MATH 2780 Chapter5 Worksheet

Summary

Steps in Model Building

Step 1: Fit all possible one-variable models of the form $E(y)=\beta_0+\beta_1 x_i$, $i=1,...,k$.

Perform the t-test H_0: $\beta_1=0$ vs. H_a: $\beta_1\neq 0$.

$t=\hat{\beta}_i / s\sqrt{W_{ii}}$, W_{ii} is the i^{th} diagonal element of $W=(X'X)^{-1}$.

Select the best one variable model (largest $|t|$ statistic). Call it x_1

Step 2: Fit all two variable models with remaining x's, $E(y)=\beta_0+\beta_1 x_1+\beta_2 x_i$, $i\neq 1$.

Perform the t-test H_0: $\beta_2=0$ vs. H_a: $\beta_2\neq 0$.

$t=\hat{\beta}_i / s\sqrt{W_{ii}}$, W_{ii} is the i^{th} diagonal element of $W=(X'X)^{-1}$.

Select the best two variable model (largest $|t|$ statistic). Call it x_2

Go back and check the t-value of $\hat{\beta}_1$ after $\hat{\beta}_2$ has been added to the model.

Step 3: Fit all three variable models with remaining x's, $E(y)=\beta_0+\beta_1 x_1+\beta_2 x_2 +\beta_3 x_i$, $i\neq 1,2$.

Perform the t-test H_0: $\beta_3=0$ vs. H_a: $\beta_3\neq 0$.

$t=\hat{\beta}_i / s\sqrt{W_{ii}}$, W_{ii} is the i^{th} diagonal element of $W=(X'X)^{-1}$.

Select the best two variable model (largest $|t|$ statistic). Call it x_2

Go back and check the t-values of $\hat{\beta}_1, \hat{\beta}_2$ after $\hat{\beta}_3$ has been added.

This procedure is continued until no further independent variables can be found that yield significant t-values (at the specified α level) in the presence of the variables already in the model.

General Form of the Multiple Regression Model:

$$E(y)=\beta_0+\beta_1 x_1+\beta_2 x_1^2+\beta_3 x_2+\beta_4 x_3+\beta_5 x_1 x_2+\beta_6 x_1 x_3+\beta_7 x_1^2 x_2+\beta_8 x_1^2 x_3$$

Coefficient and Residual Variance Estimation:

$Y=X\beta+E$ $\hat{\beta}=(X'X)^{-1}X'y$ $s^2=\dfrac{(y-X\hat{\beta})'(y-X\hat{\beta})}{n-k-1}$ $MSE=s^2,\ s=\sqrt{s^2}$	$Y=\begin{bmatrix} y_1\\ y_2\\ y_3\\ \vdots\\ y_n\end{bmatrix}, X=\begin{bmatrix}1 & x_{11} & x_{21} & \cdots & x_{kn}\\ 1 & x_{12} & x_{22} & \cdots & x_{k2}\\ 1 & x_{13} & x_{23} & \cdots & x_{k3}\\ \vdots & \vdots & \vdots & & \vdots\\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn}\end{bmatrix}, \beta=\begin{bmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \vdots\\ \beta_k\end{bmatrix}, E=\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ \varepsilon_3\\ \vdots\\ \varepsilon_n\end{bmatrix}$

Individual Coefficient Test: $t=\hat{\beta}_i / s_{\hat{\beta}_i}$, $s_{\hat{\beta}_i}=s\sqrt{W_{ii}}$, W_{ii} is the i^{th} diagonal of $W=(X'X)^{-1}$.

Two Tailed: H_0: $\beta_i=0$ $vs.$ H_a: $\beta_i\neq 0$ w/ RR $|t|>t_{\alpha/2,n-k-1}$ or $\alpha>p\text{-}value=2P(|t|>t_{\alpha,n-k-1})$.

