BE7023 Homework 1

Mike Lape

September 12, 2018

```r
library(faraway)
data("prostate")
```

1. Describe the data, include size, top ten rows, and summary statistics.

```r
# The prostate dataset contains records for 97 men with prostate cancer who
# were going to have a radical prostatectomy.
dim(prostate)
```

```
## [1] 97  9
```

```r
# The data has 97 rows/observations and 9 columns/variables
# Below are the top 10 rows
head(prostate, 10)
```

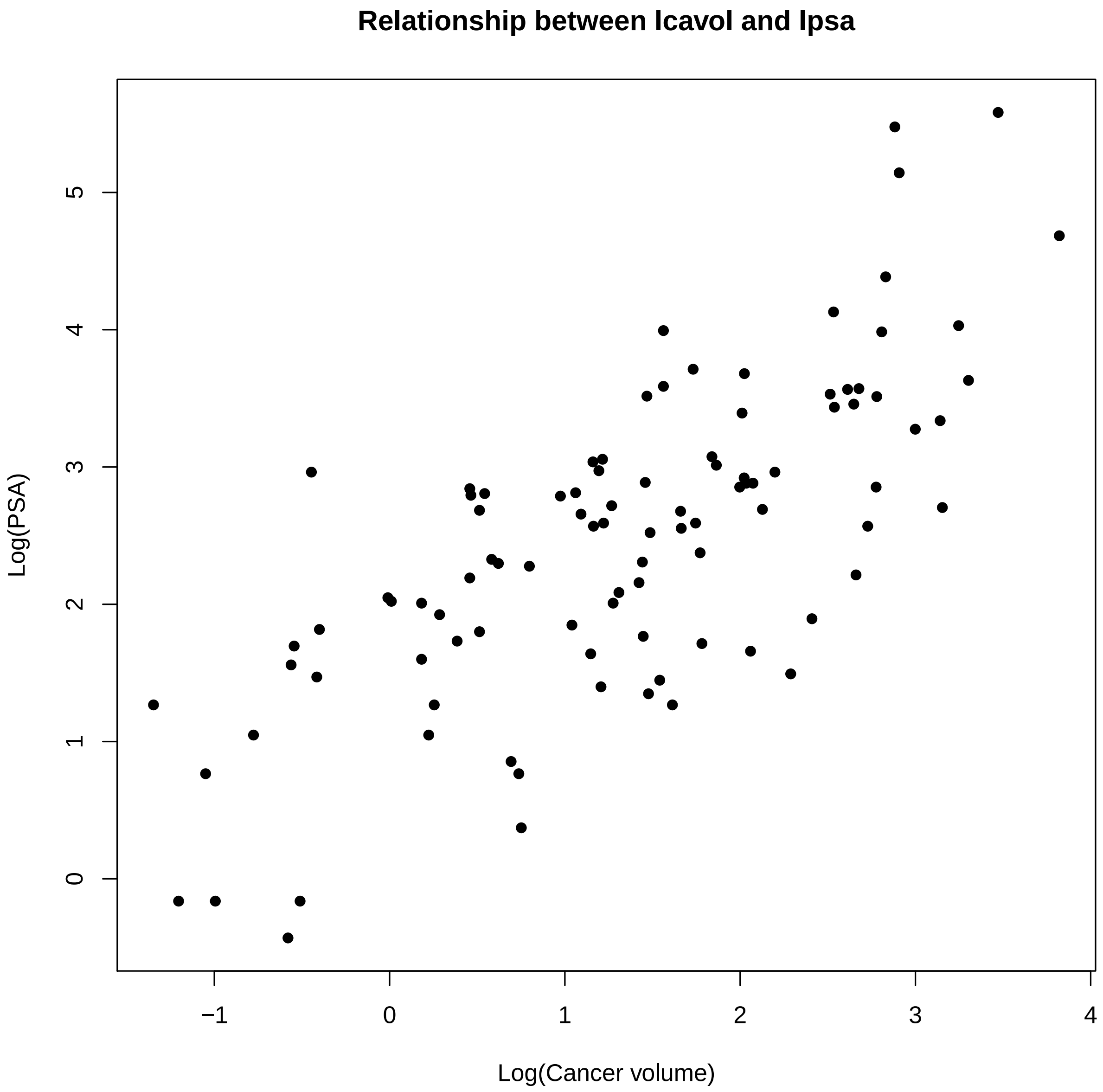
```
##        lcavol lweight age      lbph svi      lcp gleason pgg45     lpsa
## 1  -0.5798185  2.7695  50 -1.386294   0 -1.38629       6     0 -0.43078
## 2  -0.9942523  3.3196  58 -1.386294   0 -1.38629       6     0 -0.16252
## 3  -0.5108256  2.6912  74 -1.386294   0 -1.38629       7    20 -0.16252
## 4  -1.2039728  3.2828  58 -1.386294   0 -1.38629       6     0 -0.16252
## 5   0.7514161  3.4324  62 -1.386294   0 -1.38629       6     0  0.37156
## 6  -1.0498221  3.2288  50 -1.386294   0 -1.38629       6     0  0.76547
## 7   0.7371641  3.4735  64  0.615186   0 -1.38629       6     0  0.76547
## 8   0.6931472  3.5395  58  1.536867   0 -1.38629       6     0  0.85442
## 9  -0.7765288  3.5395  47 -1.386294   0 -1.38629       6     0  1.04732
## 10  0.2231436  3.2445  63 -1.386294   0 -1.38629       6     0  1.04732
```

```r
# and here are the summary statistics for the prostate dataset.
summary(prostate)
```

```
##      lcavol           lweight           age             lbph        
##  Min.   :-1.3471   Min.   :2.375   Min.   :41.00   Min.   :-1.3863  
##  1st Qu.: 0.5128   1st Qu.:3.376   1st Qu.:60.00   1st Qu.:-1.3863  
##  Median : 1.4469   Median :3.623   Median :65.00   Median : 0.3001  
##  Mean   : 1.3500   Mean   :3.653   Mean   :63.87   Mean   : 0.1004  
##  3rd Qu.: 2.1270   3rd Qu.:3.878   3rd Qu.:68.00   3rd Qu.: 1.5581  
##  Max.   : 3.8210   Max.   :6.108   Max.   :79.00   Max.   : 2.3263  
##       svi              lcp             gleason          pgg45       
##  Min.   :0.0000   Min.   :-1.3863   Min.   :6.000   Min.   :  0.00  
##  1st Qu.:0.0000   1st Qu.:-1.3863   1st Qu.:6.000   1st Qu.:  0.00  
##  Median :0.0000   Median :-0.7985   Median :7.000   Median : 15.00  
##  Mean   :0.2165   Mean   :-0.1794   Mean   :6.753   Mean   : 24.38  
##  3rd Qu.:0.0000   3rd Qu.: 1.1786   3rd Qu.:7.000   3rd Qu.: 40.00  
##  Max.   :1.0000   Max.   : 2.9042   Max.   :9.000   Max.   :100.00  
##       lpsa        
##  Min.   :-0.4308  
##  1st Qu.: 1.7317  
##  Median : 2.5915  
##  Mean   : 2.4784
##  3rd Qu.: 3.0564
##  Max.   : 5.5829
```

