Assignment 3, Statistical Modelling III

Andrew Martin

May 4, 2018

1. **Solution**

$$\begin{aligned}
&\lim_{\lambda\to 0}\frac{y^\lambda-1}{\lambda}\to\frac{1-1}{0}=\frac{0}{0}\\
&\text{using l'hôpital's rule:}\\
&=\lim_{\lambda\to 0}\frac{\frac{d}{d\lambda}\left(y^\lambda-1\right)}{\frac{d\lambda}{d\lambda}}\\
&=\lim_{\lambda\to 0}\frac{d}{d\lambda}\left(y^\lambda\right)/1\\
&=\lim_{\lambda\to 0}\frac{d}{d\lambda}\left(e^{\log y^\lambda}\right)\\
&=\lim_{\lambda\to 0}\frac{d}{d\lambda}\left(e^{\lambda\log y}\right)\\
&=\lim_{\lambda\to 0}\log(y)e^{\lambda\log y}\\
&=\log(y)e^0\\
&=\log(y)
\end{aligned}$$

As Required

2. Suppose

$$\mathbf{Y}=\eta+\mathcal{E}$$

Where $\mathcal{E}\sim N(0,\sigma^2)$ $\forall i$, independently. And let X be an $n\times p$ matrix with linearly independent columns. Show that:

$$E\left(||\mathbf{Y}-X\hat{\beta}||^2\right)=(n-p)\sigma^2+||(I-P)\eta||^2$$

where

$$P=X(X^TX)^{-1}X^T.$$

Under what condition is

$$E\left(||\mathbf{Y}-X\hat{\beta}||^2\right)=(n-p)\sigma^2$$

Solution This is a two part question. First:

$$\begin{aligned}
E\left(||\mathbf{Y}-X\hat{\beta}||^2\right)&=E\left((\mathbf{Y}-X\hat{\beta})^T(\mathbf{Y}-X\hat{\beta})\right)\\
&=E\left((\mathbf{Y}-P\mathbf{Y})^T(\mathbf{Y}-P\mathbf{Y})\right)\\
&=E\left((\mathbf{Y}^T-\mathbf{Y}^TP^T)(\mathbf{Y}-P\mathbf{Y})\right)\\
&=E\left(\mathbf{Y}^T\mathbf{Y}-\mathbf{Y}^TP^T\mathbf{Y}-\mathbf{Y}^TP\mathbf{Y}+\mathbf{Y}^TP^TP\mathbf{Y}\right)\\
&=E\left(\mathbf{Y}^T\mathbf{Y}-\mathbf{Y}^TP\mathbf{Y}-\mathbf{Y}^TP\mathbf{Y}+\mathbf{Y}^TP\mathbf{Y}\right)\\
&=E\left(\mathbf{Y}^T\mathbf{Y}-\mathbf{Y}^TP\mathbf{Y}\right)\\
&=E(\mathbf{Y}^T(I-P)\mathbf{Y})\\
&=tr((I-P)\sigma^2)+\mu^T(I-P)\mu\text{ using an identity obtained in the previous assignment}\\
&=\sigma^2(tr(I)-tr(P))+||(I-P)\eta||^2
\end{aligned}$$

Note that in his case the I is $n\times n$, i.e. $tr(I)=\sum_{i=1}^n 1=n$ and $tr(P)=tr(X(X^TX)^{-1}X^T)=tr((X^TX)^{-1}X^TX)=tr(I_{p\times p})=p$ Which gives:

$$E\left(||\mathbf{Y}-X\hat{\beta}||^2\right)=\sigma^2(n-p)+||(I-P\eta)||^2$$

The equality is true if $||(I-P)\eta||^2=0$
This condition will hold true for a 'correct' model. **As Required**

