SGupta_HW03Question3

Question 3

Fit a model with lpsa as the response and l cavol as the predictor.

```
library(faraway)
set.seed(1001)
data(prostate)
names(prostate)
```

```
[1] "lcavol"  "lweight" "age"     "lbph"    "svi"     "lcp"     "gleason"
[8] "pgg45"   "lpsa"
```

```
model1 <- lm(lpsa ~ lcavol, data = prostate)
summary1 <- summary(model1)
summary1
```

```
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
```

Record the residual standard error and the R2.

```r
rse <- summary(model1)$sigma
r2 <- summary(model1)$r.squared

print(paste("Residual Standard Error:", rse))
```

```
[1] "Residual Standard Error: 0.787499423513711"
```

```r
print(paste("Multiple R-squared:", r2))
```

```
[1] "Multiple R-squared: 0.53943190877902"
```

Now add lweight, svi, lpph, age, lcp, pgg45 and gleason to the model one at a time

```r
model2 <- lm(lpsa ~ lcavol + lweight, data = prostate)
summary2 <- summary(model2)

model3 <- lm(lpsa ~ lcavol + lweight + svi, data = prostate)
summary3 <- summary(model3)

model4 <- lm(lpsa ~ lcavol + lweight + svi + lbph, data = prostate)
summary4 <- summary(model4)

model5 <- lm(lpsa ~ lcavol + lweight + svi + lbph + age, data = prostate)
summary5 <- summary(model5)

model6 <- lm(lpsa ~ lcavol + lweight + svi + lbph + age + lcp, data = prostate)
summary6 <- summary(model6)
```

```r
model7 <- lm(lpsa ~ lcavol + lweight + svi + lbph + age + lcp + pgg45, data = prostate)
summary7 <- summary(model7)

model8 <- lm(lpsa ~ lcavol + lweight + svi + lbph + age + lcp + pgg45 + gleason, data = pros
summary8 <- summary(model8)
```

```r
rse <- c(s1 = summary1$sigma,
         s2 = summary2$sigma,
         s3 = summary3$sigma,
         s4 = summary4$sigma,
         s5 = summary5$sigma,
         s6 = summary6$sigma,
         s7 = summary7$sigma,
         s8 = summary8$sigma)

r2 <- c(s1 = summary1$r.squared,
        s2 = summary2$r.squared,
        s3 = summary3$r.squared,
        s4 = summary4$r.squared,
        s5 = summary5$r.squared,
        s6 = summary6$r.squared,
        s7 = summary7$r.squared,
        s8 = summary8$r.squared)

rse
```

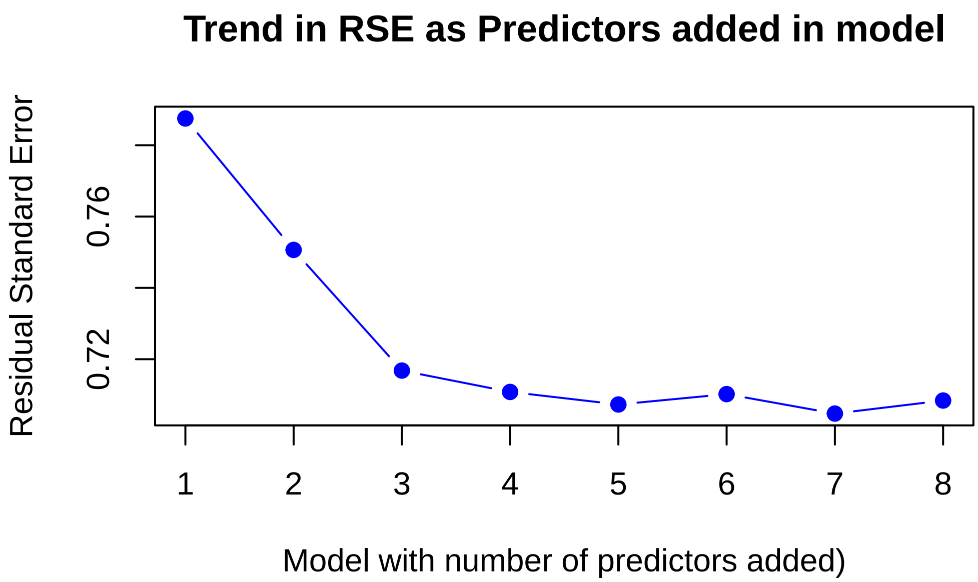
```
       s1        s2        s3        s4        s5        s6        s7        s8
0.7874994 0.7506469 0.7168094 0.7108232 0.7073054 0.7102135 0.7047533 0.7084155
```

```r
r2
```

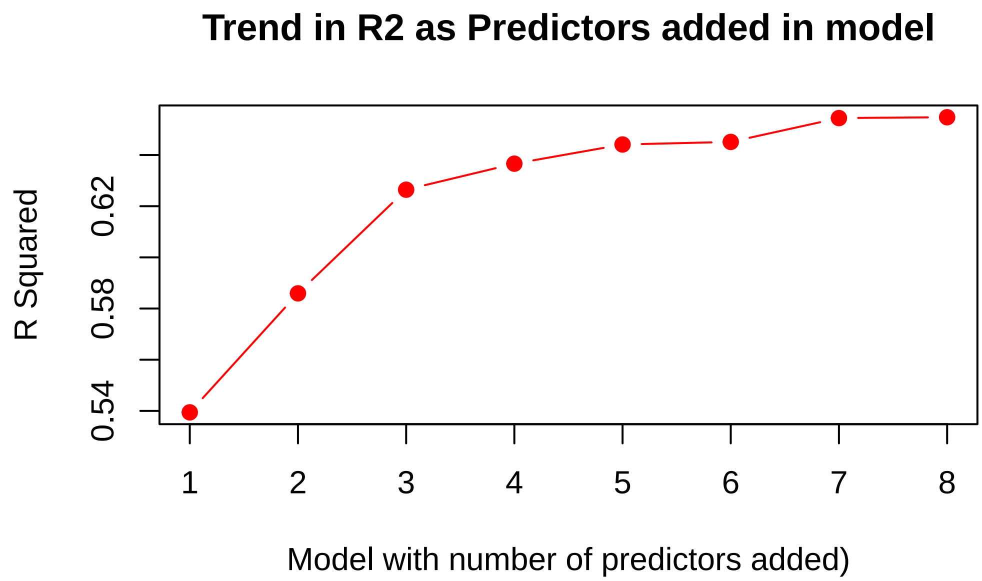
```
       s1        s2        s3        s4        s5        s6        s7        s8
0.5394319 0.5859345 0.6264403 0.6366035 0.6441024 0.6451130 0.6544317 0.6547541
```

Plot the trends in these two statistics

```
models <- 1:8
plot(models, rse, type="b", pch=19, col="blue",
     xlab="Model with number of predictors added)", ylab="Residual Standard Error",
     main="Trend in RSE as Predictors added in model")
```




```
plot(models, r2, type="b", pch=19, col="red",
     xlab="Model with number of predictors added)", ylab="R Squared",
     main="Trend in R2 as Predictors added in model")
```



Residual Standard Error (RSE): The RSE generally decreases as more predictors are added. A lower RSE indicates that the model's predictions are closer to the actual observed values.

R square : The R square generally increases as more predictors are added, because R square can never decrease when more terms are added to a linear regression model. The increase in R square shows that the added predictors explain additional variability in the response. After a certain point, each predictor addition only slightly affect R square. That means most of the significant variation is captures in the early predictors.