Bruce Campell ST 503 HW 1

Problems 1,3 Chapter 2 Faraway, Julian J.. Linear Models with R, Second Edition (Chapman & Hall/CRC Texts in Statistical Science). CRC Press.

Bruce Campbell

27 August, 2017

Sun Aug 27 15:38:15 2017

Problem 1.1

The dataset teengamb concerns a study of teenage gambling in Britain. Make a numerical and graphical summary of the data, commenting on any features that you find interesting. Limit the output you present to a quantity that a busy reader would find sufficient to get a basic understanding of the data.

This report was rendered in R Markdown with the option `echo=FALSE`. We assume a busy reader does not want to see the code.

Load and inspect the data.

When loading and inspecting the data we will note which variables are numberic, and which are strings, we'll also be on the lookout for variables that we may want to encode as factors. Here we note that gender is a candidate for such encoding.

```
##   sex status income verbal gamble
## 1   1     51   2.00      8    0.0
## 2   1     28   2.50      8    0.0
## 3   1     37   2.00      6    0.0
## 4   1     28   7.00      4    7.3
## 5   1     65   2.00      8   19.6
## 6   1     61   3.47      6    0.1
```

Check for missing data

Table 1: Number of missing elements in data set

missing.count
0

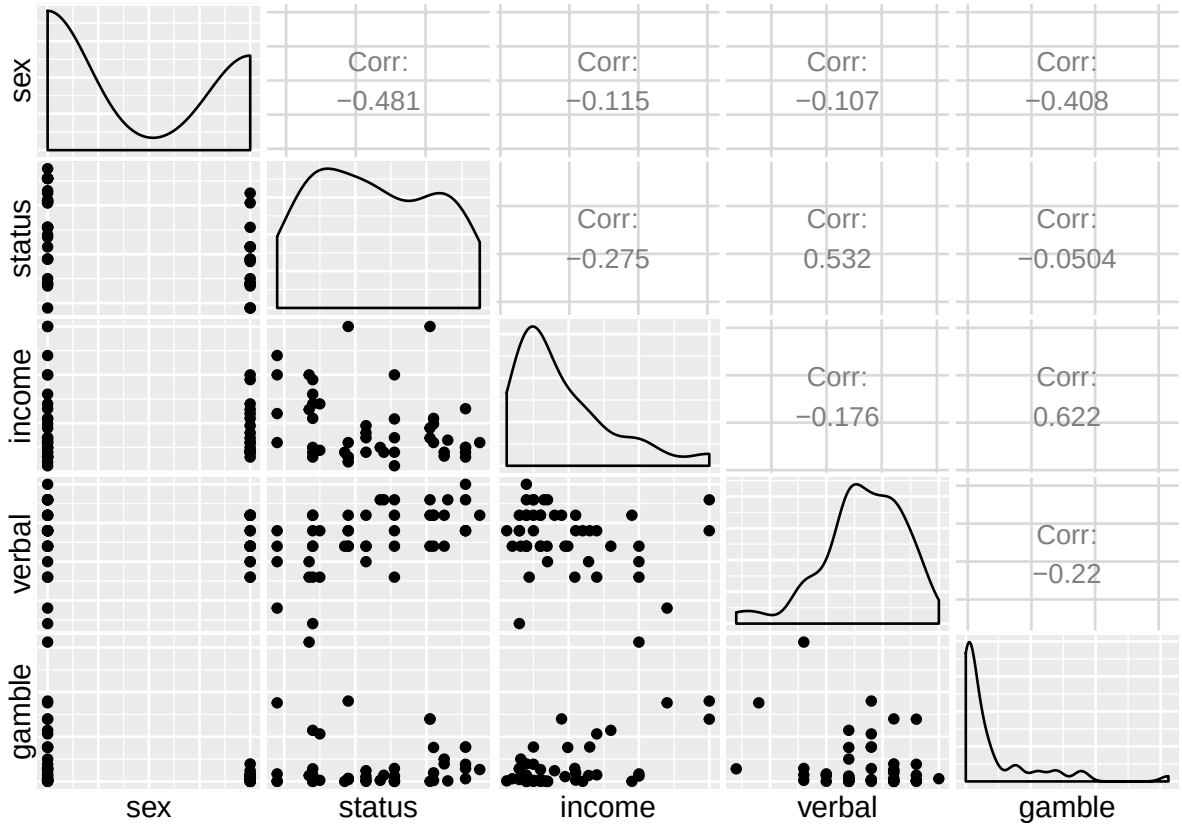
Calculate summary statistics for the variables

```
##       sex             status          income           verbal
##  Min.   :0.0000   Min.   :18.00   Min.   : 0.600   Min.   : 1.00
##  1st Qu.:0.0000   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00
##  Median :0.0000   Median :43.00   Median : 3.250   Median : 7.00
##  Mean   :0.4043   Mean   :45.23   Mean   : 4.642   Mean   : 6.66
```

