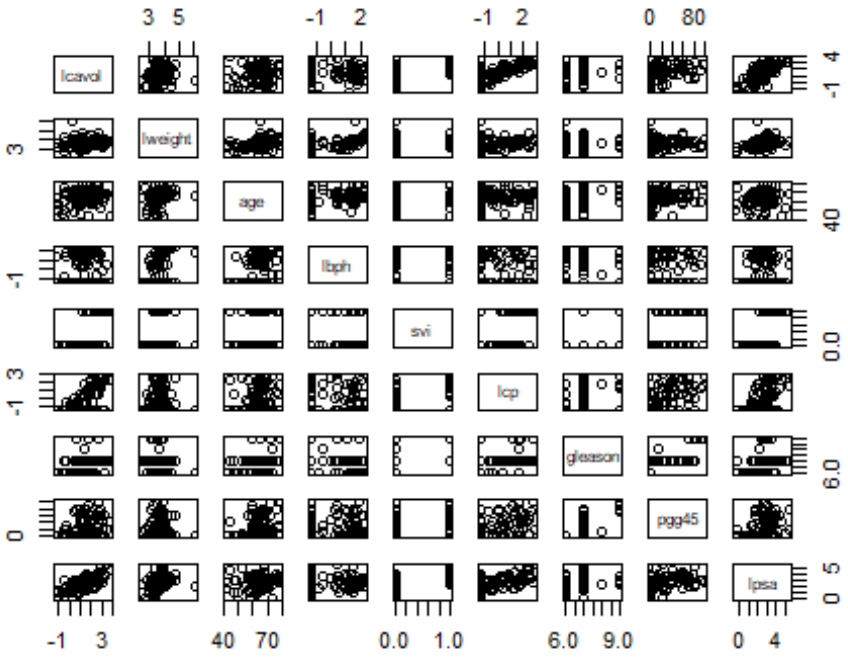
Question 1

```
library(faraway)

## Warning: package 'faraway' was built under R version 3.5.3

data("prostate")
pairs(prostate)
```




```
summary(prostate)

##      lcavol           lweight           age              lbph        
##  Min.   :-1.3471   Min.   :2.375   Min.   :41.00   Min.   :-1.3863  
##  1st Qu.: 0.5128   1st Qu.:3.376   1st Qu.:60.00   1st Qu.:-1.3863  
##  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  
##       svi              lcp              gleason          pgg45       
##  Min.   :0.0000   Min.   :-1.3863   Min.   :6.000   Min.   :  0.00  
##  1st Qu.:0.0000   1st Qu.:-1.3863   1st Qu.:6.000   1st Qu.:  0.00  
##  Median :0.0000   Median :-0.7985   Median :7.000   Median : 15.00  
##  Mean   :0.2165   Mean   :-0.1794   Mean   :6.753   Mean   : 24.38  
##  3rd Qu.:0.0000   3rd Qu.: 1.1786   3rd Qu.:7.000   3rd Qu.: 40.00  
##  Max.   :1.0000   Max.   : 2.9042   Max.   :9.000   Max.   :100.00  
##       lpsa        
```

```
##  Min.   :-0.4308
##  1st Qu.: 1.7317
##  Median : 2.5915
##  Mean   : 2.4784
##  3rd Qu.: 3.0564
##  Max.   : 5.5829
```

Observing the data, we can see that the svi and gleason variables are most likely factor variables. There also seem to be some linear relationships between the variables, such as lpsa and lcavol or lcp and lcavol. The numerical summary tells us that these are mostly older men with a minimum age of 41. The average age is 63.87. Pgg45 also tells us that the mean Gleason scores of 4 or 5 was 24.38%. Some of the other variables are hard to interpret as they are on log scale.

