Jonothan Meyer
Computational Linear Regression
03/12/21
Project 5

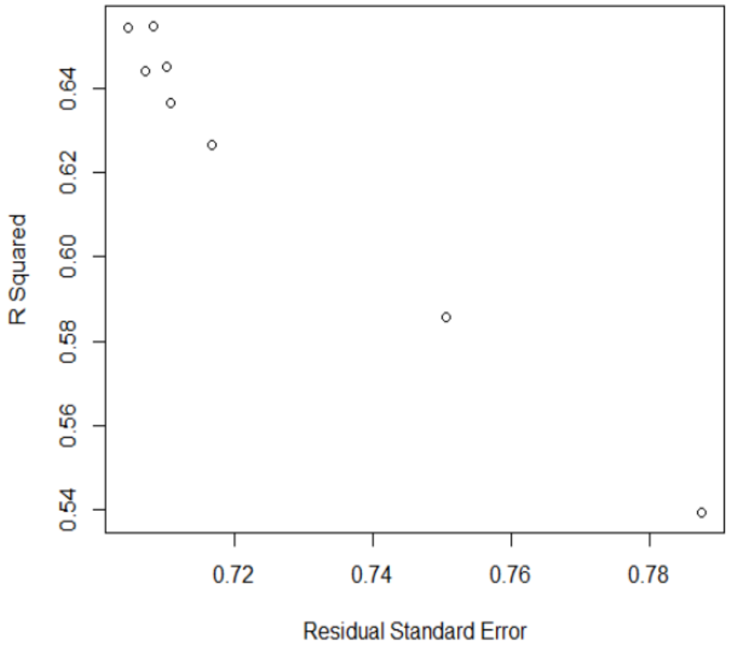
Exercise 1
(From book, pg 23, problem 4): 4) The dataset 'prostate' comes from a study on 97 men with prostate cancer who were due to receive a rradical prostatectomy. Fit a model with 'lpsa' as the response and lcavol as the predictor. Record the residual standard error and the r^2. Now add 'lweight', '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 2 stats.

```
lpsa.lcavol.lm<-lm(lpsa~lcavol, data=prostate) #First Linear Model w/1 predictor
summary(lpsa.lcavol.lm) #r^2= .5394, residual-stan-error=.7875

lpsa.lcavol.lwight.lm<-lm(lpsa~lcavol+lweight, data=prostate) #Second LM w/2 predictors
summary(lpsa.lcavol.lwight.lm) #r^2=.5859, rse=.7506

...

lpsa.lcavol.lweight.svi.lbph.age.lcp.pgg45.gleason.lm<- lm(lpsa~lcavol+lweight+svi+lbph+
                                                    age+lcp+pgg45+gleason, data=prostate) #Final LM
                                                                                  #with all predictors.
summary(lpsa.lcavol.lweight.svi.lbph.age.lcp.pgg45.gleason.lm) #r^2=.6548, rse=.7084

rse_r2_his<- data.frame(num_predictors = c(1:8),
        name_pred = c("lcavol", "+lweight", "+svi", "+lbph", "+age", "+lcp", "+pgg45", "+gleason"),
        r_squared = c(.5394, .5859, .6264, .6366, .6441, .6451, .6544, .6548),
        res_stan_error = c(.7875, .7506, .7168, .7108, .7073, .7102, .7048, .7084))
```

rse_r2:

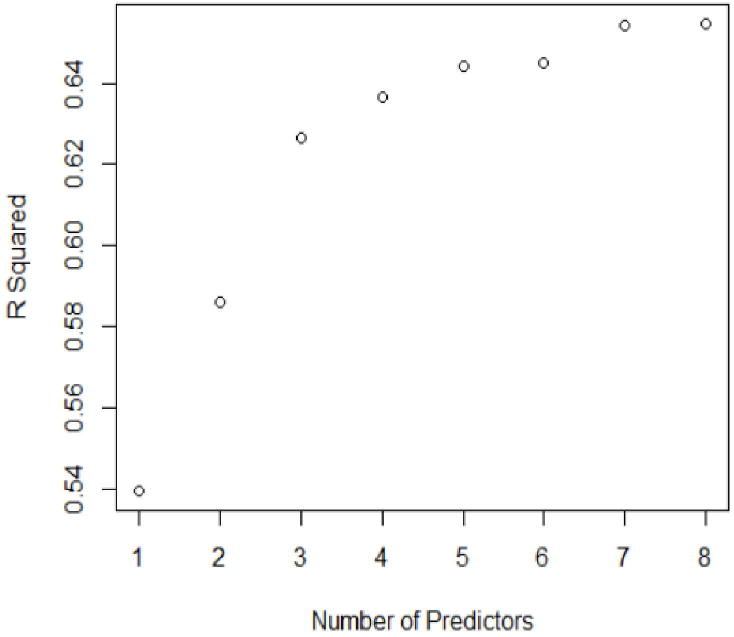
Number of Predictors	Name of Predictor	R^2	Residual Standard Error
1	lcavol	0.54	0.79
2	+lweight	0.59	0.75
3	+svi	0.63	0.72
4	+lbph	0.64	0.71
5	+age	0.64	0.71
6	+lcp	0.65	0.71
7	+pgg45	0.65	0.7
8	+gleason	0.65	0.71

Scatterplot of RSE vs. R^2



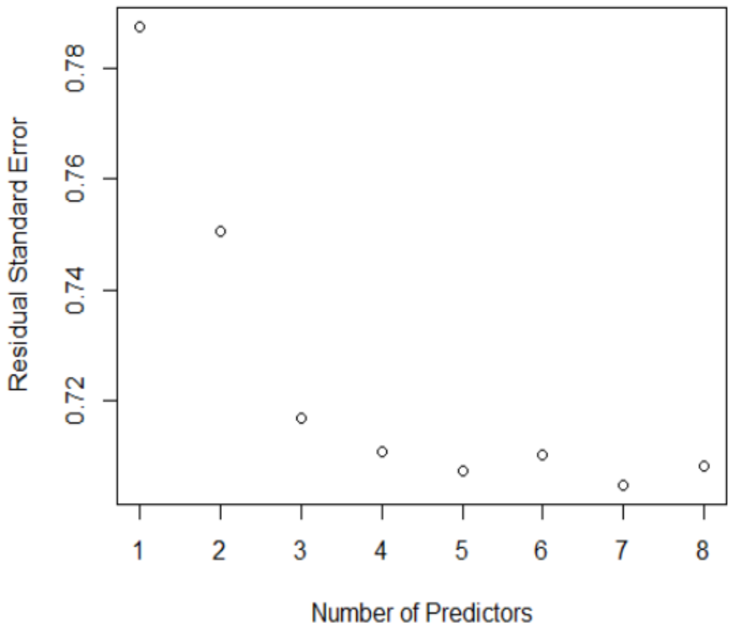
When plotting the Residual Standard Error against R^2 we can see that as residual standard error increases, r^2 decreases.

R^2 vs. Number of Predictors



As the predictors were added to the linear model R^2 increased in an apparent logarithmic way.

Number of Predictors vs. Residual Standard Error



As the number of predictors increases the residual standard error decreases, with a potential asymptote around .70.

Exercise 2 (6 points): Find five different sized vessels such as a cup or a pot or a vase in your house. Measure the weight (or width) of each vessel, the height of each vessel, and the volume of water each vessel holds as accurately as possible. Organize this data in a data frame in R, identifying volume as the response and the other two measurements as predictors.

```
Vessels <- data.frame(Vessel =
                       c("Rx_bottle", "Seltzer_can", "Sm_hand_sntizer", "Sgr_srip_case", "Vapur_wtr_btl"),
                      Volume_ml = c(47.1, 355, 59, 24.2, 500),
                      Weight_g = c(7.6, 14.6, 10.6, 9.1, 16.3),
                      Height_cm = c(7, 12.4, 9.7, 5, 19.3))
```

Vessels:

Vessel	Volume - mL	Weight - g	Height - cm
Rx Bottle	47.1	7.6	7
Seltzer Can	355	14.6	12.4
Small Hand Sanitizer	59	10.6	9.7
Glucose Strip Case	24.2	9.1	5
'Vapur' Water Bottle	500	16.3	19.3

