Stat Methods I - Homework 5

Justin Reising

November 10, 2018

7.8) Refer to Commercial Properties. Test whether both X_2 and X_3 can be dropped from the regression model given that X_1 and X_4 are retained. Use $\alpha = 0.1$. State the alternatives, decision rule, and colnclusion. What is the P-value of the test?

Table 1: Commercial Properties (Header)

Rental (Y)	Age (X_1)	Operating (X_2)	Vacancy (X_3)	Total sq. ft (X_4)
13.5	1	5.02	0.14	123000
12.0	14	8.19	0.27	104079
10.5	16	3.00	0.00	39998
15.0	4	10.70	0.05	57112
14.0	11	8.97	0.07	60000
10.5	15	9.45	0.24	101385

$$H_0 : \beta_2 = \beta_3 = 0 \quad H_1 : \beta_2 \text{ or } \beta_3 \neq 0$$

```
alpha<-0.1
p<- 2
n<-length(Age)
pf<- ncol(commercial)-1
comm.reg.full<-lm(Rental~ Age +Operating+Vacancy+Total)
comm.reg.red<-lm(Rental~Age+Total)
comm.reg.full.anova<-anova(comm.reg.full)
comm.reg.red.anova<-anova(comm.reg.red)
SSE_f<- comm.reg.full.anova$`Sum Sq`[5]
SSE_r<- comm.reg.red.anova$`Sum Sq`[3]
MSE_f<- comm.reg.full.anova$`Mean Sq`[5]
F_star<- ((SSE_r-SSE_f)/p)/MSE_f
F_star > qf(1-alpha,p,n-pf)
```

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

```
pf(F_star, p, n-pf, lower=F)
```

```
## [1] 0.00006585484
```

Since $F^* > F_{2,77}$, we reject the null hypothesis and conclude that the variables both X_2 and X_3 cannot be dropped from the regression model given that X_1 and X_4 are retained. The associated P-value for the test is 0.00006585484, which coincides with the critical value above.

7.20) A speaker stated in a workshop on applied regression analysis: "In business and the social sciences, some degree of multicolinearity in survey data is practically inevitable." Does this statement apply equally to experimental data?

This statement is true in all data sets. If by "some degree", the speaker means "not zero", then one would most definitely expect to see some degree of correlation between variables. If you really think about it, it is easy to ascertain that the probability two random variables have a correlation coefficient of 0 is quite low. In an experimental setting, scientists are interested in determining "causes" by comparing experimental and control groups. As experiments increase in complexity, as does the difficulty in controlling for confounding variables (an example of intercolinearity in and of itself). This is due to the fact that many real-world phenomena can be correlated when they are seemingly unassociated with one another like in the example below.

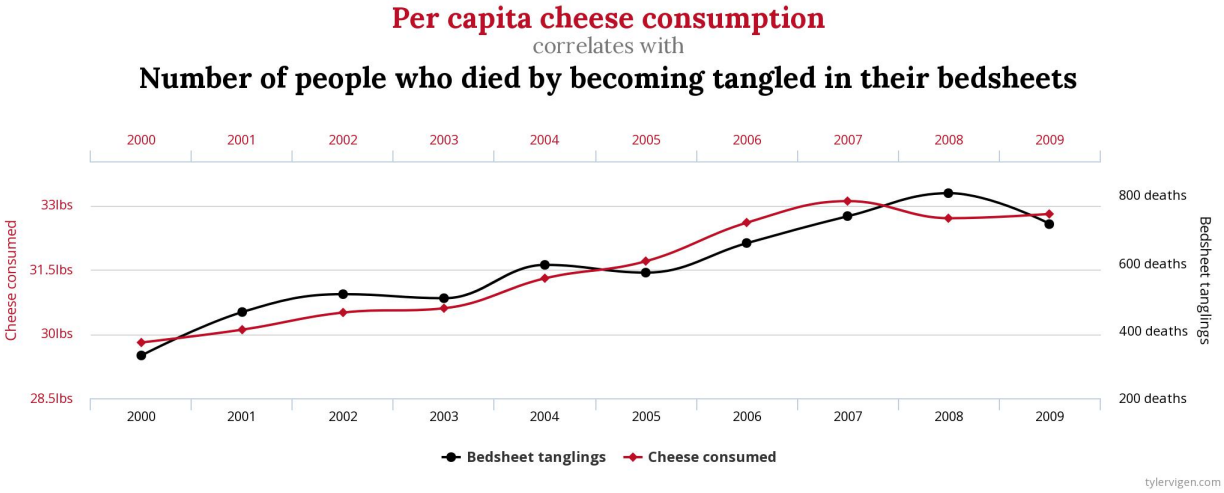


image:

7.22) The progress report of a research analyst to the supervisor stated: "All the estimated regression coefficients in our model with three predictor variables to predict sales are statistically significant. Our new preliminary model with seven predictor variables, which includes the three variables of our smaller model, is less satisfactory because only two of the seven regression coefficients are statistically significant. Yet in some initial trials the expanded model is giving more precise sales predictions than the smaller model. The reasons fOr this anomaly are now being investigated." Comment.

This would be evidence of multicolinearity between the added variables and the original three variables. It is likely that one or more of the new variables added are functions of one or more of the previous variables in the original model. It could also be attributed to there being significant outliers in the new new variables added to the model. This could maintain or increase precision of the predictions, however, the variation for the regression coefficients increases and it becomes difficult to discern which variables are contributing to the predicted value, which would be the case in the expanded model resulting in only two statistically significant regression coefficients.

7.24) Refer to Brand Preference.

(a) Fit first-order simple linear regression model (2.1) for relating brand liking (Y) to moisture content (X_1). State the fitted regression function.

$$Y_i = \beta_0 + \beta_1 X_1 + \epsilon_i \quad (2.1)$$

```
brand.reg.x1<-lm(Liking~Moisture)
brand.reg.x1$coefficients
```

```
## (Intercept)    Moisture
##      50.775       4.425
```

$$\hat{Y} = 50.775 + 4.425 X_1$$

(b) Compare the estimated regression coefficient for moisture content obtained in part (a) with the corresponding coefficient obtained in Problem 6.5b. What do you find?

```
brand.reg<- lm(Liking~Moisture + Sweetness)
brand.reg$coefficients
```

```
## (Intercept)    Moisture   Sweetness
##      37.650       4.425       4.375
```

$$\hat{Y} = 37.65 + 4.425 X_1 + 4.375 X_2$$

The regression coefficient for β_1 is the same in both models although the intercept does change. This indicates the regression coefficients of X_1 and X_2 are not dependent of one another. Thus, each regression coefficient reflects an effect of the associated predictor variable on the response variable, given that the other predictor variables are included in the model.

(c) Does $SSR(X_1)$ equal $SSR(X_1|X_2)$? If not, is the difference substaintial?

Note $SSR(X_1|X_2) = SSE(X_2) - SSE(X_1, X_2)$.

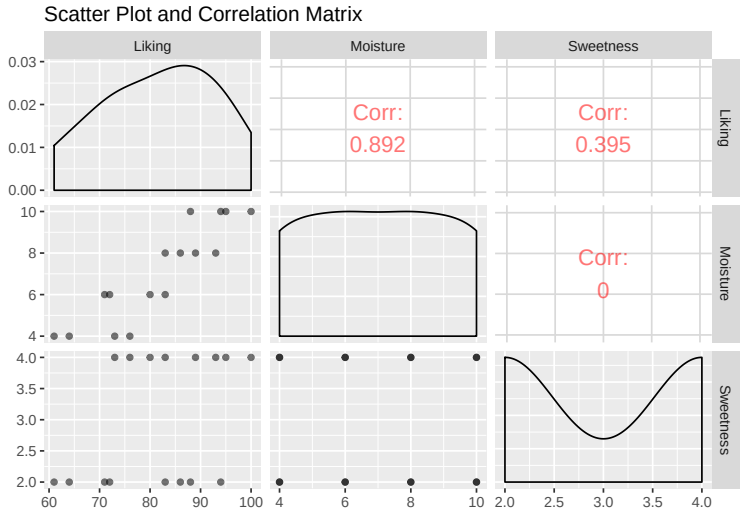
```
brand.reg.x2<-lm(Liking~Sweetness)
anova.brand.reg<-anova(brand.reg)
anova.brand.reg.x1<-anova(brand.reg.x1)
anova.brand.reg.x2<-anova(brand.reg.x2)
ssr_x1<-anova.brand.reg.x1$`Sum Sq`[1]
ssr_x2<-anova.brand.reg.x2$`Sum Sq`[1]
sse_x1<-anova.brand.reg.x1$`Sum Sq`[2]
sse_x2<-anova.brand.reg.x2$`Sum Sq`[2]
sse_x1x2<-anova.brand.reg$`Sum Sq`[3]
ssr_x1x2<-sse_x2 - sse_x1x2
ssr_x1 - ssr_x1x2
```

```
## [1] -0.0000000000004547474
```

They are practically equal.