Question 2

```
prostate$svi = as.factor(prostate$svi)
prostate$gleason = as.factor(prostate$gleason)

m1 = lm(lpsa~lcavol, data=prostate)
summary(m1)

##
## Call:
## lm(formula = lpsa ~ lcavol, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.67625 -0.41648  0.09859  0.50709  1.89673
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.50730    0.12194   12.36   <2e-16 ***
## lcavol       0.71932    0.06819   10.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7875 on 95 degrees of freedom
## Multiple R-squared:  0.5394, Adjusted R-squared:  0.5346
## F-statistic: 111.3 on 1 and 95 DF,  p-value: < 2.2e-16

rse = as.vector(0.7875)
r2 = as.vector(0.5346)
m2 = lm(lpsa~lcavol+lweight, data=prostate)
summary(m2)

##
## Call:
## lm(formula = lpsa ~ lcavol + lweight, data = prostate)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.61965 -0.50778 -0.02095  0.52291  1.89885
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.30262    0.56904  -0.532  0.59612
## lcavol       0.67753    0.06626  10.225  < 2e-16 ***
## lweight      0.51095    0.15726   3.249  0.00161 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7506 on 94 degrees of freedom
## Multiple R-squared:  0.5859, Adjusted R-squared:  0.5771
## F-statistic: 66.51 on 2 and 94 DF,  p-value: < 2.2e-16
```

```
rse = append(rse, 0.7506)
r2 = append(r2, 0.5771)
m3 = lm(lpsa~lcavol+lweight+svi, data=prostate)
summary(m3)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.72964 -0.45764  0.02812  0.46403  1.57013
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.26809    0.54350  -0.493  0.62298
## lcavol       0.55164    0.07467   7.388  6.3e-11 ***
## lweight      0.50854    0.15017   3.386  0.00104 **
## svi1         0.66616    0.20978   3.176  0.00203 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7168 on 93 degrees of freedom
## Multiple R-squared:  0.6264, Adjusted R-squared:  0.6144
## F-statistic: 51.99 on 3 and 93 DF,  p-value: < 2.2e-16
```

```
rse = append(rse, 0.7168)
r2 = append(r2, 0.6144)
m4 = lm(lpsa~lcavol+lweight+svi+lbph, data=prostate)
summary(m4)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi + lbph, data = prostate)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.82653 -0.42270  0.04362  0.47041  1.48530
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.14554    0.59747   0.244  0.80809
## lcavol       0.54960    0.07406   7.422 5.64e-11 ***
## lweight      0.39088    0.16600   2.355  0.02067 *
## svi1         0.71174    0.20996   3.390  0.00103 **
## lbph         0.09009    0.05617   1.604  0.11213
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7108 on 92 degrees of freedom
## Multiple R-squared:  0.6366, Adjusted R-squared:  0.6208
## F-statistic: 40.29 on 4 and 92 DF,  p-value: < 2.2e-16
```

```
rse = append(rse, 0.7108)
r2 = append(r2, 0.6208)
m5 = lm(lpsa~lcavol+lweight+svi+lbph+age, data=prostate)
summary(m5)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi + lbph + age, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83505 -0.39396  0.00414  0.46336  1.57888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.95100    0.83175   1.143 0.255882
## lcavol       0.56561    0.07459   7.583 2.77e-11 ***
## lweight      0.42369    0.16687   2.539 0.012814 *
## svi1         0.72095    0.20902   3.449 0.000854 ***
## lbph         0.11184    0.05805   1.927 0.057160 .
## age         -0.01489    0.01075  -1.385 0.169528
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7073 on 91 degrees of freedom
## Multiple R-squared:  0.6441, Adjusted R-squared:  0.6245
## F-statistic: 32.94 on 5 and 91 DF,  p-value: < 2.2e-16
```

```
rse = append(rse, 0.7073)
r2 = append(r2, 0.6245)
m6 = lm(lpsa~lcavol+lweight+svi+lbph+age+lcp, data=prostate)
summary(m6)
```