3. Use the result of 2 to justify that Mallow's C_p will satisfy

$$C_p \approx p$$

for a correct model

Solution Recall $E(RSS_p) \geq (n-p)\sigma^2$, with equality if the model is the 'correct' model.
So for a correct model:

$$\begin{aligned}
C_p &= \frac{RSS_p}{\hat{\sigma}^2} - (n-2p) \\
E(C_p) &= E(\frac{RSS_p}{\hat{\sigma}^2} - (n-2p)) \\
&= \frac{E(RSS_p)}{\hat{\sigma}^2} - (n-2p) \\
&\approx \frac{(n-p)\sigma^2}{\hat{\sigma}^2} - (n-2p) \\
&= n-p-(n-2p) = p
\end{aligned}$$

As Required

4. (a) Read the data

i. Plot pairwise scatterplot matrix
Solution This is the figure shown in the appendices.

As Required

ii. Create a correlation matrix of all the predictor variables
Solution The response given in R was:

```
           lcavol   lweight       age         lbph          lcp      pgg45       psa
lcavol  1.0000000 0.2805214 0.2249999  0.027349703  0.675310484 0.43365225 0.4965149
lweight 0.2805214 1.0000000 0.3479691  0.442264399  0.164537142 0.10735379 0.1899156
age     0.2249999 0.3479691 1.0000000  0.350185896  0.127667752 0.27611245 0.0165038
lbph    0.0273497 0.4422644 0.3501859  1.000000000 -0.006999431 0.07846002 0.0248382
lcp     0.6753105 0.1645371 0.1276678 -0.006999431  1.000000000 0.63152825 0.4596901
pgg45   0.4336522 0.1073538 0.2761124  0.078460018  0.631528246 1.00000000 0.2325796
psa     0.4965149 0.1899156 0.0165038  0.024838204  0.459690118 0.23257964 1.0000000
```

As Required

iii. Comment on any observed relationships between `psa` and the predictor variables
Solution `psa` seems to have relationships with lcavol, a slight relationship with svi, and a relationship with lcp
psa seems to increase when lcavol increases
psa increases when lweight increases
psa increases when svi increases
When lcp increases it appears to increase. **As Required**

iv. Comment on any observed relationships amongst the predictor variables
Solution
Using the pairwise plot matrix (in appendices) and the correlation matrix: we can see that:
The variables with any notable relationship are:
lcavol with lcp - when lcavol increases, lcp seems tot increase.
lweight with age - when lweight increases, age appears to increase a little bit
lweight with lbph - when lweight increases, lbph appears to increase
lweight with lcp - when lweight increases there is a very slight increase in lcp
age with lbph - when age increases, lbph increases
pgg45 with lcavol - The data seems to increase and variance decreases for larger lcavols
pgg45 with lweight - the variance pgg45 decreases as lweight increases
As Required

(b) Use the box-cox method the find a suitable transformation of the response variable `psa` in the context of the given model. State (with justification), the chosen λ.

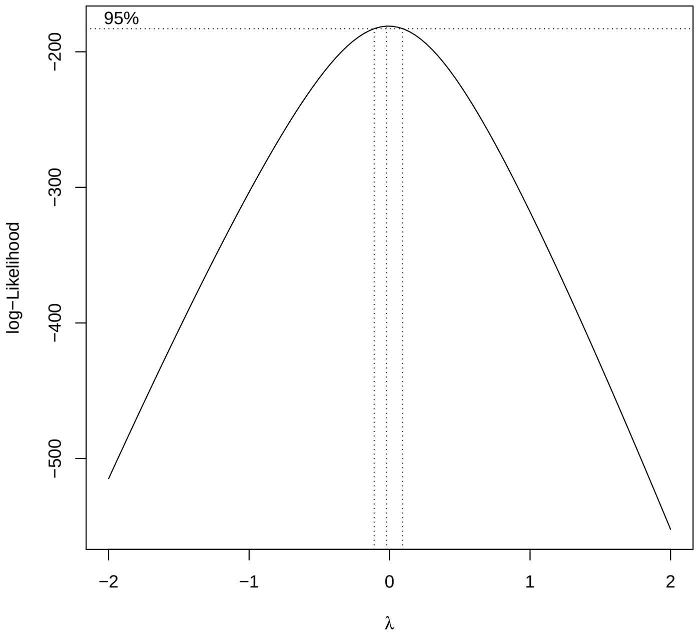
Solution The λ value is approximately 0 - as shown in the figure. Since it is so close to 0, we can simply round it to zero, as it will still approximate the transformation very closely. **As Required**

