Homework 3

Subrata Paul

9/7/2020

Problem 1

Download the `simu_hw3.txt` data from canvas and read it in R. The data has four columns `x1, x2, x3` and `y`. Print the summary of the linear regression model

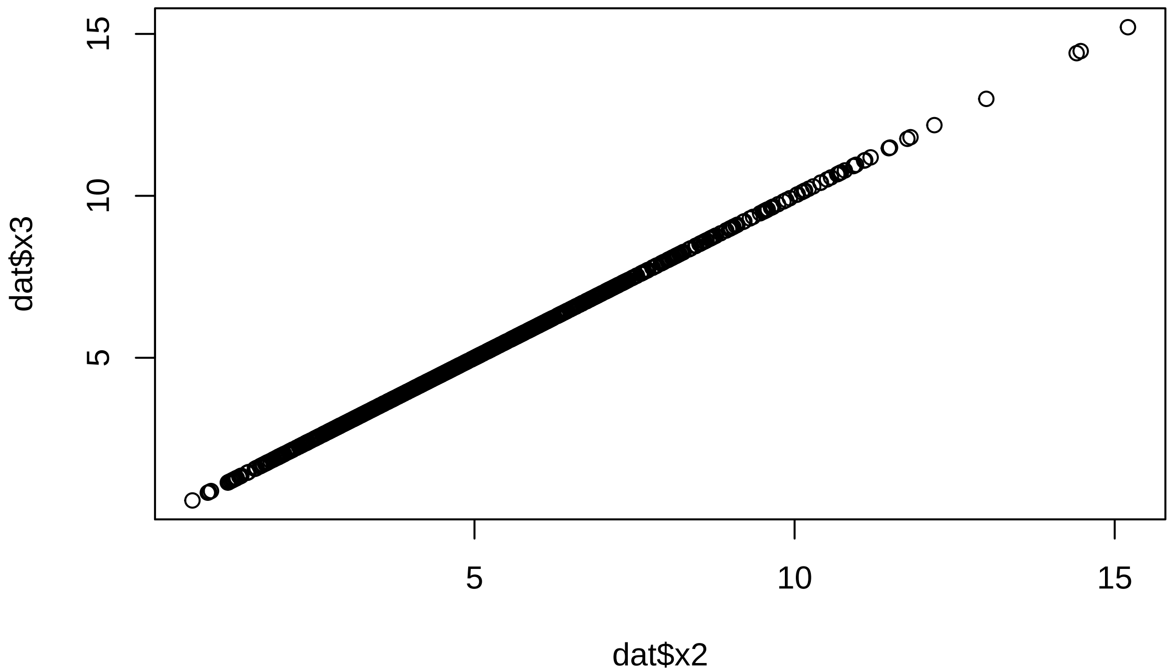
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

```r
dat = read.table('../data/simu_hw3.txt',header = T)
summary(lm(y~x1 + x2 + x3, data = dat))
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31467 -0.07056  0.00329  0.06934  0.32659
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.995e+00  8.705e-03 344.118   <2e-16 ***
## x1          -2.031e+00  5.539e-02 -36.670   <2e-16 ***
## x2           1.370e+02  3.167e+02   0.433    0.665
## x3          -1.300e+02  3.167e+02  -0.411    0.681
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1021 on 996 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 7.805e+06 on 3 and 996 DF,  p-value: < 2.2e-16
```

- Is there something that surprise you? What it is? The estimate and standard error of x_2 and x_3 are very high. A possible reason is collinearity.
- Why do you thing it might happend? Justify your answer. (You can use plots or some statistic for justification.)

```r
plot(dat$x2, dat$x3)
```



The plot of x_3 versus x_2 shows that the two variables are highly correlated. The correlation between the two variables can be calculated using the `cor` function.

```
cor(dat$x2,dat$x3)
```