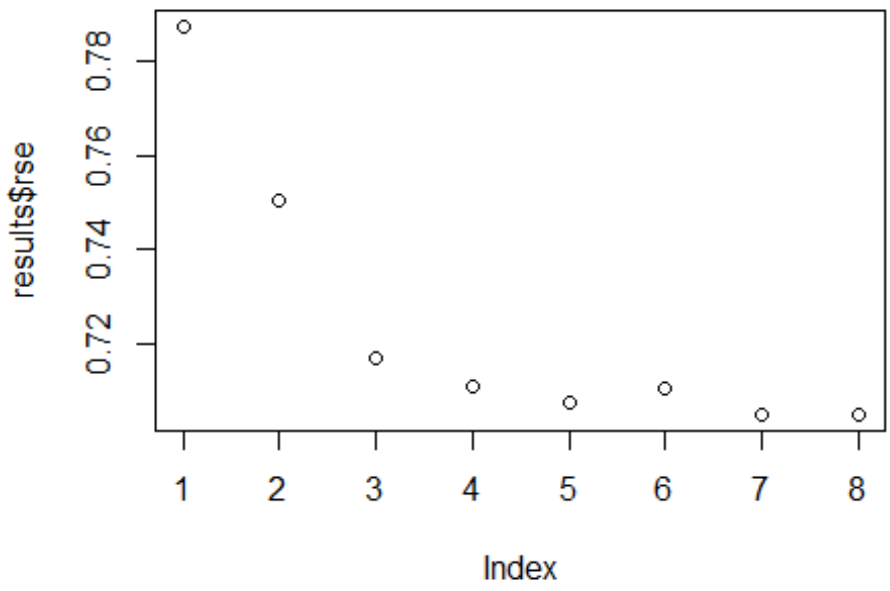
```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi + lbph + age + lcp,
##     data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.82853 -0.40741  0.01695  0.47159  1.59040
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.93487    0.83577   1.119  0.26630
## lcavol       0.58765    0.08663   6.783  1.2e-09 ***
## lweight      0.41808    0.16792   2.490  0.01462 *
## svi1         0.78256    0.24261   3.226  0.00175 **
## lbph         0.11381    0.05842   1.948  0.05452 .
## age         -0.01511    0.01081  -1.398  0.16546
## lcp         -0.04118    0.08135  -0.506  0.61392
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7102 on 90 degrees of freedom
## Multiple R-squared:  0.6451, Adjusted R-squared:  0.6215
## F-statistic: 27.27 on 6 and 90 DF,  p-value: < 2.2e-16

rse = append(rse, 0.7102)
r2 = append(r2, 0.6215)
m7 = lm(lpsa~lcavol+lweight+svi+lbph+age+lcp+pgg45, data=prostate)
summary(m7)

##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi + lbph + age + lcp +
##     pgg45, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.73117 -0.38137 -0.01728  0.43364  1.63513
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.953926   0.829439   1.150  0.25319
## lcavol       0.591615   0.086001   6.879 8.07e-10 ***
## lweight      0.448292   0.167771   2.672  0.00897 **
## svi1         0.757734   0.241282   3.140  0.00229 **
## lbph         0.107671   0.058108   1.853  0.06720 .
## age         -0.019336   0.011066  -1.747  0.08402 .
## lcp         -0.104482   0.090478  -1.155  0.25127
## pgg45        0.005318   0.003433   1.549  0.12488
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7048 on 89 degrees of freedom
## Multiple R-squared:  0.6544, Adjusted R-squared:  0.6273
## F-statistic: 24.08 on 7 and 89 DF,  p-value: < 2.2e-16
```