2. Plot data with x = lcavol and y = lpsa.

```
plot(prostate$lcavol, prostate$lpsa, xlab = "Log(Cancer volume)", ylab = "Log(PSA)",
     main = "Relationship between lcavol and lpsa", pch = 16 )
```



```
# The plot ranges from a little below -1 and almost up to 4 in the x-axis, and
# ranges from just below 0 to just above 5 in the y-axis.  It looks like these
# two features, lcavol and lpsa, have a pretty linear relationship with each other, but a linear fit wi
```

3. Fit simple linear regression model with y = lpsa and x = lcavol. Write the prediction equation. Report R2 and comment on it. Estimate population standard deviation

```r
mod <- lm(lpsa ~ lcavol, prostate)

# Get coefficients to write equation:
mod$coefficients
```

```
## (Intercept)      lcavol
##   1.5072979   0.7193201
```

```r
# Prediction Model: lpca = 1.507 + 0.719 * lcavol
summary(mod)$adj.r.squared
```

```
## [1] 0.5345838
```

```r
# The R2 value is 0.535, which suggests that this linear model doesn't fit the
# data very well.

# To get a good estimate of the population standard deviation we can calculate
# RMSE
pop_sd <- summary(mod)$sigma
paste("We thus estimate the population standard deviation to be ", round(pop_sd,3))
```

```
## [1] "We thus estimate the population standard deviation to be  0.787"
```

4. Prostate specific antigen (PSA) is an enzyme excreted from epithelial cells on the prostate. In men with normal prostates PSA is found in the blood in small quantities, but is often found at a higher level in men with prostate cancer or other prostate issues. It is therefore used as a diagnostic test for prostate cancer. By taking some blood from the man and measuring the PSA level they can determine if he has a healthy prostate or an abnormal one that requires further investigation.

5. Transform regression model back to original variables, comment on resultant model. Both variables being considered here are log transformed. So lpsa is really log(psa), while lcavol is log(cavol). To get the prediction model out of the log form we transform it as follows.

log(psa) = 1.507 + 0.719 * log(cavol)
log(psa) = log(e^(1.507)) + 0.719 * log(cavol)
log(psa) = log(4.51) + log(cavol^0.719)
log(psa) = log(4.51 * cavol^0.719)
psa = 4.51 * cavol^0.719
(psa / cavol^0.719) = 4.51

