Case Study

Pratheek Eravelli & Kanav Bhatnagar

3/26/2022

Project Setup

We will use the following packages for this homework assignment. We will also read in data from the csv file.

```
library(ggplot2)
prostate = read.table('prostate.txt', header = FALSE)
```

Now, we add labels to the columns.

```
colnames(prostate) = c('ID', 'psalevel', 'cancervolume', 'prostateweight',
                       'age', 'hyperplasia', 'svi', 'capsular', 'gleason')
head(prostate)
```

```
##   ID psalevel cancervolume prostateweight age hyperplasia svi capsular gleason
## 1  1    0.651       0.5599         15.959  50           0   0        0       6
## 2  2    0.852       0.3716         27.660  58           0   0        0       7
## 3  3    0.852       0.6005         14.732  74           0   0        0       7
## 4  4    0.852       0.3012         26.576  58           0   0        0       6
## 5  5    1.448       2.1170         30.877  62           0   0        0       6
## 6  6    2.160       0.3499         25.280  50           0   0        0       6
```

Analysis

Building the initial model

First, we build a regression model with the required variables, and look at the summary statistics.

```
prostate_lm = lm(psalevel ~ cancervolume + age + hyperplasia + svi + capsular + gleason,
                 data = prostate)
summary(prostate_lm)
```

```
##
## Call:
## lm(formula = psalevel ~ cancervolume + age + hyperplasia + svi +
##     capsular + gleason, data = prostate)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.491  -8.199  -0.080   5.923 167.267
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -14.7460    40.1894  -0.367 0.714545
## cancervolume   2.0375     0.5894   3.457 0.000836 ***
## age           -0.5327     0.4724  -1.128 0.262448
## hyperplasia    1.3518     1.1434   1.182 0.240209
## svi           19.6441    10.8303   1.814 0.073038 .
## capsular       1.0974     1.3265   0.827 0.410273
## gleason        6.9942     5.1489   1.358 0.177741
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31 on 90 degrees of freedom
## Multiple R-squared:  0.4584, Adjusted R-squared:  0.4223
## F-statistic:  12.7 on 6 and 90 DF,  p-value: 2.481e-10
```

Manual model selection

We can see in the summary for the linear model that capsular penetration, i.e. `capsular` is the least significant variable since it has the highest p-value, which is over 0.05. We fit another linear model by removing this variable.

```
prostate_lm_reduced_1 = lm(psalevel ~ cancervolume + age + hyperplasia + svi + gleason,
                           data = prostate)
summary(prostate_lm_reduced_1)
```

```
##
## Call:
## lm(formula = psalevel ~ cancervolume + age + hyperplasia + svi +
##     gleason, data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.839  -8.758   0.206   5.181 163.883
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -18.4353    39.8719  -0.462   0.6449
## cancervolume   2.2595     0.5238   4.313 4.07e-05 ***
## age           -0.5261     0.4715  -1.116   0.2674
## hyperplasia    1.3714     1.1412   1.202   0.2326
## svi           23.6477     9.6720   2.445   0.0164 *
## gleason        7.4688     5.1080   1.462   0.1471
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.94 on 91 degrees of freedom
## Multiple R-squared:  0.4543, Adjusted R-squared:  0.4243
## F-statistic: 15.15 on 5 and 91 DF,  p-value: 8.245e-11
```

Again, we notice that patient age, i.e. `age` is the least significant since it has the highest p-value, which is over 0.05. We fit another linear model by removing this variable.

```
prostate_lm_reduced_2 = lm(psalevel ~ cancervolume + hyperplasia + svi + gleason,
                           data = prostate)
summary(prostate_lm_reduced_2)
```

```
##
## Call:
## lm(formula = psalevel ~ cancervolume + hyperplasia + svi + gleason,
##     data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.262  -9.596   0.477   5.428 164.429
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -43.4846    32.9960  -1.318   0.1908
## cancervolume   2.2995     0.5233   4.394 2.97e-05 ***
## hyperplasia    0.9001     1.0616   0.848   0.3987
## svi           22.5019     9.6301   2.337   0.0216 *
## gleason        6.3942     5.0231   1.273   0.2062
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.99 on 92 degrees of freedom
## Multiple R-squared:  0.4468, Adjusted R-squared:  0.4228
## F-statistic: 18.58 on 4 and 92 DF,  p-value: 3.206e-11
```

Now, we notice that the amount of benign prostatic hyperplasia, i.e. `hyperplasia` is the least significant since it has the highest p-value, which is over 0.05. We fit another linear model by removing this variable.

```
prostate_lm_reduced_3 = lm(psalevel ~ cancervolume + svi + gleason,
                           data = prostate)
summary(prostate_lm_reduced_3)
```

```
##
## Call:
## lm(formula = psalevel ~ cancervolume + svi + gleason, data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.160  -8.338   0.651   6.014 166.891
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -44.1849    32.9358  -1.342   0.1830
## cancervolume   2.2496     0.5192   4.333 3.72e-05 ***
## svi           21.8808     9.5877   2.282   0.0248 *
## gleason        6.8982     4.9802   1.385   0.1693
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 30.94 on 93 degrees of freedom
## Multiple R-squared:  0.4425, Adjusted R-squared:  0.4245
## F-statistic:  24.6 on 3 and 93 DF,  p-value: 8.306e-12
```

Finally, we notice that the Gleason score, i.e. `gleason` is the least significant since it has the highest p-value, which is over 0.05. We fit another linear model by removing this variable.

```
prostate_lm_reduced_4 = lm(psalevel ~ cancervolume + svi,
                           data = prostate)
summary(prostate_lm_reduced_4)
```

```
##
## Call:
## lm(formula = psalevel ~ cancervolume + svi, data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.145  -7.535  -1.129   4.256 170.018
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.060      4.231   0.251   0.8027
## cancervolume    2.477      0.495   5.003 2.62e-06 ***
## svi            24.647      9.423   2.616   0.0104 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.09 on 94 degrees of freedom
## Multiple R-squared:  0.431,  Adjusted R-squared:  0.4189
## F-statistic:  35.6 on 2 and 94 DF,  p-value: 3.098e-12
```

We see that all the variables in the model have a p-value less than 0.05, which means that their slopes are all statistically significant, and they contribute to explaining the variation in the response.

Thus, the final model we have is $\texttt{psalevel} = 1.0603679 + 2.4767238 \cdot \texttt{cancervolume} + 24.6470649 \cdot \texttt{svi}$.

Unusual observations

Model assumptions