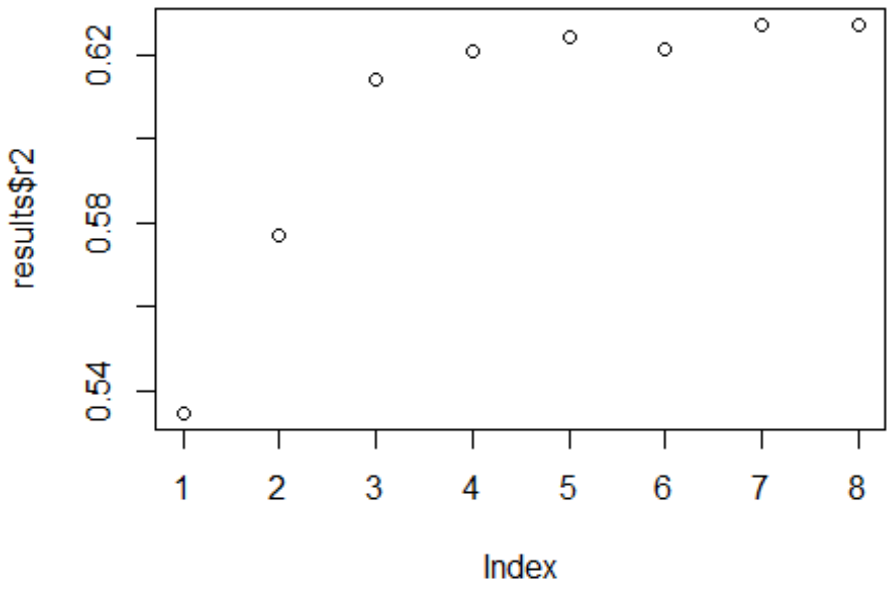
```
rse = append(rse, 0.7048)
r2 = append(r2, 0.6273)
m8 = lm(lpsa~lcavol+lweight+svi+lbph+age+lcp+pgg45+gleason, data=prostate)
summary(m8)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi + lbph + age + lcp +
##     pgg45 + gleason, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.74815 -0.35039 -0.02628  0.47655  1.70258
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.913282   0.840838   1.086  0.28044
## lcavol       0.569988   0.090100   6.326 1.09e-08 ***
## lweight      0.468791   0.169610   2.764  0.00699 **
## svi1         0.745879   0.247398   3.015  0.00338 **
## lbph         0.099685   0.058984   1.690  0.09464 .
## age         -0.021749   0.011361  -1.914  0.05890 .
## lcp         -0.125112   0.095591  -1.309  0.19408
## pgg45        0.004990   0.004672   1.068  0.28848
## gleason7     0.267607   0.219419   1.220  0.22595
## gleason8     0.496820   0.769267   0.646  0.52011
## gleason9    -0.056215   0.500196  -0.112  0.91078
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7048 on 86 degrees of freedom
## Multiple R-squared:  0.666,  Adjusted R-squared:  0.6272
## F-statistic: 17.15 on 10 and 86 DF,  p-value: < 2.2e-16
```

```
rse = append(rse, 0.7048)
r2 = append(r2, 0.6272)
results = data.frame(rse, r2)
plot(results$rse)
```



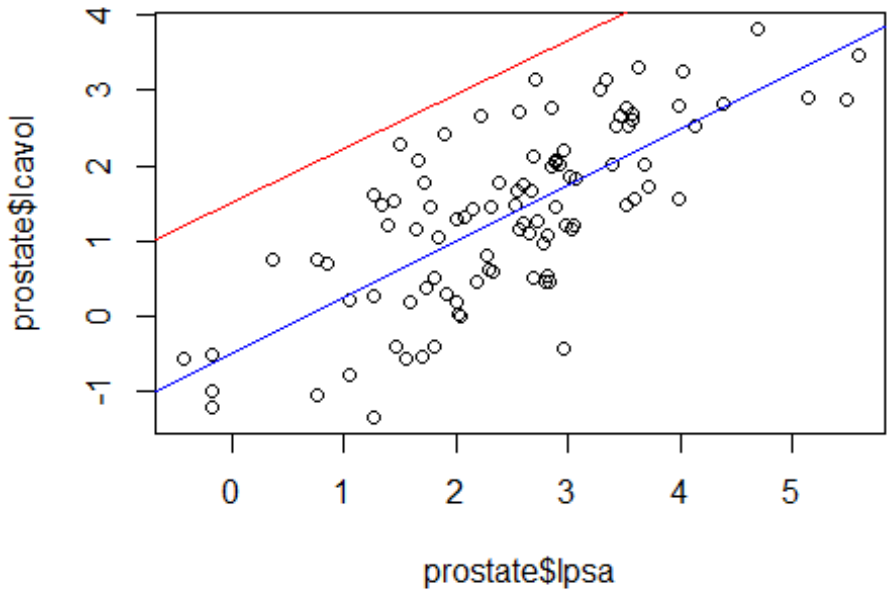

```r
plot(results$r2)
```



From the plots of the RSE and R^2, we can find the optimal number of predictors. Looking for the “elbow” in

the plot suggests that 4 predictors would be ideal, as the improvement past that is small. Those predictors are lcavol, lweight, svi, and lbph.

```
plot(prostate$lpsa,prostate$lcavol)
abline(lm(lpsa~lcavol,prostate), col="red")
abline(lm(lcavol~lpsa,prostate), col="blue")
```



The two regession lines do not intersect on the plot. They are nearly parallel.

