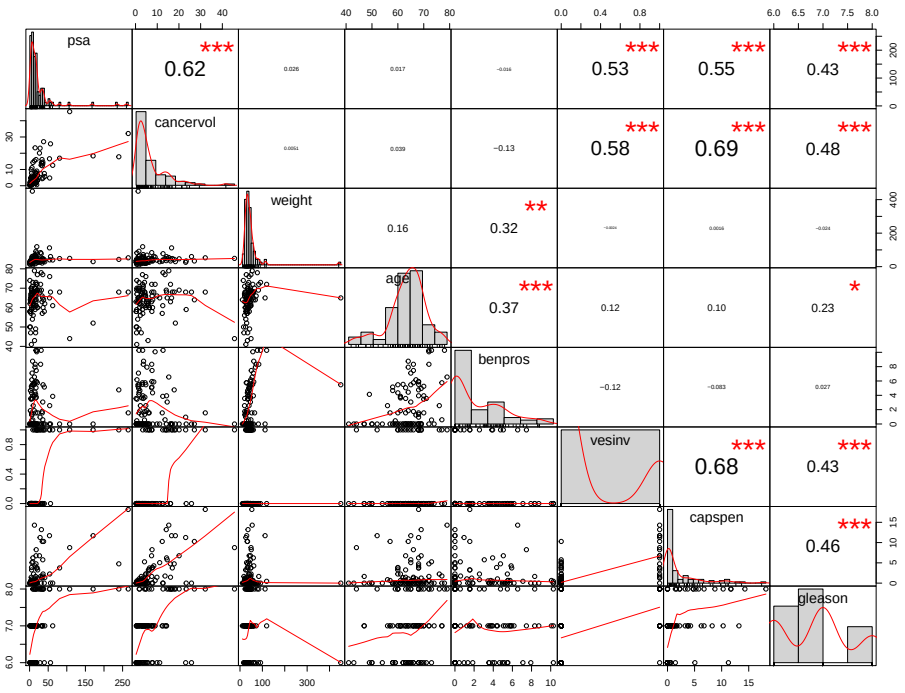


Figure 1: Scatterplots/correlation matrix before log transformation of `psa`

The scatterplots and correlation matrix in Figure 1 show that `psa` is associated with predictors `cancervol`, `vesinv`, `capspen` and `gleason`.

(b) **Is `psa` appropriate as a response variable or a transformation is necessary? In case a transformation of response is necessary, try the natural log transformation or some other transformation and use it for the rest of this problem.**

The distribution of `psa` is highly skewed to the right (see Figure 1), which suggests that an appropriate transformation is needed. We log-transform `psa` and look at its distribution and association with the other predictors.

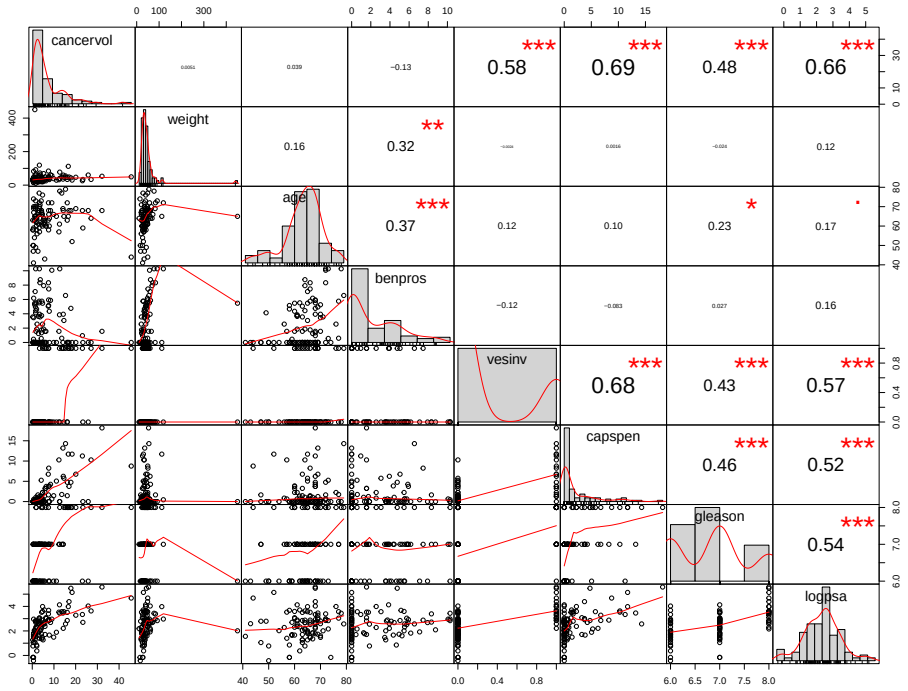


Figure 2: Scatterplots/correlation matrix after log transformation of `psa`

We see from Figure 2 that the distribution of log-transformed `psa` is close to being symmetric. So, we will use log-transformed `psa` for the model building.