(d) Refer to the correlation matrix obtained in Problem 6.5a. What bearing does this have on your findings in parts (b) and (c)?

```
ggpairs(brand, aes(alpha = 0.4),
        upper = list(continuous = wrap("cor", size = 5, color = "red")))+
  ggtitle("Scatter Plot and Correlation Matrix")
```



Note, that the correlation between Moisture (X_1) and Sweetness (X_2) is 0. This coincides with the fact that we determined in parts (b) and (c) indicating that Moisture and Sweetness were not dependent of one another. So the extra sum of squares is practically zero.

7.28 b) For a multiple regression model with five X variables, what is the relevant extra sum of squares for testing whether or not $\beta_5 = 0$ and whether or not $\beta_2 = \beta_4 = 0$?

For $\beta_5 = 0$, we would consider $SSR(X_5|X_1, X_2, X_3, X_4)$

For $\beta_2 = \beta_4 = 0$, we would consider $SSR(X_2, X_4|X_1, X_3, X_5)$

7.31) The following regression model is being considered in a water resources study:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \beta_4 \sqrt{X_{i3}} + \epsilon_i$$

State the reduced models for testing whether or not:

1. $\beta_3 = \beta_4 = 0$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

2. $\beta_3 = 0$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_4 \sqrt{X_{i3}} + \epsilon_i$$

3. $\beta_1 = \beta_2 = 5$

$$Y_i = \beta_0 + 5X_{i1} + 5X_{i2} + \beta_3 X_{i1} X_{i2} + \beta_4 \sqrt{X_{i3}} + \epsilon_i$$

4. $\beta_4 = 7$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + 7\sqrt{X_{i3}} + \epsilon_i$$

8.3) A junior investment anlayst used a polynomial regression model of relatively high order in a research seminar on municipal bonds and obtained a R^2 of 0.991 in the regression of net interest yeild of bond (Y) on industrial diversity index on municipality (X) for seven bond issues. A classmate, unimpressed, said, "You overfitted. Your curve follows the random effects in the data."

(a) Comment on the criticism.

For any polynomial regression model, it important to recognize that the size of the data set matters. In this example, the number of observations is extremely small with only 7 bond issues. Fitting a polynomial regression model (of "relatively" high order) with such a small dataset will tend to describe the variation rather than overall trend. Also, without seeing the plot, it's hard to say how bad a higher order polynomial regression model may be compared to a first or second order model. Achieving high R^2 that approaches 100% of explained variance is not the only goal of regression models; it must also be able to appropriately handle new observations in the attempt to predict the response with decent accuracy.

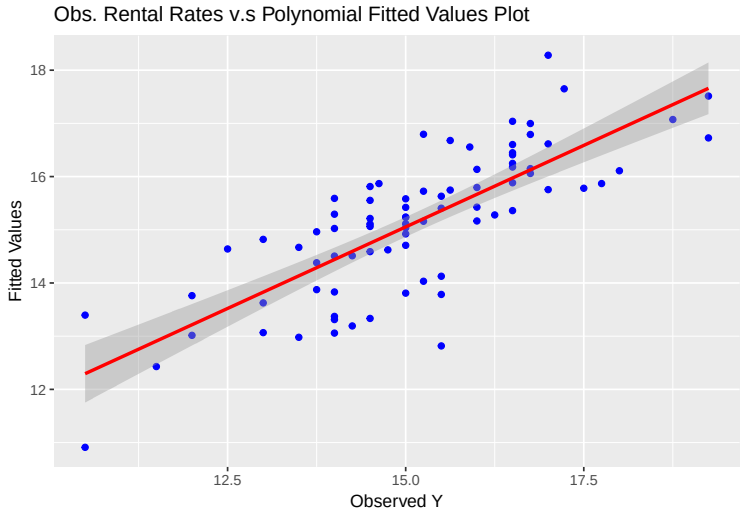
(b) Might R_a^2 defined in (6.42) be more appropriate than R^2 as a descriptive measure here?

