R_Project1_S24_

Zheng Yi Lai

2024-01-24

Question 1. The dataset **prostate** in the R package faraway 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 using R, 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. Please use R for this question.

```
#Import the library
library(faraway)
library(GGally)
```

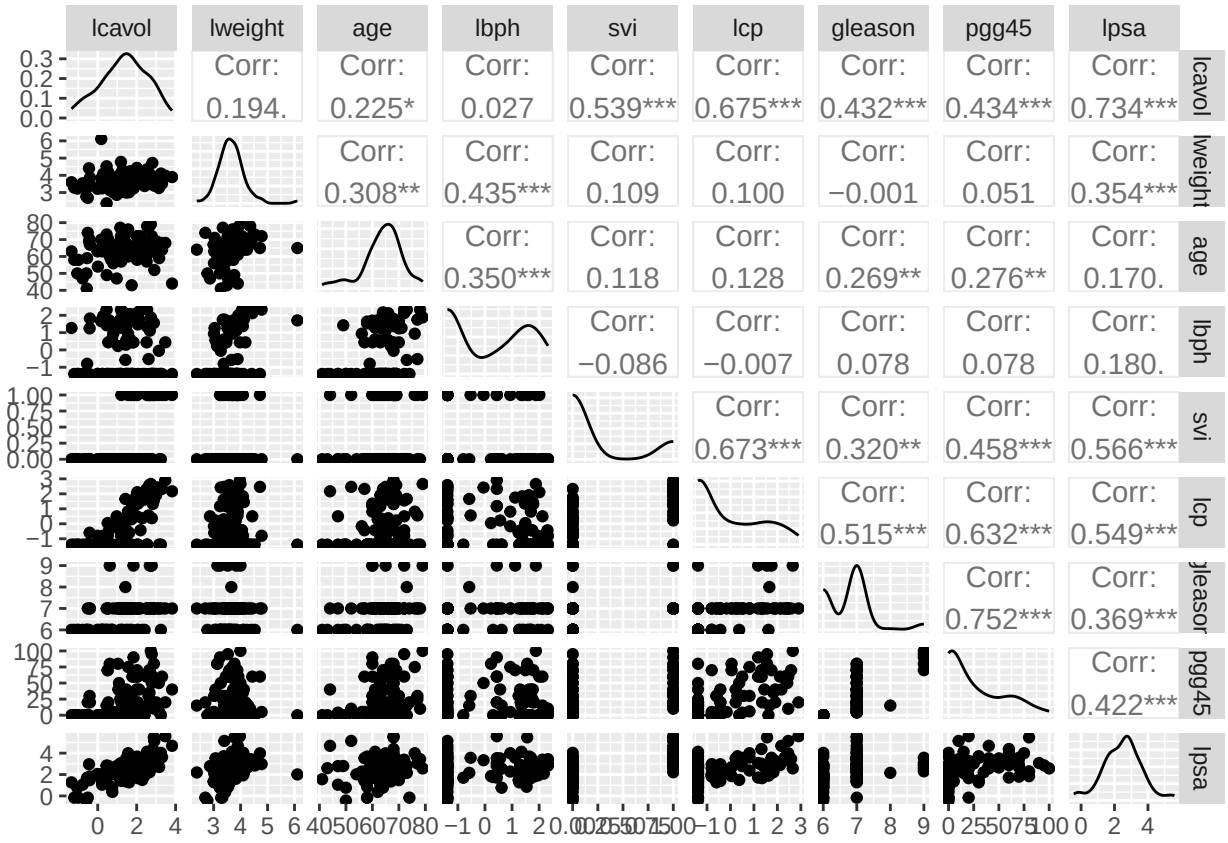
```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
##
## Attaching package: 'GGally'
```