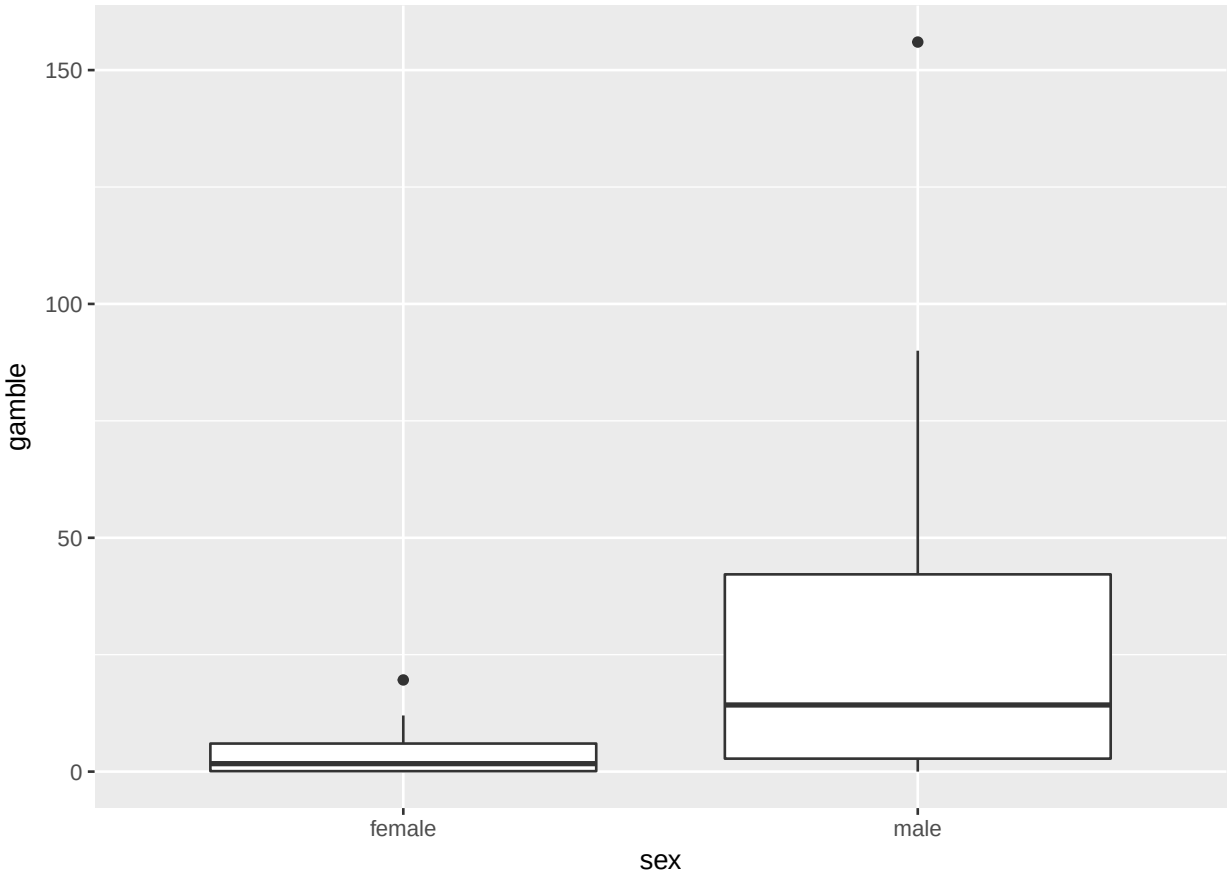
```
##  3rd Qu.:1.0000   3rd Qu.:61.50   3rd Qu.: 6.210   3rd Qu.: 8.00
##  Max.   :1.0000   Max.   :75.00   Max.   :15.000   Max.   :10.00
##      gamble
##  Min.   :  0.0
##  1st Qu.:  1.1
##  Median :  6.0
##  Mean   : 19.3
##  3rd Qu.: 19.4
##  Max.   :156.0
```

