Stat 615/415 (Regression) HW #10 – Solutions

Variable Selection and Model Building (chap. 9)

1. (**9.1**) A speaker stated: "In well-designed experiments involving quantitative explanatory variables, a procedure for reducing the number of explanatory variables after the data are obtained is not necessary." Do you agree? Discuss.

   SOLUTION. *If an experiment is designed so that all X-variables are uncorrelated, then reducing their number is not urgent, because there is no multicollinearity. On the other hand, including non-significant variables into the model negatively affects its prediction power. So, proper variable selection will still improve the model.*

2. (**9.5**) In forward stepwise regression, what advantage is there in using a relatively small α-to-enter value for adding variables? What advantage is there in using a larger α-to-enter value?

   SOLUTION. *Smaller α-to-enter is harder to overcome. Therefore, marginally significant variables will not enter the model. It will contain fewer variables overall, and each slope b_j will have a lower variance. Larger α-to-enter is easier to overcome. Therefore, marginally significant variables will enter the model. This will reduce the bias of estimated slopes b_j and consequently, the bias of predicted values $\hat{Y}$, at the expense of a higher variance.*

3. (**Stat-615 only**) Two regression models are compared. The full model uses p independent variables $X_1, \ldots, X_p$. The reduced model uses only the first q variables $X_1, \ldots, X_q$, where $q < p$. We can choose the model with the higher adjusted R^2, or we can test significance of added variables $X_{q+1}, \ldots, X_p$ with a partial F-test. These methods are related!

   Show that the full model has a higher adjusted R^2 if and only if the F-statistic for testing $H_0: \beta_{q+1} = \ldots = \beta_p = 0$ exceeds 1.

   Hint: Write explicit formulae for R^2_{adj} and F and express both of them in terms of $\frac{SSErr^{(Red)}}{SSErr^{(Full)}}$, the ratio of error sums of squares from the two models.

   SOLUTION. $R^2_{adj}(Full) = 1 - \dfrac{SSErr^{(Full)}/(n-p-1)}{SSTot/(n-1)}$ and $R^2_{adj}(Red) = 1 - \dfrac{SSErr^{(Red)}/(n-q-1)}{SSTot/(n-1)}$.

   Thus, $R^2_{adj}(Full) > R^2_{adj}(Red)$ if and only if $\dfrac{SSErr^{(Red)}}{n-q-1} < \dfrac{SSErr^{(Full)}}{n-p-1}$, or $\dfrac{SSErr^{(Red)}}{SSErr^{(Full)}} < \dfrac{n-p-1}{n-q-1}$.

   On the other hand,

   $$F = \frac{SS_{extra}/(p-q)}{SSErr^{(Full)}/(n-p-1)} = \frac{(SSErr^{(Red)} - SSErr^{(Full)})/(p-q)}{SSErr^{(Full)}/(n-p-1)} = \left(\frac{SSErr^{(Red)}}{SSErr^{(Full)}} - 1\right)\frac{n-p-1}{p-q},$$

   and $F > 1$ if and only if $\dfrac{SSErr^{(Red)}}{SSErr^{(Full)}} - 1 > \dfrac{p-q}{n-p-1}$, or $\dfrac{SSErr^{(Red)}}{SSErr^{(Full)}} > 1 + \dfrac{p-q}{n-p-1} = \dfrac{n-q-1}{n-p-1}$.

   We obtained that

   $$R^2_{adj}(Full) > R^2_{adj}(Red) \text{ if and only if } \frac{SSErr^{(Red)}}{SSErr^{(Full)}} < \frac{n-p-1}{n-q-1} \text{ if and only if } F > 1.$$

   Hence,

   $$R^2_{adj}(Full) > R^2_{adj}(Red) \text{ if and only if } F > 1.$$

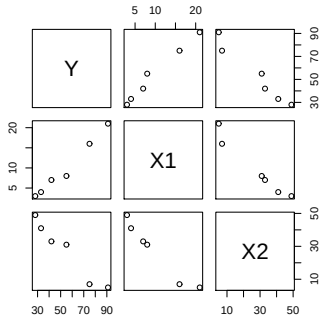
4. (Continuing 6.27 from an earlier homework) In a small-scale regression study, the following data

were obtained,

Y	$X1$	$X2$
42.0	7.0	33.0
33.0	4.0	41.0
75.0	16.0	7.0
28.0	3.0	49.0
91.0	21.0	5.0
55.0	8.0	31.0

Select the best regression equation using different model selection methods.

SOLUTION. *Conclusions are based on the R code and output below.*



1. The scatterplot matrix shows strong correlations between response Y and each predictor (so, each X_j may be significant for the prediction of Y), but also, a strong correlation between X_1 and X_2 (so, it may be unnecessary to include both predictors X_j into the regression model).