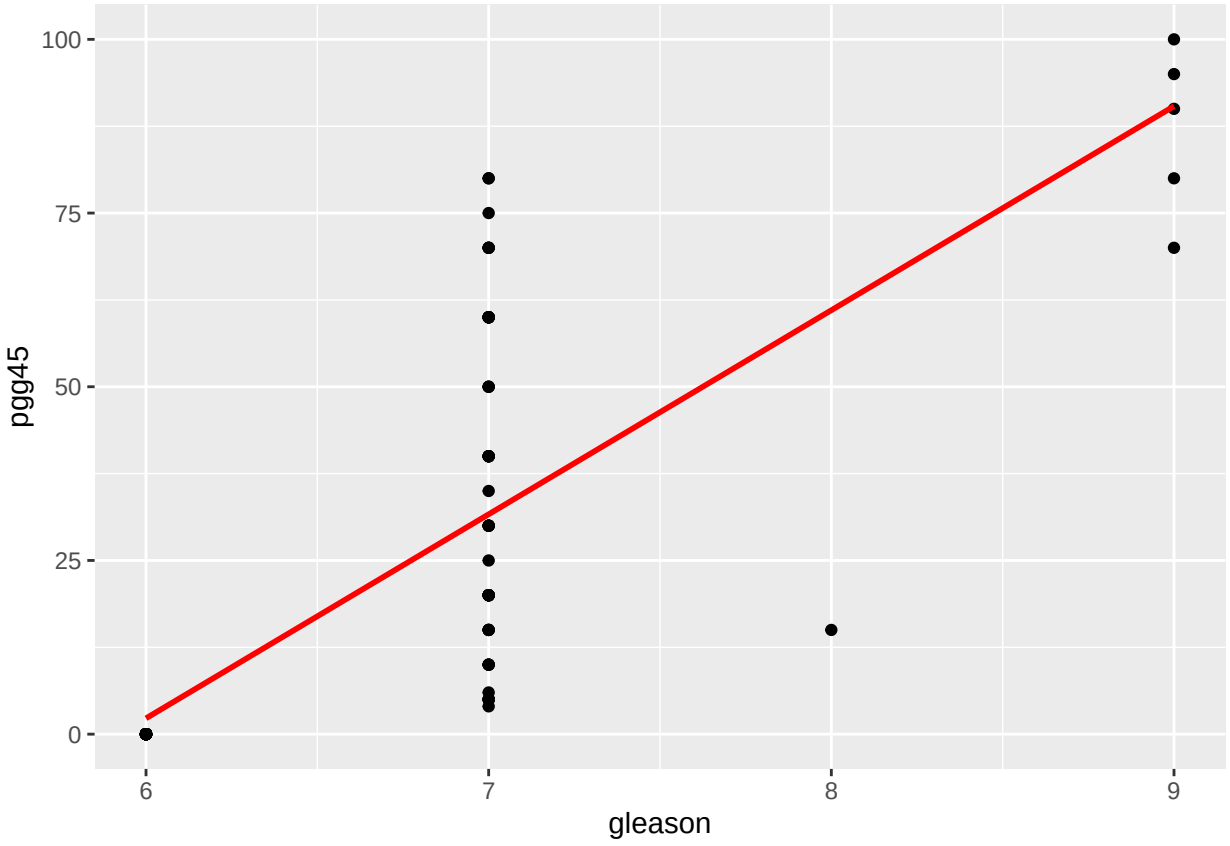
```
## The following object is masked from 'package:faraway':
##
##     happy
```

```
#Import the dataset
data(prostate)

#Plot the matrix to find the relationship between each variables
ggpairs(prostate[,1:9])
```




```
#Plot the two variables
ggplot(prostate, aes(x=gleason, y=pgg45))+
  geom_point(col="black")+
  geom_smooth(method=lm,se=FALSE, color =  "red")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
#Numerical summary of two variables
summary(prostate$pgg45)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00   15.00   24.38   40.00  100.00
```

```
summary(prostate$gleason)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.000   6.000   7.000   6.753   7.000   9.000
```

From the depicted graph, it's evident that we share identical Gleason values at 7 and 9, yet the corresponding y-values vary. My focus is particularly drawn to these two variables due to their displaying the highest correlation in the dataset. Notably, I find it intriguing that having the same x-values but different y-values leads to a notable correlation.