(c) Re-fit (b) using the transformed response

Solution This is shown in the code. The summary of the lm is given:

```
Call:
lm(formula = log(psa) ˜ lcavol + lweight + age + lbph + svi +
    lcp + gleason + pgg45, data = prostate)
```

Figure 1: Plot of λ against log likelihood



```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.77956 -0.31691 -0.04774  0.44278  1.52832 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.486274   0.885929   0.549  0.58451    
lcavol       0.549532   0.090087   6.100 2.94e-08 ***
lweight      0.623816   0.200463   3.112  0.00252 ** 
age         -0.023103   0.011282  -2.048  0.04364 *  
lbph         0.091523   0.058454   1.566  0.12108    
svi1         0.744537   0.244602   3.044  0.00310 ** 
lcp         -0.124612   0.094565  -1.318  0.19109    
gleason7     0.253874   0.216141   1.175  0.24341    
gleason8     0.480520   0.760895   0.632  0.52938    
gleason9    -0.036003   0.494866  -0.073  0.94217    
pgg45        0.004902   0.004622   1.061  0.29184    
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 0.6973 on 86 degrees of freedom
Multiple R-squared:  0.6731,Adjusted R-squared:  0.6351 
F-statistic: 17.71 on 10 and 86 DF,  p-value: < 2.2e-16
```

As Required

(d) Use Mallows' C_p and stepwise model selection to obtain the most appropriate reduced model for the transformed data

Solution Once again this is shown in the appendix. The reduced model summary is:

```
Call:
lm(formula = log(psa) ~ lcavol + lweight + age + lbph + svi, 
    data = prostate)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.86717 -0.37725  0.01257  0.40252  1.44544 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.49473    0.87652   0.564  0.57385    
lcavol       0.54400    0.07463   7.289 1.11e-10 ***
```

```
lweight      0.58821    0.19790   2.972  0.00378 **
age         -0.01644    0.01068  -1.540  0.12692
lbph         0.10122    0.05759   1.758  0.08215 .
svi1         0.71490    0.20653   3.461  0.00082 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.6988 on 91 degrees of freedom
Multiple R-squared:  0.6526,Adjusted R-squared:  0.6335
F-statistic: 34.19 on 5 and 91 DF,  p-value: < 2.2e-16
```

And note $C_p = 6.4 \approx 6$ which is what we expected for a correct model **As Required**

(e) Obtain appropriate diagnostic plots for the selected model, and present them neatly. Comment on whether the model is appropriate for the transformed data

Solution

Figure 2: Diagnostic plots: Studentised residuals

(a) (b)