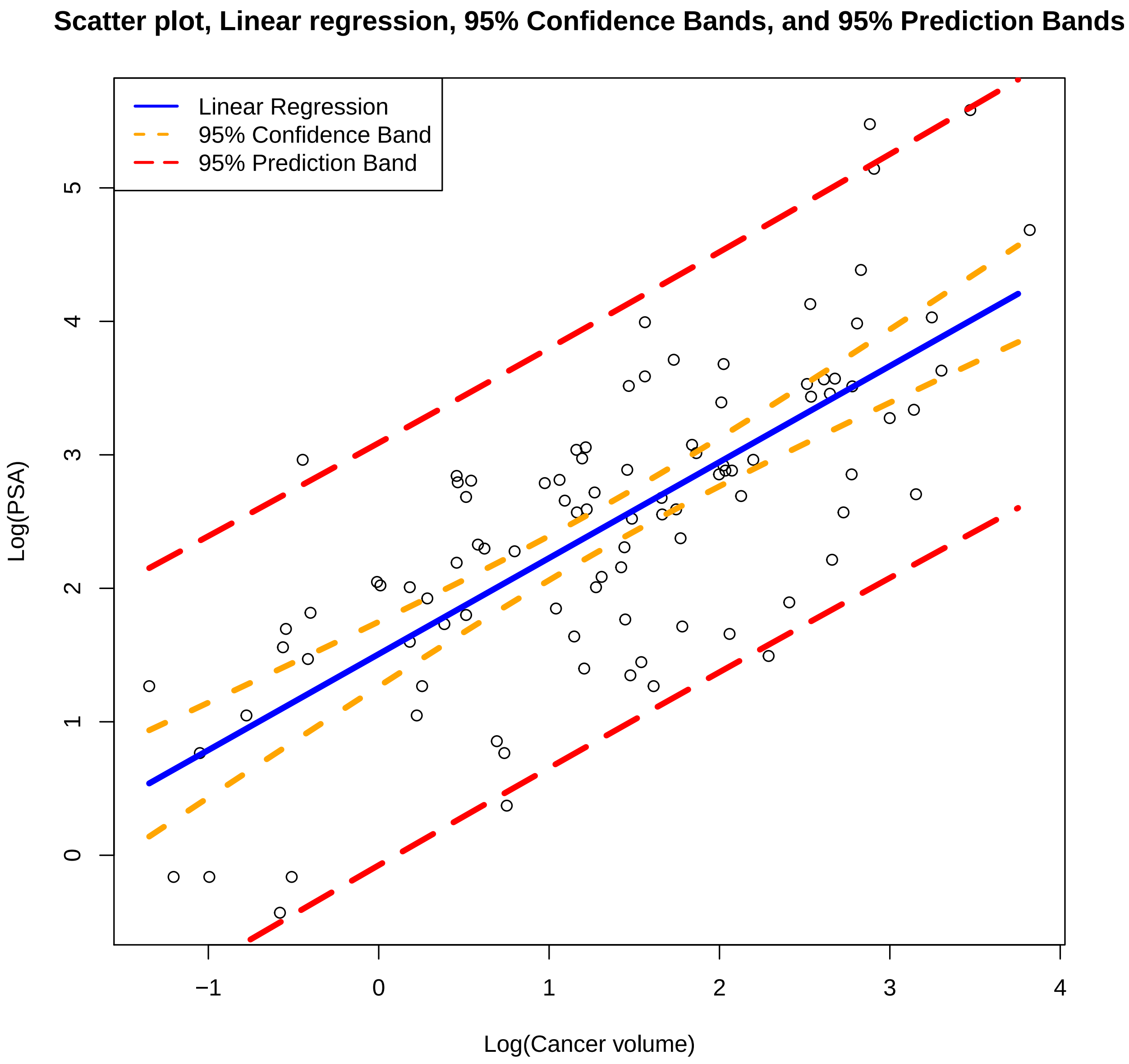
This tells us that the average ratio between psa and cancer volume to the power of 0.719 is 4.51. So if we know only the PSA or only the cancer volume then we can calculate what the average value of the other variable using this formula.

6. Build 95% confidence bands as well as prediction bands around regression line.

```r
# First we need to generate some simulated lcavol lpsa pairs.
# Using min and max values to define range
sim_lcavol <-  seq(-1.3471,3.8210,0.1)
conf <- predict(mod, list(lcavol = sim_lcavol), int = "c")
pred <- predict(mod, list(lcavol = sim_lcavol), int = "p")

# plot confidence intervals
```

```r
plot(prostate$lcavol, prostate$lpsa, xlab = "Log(Cancer volume)", ylab = "Log(PSA)", lwd = 1)
title(main = "Scatter plot, Linear regression, 95% Confidence Bands, and 95% Prediction Bands")
matlines(sim_lcavol, conf, lty = c(1,2,2), col = c("blue","orange", "orange"), lwd = 4)
matlines(sim_lcavol, pred, lty = c(1,5,5), col = c("blue", "red", "red"), lwd = 4)
legend("topleft", legend = c("Linear Regression", "95% Confidence Band", "95% Prediction Band"), lty = c
```


Scatter plot, Linear regression, 95% Confidence Bands, and 95% Prediction Bands
Linear Regression
95% Confidence Band
95% Prediction Band
Log(PSA)
Log(Cancer volume)