HA 1

Xinrui Hu

2023-02-11

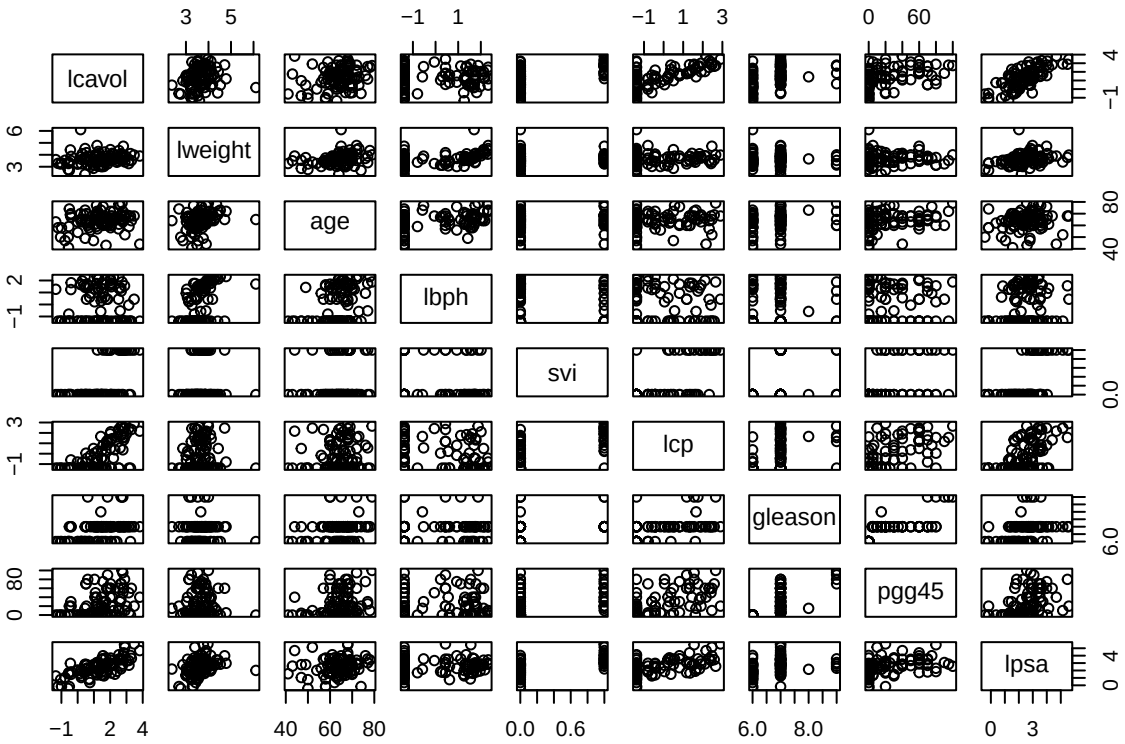
Question 1 Calculate descriptive statistics of each of the variables.

```
library("faraway")
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
```

Question 2 Create a scatter plot matrix for all the variables

```
pairs(prostate)
```



Question 3 Calculate a (Pearson) correlation matrix for all the variables.

```
cor(prostate)
```

```
##             lcavol      lweight       age        lbph         svi         lcp
## lcavol  1.00000000  0.194128387 0.2249999  0.02734971  0.53884500  0.67531058
## lweight 0.19412839  1.000000000 0.3075247  0.43493174  0.10877818  0.10023889
## age     0.22499988  0.307524741 1.0000000  0.35018592  0.11765804  0.12766778
## lbph    0.02734971  0.434931744 0.3501859  1.00000000 -0.08584327 -0.00699944
## svi     0.53884500  0.108778185 0.1176580 -0.08584327  1.00000000  0.67311122
## lcp     0.67531058  0.100238891 0.1276678 -0.00699944  0.67311122  1.00000000
## gleason 0.43241705 -0.001283003 0.2688916  0.07782044  0.32041222  0.51482991
## pgg45   0.43365224  0.050846195 0.2761124  0.07846000  0.45764762  0.63152807
## lpsa    0.73446028  0.354121818 0.1695929  0.17980950  0.56621818  0.54881316
##              gleason     pgg45      lpsa
## lcavol   0.432417052 0.4336522 0.7344603
## lweight -0.001283003 0.0508462 0.3541218
## age      0.268891599 0.2761124 0.1695929
## lbph     0.077820444 0.0784600 0.1798095
## svi      0.320412221 0.4576476 0.5662182
## lcp      0.514829912 0.6315281 0.5488132
## gleason  1.000000000 0.7519045 0.3689867
## pgg45    0.751904512 1.0000000 0.4223157
## lpsa     0.368986693 0.4223157 1.0000000
```

Question 4 Show the same matrix again, but round the correlations (use three decimal places). Which variable has the highest correlation with lcavol (The variable represents the log(cancer volume in cm3)?

```
round(cor(prostate),digits = 2)
```

```
##         lcavol lweight  age  lbph   svi   lcp gleason pgg45 lpsa
## lcavol    1.00    0.19 0.22  0.03  0.54  0.68    0.43  0.43 0.73
```