Plot the features

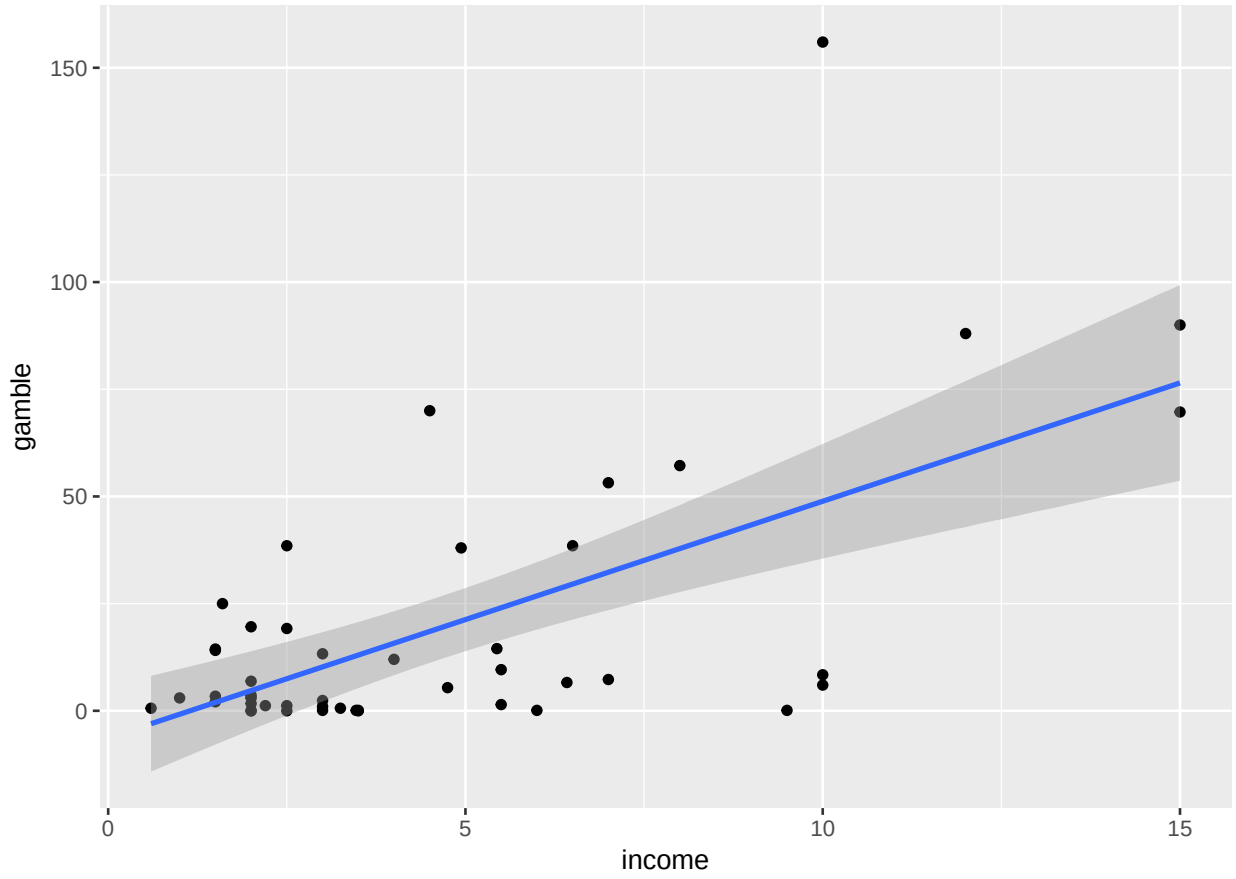
This data come from a study of teenage gambpling in Britan. The response vaiable in this case is gamble, and the other variables in the data set are candidates for predictors in any modelling we do. When creating plots, we'll be interested in how the predictors relate to the response. We'll also be on the lookout for outliers.



We note that gender seems to vary with gamble. We can see this better in a box plot.