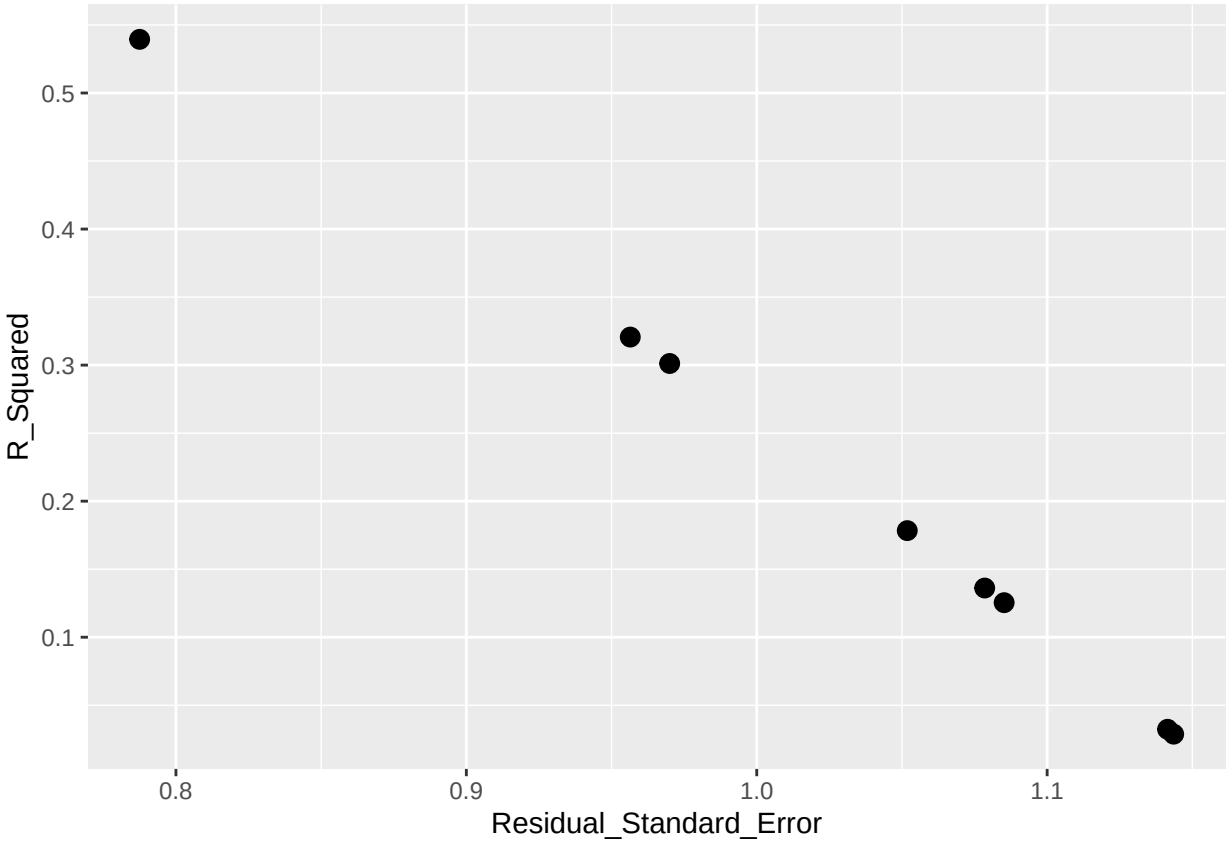
Question 2. Continue using the dataset **prostate**. Fit a model with *lpsa* as the response and *lcavol* as the predictor. Record the residual standard error and the R^2. Now replace *lcavol* with *lweight*, *svi*, *lbph*, *age*, *lcp*, *pgg45* and *gleason* to the model one at a time. For each model record the residual standard error and the R^2. Plot the trends in these two statistics. Please use R for this question

```
#Fit all the requirements into linear regression model
fit <- lm(lpsa ~ lcavol, data=prostate)
fit1 <- lm(lpsa ~ lweight, data=prostate)
fit2 <- lm(lpsa ~ svi, data=prostate)
```

```r
fit3 <- lm(lpsa ~ lbph, data=prostate)
fit4 <- lm(lpsa ~ age, data=prostate)
fit5 <- lm(lpsa ~ lcp, data=prostate)
fit6 <- lm(lpsa ~ pgg45, data=prostate)
fit7<- lm(lpsa ~ gleason, data=prostate)

#Put it into a dataframe
variable <- c("lcavol","lweight", "svi", "lbph", "age", "lcp", "pgg45", "gleason")
Residual_Standard_Error<-c(sigma(fit),sigma(fit1),sigma(fit2),sigma(fit3),sigma(fit4),
                           sigma(fit5),sigma(fit6),sigma(fit7))
R_Squared <-c(summary(fit)$r.squared,summary(fit1)$r.squared,summary(fit2)$r.squared,
              summary(fit3)$r.squared,summary(fit4)$r.squared,summary(fit5)$r.squared,
              summary(fit6)$r.squared,summary(fit7)$r.squared)
df <- data.frame(variable,Residual_Standard_Error,R_Squared)
df
```

```
##   variable Residual_Standard_Error  R_Squared
## 1   lcavol               0.7874994 0.53943191
## 2  lweight               1.0851943 0.12540226
## 3      svi               0.9564563 0.32060303
## 4     lbph               1.1414756 0.03233146
## 5      age               1.1435791 0.02876175
## 6      lcp               0.9700208 0.30119588
## 7    pgg45               1.0518325 0.17835059
## 8  gleason               1.0785051 0.13615118
```

```r
#Plot the points using ggplot2
ggplot(df, aes(Residual_Standard_Error, R_Squared)) +
  geom_point(color = "black", size = 3)
```