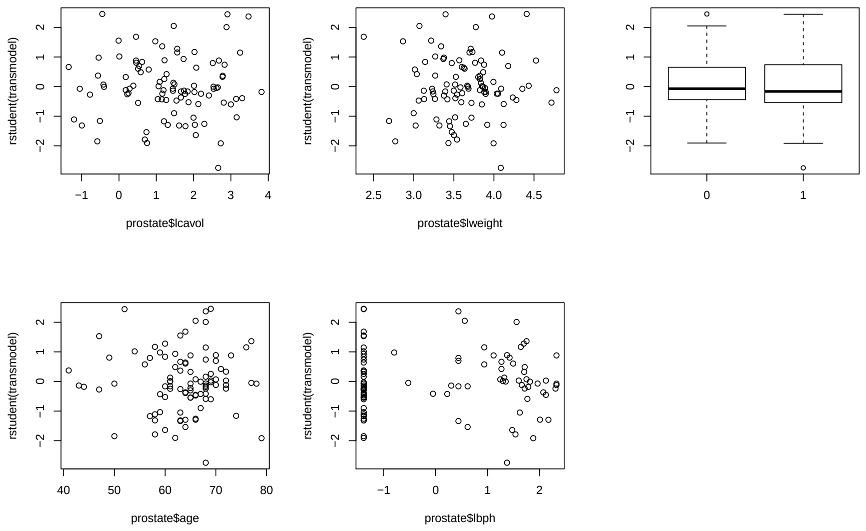


Figure 3: Diagnostic plots

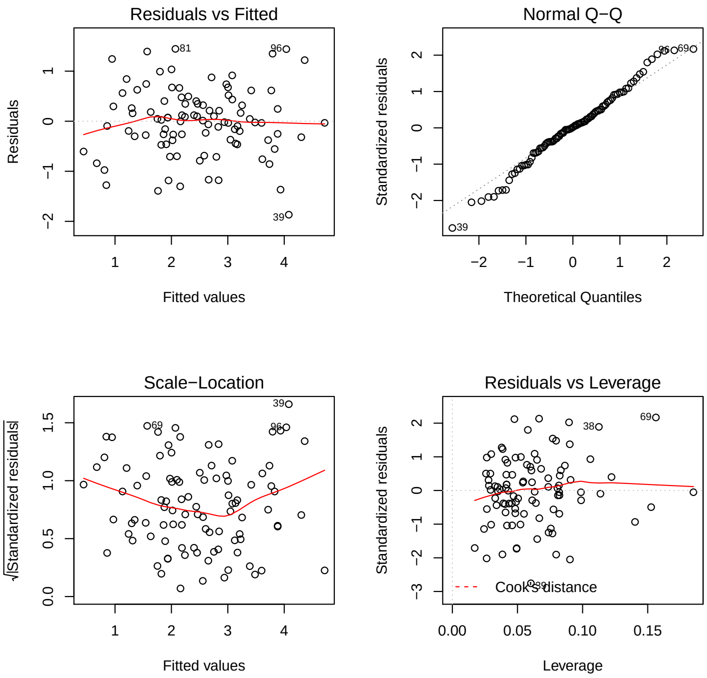


Figure 2 shows the studentised residuals, and Figure 3 shows the regular diagnostic plots for the model.

Assumptions:

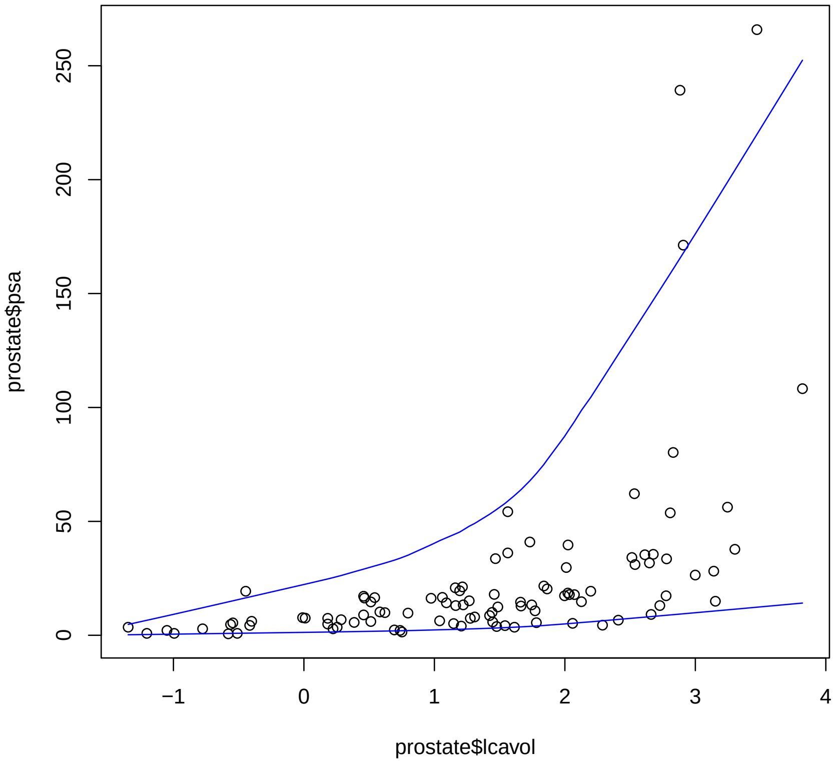
- e_i have zero mean: Observing the studentised residual plots - the points appear to average to a mean of 0, however the lbph residual plot indicates a decreasing mean towards the right of the plot. This is a fairly negligible effect however. (so this holds)
- e_i have equal variance: The spread of the e_i in the studentised residual plots appears equal throughout . The Scale location indicates a dip in the centre indicating a decrease in variance nearing the centre.
- e_i are normally distributed: observing the normal q-q plot shows this, however for ends of the plot, there seems to be some departure from normality.
- Observing the residuals vs leverage plot indicates that there are no points of high leverage or influence (no points outside the Cook's distance contours).
- e_i are independent: No direct information is given on this, so assume independence holds as it is a design assumption, and there are no clear indicators in the construction of the problem. i.e. we assume the e_i are independent.