(c) **For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.**

| | Predictor | t-test p-value | Significant |
|---|---|---|---|
| 1 | cancervol | 8.47e-12 | Yes |
| 2 | weight | 7.99e-01 | No |
| 3 | age | 8.67e-01 | No |
| 4 | benpros | 8.73e-01 | No |
| 5 | vesinv | 2.61e-08 | Yes |
| 6 | capspen | 5.06e-09 | Yes |
| 7 | gleason | 1.13e-05 | Yes |

Table 1: t-test p-values for each simple linear regression model

Table 1 shows that the predictors `weight`, `age` and `benpros` are not statistically significant in their association with the response `log(psa)`. These findings are in line with the scatterplots

shown in Figure 2.

(d) **Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?**

```
lm(formula = logpsa ~ ., data = data)

Residuals:
Min      1Q  Median      3Q     Max
-1.8831 -0.4663  0.0804  0.4738  1.5322

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.68580    0.99875   -0.69   0.4941
cancervol    0.06945    0.01462    4.75  7.8e-06 ***
weight       0.00138    0.00182    0.76   0.4508
age         -0.00280    0.01172   -0.24   0.8119
benpros      0.08747    0.02961    2.95   0.0040 **
vesinv1      0.78262    0.26834    2.92   0.0045 **
capspen     -0.02652    0.03286   -0.81   0.4218
gleason      0.35815    0.12798    2.80   0.0063 **
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 0.768 on 89 degrees of freedom
Multiple R-squared:  0.589,Adjusted R-squared:  0.557
F-statistic: 18.2 on 7 and 89 DF,  p-value: 7.69e-15
```

Based on the summary, we reject the null hypothesis $H_0 : \beta_j = 0$ for the predictors `cancervol`, `benpros`, `vesinv`, and `gleason`.

(e) **Build a "reasonably good" multiple regression model for these data. Carefully justify all the choices you make in building the model. Be sure to verify the model assumptions.**

The results of part (d) suggest that a reasonable model to start with is the model with `log(psa)` as the response and `cancervol`, `benpros`, `vesinv`, and `gleason` as the predictors. Here, we will use 0.05 as cutoff for significance but other choices are also possible.

```
lm(formula = logpsa ~ cancervol + benpros + vesinv + gleason,
data = data)

Residuals:
Min      1Q  Median      3Q     Max
-1.8853 -0.5028  0.0989  0.5369  1.5662

Coefficients:
Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  -0.6501     0.8100   -0.80  0.42425
cancervol     0.0649     0.0128    5.05  2.2e-06 ***
benpros       0.0914     0.0261    3.51  0.00071 ***
vesinv1       0.6842     0.2364    2.89  0.00475 **
gleason       0.3338     0.1233    2.71  0.00810 **
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 0.761 on 92 degrees of freedom
Multiple R-squared:  0.583,Adjusted R-squared:  0.565
F-statistic: 32.2 on 4 and 92 DF,  p-value: <2e-16
```

The above summary shows that each predictor is significant in the presence of the other three. Next, we perform the partial F-test to check if the predictors `weight`, `age` and `capspen` can be jointly dropped for the full model.

```
Analysis of Variance Table

Model 1: logpsa ~ cancervol + weight + age + benpros + vesinv + capspen +
gleason
Model 2: logpsa ~ cancervol + benpros + vesinv + gleason
Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     89 52.5
2     92 53.2 -3    -0.752 0.43   0.74
```