From the graphical representation, it's evident that there exists a negative relationship between the residual standard error (RSE) and R-squared (R^2). This implies that as the R^2 value increases, the corresponding residual standard error tends to decrease. In theory, a higher R^2 indicates a better model fit, signifying that the observed data points are closer to the regression line. Consequently, this alignment results in a smaller residual standard error, reinforcing the notion that the model's predictions are more accurate and closely match the actual data points.

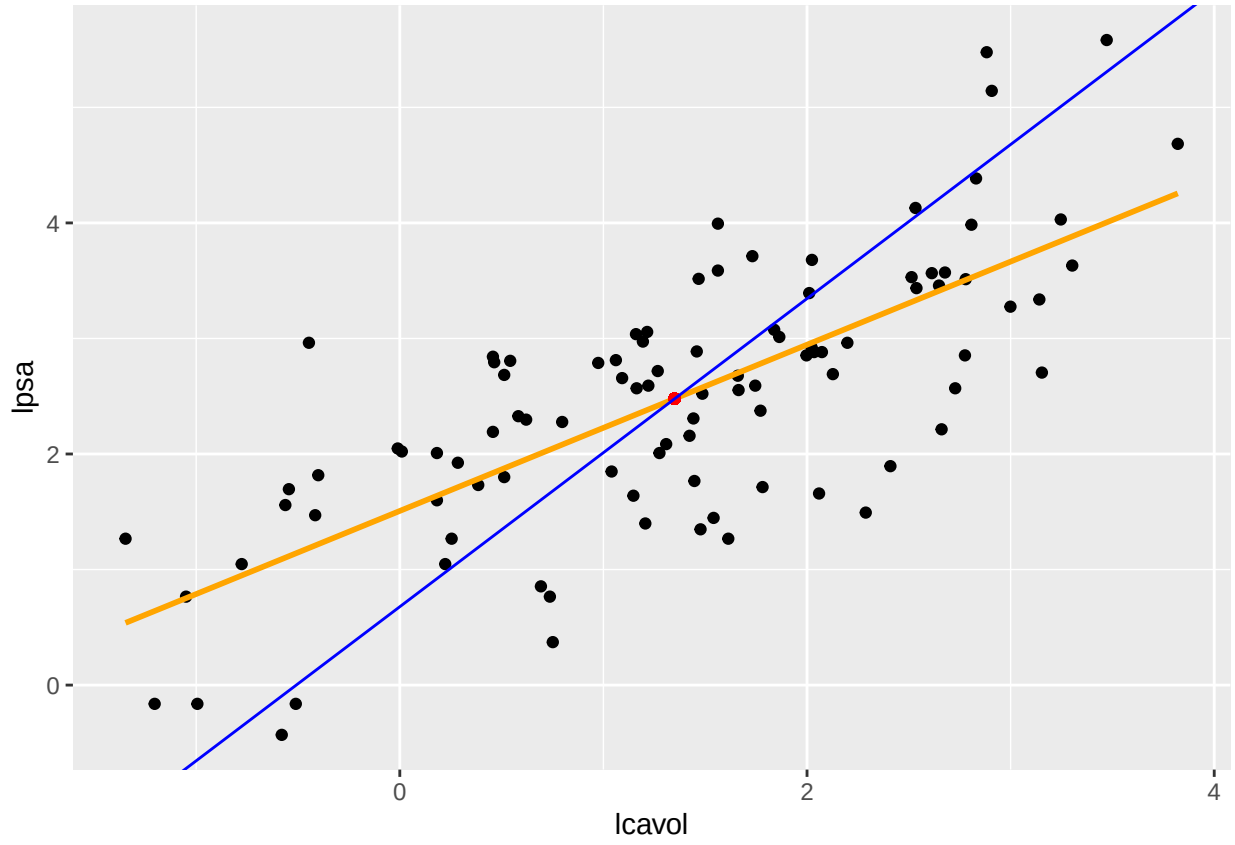
Question 3. Continue using the dataset **prostate**, plot *lpsa* against *lcavol*. Fit the regressions of *lpsa* on *lcavol* and *lcavol* on *lpsa*. Display both regression lines on the plot. At what point do the two lines intersect? Please use R for this question. (Hint: Rewrite the equation as mentioned in the question).

```r
#Fit in a regression model using lcoval on lspa and lspa on lcavol
lcavol_lpsa <-lm(lcavol ~ lpsa, data=prostate)