We also note that income seems to have an association with gambling.



We observe a data element with a large value of gamble. This needs to be noted and considered when we evaluate any models that we fit with this data.

Problem 1.3

The dataset prostate is from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. Make a numerical and graphical summary of the data as in the first question.

Load and inspect the data from the

```
##       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
```

We note that the documentation provides the following details the meaning of the features ;

- lcavol=log(cancer volume)
- lweight=log(prostate weight)
- age=age
- lbph=log(benign prostatic hyperplasia amount)

- svi=seminal vesicle invasion
- lcp=(capsular penetration)
- gleason=leason score
- pgg45=percentage Gleason scores 4 or 5
- lpsa=log(prostate specific antigen)

Check for missing data

Table 2: Number of missing elements in data set

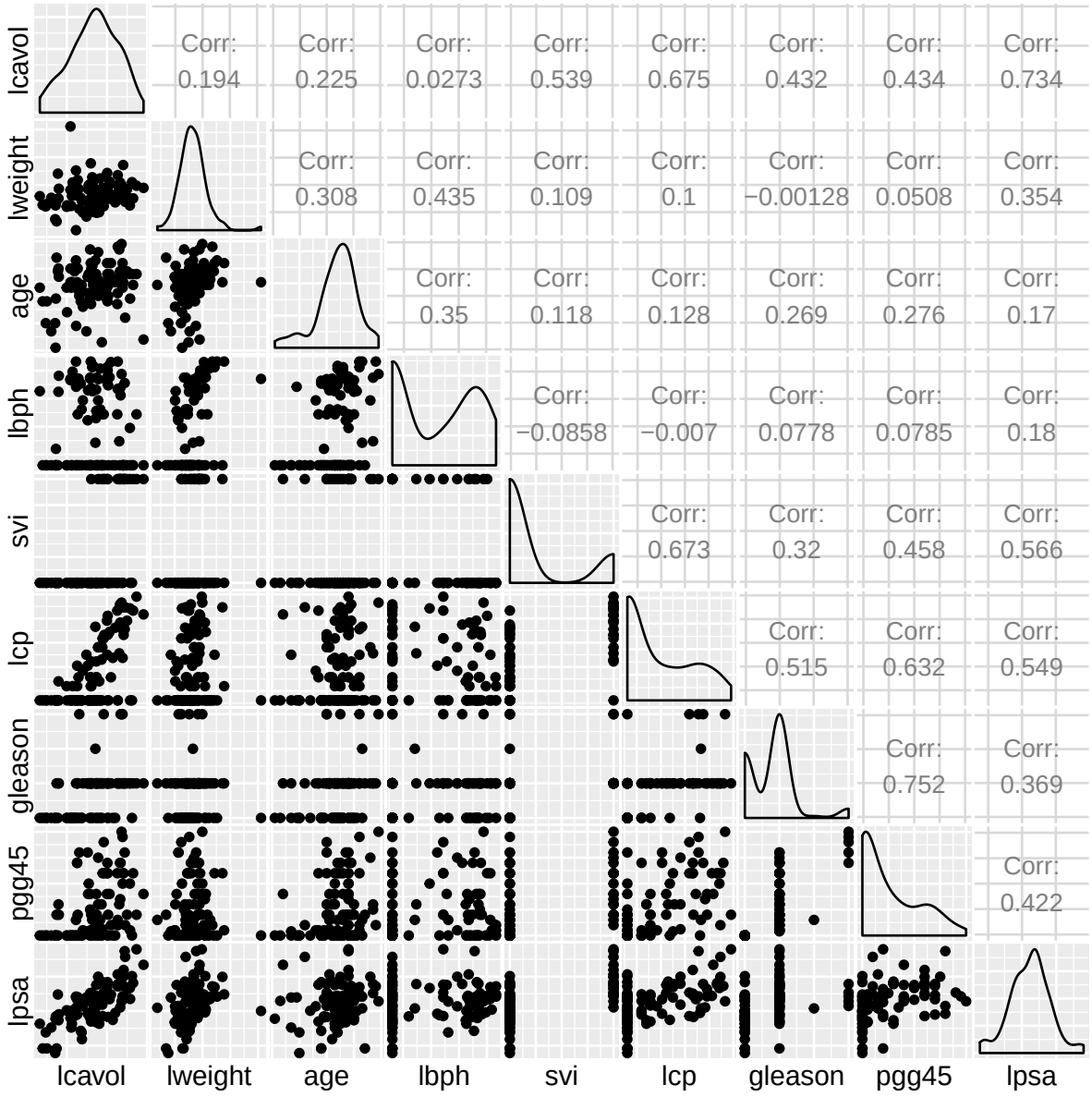
missing.count
0

We note that the gleason score might be a variable that is a candidate for encoding as a factor variable.

