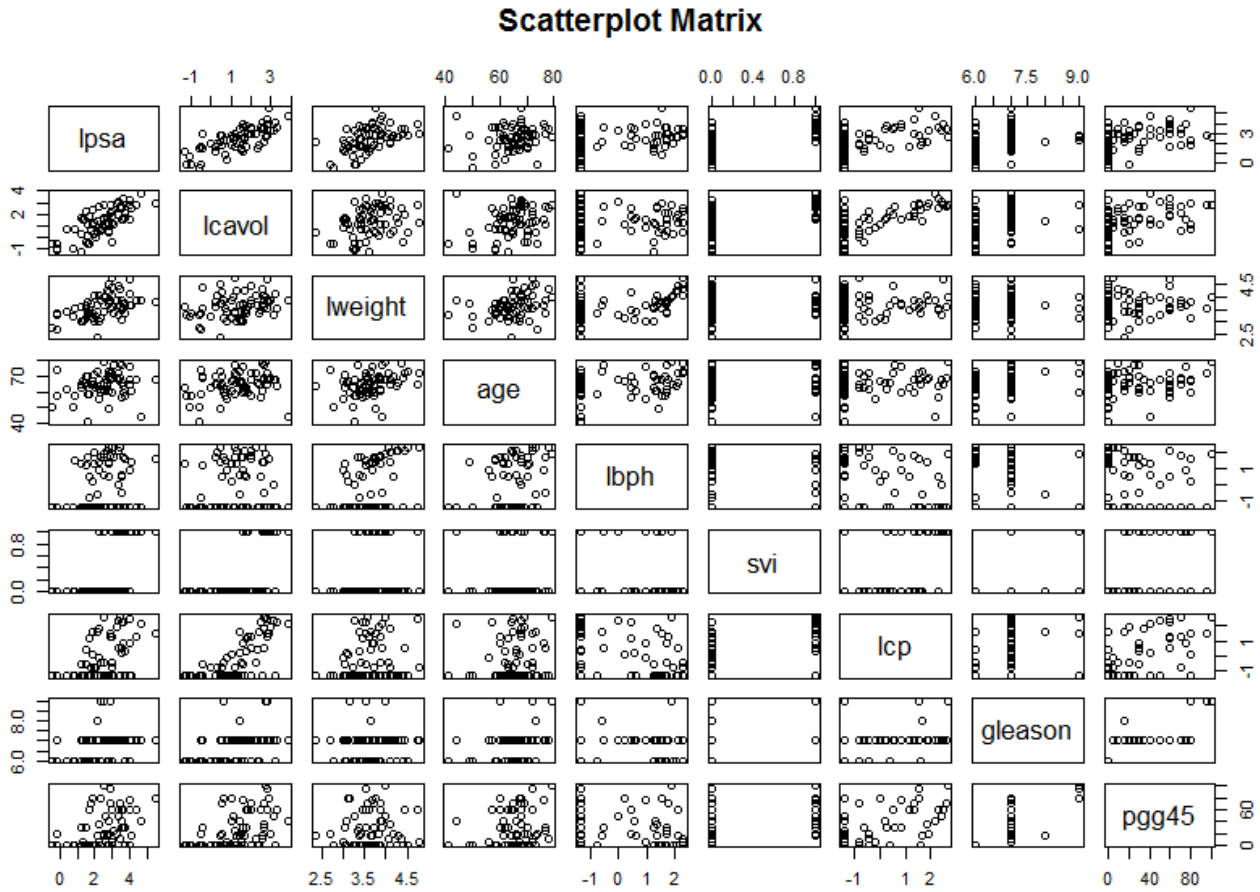
High-dimensional data analysis HW2

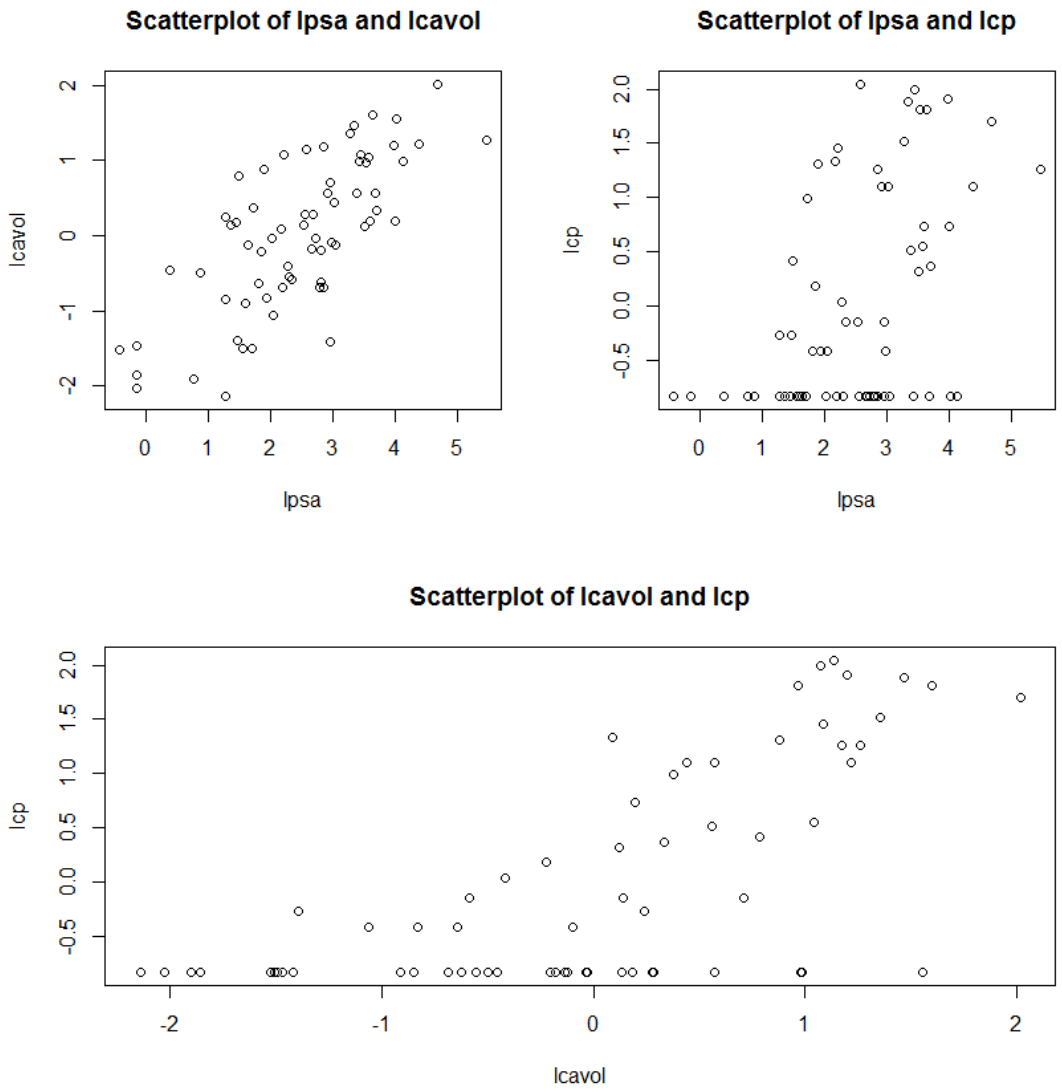
Produce all result of **3.2.1** by your own R code. Put R output & R code in Appendix.

Correlations of predictors in the prostate cancer data

	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45
lcavol	1.000							
lweight	0.300	1.000						
age	0.286	0.317	1.000					
lbph	0.063	0.437	0.287	1.000				
svi	0.593	0.181	0.129	-0.139	1.000			
lcp	**0.692**	0.157	0.173	-0.089	**0.671**	1.000		
gleason	0.426	0.024	0.366	0.033	0.307	0.476	1.000	
pgg45	0.483	0.074	0.276	-0.030	0.481	0.663	0.757	1.000



We see that svi is a binary variable, and gleason is an ordered categorical variable.



We find that both lcavol and lcp show a strong relationship with the response lpsa, and with each other.

Linear model fit to the prostate cancer data

Term	Coefficient	Std.Error	Z Score
Intercept	2.46493	0.08931	27.593
Lcavol	0.67953	0.12663	**5.366**
Lweight	0.26305	0.09563	**2.751**
Age	-0.14146	0.10134	-1.396
Lbph	0.21015	0.10222	2.056