Coefficient of determination R^2 and R^2_a. Want simple model with large R^2 and R^2_a close full model.

$$R^2=1-SSE/SS_{yy},\quad 0\le R^2\le 1,\ SSE=\sum(y_i-\hat{y}_i)^2,\ SS_{yy}=\sum(y_i-\bar{y})^2$$

$$R_a^2=1-[SSE/(n-k-1)]/[SS_{yy}/(n-1)]=1-[(n-1)/(n-k-1)](1-R^2),\quad R_a^2\le R^2$$

Mallow's C_p Statistic: Prefer a small value of C_p near p+1. Addresses the issue of overfitting.

$$C_p=SSE_p/MSE_k+2(p+1)-n$$

Predictive Sum of Squares Statistic: Desire a model with a small *PRESS*.

$$PRESS=\sum_{i=1}^{n}[y_i-\hat{y}_{(i)}]^2$$

***F*-Test for Comparing Nested Models**

Reduced Model: $E(y|x's)=\beta_0+\beta_1 x_1+\ldots+\beta_g x_g$

Complete Model: $E(y|x's)=\beta_0+\beta_1 x_1+\ldots+\beta_g x_g+\beta_{g+1}x_{g+1}+\ldots+\beta_k x_k$

H_0: $\beta_{g+1}=\ldots=\beta_k=0$ $vs.$ H_a: At least one tested $\beta_i\neq 0$.

$F=\dfrac{(SSE_R-SSE_C)/(k-g)}{SSE_C/(n-k-1)}$, Reject if $F>F_{\alpha,k-g,n-k-1}$ or $\alpha>p\text{-}value=P(F>F_{\alpha,k-g,n-k-1})$.

MATH 2780 Chapter5 Worksheet

Exercise 6.8: Clerical staff work hours.

In any production process in which one or more workers are engaged in a variety of tasks, the total time spent in production varies as a function of the size of the work pool and the level of output of the various activities. For example, in a large metropolitan department store, the number of hours worked (y) per day by the clerical staff may depend on the following variables:

x_1 = Number of pieces of mail processed (open, sort, etc.)
x_2 = Number of money orders and gifts certificates sold
x_3 = Number of window payments (customer charge accounts)
x_4 = Number of change order transactions processed
x_5 = Number of checks cashed
x_6 = Number of pieces of miscellaneous mail processed on an "as available" basis
x_7 = Number of bus tickets sold

The output counts for these activities on each of 52 working days were recorded, and the data saved in the CLERICAL file.

a. Conduct a stepwise regression analysis of the data using an available statistical software package.
R code output. Which variables should be included.

b. Interpret the β estimates in the resulting stepwise model.
R ode output. What do coefficients mean?

c. What are the dangers associated with drawing inferences from the stepwise model?
High probabilities of Type I and Type II errors
No higher-order terms or interactions
Nonsensical terms in the model
Important independent variables omitted that interact with other x's

Exercise 6.9: For this exercise, consider only the independent variables x_1, x_2, x_3, x_4 in an all-possible-regressions select procedure.

a. How many models for $E(y)$ are possible, if the model includes (i) one variable, (ii) two variables, (iii) three variables, and (iv) four variables?
R code output.

b. For each case in part a, use a statistical software package to find the maximum R^2, minimum MSE, minimum C_p, and minimum $PRESS$.
R code output.

c. Plot each of the quantities R^2, MSE, C_p, and PRESS in part b against p, the number of predictors in the subset model.
R code output.

d. Based on the plots in part c, which variables would you select for predicting total hours worked, y?
R code output.

MATH 2780 Chapter5 Worksheet

```r
# Worksheet #
install.packages("olsrr")

# read data
mydata <- read.delim("CLERICAL.txt",header = TRUE)

y  <- c(mydata[, 3]) #ln salary
x1 <- c(mydata[, 4]) #x1
x2 <- c(mydata[, 5]) #x2
x3 <- c(mydata[, 6]) #x3
x4 <- c(mydata[, 7]) #x4
x5 <- c(mydata[, 8]) #x5
x6 <- c(mydata[, 9]) #x6
x7 <- c(mydata[, 10]) #x7

df<- data.frame(cbind(x1,x2,x3,x4,x5,x6,x7))
names(df) <- c("x1","x2","x3","x4","x5","x6","x7")

cor(df)

# Ex 6.8
library(olsrr)
fullmodel <- lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, data = df)
k <- ols_step_all_possible(fullmodel,max_order = 7)
k
plot(k)



finalmodel <- lm(y ~ x2 + x3 + x4 + x5 + x6 + x7, data = df)
finalmodel$coefficients

# Ex 6.9
df2 <- data.frame(cbind(x1,x2,x3,x4))
names(df2) <- c("x1","x2","x3","x4")

choose(4,1)
choose(4,2)
choose(4,3)
choose(4,4)

model <- lm(y ~ x1 + x2 + x3 + x4, data = df)
k2 <- ols_step_all_possible(model,max_order = 4)
k2
plot(k2)
finalmodel <- lm(y ~ x2 + x3 + x4, data = df)
finalmodel$coefficients
```