This would be more appropriate as it would take into account the number of terms included as predictors. In this case, the higher order terms. An even better descriptive measure would be the predicted R^2.

8.8) Refer to Commercial Properties. The vacancy rate predictor (X_3) does not appear to be needed when property age (X_1), operating expenses (X_2), and total square footage (X_4) are included in the model as predictors of rental rates (Y).

(a) The age of the property (X_1) appears to exhibit some curvature when plotted against the rental rates (Y). Fit a polynomial regression model with centered property age (x_1), square of centered property age (x_1^2), operating expenses (X_2), and total square footage (X_4). Plot the Y observations against the fitted values. Does the response function provide a good fit?

```
cAge<-Age-mean(Age)
sqAge<-cAge**2
comm.reg.centered<- lm(Rental~cAge+sqAge+Operating+Total)
comm_poly_df<-data.frame(Rental,comm.reg.centered$fitted.values)
```



Since there are significant deviations from the regression line near the mean as oppose to the upper and lower bounds. To me, this indicates that model is not a good fit.

(b) Calculate R_a^2. What information does this measure provide?

```
comm.reg.cen.sum<-summary(comm.reg.centered)
comm.reg.cen.sum$adj.r.squared
```

```
## [1] 0.5926885
```

This measure informs us that approximately 59.27% of the variance is explained after taking into account the number of terms in the regression model. This is not particularly high, and as we can see in the plot, a lot of the points are falling outside of the bands around the regression line, especially around the center of the plot where the mean is.

(c) Test whether or not the square of centered property age (x_1^2) can be dropped from the model. Use $\alpha = 0.05$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

$$H_0 : \beta_2 = 0 \quad H_1 : \beta_2 \neq 0$$

Note, alternatively, we are testing $F^* = \frac{SSR(x_2^2|x_1,X_2,X_4)}{MSE}$

```
aplha<-.05
n<- length(cAge)
p<- 1
anova.comm.reg.centered<-anova(comm.reg.centered)
F_star<- anova.comm.reg.centered$`Sum Sq`[2]/anova.comm.reg.centered$`Mean Sq`[5]
F_star > qf(1-alpha,p,n-p)
```

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

```
pf(F_star, p, n-p, lower=F)
```

```
## [1] 0.9514976
```

Since $F^* \leq F_{(1,79)}$, we fail to reject the null hypothesis and conclude that the square of centered property age (x_1^2) can be dropped from the model. Also, the P-value is 0.996284.

8.11) Refer to Brand Preferences.

(a) Fit regression model (8.22)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1} X_{i2} + \beta_5 X_{i1} X_{i3} + \epsilon_i \quad (8.22)$$

```
brand.reg.int<- lm(Liking~Moisture+Sweetness+Moisture:Sweetness)
brand.reg.int$coefficients
```

```
##        (Intercept)           Moisture          Sweetness
##             27.150              5.925              7.875
## Moisture:Sweetness
##             -0.500
```

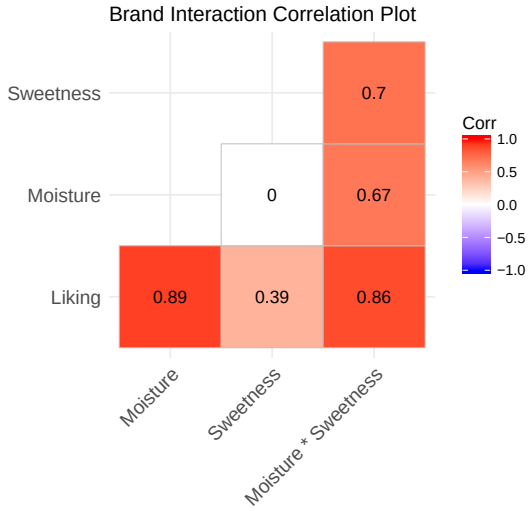
$$\hat{Y} = 27.15 + 5.925 X_1 + 7.875 X_2 - 0.5 X_1 X_2$$

(b) Test whether or not the interaction term can be dropped from the model. Use $\alpha = 0.05$. State the alternatives, decision rule, and conclusion.