2. Exhaustive search shows that if we use only one of the two variables, then it should be X_1. This model with 1 predictor has a higher R^2_{adj}, a lower BIC, and Mallows C_p that is closer to p. Hence, exhaustive search chooses the model $Y = \beta_0 + \beta_1 X_1 + \varepsilon$.

3. Sequential search by AIC returns the same model $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, regardless of the search direction. Indeed, forward and backward searches can only go along the same path, through models with both X_1 and X_2 and with only X_1.

4. Sequential search by partial F-tests returns the same model. ANOVA also shows that X_2 is not significant in presence of X_1, with a p-value of 0.64. Thus, X_2 will not enter the model in a forward search when X_1 is already included, and X_2 will be removed from the full model in a backward search.

Our model selection results in the regression equation $\hat{Y} = 20.236 + 3.434 X_1$.

5. (**9.10–9.11, 9.18, 9.21–9.22**) A personnel officer in a governmental agency administered four newly developed aptitude tests to each of 25 applicants for entry-level clerical positions in the agency. For 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, and scores of the four tests (X_1, X_2, X_3, X_4) and the job proficiency score (Y) were recorded.

   The resulting **Job Proficiency** data set is available on our Blackboard in "Data sets" and on the next page of this homework assignment.

   (a) Obtain the scatter plot matrix of these data. What do the scatter plots suggest about the nature of the functional relationship between the response variable and each of the predictor variables? Do you notice any serious multicollinearity problems?

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

   (c) Using only first-order terms for the predictor variables in the pool of potential X-variables, find the best regression models according to different criteria - adjusted R^2, C_p, and BIC.

   (d) Using forward stepwise selection, find the best subset of predictor variables to predict job proficiency. Use the α-to-enter limit of 0.05.

   (e) Repeat the previous question using the backward elimination method and the α-to-remove limit of 0.10.

(f) To assess and compare internally the predictive ability of our models, split the data into training and testing subsets and estimate the mean squared prediction error MSPE for all regression models identified in (b–e).

(g) To assess and compare externally the validity of our models, 25 a1dditional applicants for entry level clerical positions were similarly tested and hired. Their data are below, in the table on the right. Use these data as the testing set and estimate MSPE for all regression models identified in (b–e).

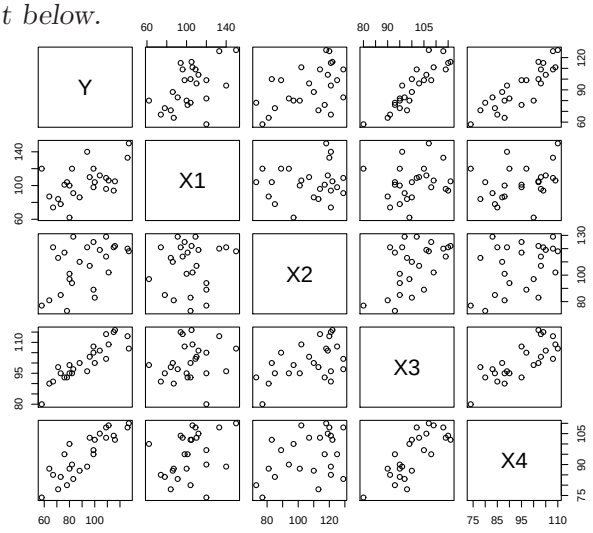
Original Job Proficiency data

Y	X_1	X_2	X_3	X_4
88.0	86.0	110.0	100.0	87.0
80.0	62.0	97.0	99.0	100.0
96.0	110.0	107.0	103.0	103.0
76.0	101.0	117.0	93.0	95.0
80.0	100.0	101.0	95.0	88.0
73.0	78.0	85.0	95.0	84.0
58.0	120.0	77.0	80.0	74.0
116.0	105.0	122.0	116.0	102.0
104.0	112.0	119.0	106.0	105.0
99.0	120.0	89.0	105.0	97.0
64.0	87.0	81.0	90.0	88.0
126.0	133.0	120.0	113.0	108.0
94.0	140.0	121.0	96.0	89.0
71.0	84.0	113.0	98.0	78.0
111.0	106.0	102.0	109.0	109.0
109.0	109.0	129.0	102.0	108.0
100.0	104.0	83.0	100.0	102.0
127.0	150.0	118.0	107.0	110.0
99.0	98.0	125.0	108.0	95.0
82.0	120.0	94.0	95.0	90.0
67.0	74.0	121.0	91.0	85.0
109.0	96.0	114.0	114.0	103.0
78.0	104.0	73.0	93.0	80.0
115.0	94.0	121.0	115.0	104.0
83.0	91.0	129.0	97.0	83.0