Conclusion: Apart from a few small deviations from the assumptions, the regression model appears reasonable. **As Required**

(f) Obtain a scatter plot of `psa` verses `lcavol` showing the back-transformed 95% prediction bands. Comment on whether or not they appear appropriate.

Solution **As Required**

Figure 4: Scatter plot with prediction bands



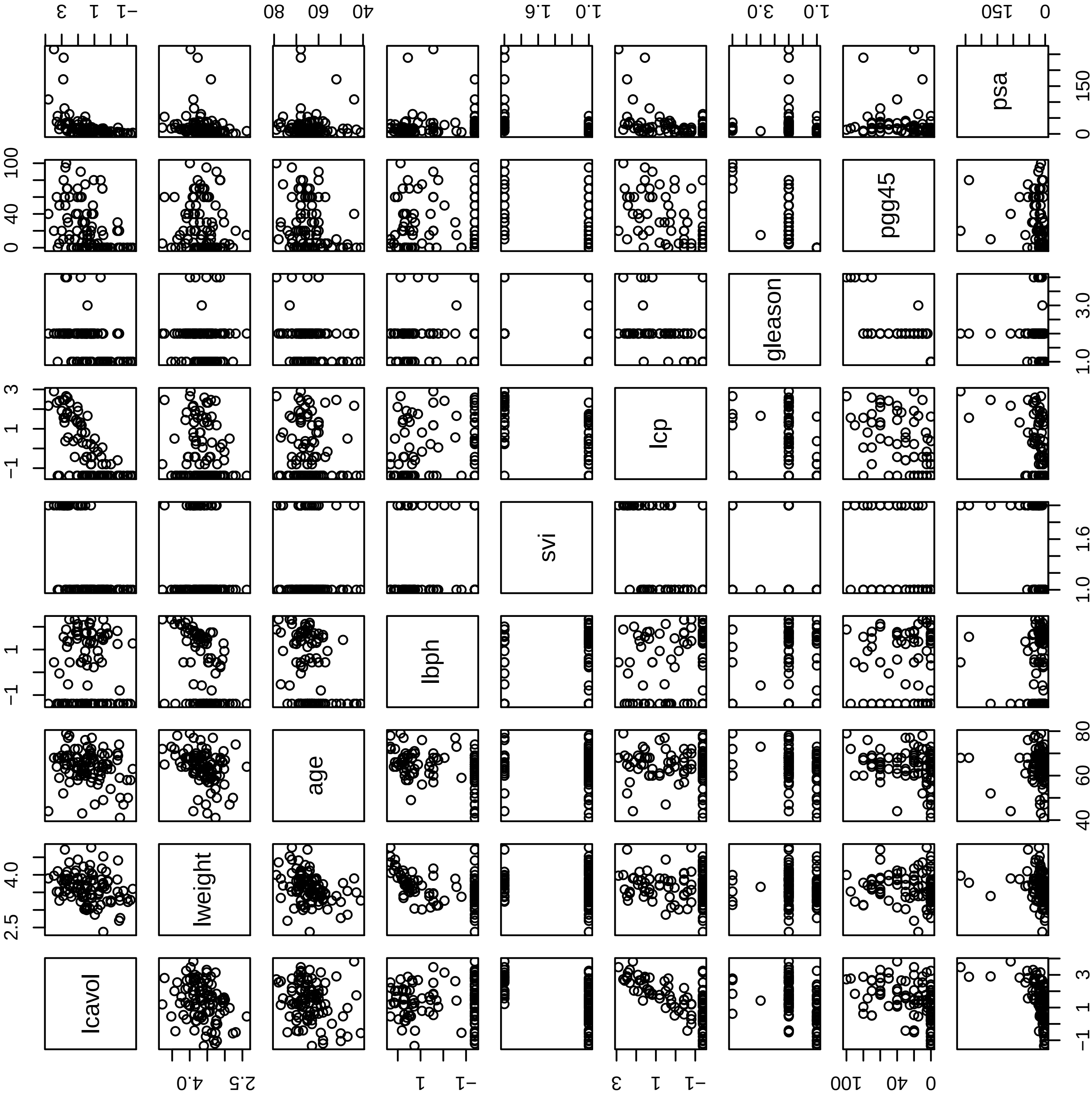
Since most of the data (excluding a few outliers) fits nicely within the prediction bands, it is clear that they appear appropriate