Note, $F^* = \frac{SSE_{red} - SSE_{full}}{MSE_{full}}$

$$H_0 : \beta_3 = 0 \quad H_1 : \beta_3 \neq 0$$

Note, the high correlation between the predictor variables and interaction terms. We will center the variables in the regression model.




```
cMoisture<- Moisture-mean(Moisture)
cSweetness<- Sweetness-mean(Sweetness)
cM_S<- cMoisture*cSweetness
brand.int.cen.reg<-lm(Liking~cMoisture+cSweetness+cM_S)
anova.brand.int.reg<-anova(brand.int.cen.reg)
brand.cen.reg<- lm(Liking~cMoisture+cSweetness)
anova.brand.cen.reg<-anova(brand.cen.reg)
alpha<-0.05
n<-length(Liking)
p<-1
sse_r<-anova.brand.cen.reg$`Sum Sq`[3]
```

```r
sse_f<-anova.brand.int.reg$`Sum Sq`[4]
mse_f<-anova.brand.int.reg$`Mean Sq`[4]
F_star<-(sse_r-sse_f)/mse_f
F_star>qf(1-alpha,p,n-p)
```

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

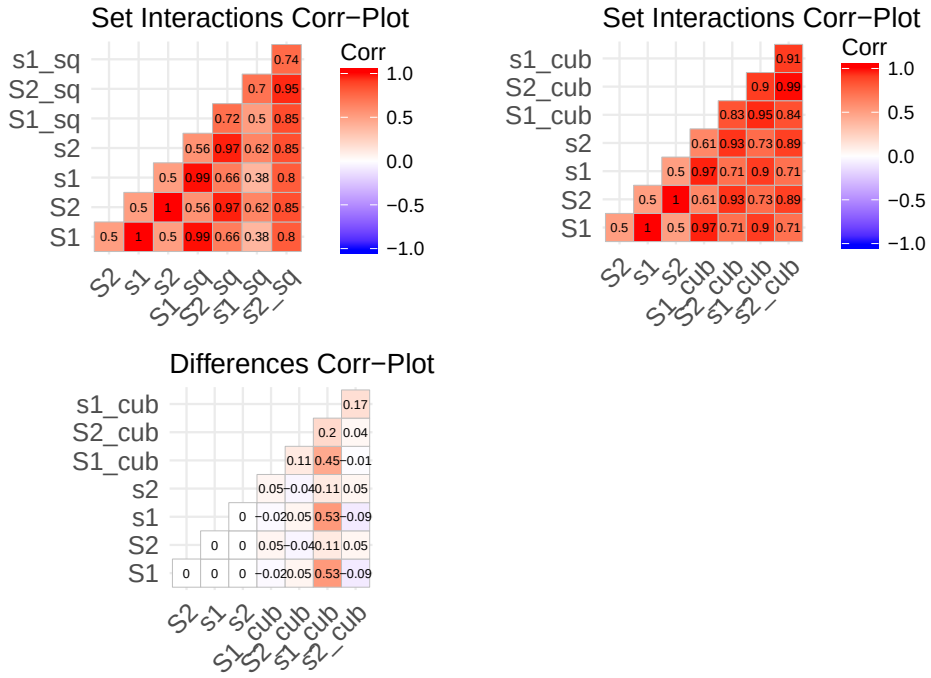
Since $F^* \leq F_{(1,15)}$, we fail to reject the null hypothesis and conclude that the interaction term can be dropped from the model.

8.29) Consider the second-order regression model with one predictor variable in (8.2) and the following two sets of X values.

$$Y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \epsilon_i \qquad (8.2)$$

Set 1:	1.0	1.5	1.1	1.3	1.9	0.8	1.2	1.4
Set 2:	12	1	123	17	415	71	283	38

For each set, calculate the coefficient correlation between X and X^2, then between x and x^2. Also calculate the coefficients of correlation between X and X^3 and between x and x^3. What generalizations are suggested by your results?



