Prostate Cancer Study

Tyler Frankenberg

2/13/2022

Import packages

```
library(tidyverse)
```

Import data

```
url <- "https://raw.githubusercontent.com/curdferguson/data621/main/datasets/prostate.txt"

prostate <- read_tsv(url, skip = 1, col_names = c("index", "lcavol", "lweight", "age", "lbph",
"svi", "lcp", "gleason", "pgg45", "lpsa"), show_col_types=FALSE)

prostate <- prostate[,2:10]
prostate$svi <- factor(prostate$svi)
```

Glimpse dataset structure and each column’s summary statistics

```
prostate %>% head(5)
```

```
## # A tibble: 5 x 9
##   lcavol lweight   age  lbph svi     lcp gleason pgg45   lpsa
##    <dbl>   <dbl> <dbl> <dbl> <fct> <dbl>   <dbl> <dbl>  <dbl>
## 1 -0.580    2.77    50 -1.39 0     -1.39       6     0 -0.431
## 2 -0.994    3.32    58 -1.39 0     -1.39       6     0 -0.163
## 3 -0.511    2.69    74 -1.39 0     -1.39       7    20 -0.163
## 4 -1.20     3.28    58 -1.39 0     -1.39       6     0 -0.163
## 5  0.751    3.43    62 -1.39 0     -1.39       6     0  0.372
```

```
prostate[,1:5] %>% summary()
```

```
##      lcavol           lweight           age             lbph          svi
##  Min.   :-1.3471   Min.   :2.375   Min.   :41.00   Min.   :-1.3863   0:76
##  1st Qu.: 0.5128   1st Qu.:3.376   1st Qu.:60.00   1st Qu.:-1.3863   1:21
##  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
```