Additional Data for (f – g)

Y	X_1	X_2	X_3	X_4
58.0	65.0	109.0	88.0	84.0
92.0	85.0	90.0	104.0	98.0
71.0	93.0	73.0	91.0	82.0
77.0	95.0	57.0	95.0	85.0
92.0	102.0	139.0	101.0	92.0
66.0	63.0	101.0	93.0	84.0
61.0	81.0	129.0	88.0	76.0
57.0	111.0	102.0	83.0	72.0
66.0	67.0	98.0	98.0	84.0
75.0	91.0	111.0	96.0	84.0
98.0	128.0	99.0	98.0	89.0
100.0	116.0	103.0	103.0	103.0
67.0	105.0	102.0	88.0	83.0
111.0	99.0	132.0	109.0	105.0
97.0	93.0	95.0	106.0	98.0
99.0	99.0	113.0	104.0	95.0
74.0	110.0	114.0	91.0	78.0
117.0	128.0	134.0	108.0	98.0
92.0	99.0	110.0	96.0	97.0
95.0	111.0	113.0	101.0	91.0
104.0	109.0	120.0	104.0	106.0
100.0	78.0	125.0	115.0	102.0
95.0	115.0	119.0	102.0	94.0
81.0	129.0	70.0	94.0	95.0
109.0	136.0	104.0	106.0	104.0

SOLUTION. *Conclusions are based on the R code and output below.*



(a) *The scatterplot matrix shows that variables X_3 and X_4 are in a linear relationship with response Y, and they are also strongly correlated with each other. Predictors X_1 and X_2 seem weakly correlated with Y.*

(b) *The estimated regression equation is* $\hat{Y} = -124.38 + 0.30X_1 + 0.05X_2 + 1.31X_3 + 0.52X_4$. *T-tests show that variable X_2 is not significant in presence of X_1, X_3, X_4. The p-value for $H_0 : \beta_2 = 0$ is 0.4038. Thus, it can be dropped from the model.*

(c) *Exhaustive search shows that the model with $p = 3$ variables has the highest $R^2_{adj} = 0.956$, the lowest BIC=−68.6, and Mallows $C_p = 3.7$ that is closest to p.*

Among the models with $p = 3$ variables, the best model includes X_1, X_3, X_4, according to any of these criteria, as well as the lowest SSErr and the highest SSReg and R^2.

(d) *Forward selection returns the model with 3 variables - X_1, X_3, X_4. The estimated regression equation is $\hat{Y} = -124.2 + 0.30X_1 + 1.36X_3 + 0.52X_4$. As we see, removing X_2 practically does not change the slopes for other variables.*

(e) *Backward elimination gives the same result.*

(f) *We identified two regression models - the full model in (a) and the model with X_1, X_3, X_4 in (c-e).*

I split the $n = 25$ sampling units into a training subsample of size 15 and a training subsample of size 10.

The model with 3 predictors returns an estimated MSPE = 23.42.

The model with all 4 predictors returns an estimated MSPE = 25.60.

Hence, the reduced model with 3 predictors has a higher prediction power (estimated internally), and probably, underestimated. This is based on random subsampling, so your results may be different.

(g) *Using the new 25 applicants as the testing set, we estimate MSPE externally.*

The model with 3 predictors returns an estimated MSPE = 15.71.

The model with all 4 predictors returns an estimated MSPE = 13.96.

Hence, the full model with all 4 predictors has a higher prediction power (estimated internally).

There is no randomness involved here, and you should get the same result.

R Code and Output for Problem #4

```
 # Enter the data.
> Y = c(42,33,75,28,91,55); X1 = c(7,4,16,3,21,8); X2 = c(33,41,7,49,5,31);
> A = data.frame(Y,X1,X2); attach(A);
> A
   Y X1 X2
1 42  7 33
2 33  4 41
3 75 16  7
4 28  3 49
5 91 21  5
6 55  8 31

 # Scatterplot matrix
> par(mfrow=c(3,3))
> plot(A)

 # Exhaustive search
> install.packages("leaps")
> library(leaps)
> exhaustive.search = regsubsets( Y ~ X1 + X2, data=A )
> summary(exhaustive.search)
         X1  X2
1  ( 1 ) "*" " "
2  ( 1 ) "*" "*"
> exhaustive.search = regsubsets( Y ~ X1 + X2, data=A )
> summary(exhaustive.search)$adjr2
[1] 0.9724995 0.9663230
```

```
> summary(exhaustive.search)$cp
[1] 1.266385 3.000000
> summary(exhaustive.search)$bic
[1] -19.31664 -18.03531

> # Sequential search
> null = lm( Y ~ 1, data=A )
> full = lm( Y ~ X1 + X2, data=A )
> step( null, scope=list( lower=null, upper=full ), direction="forward" )
> step( full, scope=list( lower=null, upper=full ), direction="backward" )
> step( full, scope=list( lower=null, upper=full ), direction="both" )
(Intercept)           X1
     20.236        3.434

 # Sequential F-tests
> install.packages("SignifReg")
> library(SignifReg)
> full.model = Y ~ X1 + X2;
> SignifReg( scope = full.model, data=A, direction="forward" )
Coefficients:
(Intercept)           X1
     20.236        3.434

 # How significant is X2 in presence of X1
> anova(full)
          Df  Sum Sq Mean Sq  F value  Pr(>F)
X1         1 3004.41 3004.41 145.2027 0.00123 **
X2         1    5.51    5.51   0.2664 0.64140
```