The anova table indicates that our model is as good as the full model (p-value $= 0.74 \not< 0.05$). Now, we test if the two-way interactions terms are jointly significant.

```
Analysis of Variance Table

Model 1: logpsa ~ cancervol + benpros + vesinv + gleason
Model 2: logpsa ~ (cancervol + benpros + vesinv + gleason)^2
Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     92 53.2
2     86 46.8  6      6.45 1.98  0.078 .
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

From the anova table we can see that the two-way interaction terms do not need to included into our model (p-value $= 0.078 \not< 0.05$). Next we perform a diagnostics of the model assumptions. The plots in Figure 3 show that there are no serious violations of the model assumptions in terms of homoscedasticity and normality. But there is a trend in the residual plot that seems problematic. Moreover, there is evidence of dependence in errors over time, but we will ignore this issue. To get an improved model, we modify our model by log-transforming the predictor `cancervol` as its distribution is also right-skewed. After fitting the model and performing similar analysis as above, we can conclude that the new modified model is better than the previous one. Below is the summary for the modified model.

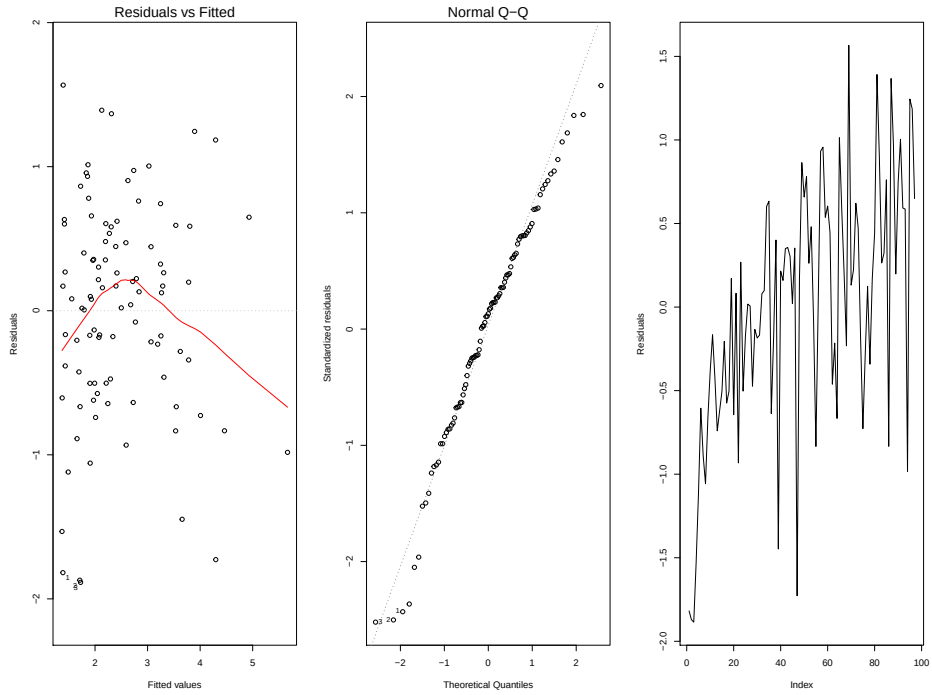


Figure 3: Diagnostics plots

```
lm(formula = logpsa ~ log(cancervol) + benpros + vesinv + gleason,
data = data)

Residuals:
Min      1Q  Median      3Q     Max
-1.6754 -0.3803  0.0392  0.5148  1.9260

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.3163     0.7695   -0.41   0.6820
log(cancervol)   0.5050     0.0796    6.35  8.2e-09 ***
benpros          0.0642     0.0245    2.62   0.0102 *
vesinv1          0.6588     0.2175    3.03   0.0032 **
gleason          0.2629     0.1180    2.23   0.0283 *
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 0.717 on 92 degrees of freedom
Multiple R-squared:  0.63,Adjusted R-squared:  0.614
F-statistic: 39.1 on 4 and 92 DF,  p-value: <2e-16
```

As seen in Figure 4, there is a visible improvement in the residual plot as there is little evidence of a trend. Dependence in errors over time continues to be an issue, but this will be ignored.

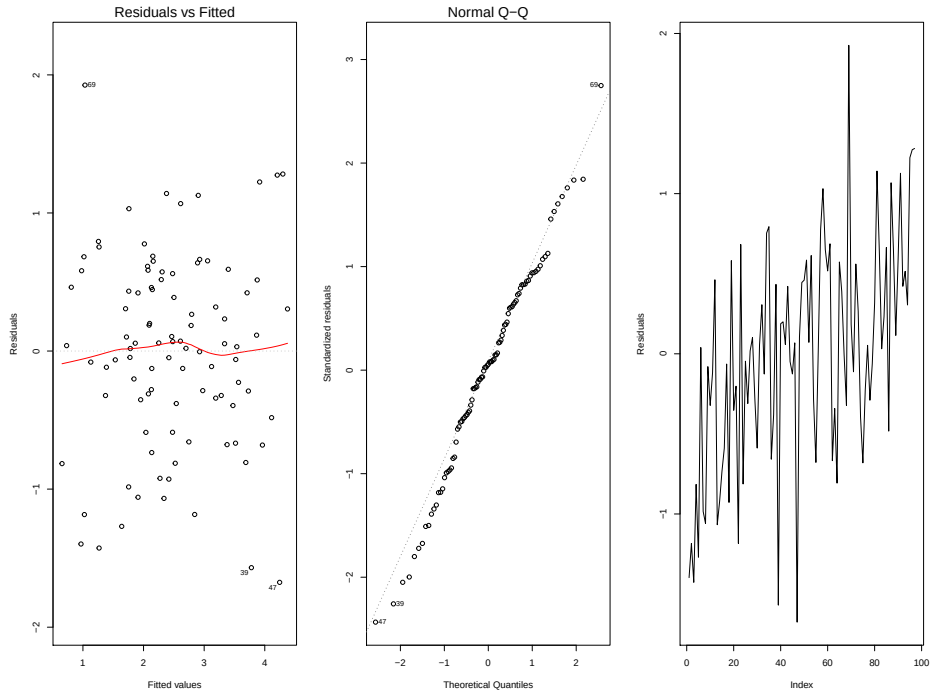


Figure 4: Diagnostics plots

(f) **Write the final model in equation form, being careful to handle qualitative predictors (if any) properly.**

$$log(psa) = -0.3163 + 0.5050 \cdot log(cancervol) + 0.0642 \cdot benpros + 0.6588 \cdot vesinv + 0.2629 \cdot gleason$$

(g) **Use the final model to predict the `PSA` level for a patient whose quantitative predictors are at the sample means of the variables and qualitative predictors (if any) are at the most frequent category.**

$\exp(2.64) = 14$