One thing to note is that both sets contain strictly positive values, which gives us strictly positive correlations. Here, we can see that for both sets, $\rho_{X,x} = 1$, while between sets, $\rho_{x_1,x_2} = 0.5$. Also, these remain unchanged between the squared copmarisons and the cubed comparisons. In the differences plot, we can see that increasing the order of the generally increases the correlation and quite significantly for Set 1 as seen in column 6. Also note that the variance for Set one is very low (0.11) and the correlation between the centered sets 1 and 2 and other terms quickly approaches 1 by increasing the order.

9.4) In forward stepwise regression, what advantage is there in using relatively small α-to-enter value for adding variables? What andvatage is there for using a larger α-to-enter value?

The choice of a relatively small α-to-enter value will help suppress the "picking procedure" in picking too many predictor variables, although too small of an α-to-enter value can be too conservative and cause σ^2 to be severely overestimated. The advantage for a larger α-to-enter value is that it could include more predictor variables that may end up having more of an effect and allow previously entered predictor variables to drop out.

9.10) Job Proficiency. A personnel officer in a government agency administered four newly developed aptitude tests to each of 25 applicants for entry-level clerical ositions in the agency. For the purpose of the study, all 25 applicants were accepted for positions irrespective of their test scores. After a probationary period, each applicant was rated for proficiency on the job. The scores on the four tests (X_1, X_2, X_3, X_4) and the job proficiency score (Y) for the 25 employees were as follows:

Table 2: Job Proficiency (Header)

Score	X1	X2	X3	X4
88	86	110	100	87
80	62	97	99	100
96	110	107	103	103
76	101	117	93	95
80	100	101	95	88
73	78	85	95	84

(c) Fit the multiple regression function containing all four predictor variables as first-order terms. Does it appear that all predictor variables should be retained?

```
job.reg<- lm(Score~X1+X2+X3+X4)
coef(summary(job.reg))
```

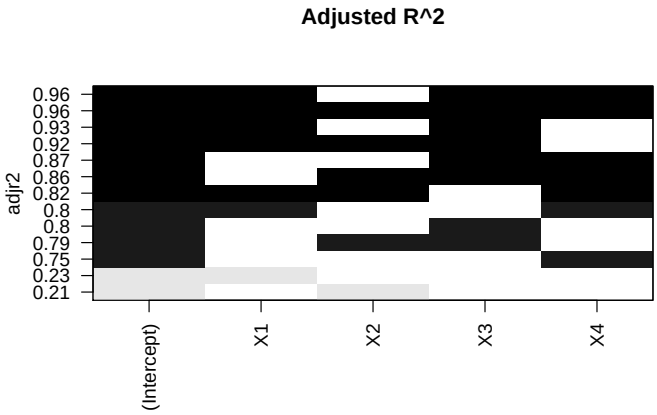
```
##                  Estimate Std. Error     t value            Pr(>|t|)
## (Intercept) -124.38182058 9.94106316 -12.5119234 0.00000000006475642
## X1             0.29572537 0.04397141   6.7254011 0.00000152371029639
## X2             0.04828772 0.05661745   0.8528769 0.40382624459919270
## X3             1.30601100 0.16409129   7.9590514 0.00000012617258157
## X4             0.51981909 0.13194285   3.9397290 0.00080997138219049
```

Note the p-value for X_2 is relatively high and is indicative for consideration that the predictor variable be removed.

9.11) Refer to Job Proficiency.

(a) Using only first-order terms for the predictor variables in the pool of potential X variables, find the four best subset regression model according to the $R^2_{a,p}$ criterion.

```
# From the "leaps"" package
job.reg.subsets<-regsubsets(Score~X1+X2+X3+X4,data=job,nbest=4)
plot(job.reg.subsets, scale = "adjr2", main = "Adjusted R^2")
```



This plot was constructed using the best subsets algorithm in the "leaps" package in R. The way to read this plot in our model selection is there is a black block where the predictor variable being selected for the model with the corresponding adjusted R squared on the y-axis. So the 4 best models are as follows:

1. For $R^2_a = 0.96$, the model is $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4$
2. For $R^2_a = 0.96$, the model is $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + +\beta_3 X_3 + \beta_4 X_4$
3. For $R^2_a = 0.93$, the model is $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_3 X_3$
4. For $R^2_a = 0.92$, the model is $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

(b) Since there is relatively little difference in $R^2_{a,p}$ for the four best subset models, what other criteria would you use to help in the selection of the best model? Discuss.