lpsa_lcavol <-lm(lpsa ~ lcavol, data=prostate)

#Matrix_A constructed by the hint from the question
matrix_A<- c(-coef(lcavol_lpsa)[1]/coef(lcavol_lpsa)[2],slope=1/coef(lcavol_lpsa)[2])
#Matrix_B using the coefficient matrix by the lpsa_lcavol model
matrix_B<-coef(lpsa_lcavol)

#Slope and intercept rearranged by the hint of the question
slope_1 <- 1/coef(lcavol_lpsa)[2]
intercept_1 <- -coef(lcavol_lpsa)[1]/coef(lcavol_lpsa)[2]

#Finding the intercept by combining two matrix
cm <- rbind(matrix_A,matrix_B) # Coefficient matrix
c(-solve(cbind(cm[,2],-1)) %*% cm[,1])
```

```
## [1] 1.350010 2.478387
```

```
#Plot the graph by using ggplot2
ggplot(prostate, aes(x=lcavol, y=lpsa))+
  geom_point(col="black")+
  geom_smooth(method=lm,se=FALSE, color =  "orange")+
  geom_point(aes(x=1.350010,y=2.478387),colour="red")+
  geom_abline(slope=slope_1, intercept=intercept_1,color="blue")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Referring to the graph above, I have highlighted the intersection point in red, and its coordinates are (x = 1.350010, y = 2.478387)."