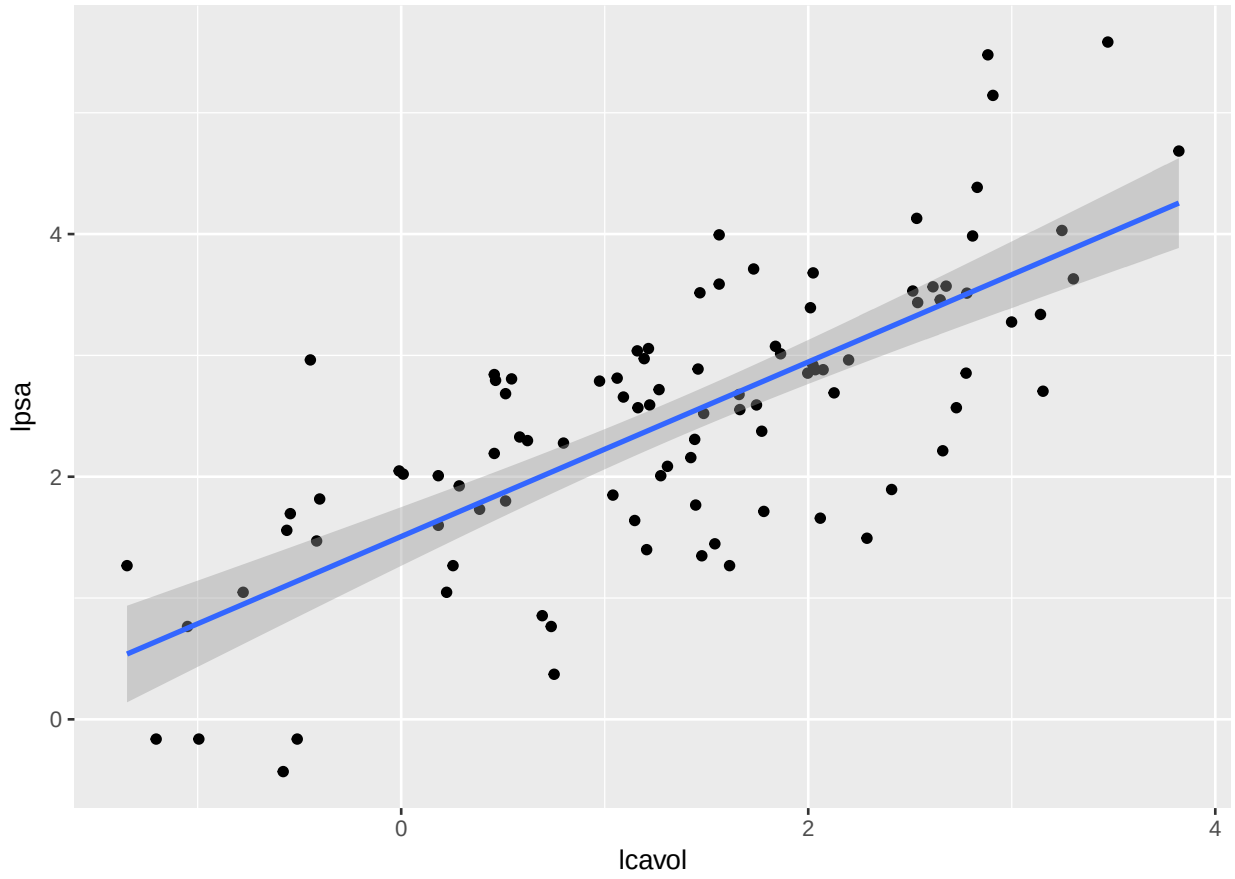
Summary statistics for the prostate data