Question 3

```
lmod1 = lm(lpsa~., prostate)
summary(lmod1)

##
## Call:
## lm(formula = lpsa ~ ., data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.74815 -0.35039 -0.02628  0.47655  1.70258
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.913282   0.840838   1.086  0.28044
## lcavol       0.569988   0.090100   6.326 1.09e-08 ***
## lweight      0.468791   0.169610   2.764  0.00699 **
```

```
## age         -0.021749   0.011361  -1.914  0.05890 .
## lbph         0.099685   0.058984   1.690  0.09464 .
## svi1         0.745879   0.247398   3.015  0.00338 **
## lcp         -0.125112   0.095591  -1.309  0.19408
## gleason7     0.267607   0.219419   1.220  0.22595
## gleason8     0.496820   0.769267   0.646  0.52011
## gleason9    -0.056215   0.500196  -0.112  0.91078
## pgg45        0.004990   0.004672   1.068  0.28848
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7048 on 86 degrees of freedom
## Multiple R-squared:  0.666,  Adjusted R-squared:  0.6272
## F-statistic: 17.15 on 10 and 86 DF,  p-value: < 2.2e-16
```

```
confint(lmod1, "age", level = 0.9)
```

```
##             5 %         95 %
## age -0.0406403 -0.002858458
```

```
confint(lmod1, "age", level = 0.95)
```

```
##           2.5 %       97.5 %
## age -0.04433459 0.0008358327
```

The R^2 of this model is 0.666. This means that the predictors explain 66.6% of the variability in the response, lpsa. Lcavol, lweight, age, lbph, and svi1 are significant at 90%. A one unit increase in lcavol results in a 0.569988 increase in lpsa, holding all other predictors constant. A one unit increase in lweight results in a 0.4688 increase in lpsa, holding all other predictors constant. An svi of 1 increases lpsa by 0.7459 compared to svi of 0, holding all othe predictors constant. The confidence intervals for the coefficient of age are both near to zero. This suggests the p-value for age is not significant. The summary shows an actual p-value of 0.0589, meaning it is not significant at 95% confidence.

```
noint = lm(lpsa~.-1, prostate)
summary(noint)
```

```
##
## Call:
## lm(formula = lpsa ~ . - 1, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.74815 -0.35039 -0.02628  0.47655  1.70258
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## lcavol    0.569988   0.090100   6.326 1.09e-08 ***
## lweight   0.468791   0.169610   2.764  0.00699 **
## age      -0.021749   0.011361  -1.914  0.05890 .
## lbph      0.099685   0.058984   1.690  0.09464 .
```

```
## svi0      0.913282   0.840838   1.086  0.28044
## svi1      1.659161   0.891998   1.860  0.06630 .
## lcp      -0.125112   0.095591  -1.309  0.19408
## gleason7  0.267607   0.219419   1.220  0.22595
## gleason8  0.496820   0.769267   0.646  0.52011
## gleason9 -0.056215   0.500196  -0.112  0.91078
## pgg45     0.004990   0.004672   1.068  0.28848
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7048 on 86 degrees of freedom
## Multiple R-squared:  0.941,  Adjusted R-squared:  0.9334
## F-statistic: 124.6 on 11 and 86 DF,  p-value: < 2.2e-16
```