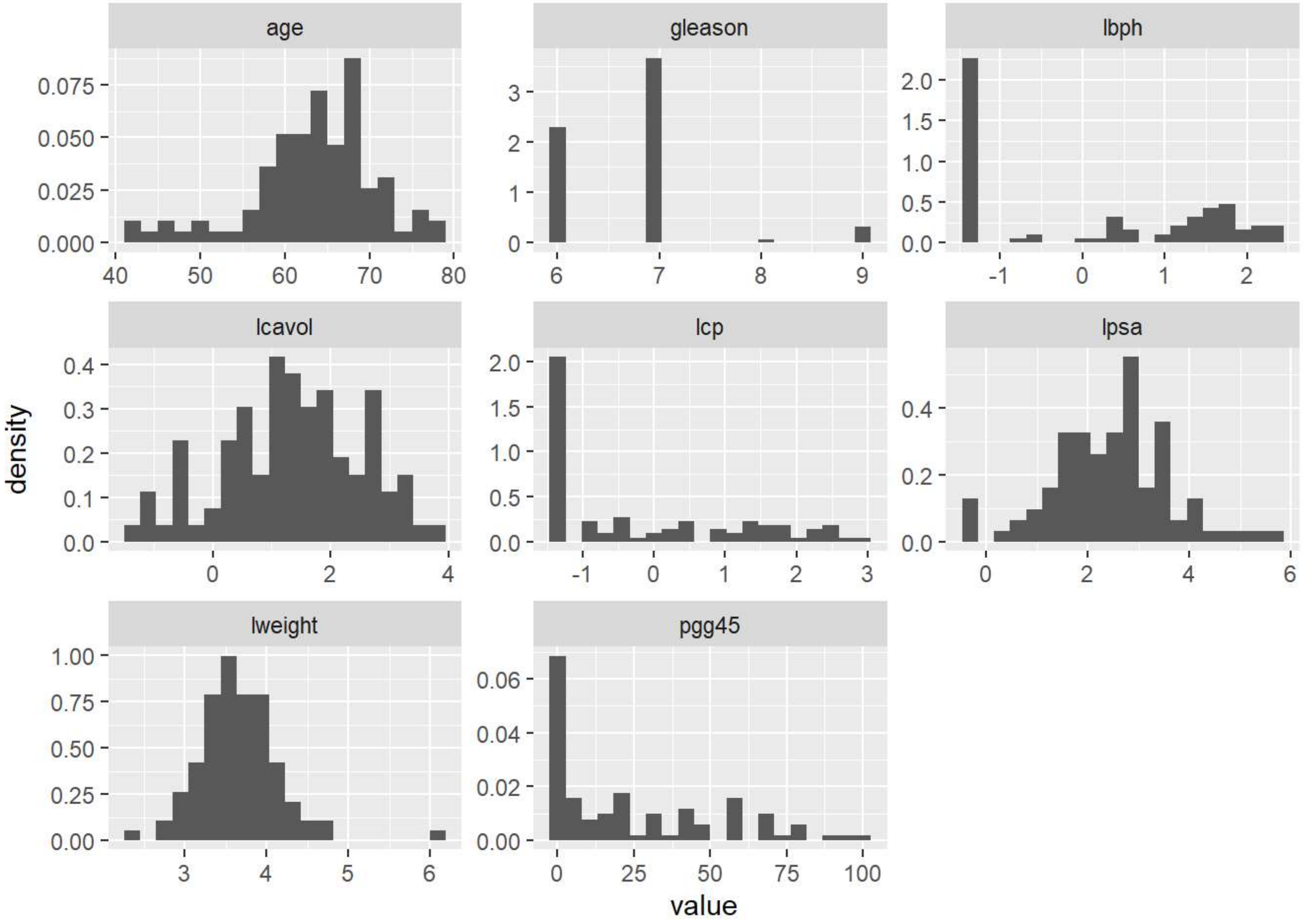
```
cat("\n")
```

```
prostate[,6:9] %>% summary()
```

```
##       lcp              gleason          pgg45             lpsa        
##  Min.   :-1.3863   Min.   :6.000   Min.   :  0.00   Min.   :-0.4308  
##  1st Qu.:-1.3863   1st Qu.:6.000   1st Qu.:  0.00   1st Qu.: 1.7317  
##  Median :-0.7985   Median :7.000   Median : 15.00   Median : 2.5915  
##  Mean   :-0.1794   Mean   :6.753   Mean   : 24.38   Mean   : 2.4784  
##  3rd Qu.: 1.1786   3rd Qu.:7.000   3rd Qu.: 40.00   3rd Qu.: 3.0564  
##  Max.   : 2.9042   Max.   :9.000   Max.   :100.00   Max.   : 5.5829  
```