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

R reported that the correlation between the two variables is one but in such case R should automatically remove one of the two predictors from the model due to singularity of $X^T X$. In this case it did not do that because the correlation is not exactly equal to 1.

```
cor(dat$x2,dat$x3)==1
```

```
## [1] FALSE
```

Printing more decimal points would make it clear.

```
sprintf("%.20f", cor(dat$x2, dat$x3))
```

```
## [1] "0.99999999998956445868"
```

So, x_2 and x_3 are not exactly equal rather one is a perturbed version of another that's why the $X^T X$ matrix is not singular but ill-conditioned.

```
lmod = lm(y~x1 + x2 + x3, data = dat)
X = model.matrix(lmod) # Design matrix (with one on the first column)
kappa(t(X)%*%X) # Condition number
```

```
## [1] 1.201822e+12
```

- What model do you recommend? Run the recommended model and print the summary.

I would recommend any of the two variables.

```
summary(lm(y~x1+x2, data = dat))
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = dat)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31617 -0.07116  0.00342  0.06945  0.32693
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.995438   0.008701  344.26   <2e-16 ***
## x1          -2.031443   0.055365  -36.69   <2e-16 ***
## x2           7.001023   0.001446 4840.32   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1021 on 997 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.172e+07 on 2 and 997 DF,  p-value: < 2.2e-16
```

Problem 2

Fit $y = \beta_0 + \beta_1 x_1$ model and populate the following table without using the `anova` function.

```
lmod1 = lm(y~x1, data = dat)
RSS = sum(lmod1$residuals^2)
ybar = mean(dat$y)
TSS = sum((dat$y - ybar)^2)
SS_reg = TSS - RSS
c(SS_reg, RSS, TSS) #SS
```

```
## [1]     56.3539 244215.0362 244271.3901
```

```
c(SS_reg, RSS, TSS)/c(1,998,999) #MS
```

```
## [1]  56.3539 244.7044 244.5159
```

```
library(xtable)
options(xtable.comment = FALSE)
tab = data.frame(Source = c("$SS_{reg}(X_1)$", "$RSS(X_1)$", "TSS" ),
                 SS = c(SS_reg, RSS, TSS),
                 df = c(1,998,999),
                 MS = c(SS_reg, RSS, TSS)/c(1,998,999))
print(xtable(tab), sanitize.text.function = function(x) {x})
```

	Source	SS	df	MS
1	$SS_{reg}(X_1)$	56.35	1.00	56.35
2	$RSS(X_1)$	244215.04	998.00	244.70
3	TSS	244271.39	999.00	244.52

`SS`, `df`, and `MS` represent the sum of squares, degrees of freedom, and mean sum of squares, respectively. `MS` = `SS/df`.

Problem 3

Fit $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ model. Now, if you want, you can use the `anova` function.

```
lmod2 = lm(y~x1+x2, data = dat)
summary(lmod2)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31617 -0.07116  0.00342  0.06945  0.32693
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.995438   0.008701  344.26   <2e-16 ***
## x1          -2.031443   0.055365  -36.69   <2e-16 ***
## x2           7.001023   0.001446 4840.32   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1021 on 997 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.172e+07 on 2 and 997 DF,  p-value: < 2.2e-16
```

```r
RSS = sum(lmod2$residuals^2)
ybar = mean(dat$y)
TSS = sum((dat$y - ybar)^2)
SS_reg = TSS - RSS
c(SS_reg, RSS, TSS) #SS
```

```
## [1] 244260.99809     10.39203 244271.39012
```

```r
c(SS_reg, RSS, TSS)/c(2,997,999) #MS
```

```
## [1] 1.221305e+05 1.042330e-02 2.445159e+02
```

```r
library(xtable)
tab = data.frame(Source = c("$SS_{reg}(X_1, X_2)$", "$RSS(X_1, X_2)$", "TSS" ),
                 SS = c(SS_reg, RSS, TSS),
                 df = c(2,997,999),
                 MS = c(SS_reg, RSS, TSS)/c(2,997,999))
print(xtable(tab), sanitize.text.function = function(x) {x})
```