THe p-value of the intercept is .28044, thus it is not significant at alpha = 0.05. The R^2 of the model without the intercept is 0.941. This is much higher than the model with the intercept and this should be preferred as it explains more of the variability in the response.

```
lmod2 = lm(lpsa~lcavol+lweight+svi, prostate)
summary(lmod2)

##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.72964 -0.45764  0.02812  0.46403  1.57013
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.26809    0.54350  -0.493  0.62298
## lcavol       0.55164    0.07467   7.388  6.3e-11 ***
## lweight      0.50854    0.15017   3.386  0.00104 **
## svi1         0.66616    0.20978   3.176  0.00203 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7168 on 93 degrees of freedom
## Multiple R-squared:  0.6264, Adjusted R-squared:  0.6144
## F-statistic: 51.99 on 3 and 93 DF,  p-value: < 2.2e-16
```

Removing all non-significant predictors, the new model has a slightly lower R^2 and slightly higher RSE. However, the differences are small and I would prefer the new model as it is less complex.

```
new = data.frame(lcavol = 1.44692, lweight = 3.62301, age = 65, lbph =
0.30010, svi = 0, lcp = -0.79851, gleason = 7, pgg45 = 15)
new$svi = as.factor(new$svi)
new$gleason = as.factor(new$gleason)
```

```
pred1 = predict(lmod1, new, interval = "prediction", level=0.95)
pred1
```

```
##        fit      lwr      upr
## 1 2.495014 1.069834 3.920195
```

```
new2 = new
new2$age = 20
new2
```

```
##    lcavol lweight age   lbph svi      lcp gleason pgg45
## 1 1.44692 3.62301  20 0.3001   0 -0.79851       7    15
```

```
pred2 = predict(lmod1, new2, interval = "prediction", level=0.95)
pred2
```

```
##        fit      lwr      upr
## 1 3.473736 1.729198 5.218275
```

```
pred3 = predict(lmod2, new2, interval = "prediction", level=0.95)
pred3
```

```
##        fit       lwr      upr
## 1 2.372534 0.9383436 3.806724
```

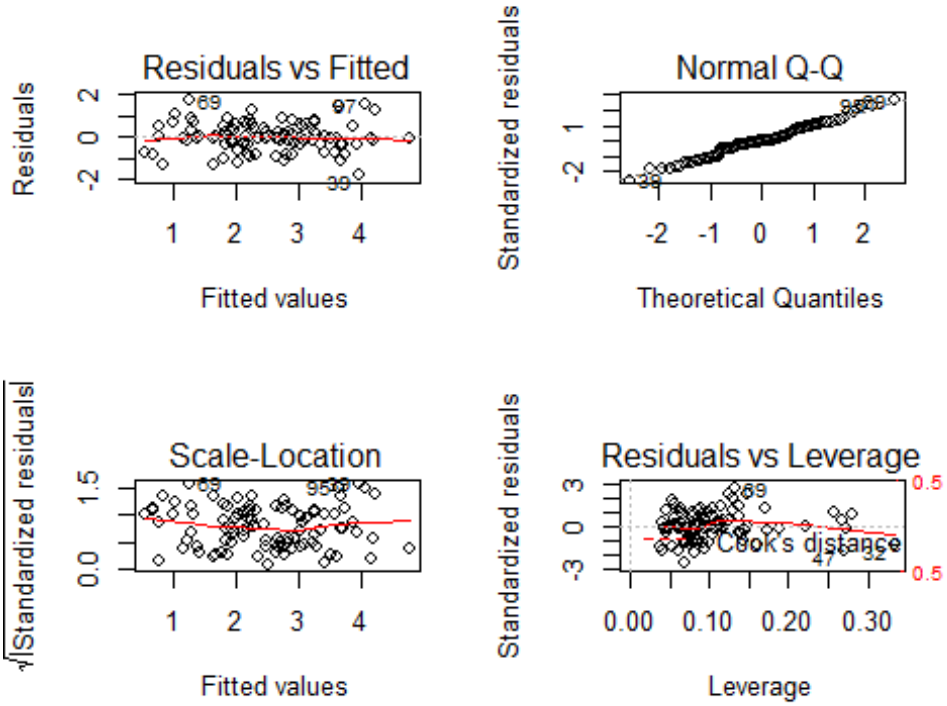
The confidence interval is wider when the age is 20 because that is outside the range of ages the model was trained on. The minimum age of the data was 41, thus an age of 20 is harder to predict and the error is bigger. The confidence interval using lmod2 is smaller and thus I would prefer that one as it is more confident of the prediction.