R Code and Output for Problem #5

```
> setwd("C:\\Teach\\615 Regression\\Book data")
> A1 = read.table("CH09PR10.txt")
> A2 = read.table("CH09PR22.txt")
> A1$Y=V1; A1$X1=V2; A1$X2=V3; A1$X3=V4; A1$X4=V5;
> A1 = data.frame(Y,X1,X2,X3,X4)

 # Scatterplot matrix for (a)
> par(mfrow=c(5,5))
> plot(A1)

 # Full model for (b)
> full = lm( Y ~ ., data=A1 )
> summary(full)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -124.38182    9.94106 -12.512 6.48e-11 ***
X1             0.29573    0.04397   6.725 1.52e-06 ***
X2             0.04829    0.05662   0.853  0.40383
X3             1.30601    0.16409   7.959 1.26e-07 ***
X4             0.51982    0.13194   3.940  0.00081 ***

 # Exhaustive search for (c)
```

```
> exhaustive = regsubsets( Y ~ X1+X2+X3+X4, data=A1 )
> summary(exhaustive)
         X1  X2  X3  X4
1  ( 1 ) " " " " "*" " "
2  ( 1 ) "*" " " "*" " "
3  ( 1 ) "*" " " "*" "*"
4  ( 1 ) "*" "*" "*" "*"
> summary(exhaustive)$adjr2
[1] 0.7962344 0.9269043 0.9560482 0.9554702
> summary(exhaustive)$cp
[1] 84.246496 17.112978  3.727399  5.000000
> summary(exhaustive)$bic
[1] -34.39587 -57.91831 -68.57933 -66.25356

 # Forward selection with partial F-tests for (d)
> SignifReg( scope = Y ~ X1+X2+X3+X4, data=A1, direction="forward", alpha=0.05 )
(Intercept)           X3           X1           X4
  -124.2000       1.3570       0.2963       0.5174

 # Backward elimination with partial F-tests for (e)
> SignifReg( scope = Y ~ X1+X2+X3+X4, data=A1, direction="backward", alpha=0.10 )
(Intercept)           X1           X3           X4
  -124.2000       0.2963       1.3570       0.5174

 # Internal estimation of MSPE for (f)
> n = length(Y)
> testing = sample( n,10 )
> training = -testing;
>
> model3 = lm( Y ~ X1 + X3 + X4, data=A1, subset=training )
> Yhat3 = predict(model3, A1)
> MSPE3 = mean((Y[testing] - Yhat3[testing])^2)
> MSPE3
[1] 23.41918
>
> model4 = lm( Y ~ X1 + X2 + X3 + X4, data=A1, subset=training )
> Yhat4 = predict(model4, A1)
> MSPE4 = mean((Y[testing] - Yhat4[testing])^2)
> MSPE4
[1] 25.59962

 # External estimation of MSPE for (g)
> A2$Y=A2$V1; A2$X1=A2$V2; A2$X2=A2$V3; A2$X3=A2$V4; A2$X4=A2$V5;
> head(A2)
  V1  V2  V3  V4 V5  Y  X1  X2  X3 X4
1 58  65 109  88 84 58  65 109  88 84
2 92  85  90 104 98 92  85  90 104 98
3 71  93  73  91 82 71  93  73  91 82
4 77  95  57  95 85 77  95  57  95 85
5 92 102 139 101 92 92 102 139 101 92
6 66  63 101  93 84 66  63 101  93 84
>
```

```
> model3 = lm( Y ~ X1 + X3 + X4, data=A1 )
> Yhat3 = predict(model3, A2)
> MSPE3 = mean((A2$Y - Yhat3)^2)
>
> model4 = lm( Y ~ X1 + X2 + X3 + X4, data=A1 )
> Yhat4 = predict(model4, A2)
> MSPE4 = mean((A2$Y - Yhat4)^2)
> MSPE3
[1] 15.70972
> MSPE4
[1] 13.95808
```