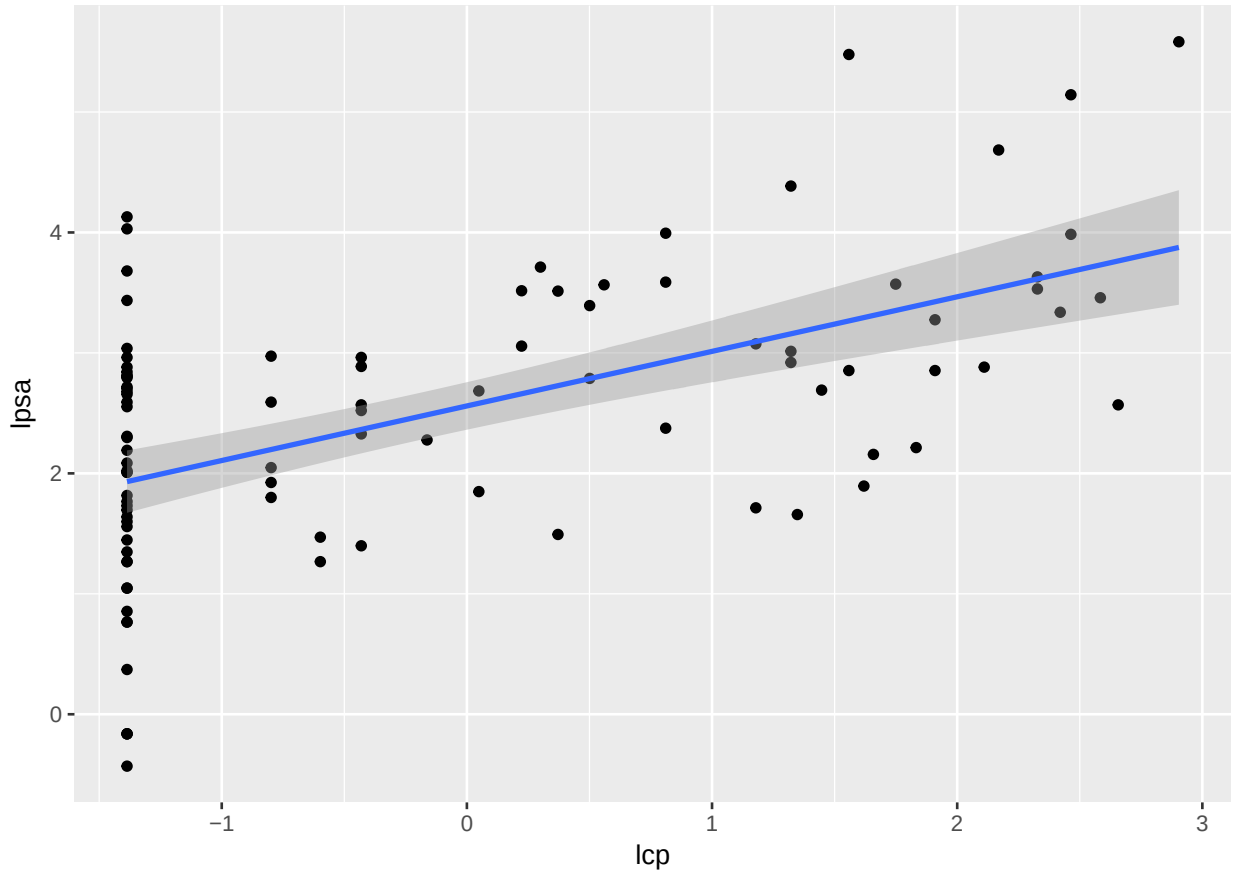
```
##      lcavol           lweight           age             lbph
##  Min.   :-1.3471   Min.   :2.375   Min.   :41.00   Min.   :-1.3863
##  1st Qu.: 0.5128   1st Qu.:3.376   1st Qu.:60.00   1st Qu.:-1.3863
##  Median : 1.4469   Median :3.623   Median :65.00   Median : 0.3001
##  Mean   : 1.3500   Mean   :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
```

Plot of the variables for the prostate data



We note a relationship between lcavol and lpsa, and between lcp and lpsa. We also note that there is a replationship between the gleason score and lpsa and pgg45.


lpsa
4
2
0
0
2
4
lcavol



We note that there are a number of lcp values at -1.39. We should follow up with the study authors to undersand this better. It may be due to limitations or constriants in instrumentation that was used to make the measurements.