	Source	SS	df	MS
1	$SS_{reg}(X_1, X_2)$	244261.00	2.00	122130.50
2	$RSS(X_1, X_2)$	10.39	997.00	0.01
3	TSS	244271.39	999.00	244.52

Problem 4

Define $SS_{reg}(X_2|X_1) = RSS(X_1) - RSS(X_1, X_2)$. $SS_{reg}(X_2|X_1)$ is called the extra sum of squares. Calculate $SS_{reg}(X_2|X_1)$. Can you write $SS_{reg}(X_2|X_1)$ in terms of SS_{reg} of the above models?

$$SS_{reg}(X_2|X_1) = 244215.036 - 10.39203 = 244204.6$$

$$SS_{reg}(X_2|X_1) = SS_{reg}(X_1, X_2) - SS_{reg}(X_1) = 245634.72687 - 1430.083 = 244204.6$$

Problem 5

The dataset `teengamb` from `faraway` package concerns a study of teenage gambling in Britain. Fit a regression model with the expenditure on gambling as the response and the sex, status, income and verbal score as predictors. Present the output.

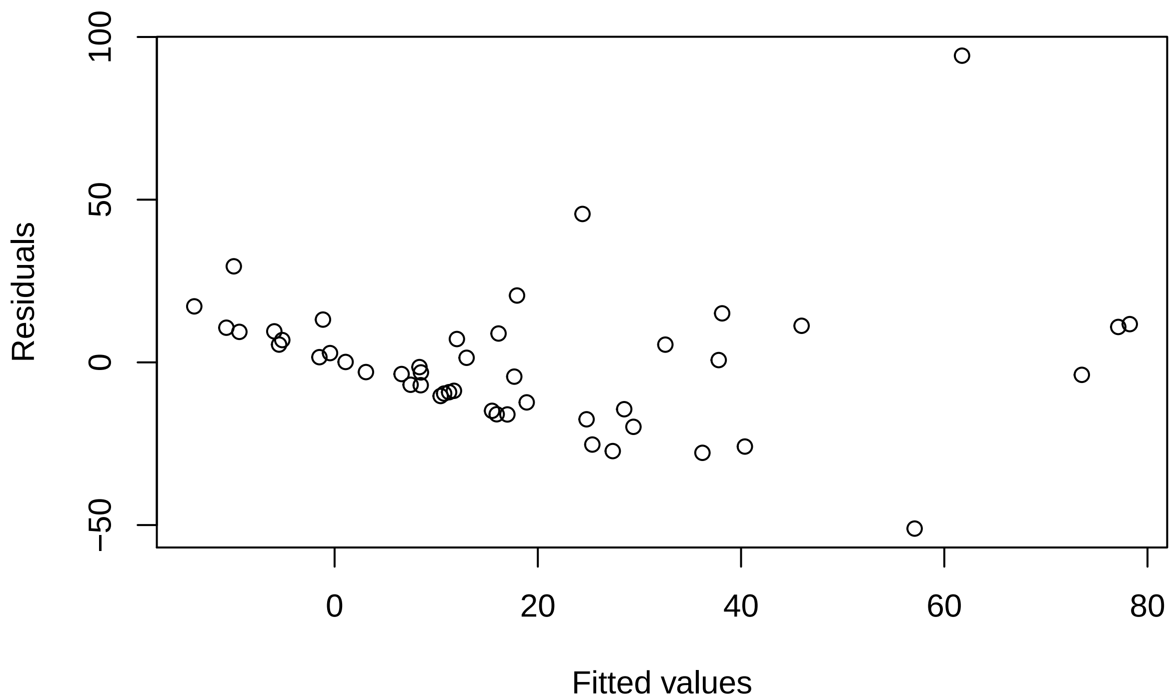
```
data(teengamb, package = 'faraway')
lmod = lm(gamble ~., data = teengamb)
summary(lmod)
```

```
##
## Call:
## lm(formula = gamble ~ ., data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.55565   17.19680   1.312   0.1968
## sex         -22.11833    8.21111  -2.694   0.0101 *
## status        0.05223    0.28111   0.186   0.8535
## income        4.96198    1.02539   4.839 1.79e-05 ***
## verbal       -2.95949    2.17215  -1.362   0.1803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.69 on 42 degrees of freedom
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.4816
## F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

(a) What percentage of variation in the response is explained by these predictors?

The percentage of variation explained is 52.67%.