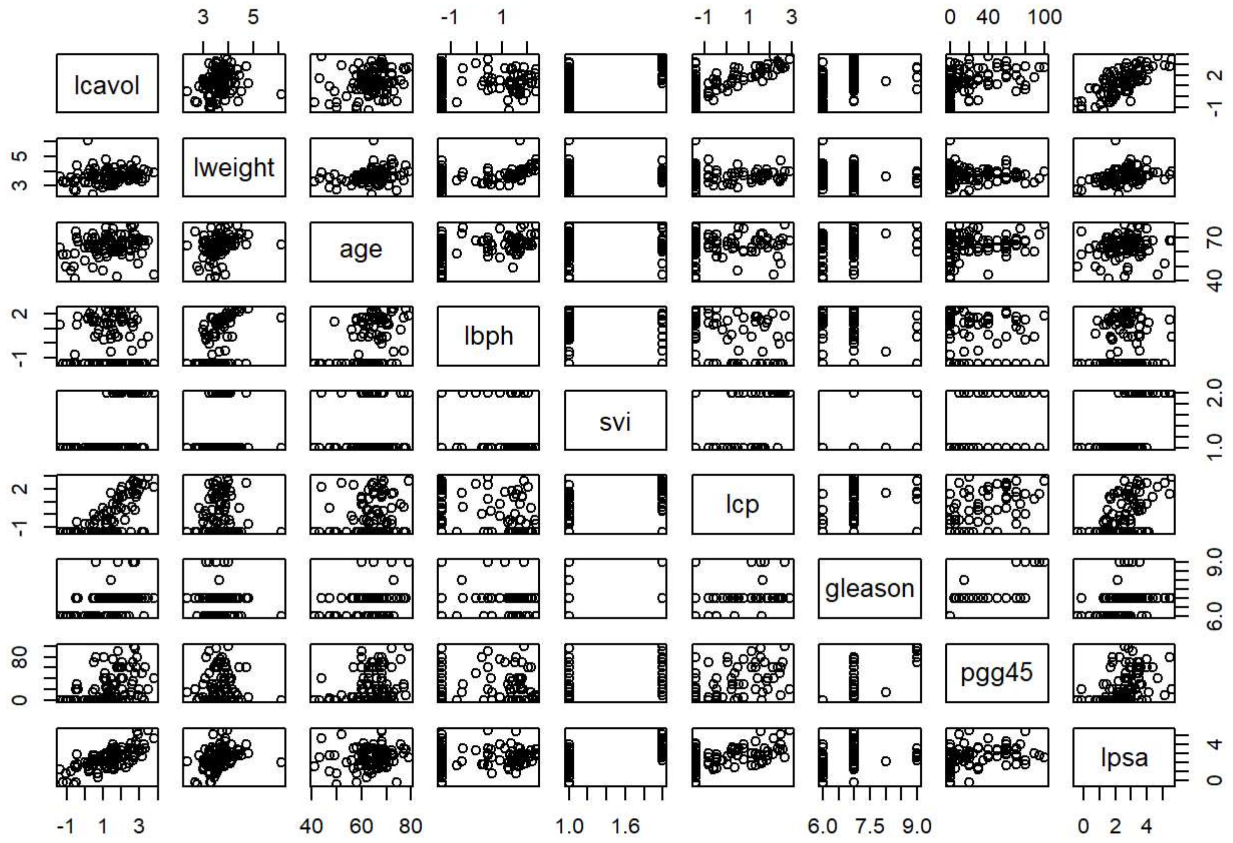
View histograms of numerical variables' distribution

```
prostate_numeric <- prostate %>% select(where(is.numeric))
prostate_numeric_long <- prostate_numeric %>% pivot_longer(colnames(prostate_numeric)) %>% as.data.frame()

ggplot(data=prostate_numeric_long, aes(x=value)) +
  geom_histogram(aes(y=..density..), bins=20) +
  facet_wrap(~ name, scales = "free")
```