	Index	N	Predictors	R-Square	Adj. R-Square	Mallow's Cp
5	**1**	1	x5	3.449436e-01	0.33184249	18.775229
6	2	1	x6	2.490136e-01	0.23399391	28.554154
3	3	1	x3	2.129999e-01	0.19725986	32.225330
7	4	1	x7	2.021349e-01	0.18617758	33.332886
2	5	1	x2	8.574310e-02	0.06745797	45.197648
4	6	1	x4	7.190738e-03	-0.01266545	53.205130
1	7	1	x1	5.852407e-05	-0.01994031	53.932174
16	**8**	2	x2 x5	4.362622e-01	0.41325254	11.466378
26	9	2	x5 x6	4.331875e-01	0.41005230	11.779811
23	10	2	x4 x5	3.831555e-01	0.35797821	16.879978
20	11	2	x3 x5	3.806171e-01	0.35533621	17.138737
11	12	2	x1 x5	3.708897e-01	0.34521168	18.130339
27	13	2	x5 x7	3.651318e-01	0.33921880	18.717285
28	14	2	x6 x7	3.485945e-01	0.32200655	20.403062
21	15	2	x3 x6	3.430647e-01	0.31625099	20.966765
22	16	2	x3 x7	2.752420e-01	0.24566008	27.880478
17	17	2	x2 x6	2.662988e-01	0.23635182	28.792135
18	18	2	x2 x7	2.595659e-01	0.22934411	29.478474
24	19	2	x4 x6	2.490153e-01	0.21836290	30.553980
12	20	2	x1 x6	2.490137e-01	0.21836122	30.554144
14	21	2	x2 x3	2.473380e-01	0.21661705	30.724969
19	22	2	x3 x4	2.368687e-01	0.20572053	31.792181
13	23	2	x1 x7	2.215867e-01	0.18981474	33.350003
9	24	2	x1 x3	2.140884e-01	0.18201035	34.114370
25	25	2	x4 x7	2.037698e-01	0.17127057	35.166229
15	26	2	x2 x4	9.122722e-02	0.05413445	46.638607
8	27	2	x1 x2	8.586310e-02	0.04855139	47.185415
10	28	2	x1 x4	7.206718e-03	-0.03331546	55.203501
48	**29**	3	x2 x4 x5	4.805843e-01	0.44812080	8.948273
51	30	3	x2 x5 x6	4.756979e-01	0.44292902	9.446381
60	31	3	x4 x5 x6	4.711309e-01	0.43807661	9.911929
31	32	3	x1 x2 x5	4.615165e-01	0.42786126	10.892008
54	33	3	x3 x4 x5	4.547233e-01	0.42064349	11.584494
57	34	3	x3 x5 x6	4.510827e-01	0.41677538	11.955607
41	35	3	x1 x5 x6	4.505365e-01	0.41619505	12.011286
45	36	3	x2 x3 x5	4.471967e-01	0.41264644	12.351745
63	37	3	x5 x6 x7	4.455568e-01	0.41090415	12.518904
52	38	3	x2 x5 x7	4.453209e-01	0.41065344	12.542957
38	39	3	x1 x4 x5	4.154999e-01	0.37896866	15.582851
61	40	3	x4 x5 x7	4.052444e-01	0.36807216	16.628281
42	41	3	x1 x5 x7	4.022445e-01	0.36488483	16.934079
35	42	3	x1 x3 x5	3.949732e-01	0.35715908	17.675302
58	43	3	x3 x5 x7	3.883081e-01	0.35007739	18.354731
59	44	3	x3 x6 x7	3.823476e-01	0.34374436	18.962334
55	45	3	x3 x4 x6	3.670196e-01	0.32745836	20.524842
53	46	3	x2 x6 x7	3.639552e-01	0.32420236	20.837228
43	47	3	x1 x6 x7	3.601506e-01	0.32016004	21.225056
62	48	3	x4 x6 x7	3.544403e-01	0.31409279	21.807158
46	49	3	x2 x3 x6	3.508497e-01	0.31027782	22.173173
36	50	3	x1 x3 x6	3.434446e-01	0.30240994	22.928031
47	51	3	x2 x3 x7	3.098577e-01	0.26672385	26.351818
56	52	3	x3 x4 x7	3.025300e-01	0.25893808	27.098799
37	53	3	x1 x3 x7	2.803393e-01	0.23536047	29.360877
33	54	3	x1 x2 x7	2.757179e-01	0.23045028	29.831969
49	55	3	x2 x4 x6	2.663158e-01	0.22046053	30.790404
32	56	3	x1 x2 x6	2.663029e-01	0.22044685	30.791716
44	57	3	x2 x3 x4	2.662816e-01	0.22042417	30.793892
50	58	3	x2 x4 x7	2.612671e-01	0.21509626	31.305061
39	59	3	x1 x4 x6	2.490155e-01	0.20207893	32.553967
29	60	3	x1 x2 x3	2.483994e-01	0.20142436	32.616768
34	61	3	x1 x3 x4	2.389245e-01	0.19135733	33.582616
40	62	3	x1 x4 x7	2.237953e-01	0.17528247	35.124867
30	63	3	x1 x2 x4	9.128705e-02	0.03449250	48.632508
90	**64**	4	x2 x4 x5 x6	5.182013e-01	0.47719717	7.113662
94	65	4	x3 x4 x5 x6	5.156128e-01	0.47438838	7.377528
84	66	4	x2 x3 x4 x5	5.143129e-01	0.47297783	7.510039
68	67	4	x1 x2 x4 x5	5.126295e-01	0.47115120	7.681639