```
plot(lmod$fitted.values,lmod$residuals, xlab = 'Fitted values', ylab = 'Residuals')
```



The plot of the responses versus the fitted value doesn't look terribly linear, so R^2 may not be a very useful measure of model fit.

(b) Which observation has the largest (positive) residual? Give the case number.

```
which.max(lmod$residuals)
```

```
## 24
## 24
```

```
lmod$residuals[which.max(lmod$residuals)]
```

```
##       24
## 94.25222
```

The largest residual is obtained for 24th observation and it is 94.25222.

(c) Compute the mean and median of the residuals.

Mean of the residuals should be zero as the OLS method is used for model fit. Why?

```
mean(lmod$residuals)
```

```
## [1] -3.065293e-17
```

```
median(lmod$residuals)
```

```
## [1] -1.451392
```

(d) Compute the correlation of the residuals with the fitted values.

According to the theory the correlation between residuals and the fitted values should be zero.

```
cor(lmod$residuals, lmod$fitted.values)
```

```
## [1] -1.070659e-16
```

(e) Compute the correlation of the residuals with the income.

```r
cor(lmod$residuals, teengamb$income)
```

```
## [1] -7.242382e-17
```

(f) For all other predictors held constant, what would be the difference in predicted expenditure on gambling for a male compared to a female?

```r
coef(lmod)[2]
```

```
##       sex
## -22.11833
```

Females are estimated to spend about 22 pounds less per year on gambling than their male counterparts, holding other factors constant.

Problem 6

In this question, we investigate the relative merits of methods for computing the coefficients. Generate some artificial data by:

```r
x<-1:20
y <- x+ rnorm(20)
```

Fit a polynomial in x for predicting y. Compute $\hat{\beta}$ in two ways — by `lm()` and by using the direct calculation described in the chapter. At what degree of polynomial does the direct calculation method fail? (Note the need for the `I()` function in fitting the polynomial, that is, `lm(y ~ x + I(x^2))`.

The direct solve method fails at degree 6 (for me).

```r
set.seed(1)
x <- 1:20
y <- x + rnorm(20)
# Compute betahat using lm function for polynomials up to 5
lmod1 <- lm(y ~ x)
lmod2 <- lm(y ~ x + I(x^2))
lmod3 <- lm(y ~ x + I(x^2) + I(x^3))
lmod4 <- lm(y ~ x + I(x^2) + I(x^3) + I(x^4))
lmod5 <- lm(y ~ x + I(x^2) + I(x^3) + I(x^4) + I(x^5))
lmod6 <- lm(y ~ x + I(x^2) + I(x^3) + I(x^4) + I(x^5) + I(x^6))
X = model.matrix(lmod1) # extract X matrix from model
solve(t(X) %*% X, t(X) %*% y) # compute coefficients manually
```

```
##                    [,1]
## (Intercept) -0.03609284
## x            1.02158254
```

```r
coef(lmod1) # pull out coefficients estimated by lm function
```

```
## (Intercept)           x
## -0.03609284  1.02158254
```

```r
X = model.matrix(lmod6)
solve(t(X) %*% X, t(X) %*% y)
```

```
## Error in solve.default(t(X) %*% X, t(X) %*% y): system is computationally singular: reciprocal condi
```