As a rule of thumb, we want to consider the model with the least number of predictors that maintain high adjusted R squared. As these are are quite similar in their adjusted R squared values, we may want to consider the Mallows' Cp criterion in addition to the adjusted R squared to investigate which model demonstrates the least amount of bias.

9.18) Refer to Job Proficiency.

(a) Using forward stepwise regression, find the best subset of predictor variables to predict job proficiency. Use α limits of 0.05 and 0.1 for adding or deleting a variable respectively.

```
null<-lm(Score~1)
full<-job.reg
step(null,scope=list(lower=null,upper=full),direction="forward")
```

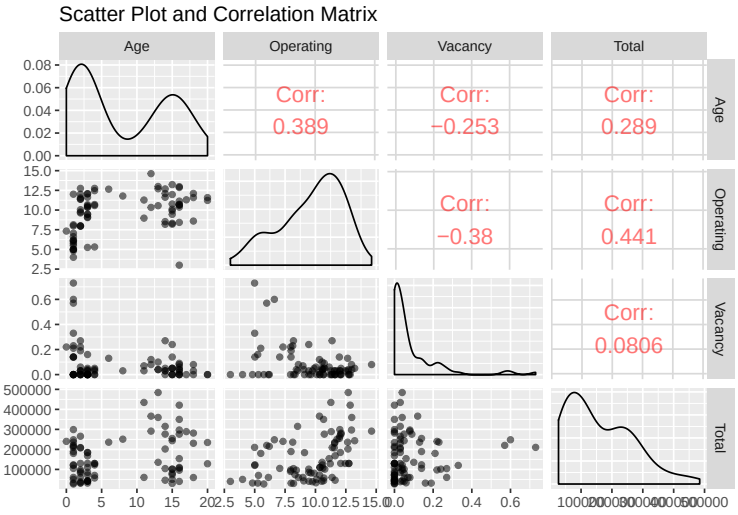
```
## Start:  AIC=149.3
## Score ~ 1
##
##        Df Sum of Sq    RSS    AIC
## + X3    1    7286.0 1768.0 110.47
## + X4    1    6843.3 2210.7 116.06
## + X1    1    2395.9 6658.1 143.62
## + X2    1    2236.5 6817.5 144.21
## <none>               9054.0 149.30
##
## Step:  AIC=110.47
## Score ~ X3
##
##        Df Sum of Sq     RSS     AIC
## + X1    1   1161.37  606.66  85.727
## + X4    1    656.71 1111.31 100.861
## <none>               1768.02 110.469
## + X2    1     12.21 1755.81 112.295
##
## Step:  AIC=85.73
## Score ~ X3 + X1
##
##        Df Sum of Sq    RSS    AIC
## + X4    1   258.460 348.20 73.847
## <none>              606.66 85.727
## + X2    1     9.937 596.72 87.314
##
## Step:  AIC=73.85
## Score ~ X3 + X1 + X4
##
##        Df Sum of Sq    RSS    AIC
## <none>              348.20 73.847
## + X2    1     12.22 335.98 74.954
```

```
##
## Call:
## lm(formula = Score ~ X3 + X1 + X4)
##
## Coefficients:
## (Intercept)           X3           X1           X4
##   -124.2000       1.3570       0.2963       0.5174
```

As it is seems to be a general consensus among those in the field to avoid forward stepwise regression in practice, the 'step' function above uses AIC rather than p-values. However, for a single variable at a time, AIC does correspond to using a p-value of 0.15. Using this function still yeilds the same selction as subsetting and choosing the model with the highest adjusted R squared. Also, I'm stubborn and do not want to use SAS since I'm composing in R Markdown.

10.18) Refer to Commercial Properties.

(a) What do the scatter plot matrix and the correlation matrix show about pairwise linear associations among predictor variables?



None of the predictor variables appear to be significantly correlated, although as mentioned before, the Age variables has a bi-modal distribution and Vacancy is severely skewed due to the amount of zeros.

(b) Obtain the four variance inflation factors. Do they indicate that a serious multicolinearity problem exists here?

```
##       Age Operating   Vacancy     Total
##  1.240348  1.648225  1.323552  1.412722
```

Note that each of the VIF values are less than 10 and indicate that we do not have a problem with multicoliearity.