71	68	4	x1 x2 x5 x6	4.949149e-01	0.45192894	9.487435
80	69	4	x1 x4 x5 x6	4.937641e-01	0.45068022	9.604744
91	70	4	x2 x4 x5 x7	4.907041e-01	0.44735980	9.916674
98	71	4	x4 x5 x6 x7	4.850083e-01	0.44117926	10.497292
87	72	4	x2 x3 x5 x6	4.831123e-01	0.43912190	10.690567
93	73	4	x2 x5 x6 x7	4.829555e-01	0.43895174	10.706552
72	74	4	x1 x2 x5 x7	4.782875e-01	0.43388647	11.182399
83	75	4	x1 x5 x6 x7	4.706603e-01	0.42561007	11.959909
74	76	4	x1 x3 x4 x5	4.701516e-01	0.42505817	12.011755
65	77	4	x1 x2 x3 x5	4.660953e-01	0.42065664	12.425248
77	78	4	x1 x3 x5 x6	4.620261e-01	0.41624108	12.840060
95	79	4	x3 x4 x5 x7	4.592229e-01	0.41319931	13.125812
97	80	4	x3 x5 x6 x7	4.566753e-01	0.41043487	13.385512
88	81	4	x2 x3 x5 x7	4.519096e-01	0.40526362	13.871315
81	82	4	x1 x4 x5 x7	4.509865e-01	0.40426191	13.965419
78	83	4	x1 x3 x5 x7	4.114179e-01	0.36132577	17.998968
96	84	4	x3 x4 x6 x7	4.090382e-01	0.35874352	18.241552
89	85	4	x2 x3 x6 x7	3.918721e-01	0.34011654	19.991426
79	86	4	x1 x3 x6 x7	3.864186e-01	0.33419895	20.547343
73	87	4	x1 x2 x6 x7	3.748137e-01	0.32160641	21.730324
85	88	4	x2 x3 x4 x6	3.724009e-01	0.31898824	21.976282
92	89	4	x2 x4 x6 x7	3.692869e-01	0.31560923	22.293716
75	90	4	x1 x3 x4 x6	3.680325e-01	0.31424800	22.421594
82	91	4	x1 x4 x6 x7	3.666668e-01	0.31276611	22.560808
66	92	4	x1 x2 x3 x6	3.512598e-01	0.29604782	24.131374
86	93	4	x2 x3 x4 x7	3.318551e-01	0.27499171	26.109448
67	94	4	x1 x2 x3 x7	3.150541e-01	0.25676088	27.822106
76	95	4	x1 x3 x4 x7	3.061685e-01	0.24711898	28.727896
70	96	4	x1 x2 x4 x7	2.779471e-01	0.21649580	31.604729
64	97	4	x1 x2 x3 x4	2.681887e-01	0.20590688	32.599484
69	98	4	x1 x2 x4 x6	2.663193e-01	0.20387834	32.790050
114	**99**	5	x2 x3 x4 x5 x6	5.449760e-01	0.49551685	6.384302
105	100	5	x1 x2 x4 x5 x6	5.434443e-01	0.49381870	6.540437
99	101	5	x1 x2 x3 x4 x5	5.341235e-01	0.48348471	7.490588
106	102	5	x1 x2 x4 x5 x7	5.321800e-01	0.48132997	7.688704
109	103	5	x1 x3 x4 x5 x6	5.276655e-01	0.47632478	8.148901