Svi	0.30520	0.12360	**2.469**
Lcp	-0.28849	0.15453	-1.867
Gleason	-0.02131	0.14525	-0.147
Pgg45	0.26696	0.15361	1.738

Since $t_{67-9}(0.025) = 2.001717$, a Z-score greater than 2 in absolute value is approximately significant at the 5% level. The covariates lcavol, lweight and svi shows the strongest effect. But that lcp is not significant.

Linear model fit to the prostate cancer data without lcavol and svi

Term	Coefficient	Std.Error	Z Score
Intercept	2.46205	0.11648	21.137
Lweight	0.41210	0.12124	3.399
Age	-0.07451	0.13143	-0.567
Lbph	0.17346	0.13110	1.323
Lcp	0.34861	0.15655	**2.227**
Gleason	0.07068	0.18649	0.379
Pgg45	0.23994	0.19818	1.211

That lcp becomes strongly significant($2.227 > t_{67-7}(0.025)$), because we dropped out the covariates of lcavol and svi which have high correlationship with lcp.

We consider dropping all the non-significant terms.

Therefore,

$$F = \frac{(32.815 - 29.426)/(9-5)}{29.426/(67-9)} = 1.6698$$

which has a $p-value$ of $P(F_{4,58} > 1.6698) = 0.1693$, is not significant.