(1) Identify X, Y, Yhat, and e in this context; include the, R code you use.

```
X <- matrix(c(1,1,1,1,1, 7.6,14.6,10.6,9.1,16.3, 7.0,12.4,9.7,5.0,19.3), nrow=5, ncol=3)
X.another <- cbind(rep(1,5), Vessels$Weight_g, Vessels$Height_cm) #Same Result
X.matrix.model <- model.matrix(Vessels.lm) #Also Same Result
```

X Matrix:

1	7.6	7
1	14.6	12.4
1	10.6	9.7
1	9.1	5
1	16.3	19.3

```
Y <- matrix(c(47.1,355,59,24.2,500), nrow=5, ncol=1)
```

Y Matrix:

47.1
355
59
24.2
500

```
m.beta  <- matrix(c(-377.10, 35.50, 15.07), ncol=1)
```

Beta Matrix:

-377.1
35.5
15.07

```
Yhat <- fitted(Vessels.lm)
```

Yhat:

1	2	3	4	5
-1.82	328.06	145.37	21.29	492.4

```
e <- Y-Yhat
```

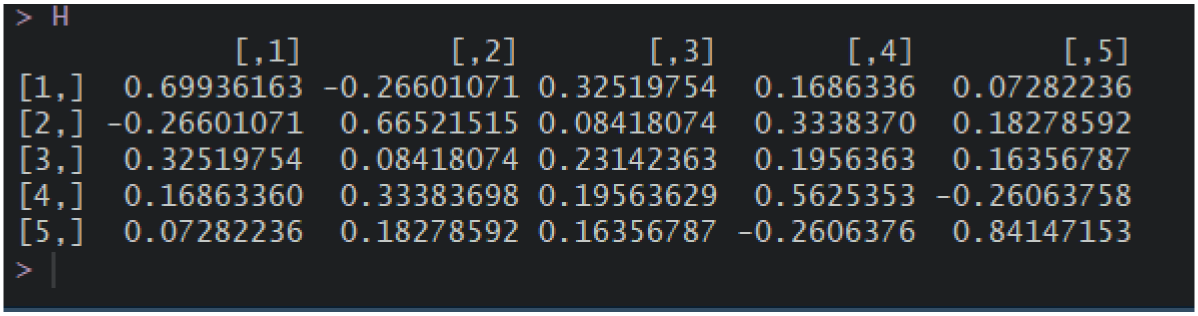
e:

48.92
26.94
-86.37
2.9
7.6

(2) Identify H, SSR, SSE, SST, MSR, MSE, s2 and R2 in this context; include the R code you use.

H<-(X%*%inv(t(X)%*%X))%*%t(X)

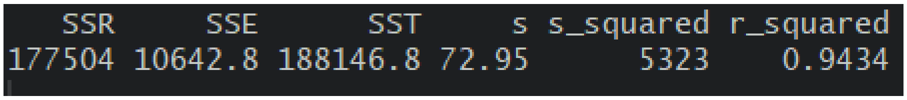
H:



```
> H
           [,1]        [,2]       [,3]       [,4]        [,5]
[1,]  0.69936163 -0.26601071 0.32519754  0.1686336  0.07282236
[2,] -0.26601071  0.66521515 0.08418074  0.3338370  0.18278592
[3,]  0.32519754  0.08418074 0.23142363  0.1956363  0.16356787
[4,]  0.16863360  0.33383698 0.19563629  0.5625353 -0.26063758
[5,]  0.07282236  0.18278592 0.16356787 -0.2606376  0.84147153
> 
```

nsmp = 5
J<-matrix(1, nrow=nsmp, ncol=nsmp)
SSR<-t(Y)%*%(H-1/nsmp*J)%*%Y
SSE<-t(Y)%*%(diag(nsmp)-H)%*%Y
SST<-t(Y)%*%(diag(nsmp)-1/nsmp*J)%*%Y
anova(Vessels.lm) #Used for calculating MSR and MSE
summary(Vessels.lm) #Used for finding s, used to calculate s^2. Also used to find R^2
dataframe <- data.frame(SSE = c(SSE), SST = c(SST), s = c(72.95), s_squared = c(5323), r_squared = c(.9434))

dataframe



```
   SSR     SSE      SST     s s_squared r_squared
177504 10642.8 188146.8 72.95      5323    0.9434
```

s_sqr_b.matrix.Vessels <- 5323*inv(t(X)%*%X)

(3) Interpret R2 in terms of the real world variables.