Question 4

```
par(mfrow=c(2,2))
plot(lmod1)
```

```
## Warning: not plotting observations with leverage one:
##   37
```

```
## Warning: not plotting observations with leverage one:
##   37
```



Using these plots, we can see there seems to be constant variance, linearity, and normality. There do not appear to be outliers as no point has a Cook's distance greater than 0.5. The plot from question one shows the response vs. predictors. The relationships appear to be linear, though not strong except for lcavol.

Question 5

```
prostate$svi = as.numeric(prostate$svi)
prostate$gleason = as.numeric(prostate$gleason)
cor(prostate)
```

```
##              lcavol      lweight       age        lbph         svi
## lcavol  1.00000000  0.194128387 0.2249999  0.02734971  0.53884500
## lweight 0.19412839  1.000000000 0.3075247  0.43493174  0.10877818
## age     0.22499988  0.307524741 1.0000000  0.35018592  0.11765804
## lbph    0.02734971  0.434931744 0.3501859  1.00000000 -0.08584327
## svi     0.53884500  0.108778185 0.1176580 -0.08584327  1.00000000
## lcp     0.67531058  0.100238891 0.1276678 -0.00699944  0.67311122
## gleason 0.43241705 -0.001283003 0.2688916  0.07782044  0.32041222
## pgg45   0.43365224  0.050846195 0.2761124  0.07846000  0.45764762
## lpsa    0.73446028  0.354121818 0.1695929  0.17980950  0.56621818
##                 lcp      gleason     pgg45      lpsa
## lcavol   0.67531058  0.432417052 0.4336522 0.7344603
## lweight  0.10023889 -0.001283003 0.0508462 0.3541218
## age      0.12766778  0.268891599 0.2761124 0.1695929
## lbph    -0.00699944  0.077820444 0.0784600 0.1798095
```

```
## svi      0.67311122  0.320412221 0.4576476 0.5662182
## lcp      1.00000000  0.514829912 0.6315281 0.5488132
## gleason  0.51482991  1.000000000 0.7519045 0.3689867
## pgg45    0.63152807  0.751904512 1.0000000 0.4223157
## lpsa     0.54881316  0.368986693 0.4223157 1.0000000
```

```
vif(lmod1)
```

```
##   lcavol  lweight      age     lbph     svi1      lcp gleason7 gleason8
## 2.179199 1.371096 1.382570 1.415080 2.027131 3.452249 2.293919 1.178914
## gleason9    pgg45
## 2.388330 3.355597
```

The correlation matrix shows some strong correlations between predictors. However, the VIF does not show multicollinearity as no predictor is over 10.

```
prostate$svi = as.factor(prostate$svi)
prostate$gleason = as.factor(prostate$gleason)
lmod1_step = step(lmod1, direction = "both", trace = F)
AIC(lmod1_step)
```

```
## [1] 215.8997
```

```
summary(lmod1_step)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83505 -0.39396  0.00414  0.46336  1.57888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.95100    0.83175   1.143 0.255882
## lcavol       0.56561    0.07459   7.583 2.77e-11 ***
## lweight      0.42369    0.16687   2.539 0.012814 *
## age         -0.01489    0.01075  -1.385 0.169528
## lbph         0.11184    0.05805   1.927 0.057160 .
## svi2         0.72095    0.20902   3.449 0.000854 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7073 on 91 degrees of freedom
## Multiple R-squared:  0.6441, Adjusted R-squared:  0.6245
## F-statistic: 32.94 on 5 and 91 DF,  p-value: < 2.2e-16
```

The AIC of the stepwise mddel is 215.8997. The model removed lcp, gleason, and pgg45. The R^2 is 0.644 which is lower than lmod1. Lweight, age, and lbph are not significant at

95%. I would choose the stepwise model as it is less complex and the differences in RSE and R^2 are small.