```r
coef(lmod6)
```

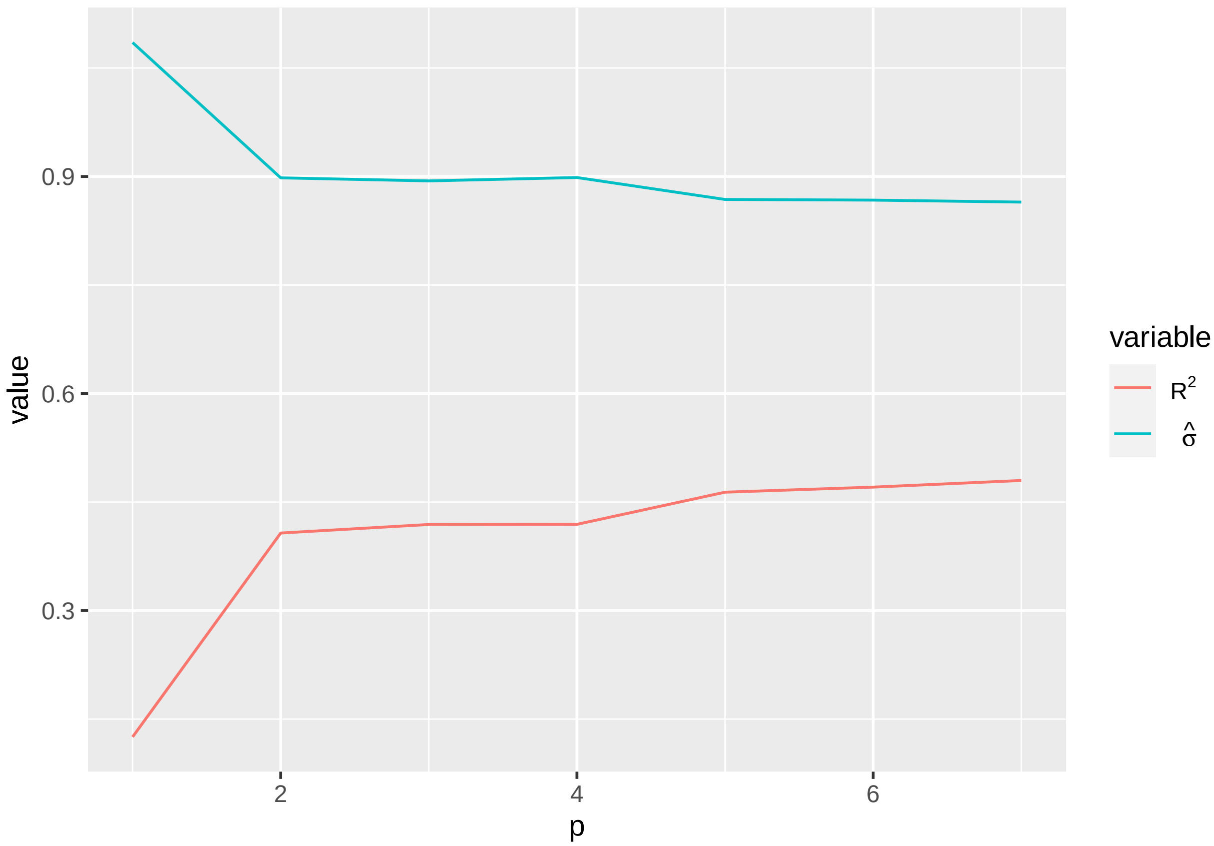
```
##   (Intercept)             x        I(x^2)        I(x^3)        I(x^4)
## -2.587888e+00  3.816223e+00 -1.076985e+00  1.966281e-01 -1.792915e-02
```