Appendix

The code used is below:

```
###Question 4 A3
library(MASS)
setwd("~/Uni/2018/Statistical Modelling")

pdf(file="PairsPlotA3.pdf")
##a --
#read in the data
prostate = read.csv("prostate-a3.csv",header=T)
#convert to factors

prostate$svi = as.factor(prostate$svi)
prostate$gleason=as.factor(prostate$gleason)
#remove lpsa and train as they aren't defined for the given model
prostate$lpsa = NULL
prostate$train = NULL
##i
#plot pairwise scatterplots
pairs(prostate)
##ii
#correlation matrix
cor(prostate[sapply(prostate,is.numeric)])
##iii
#no code
##iv
#no code
##b --
#Use boxcox to find lambda
basemodel = lm(psa~lcavol+lweight+age+lbph+svi+lcp+gleason+pgg45,data=prostate)
boxcox(basemodel)
#Lambda contains 0, and 0 is the easiest - effectively take the log transform
##c --
#Refit the model, using transformed variables
transmodel= lm(log(psa)~lcavol+lweight+age+lbph+svi+lcp+gleason+pgg45,data=prostate)
##d --
#use stepwise model selection
#minimising AIC is the same as minimising the Cp, so just doing step is good
s2=sum((transmodel$residuals)^2)/transmodel$df
newnewmodel= step(object = transmodel,scale= s2)
#note aic approx Cp
##e --
#Obtain diagnostic plots
tmp = par(mfrow=c(2,2))
plot(newnewmodel)
par(tmp)
#residual plots
tmp = par(mfrow=c(2,2))
plot(prostate$lcavol,rstudent(transmodel))
plot(prostate$lweight,rstudent(transmodel))
plot(prostate$age,rstudent(transmodel))
plot(prostate$lbph,rstudent(transmodel))
plot(prostate$svi,rstudent(transmodel))
par(tmp)

##f --
#Obtain a scatter of psa vs lcavol showing 95% prediction bounds

plot(y=prostate$psa,x=prostate$lcavol)
pred= predict(newnewmodel, interval="prediction")
lines(lowess(prostate$lcavol,exp(pred[,3])),col="blue")
lines(lowess(prostate$lcavol,exp(pred[,2])),col="blue")
dev.off()
```

Figure 5: Appendix: Pairs plots


lcavol
lweight
age
lbph
svi
lcp
gleason
pgg45
psa