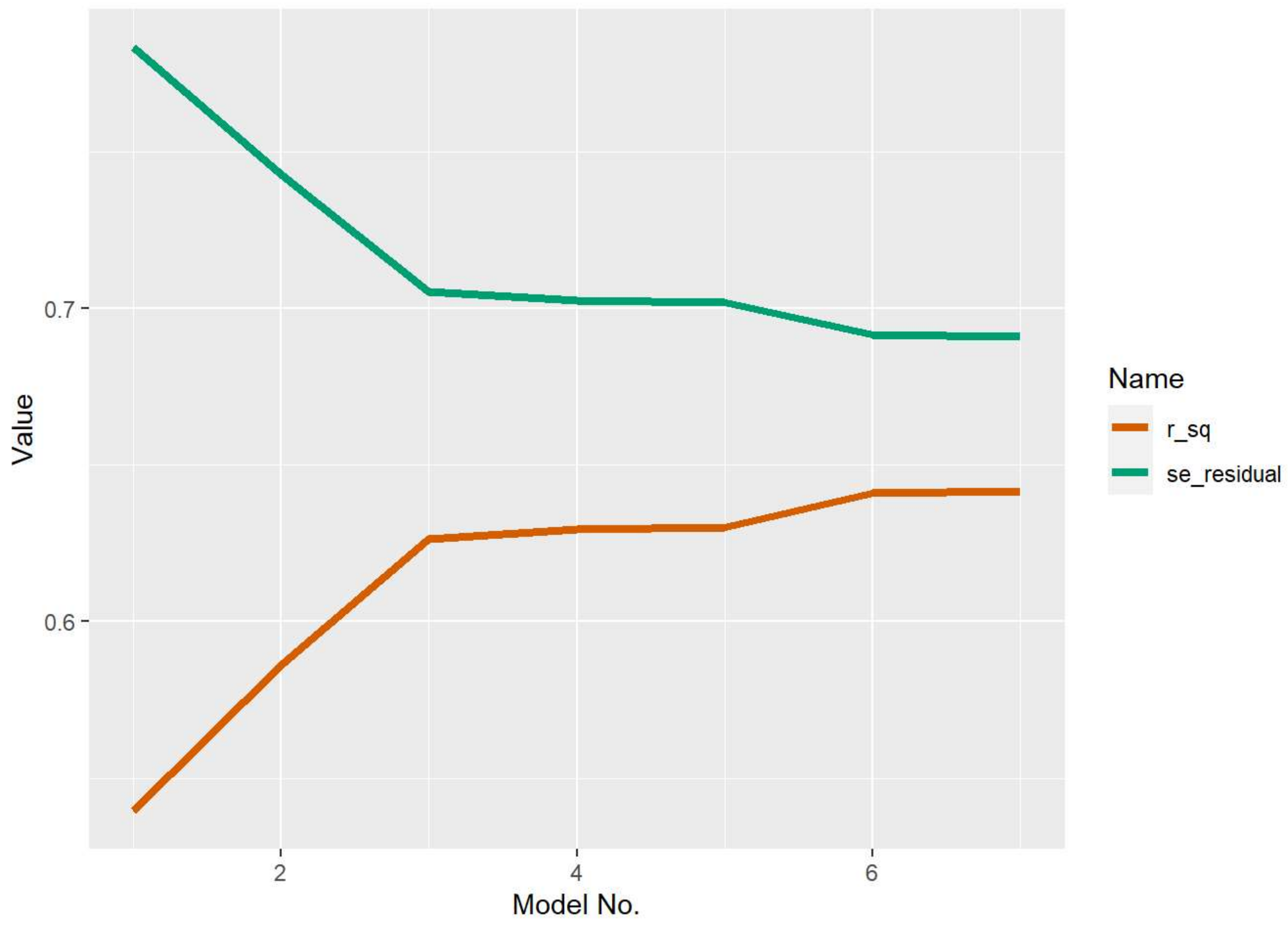
View relationships between each pair of variables in the model

```
pairs(prostate)
```