```
##        I(x^5)        I(x^6)
##  7.822853e-04 -1.297061e-05
```

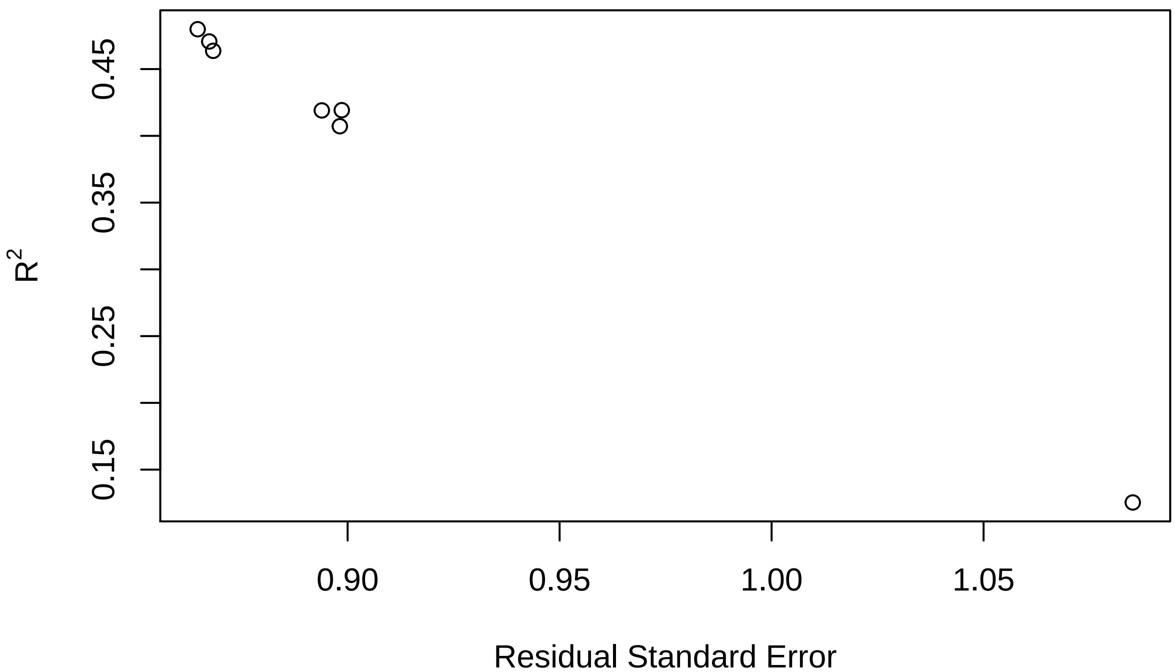
Problem 7

The dataset **prostate** in the **faraway** package comes from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. Fit a model with **lpsa** as the response and **lcavol** as the predictor. Record the residual standard error and the R^2. Now add **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.

```r
library(ggplot2)
data(prostate, package = 'faraway')
pred = c('lweight', 'svi','lbph','age','lcp','pgg45','gleason')
r_squared<-c()
rse <- c()
for(i in 1:length(pred)){
  model_formula = as.formula(paste0('lpsa ~ ', paste(pred[1:i], collapse = '+')))
  lmod = lm(model_formula, data = prostate)
  r_squared[i]<-summary(lmod)$r.squared
  rse[i]<-summary(lmod)$sigma
}
plot_dat = reshape2::melt(data.frame(p = seq(1,length(pred)),rsq = r_squared, rse = rse), id.vars = 'p'
ggplot(data = plot_dat, aes(x = p, y = value, color = variable))+
  geom_line()+
  scale_color_discrete(labels = c(expression(R^2), expression(hat(sigma))))
```



```r
plot(r_squared ~ rse, xlab = 'Residual Standard Error', ylab = expression(R^2))
```



$R^2 = 1 - \frac{RSS}{TSS}$, and the residual standard error is $s\hat{igma} = \sqrt{RSS/(n-p)}$. As we add more regressor variables to our regression model, RSS will decrease. When RSS decreases, we EXPECT $\hat{\sigma}$ to decrease and R^2 must increase (since TSS is a constant). Note: I say we expect $\hat{\sigma}$ to decrease because $n - p$ in the denominator of $\hat{\sigma}$ also changes as we add regressors. It is possible that $\hat{\sigma}$ could in fact increase, though this is generally not the case for well-chosen models.