R2 represents how much the linear model we built encapsulates the real data. In terms of the variables we are looking at this is to say that when looking at the comparing the weight and height of our containers and how those variables relate to the volume of our container we can say our mathematical model "captures" the relationship by 94.34% (R2 = .9434).

Exercise 3 (10 points): Complete Textbook Exercise 7 from Chapter 2 (page 31), stated as follows: An experiment was conducted to determine the effect of four factors on the resistivity of a semiconductor wafer. The data is found in wafer where each of the four factors is coded as □ or + depending on whether the low or the high setting for that factor was used. Fit the linear model resist ~ x1 + x2 + x3 + x4.

library(faraway)
head(wafer)
wafer.lm <- lm(resist~x1+x2+x3+x4, data =wafer)

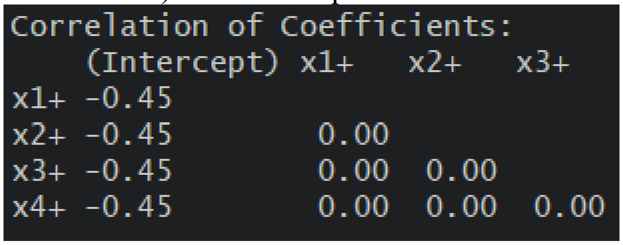
(1) Extract the X matrix using the model.matrix function. Examine this to determine how the low and high levels have been coded in the model.

X <- model.matrix(wafer.lm)

low and high levels were coded as either 0/1, 0 for low, 1 for high

(2) Compute the correlation in the X matrix. Why are there some missing values in the matrix?

summary(wafer.lm, correlation = TRUE) #All intercept correlations are -.45 for each predictor.



```
Correlation of Coefficients:
    (Intercept) x1+   x2+   x3+
x1+ -0.45
x2+ -0.45        0.00
x3+ -0.45        0.00  0.00
x4+ -0.45        0.00  0.00  0.00
```

There are some missing values in this matrix simply because it would be redundant and a bit less reable to include them. If they were to be filled in they would just be a mirror image of the info we already have, so better to leave them off to keep the presentation more concise and readable.

(3) What difference in resistance is expected when moving from the low to the high level of x1?

When moving from low to high level of x1 the difference in resistance is calculated by (1 * 25.76) - (0 * 25.76) = 25.76. When moving from low level to high level for x1 we can expect 25.76 resistance according to our linear model. This is because in our linear model the coefficient of x1 is 25.76.

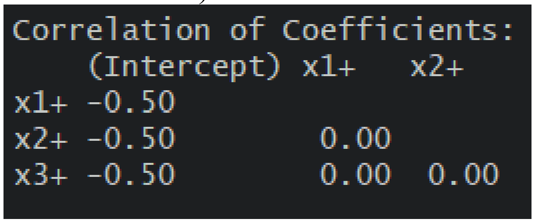
(4) Refit the model without x4 and examine the regression coefficients and standard errors? What stayed the the same as the original fit and what changed?

wafer_nox4.lm <- lm(resist~x1+x2+x3, data = wafer)
summary(wafer_nox4.lm, correlation = TRUE)

x1, x2, x3 coefficients all stayed the same, however the intercept was lower by 7 units.

(5) Explain how the change in the regression coefficients is related to the correlation matrix of X.

summary(wafer_nox4.lm, correlation = TRUE)



```
Correlation of Coefficients:
    (Intercept) x1+   x2+
x1+ -0.50
x2+ -0.50        0.00
x3+ -0.50        0.00  0.00
```

Since there is 0 correlation between the any of the predictor variables we wouldn't expect there to be any change in the coefficients (other than the intercept) by removing x4 from the linear model. If the data set had not been binary, and we observed values for the correlation of coefficients then we would expect to see a change in the coefficients, since those variables correlate with