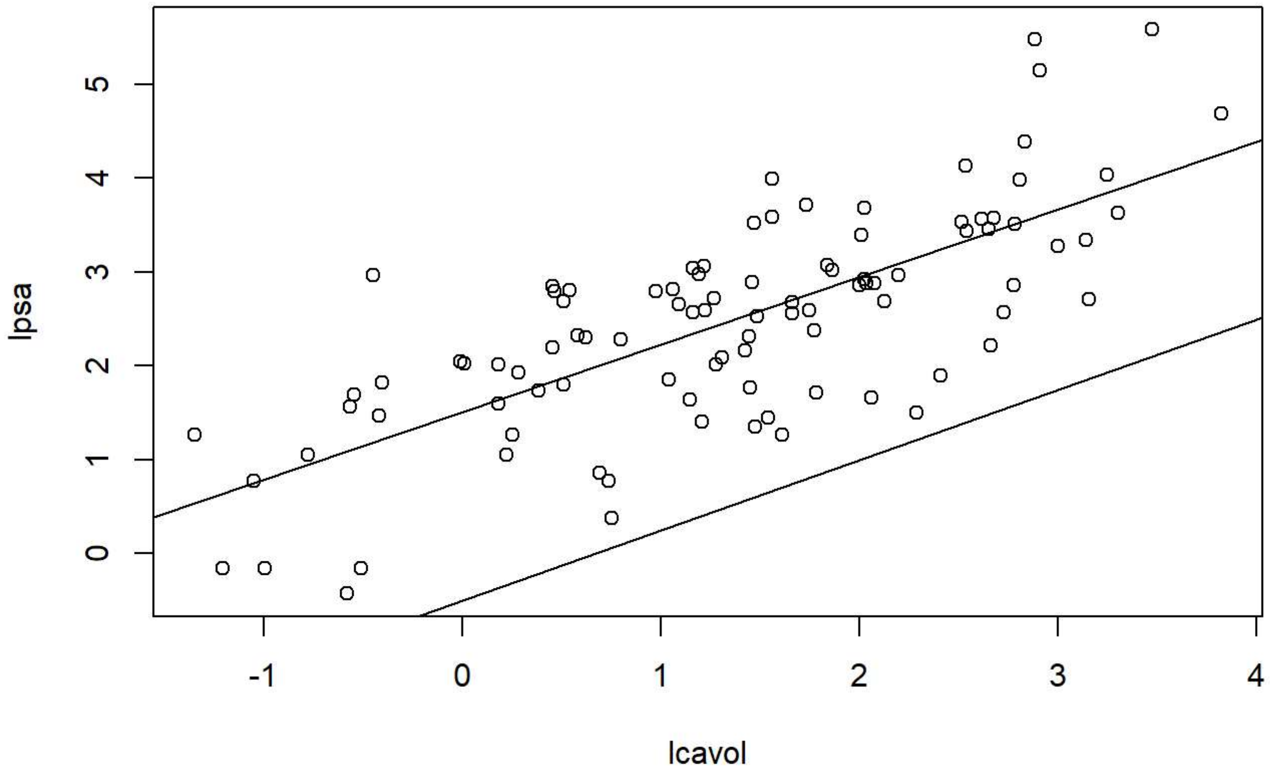
Create a model through forward stepwise selection and graph change in R^2, Residual Standard Error

```
lm1 <- lm(lpsa ~ lcavol, data=prostate)
lm1_sum <- summary(lm1, cor=TRUE)

lm2 <- lm(lpsa ~ lcavol + lweight, data=prostate)
lm2_sum <- summary(lm2, cor=TRUE)

lm3 <- lm(lpsa ~ lcavol + lweight + svi, data=prostate)
lm3_sum <- summary(lm3, cor=TRUE)

lm4 <- lm(lpsa ~ lcavol + lweight + svi + age, data=prostate)
lm4_sum <- summary(lm4, cor=TRUE)

lm5 <- lm(lpsa ~ lcavol + lweight + svi + age + lcp, data=prostate)
lm5_sum <- summary(lm5, cor=TRUE)

lm6 <- lm(lpsa ~ lcavol + lweight + svi + age + lcp + pgg45, data=prostate)
lm6_sum <- summary(lm6, cor=TRUE)

lm7 <- lm(lpsa ~ lcavol + lweight + svi + age + lcp + pgg45 + gleason, data=prostate)
lm7_sum <- summary(lm7, cor=TRUE)

lm_tibble <- tibble(
       model = c(1, 2, 3, 4, 5, 6, 7),
       se_residual = c(sd(lm1$residuals), sd(lm2$residuals), sd(lm3$residuals), sd(lm4$residuals),
                       sd(lm5$residuals), sd(lm6$residuals), sd(lm7$residuals)),
       r_sq = c(lm1_sum$`r.squared`, lm2_sum$`r.squared`, lm3_sum$`r.squared`,
                 lm4_sum$`r.squared`, lm5_sum$`r.squared`, lm6_sum$`r.squared`,
                 lm7_sum$`r.squared`))

lm_tibble_long <- pivot_longer(lm_tibble, cols= c(se_residual, r_sq), names_to= "Name", values_to = "Value")

ggplot(data=lm_tibble_long, aes(x=model, y=Value, group=Name, fill=Name)) + geom_line(aes(col=Name), size=1.5) + scale_color_manual(values=c("#D55E00", "#009E73")) + xlab("Model No.") + ylab("Value")
```



Plot the relationship between `lpsa` and `lcavol`

```
lm1.5 <- lm(lcavol ~ lpsa, data=prostate)

plot(lpsa ~ lcavol, data=prostate)
abline(lm1)
abline(lm1.5)
```


lpsa
0
1
2
3
4
5
-1
0
1
2
3
4
lcavol