```
library(PerformanceAnalytics)
data=read.csv('prostate_cancer.csv')[,-1] # read data .csv file omiting IDs

# (a)
chart.Correlation(data) # matrix of scatterplots/correlations

# (b)
data$logpsa=log(data$psa) # log transform psa
data=data[,-1] # drop psa
chart.Correlation(data) # matrix of scatterplots/correlations after transformation

data$vesinv=as.factor(data$vesinv) # treat vesinv as qualitative
attach(data)
```

```
# (c)
options(digits = 3)
pval=c()
# regress log transformed psa against each predictor separately
for (i in 1:(ncol(data)-1))
{
result=summary(lm(logpsa~data[,i]))
pval=c(pval,result$coefficients[2,4]) # record p-values of t-test
}
M=data.frame(pval)
M=cbind(names(data)[1:(ncol(data)-1)],M,ifelse(pval<0.05,'Yes','No'))
names(M)=c('Predictor','T-test p-value','Significant')

print(M) # print results

# (d)

# regress log(psa) against all predictors
mod1=lm(logpsa~.,data = data)
summary(mod1)

# (e)

#  regress log(psa) against cancervol, benpros, vesinv and  gleason
mod2=lm(logpsa~cancervol+benpros+vesinv+gleason,data=data)

# test if remaining predictors (weight,age,capspen) are significant
anova(mod1,mod2)

# regress log(psa) against cancervol, benpros, vesinv and  gleason
# including all two-way interactions
mod3=lm(logpsa~(cancervol+benpros+vesinv+gleason)^2,data=data)

# test if two-way interactions are significant
anova(mod2,mod3)

# summary of mod2
summary(mod2)

# Check model assumptions
# residuals vs fitted values plot and qq-plot of standardized residuals
par(mfrow=c(1,3))
plot(mod2,1:2)
plot(mod2$residuals,ylab='Residuals',type='l')

# mod2 may be considered acceptable, but let's see if we can find a better model

# regress log(psa) against all predictors where chancervol is log-transformed
```

```
mod4=lm(logpsa~log(cancervol)+weight+age+capspen+benpros+vesinv+gleason,data = data)
summary(mod4)

# regress log(psa) against log(cancervol), benpros, vesinv and  gleason
mod5=lm(logpsa~log(cancervol)+benpros+vesinv+gleason,data=data)

# test if remaining predictors (weight,age,capspen) are significant
anova(mod4,mod5)

# regress log(psa) against cancervol, benpros, vesinv and  gleason
# including all two-way interactions
mod6=lm(logpsa~(log(cancervol)+benpros+vesinv+gleason)^2,data=data)

# test if two-way interactions are significant
anova(mod5,mod6)

# mod5 - final model
summary(mod5)

# Check model assumptions
# residuals vs fitted values plot and qq-plot of standardized residuals
plot(mod5,1:2)
plot(mod5$residuals,ylab='Residuals',type='l')

# (f)

# predict psa for a patient whose quantitative predictors are at the sample means
# of the variables and qualitative predictors are at the most frequent category

exp(predict(mod5,data.frame(cancervol=mean(cancervol),
benpros=mean(benpros),
vesinv=levels(data$vesinv)[which.max(table(data$vesinv))],
gleason=mean(gleason))))
```