Then, base error $= E(Y_0 - \bar{Y})^2 \approx 1.056733$, where Y_0 is testing set.

And, mean prediction error $= E(Y_0 - \hat{Y})^2 \approx 0.521274$.

Hence the linear model reduces the base error rate by about 50%.

Output

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.46205    0.11648  21.137  < 2e-16 ***
X2lweight    0.41210    0.12124   3.399  0.00121 **
X2age       -0.07451    0.13143  -0.567  0.57289
X2lbph       0.17346    0.13110   1.323  0.19080
X2lcp        0.34861    0.15655   2.227  0.02972 *
X2gleason    0.07068    0.18649   0.379  0.70602
X2pgg45      0.23994    0.19818   1.211  0.23075

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.46493    0.08931  27.598  < 2e-16 ***
Xlcavol      0.67953    0.12663   5.366 1.47e-06 ***
Xlweight     0.26305    0.09563   2.751  0.00792 **
Xage        -0.14146    0.10134  -1.396  0.16806
Xlbph        0.21015    0.10222   2.056  0.04431 *
Xsvi         0.30520    0.12360   2.469  0.01651 *
Xlcp        -0.28849    0.15453  -1.867  0.06697 .
Xgleason    -0.02131    0.14525  -0.147  0.88389
Xpgg45       0.26696    0.15361   1.738  0.08755 .

          lcavol lweight   age    lbph     svi     lcp gleason  pgg45
lcavol     1.000   0.300 0.286  0.063   0.593   0.692   0.426  0.483
lweight    0.300   1.000 0.317  0.437   0.181   0.157   0.024  0.074
age        0.286   0.317 1.000  0.287   0.129   0.173   0.366  0.276
lbph       0.063   0.437 0.287  1.000  -0.139  -0.089   0.033 -0.030
svi        0.593   0.181 0.129 -0.139   1.000   0.671   0.307  0.481
lcp        0.692   0.157 0.173 -0.089   0.671   1.000   0.476  0.663
gleason    0.426   0.024 0.366  0.033   0.307   0.476   1.000  0.757
pgg45      0.483   0.074 0.276 -0.030   0.481   0.663   0.757  1.000

Analysis of Variance Table

Model 1: Y ~ lcavol + lweight + lbph + svi
Model 2: Y ~ X
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     62 32.815
2     58 29.426  4    3.3886 1.6698 0.1693

> ###critical value
> qt(0.975,67-9)
[1] 2.001717
> qt(0.975,67-7)
[1] 2.000298
```

R-Code

```
library(ElemStatLearn)
data(prostate)

attach(prostate)
lcavol=(lcavol-mean(lcavol))/sd(lcavol)
lweight=(lweight-mean(lweight))/sd(lweight)
age=(age-mean(age))/sd(age)
lbph=(lbph-mean(lbph))/sd(lbph)
svi=(svi-mean(svi))/sd(svi)
lcp=(lcp-mean(lcp))/sd(lcp)
gleason=(gleason-mean(gleason))/sd(gleason)
pgg45=(pgg45-mean(pgg45))/sd(pgg45)
data=cbind(lpsa,lcavol,lweight,age,lbph,svi,lcp,gleason,pgg45,train)

###Correlations of predictors in the prostate cancer data
predictor_data=subset(data,select=c(lcavol:pgg45),train==TRUE)
B=round(cor(predictor_data),digits=3)

###scatterplot
pairs(~lpsa+lcavol+lweight+age+lbph+svi+lcp+gleason+pgg45,data=data,
      main="Scatterplot Matrix")
par(mfrow=c(1,2))
plot(lpsa,lcavol,main="Scatterplot of lpsa and lcavol")
plot(lpsa,lcp,main="Scatterplot of lpsa and lcp")
plot(lcavol,lcp,main="Scatterplot of lcavol and lcp")
```

```
###linear model
X=predictor_data
Y=subset(data,select=c(lpsa),train==TRUE)
model=lm(Y~X)
summary(model)

###linear model without lcavol and svi
X2=subset(predictor_data,select=c(lweight,age,lbph,lcp,gleason,pgg45))
model2=lm(Y~X2)
summary(model2)

###critical value
qt(0.975,67-9)
qt(0.975,67-7)

###dropping all the non-significant terms
dataX=as.data.frame(X)
model3=lm(Y~lcavol+lweight+lbph+svi,data=dataX)
anova(model3,model)

###base error rate
y0=subset(data,train==FALSE,select=c(lpsa))
c=as.matrix(y0-mean(Y))
base_error=mean(c^2)
```

```
###mean prediction error

x0=subset(data,train==FALSE,select=c(lcavol:pgg45))

X=cbind(1,X)

x0=cbind(1,x0)

beta=solve(t(X)%*%X)%*%t(X)%*%Y

beta=as.vector(beta)

Y_hat=x0%*%beta

mean((y0-Y_hat)^2)
```