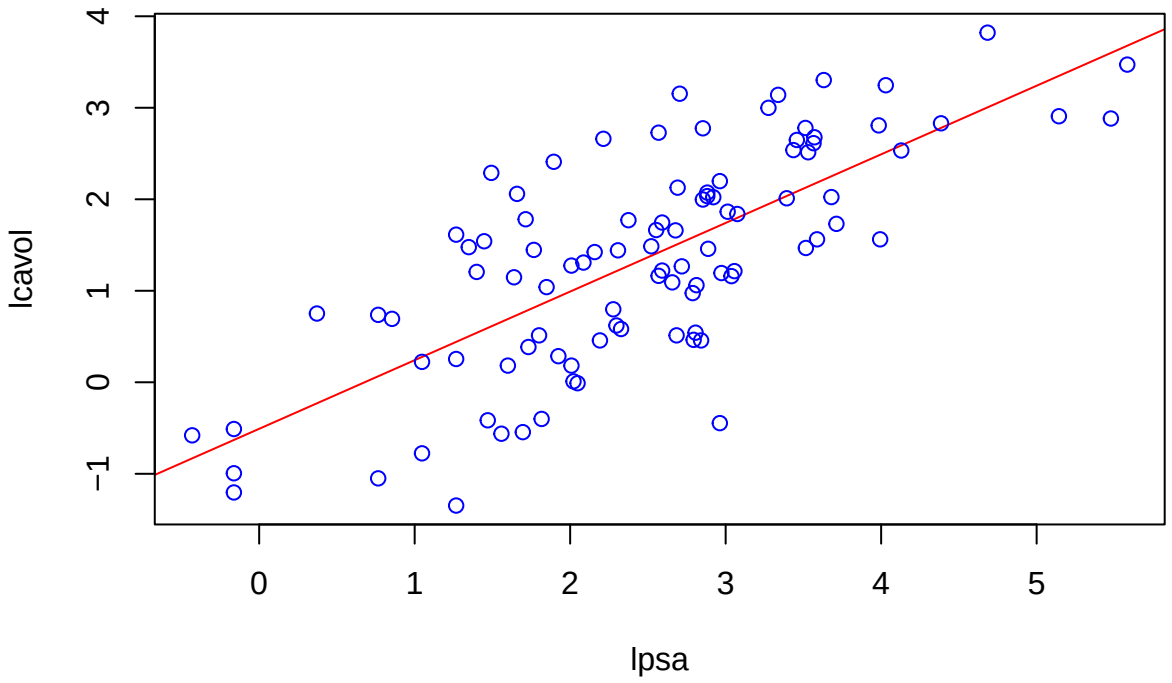
```
## lweight   0.19    1.00 0.31  0.43  0.11  0.10    0.00  0.05 0.35
## age       0.22    0.31 1.00  0.35  0.12  0.13    0.27  0.28 0.17
## lbph      0.03    0.43 0.35  1.00 -0.09 -0.01    0.08  0.08 0.18
## svi       0.54    0.11 0.12 -0.09  1.00  0.67    0.32  0.46 0.57
## lcp       0.68    0.10 0.13 -0.01  0.67  1.00    0.51  0.63 0.55
## gleason   0.43    0.00 0.27  0.08  0.32  0.51    1.00  0.75 0.37
## pgg45     0.43    0.05 0.28  0.08  0.46  0.63    0.75  1.00 0.42
## lpsa      0.73    0.35 0.17  0.18  0.57  0.55    0.37  0.42 1.00
```

```
# lpsa has the highest correlation with lcavol
```

Question 5 Show a scatter plot. Put lcavol in the y-axis and the variable you found in question 4 in the x-axis.

Include a regression line and label the axis.

```
lg <- lm(lcavol ~ lpsa, data = prostate)
plot(prostate$lpsa, prostate$lcavol,
     col = "blue",
     xlab = "lpsa", ylab = "lcavol",
     abline(lg, col = "red"))
```



Question 6 Update the regression model you created for the regression line in question 5 by adding a second predictor: age.

Show the regression model output. What percentage of the variance of the outcome variables is explained by the two predictors?

```
lg1 <- lm(lcavol ~ lpsa + age, data = prostate)
summary(lg1)
```

```
##
## Call:
## lm(formula = lcavol ~ lpsa + age, data = prostate)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -2.23486 -0.62468  0.02114  0.54421  1.71757
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.50978    0.70670  -2.136   0.0352 *
## lpsa         0.73201    0.07170  10.210   <2e-16 ***
## age          0.01637    0.01112   1.473   0.1442
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7992 on 94 degrees of freedom
## Multiple R-squared:  0.5498, Adjusted R-squared:  0.5402
## F-statistic:  57.4 on 2 and 94 DF,  p-value: < 2.2e-16
```

```
# 0.5498
```

Question 7 What is the Residual Standard Error of the model you created in Question 6?

```
sqrt(deviance(lg1) / df.residual(lg1))
```

```
## [1] 0.7991732
```

```
# 0.7991732
```