118	104	5	x2 x4 x5 x6 x7	5.264354e-01	0.47496096	8.274297
119	105	5	x3 x4 x5 x6 x7	5.187853e-01	0.46647933	9.054134
113	106	5	x1 x4 x5 x6 x7	5.172863e-01	0.46481737	9.206941
115	107	5	x2 x3 x4 x5 x7	5.168036e-01	0.46428229	9.256139
108	108	5	x1 x2 x5 x6 x7	5.084437e-01	0.45501364	10.108337
102	109	5	x1 x2 x3 x5 x6	4.979120e-01	0.44333720	11.181918
117	110	5	x2 x3 x5 x6 x7	4.871904e-01	0.43145018	12.274861
110	111	5	x1 x3 x4 x5 x7	4.818079e-01	0.42548267	12.823539
103	112	5	x1 x2 x3 x5 x7	4.788800e-01	0.42223652	13.122004
112	113	5	x1 x3 x5 x6 x7	4.742651e-01	0.41712004	13.592434
116	114	5	x2 x3 x4 x6 x7	4.157940e-01	0.35229330	19.552877
111	115	5	x1 x3 x4 x6 x7	4.118316e-01	0.34790029	19.956790
104	116	5	x1 x2 x3 x6 x7	3.960782e-01	0.33043448	21.562669
107	117	5	x1 x2 x4 x6 x7	3.807793e-01	0.31347268	23.122208
100	118	5	x1 x2 x3 x4 x6	3.734175e-01	0.30531069	23.872656
101	119	5	x1 x2 x3 x4 x7	3.357074e-01	0.26350163	27.716757
120	**120**	6	x1 x2 x3 x4 x5 x6	5.608627e-01	0.50231110	6.764836
124	121	6	x1 x2 x4 x5 x6 x7	5.595567e-01	0.50083089	6.897974
126	122	6	x2 x3 x4 x5 x6 x7	5.471085e-01	0.48672291	8.166922
121	123	6	x1 x2 x3 x4 x5 x7	5.430458e-01	0.48211863	8.581056
125	124	6	x1 x3 x4 x5 x6 x7	5.362805e-01	0.47445125	9.270702
123	125	6	x1 x2 x3 x5 x6 x7	5.086947e-01	0.44318732	12.082750
122	126	6	x1 x2 x3 x4 x6 x7	4.187414e-01	0.34124029	21.252417
127	**127**	7	x1 x2 x3 x4 x5 x6 x7	5.683657e-01	0.49969658	8.000000

MATH 2780 Chapter5 Worksheet

	Index	N	Predictors	R-Square	Adj. R-Square	Mallow's Cp
3	**1**	1	x3	2.129999e-01	0.19725986	2.544459
2	2	1	x2	8.574310e-02	0.06745797	10.717423
4	3	1	x4	7.190738e-03	-0.01266545	15.762386
1	4	1	x1	5.852407e-05	-0.01994031	16.220447
8	**5**	2	x2 x3	2.473380e-01	0.21661705	2.339122
10	6	2	x3 x4	2.368687e-01	0.20572053	3.011499
6	7	2	x1 x3	2.140884e-01	0.18201035	4.474550
9	8	2	x2 x4	9.122722e-02	0.05413445	12.365210
5	9	2	x1 x2	8.586310e-02	0.04855139	12.709716
7	10	2	x1 x4	7.206718e-03	-0.03331546	17.761360
14	**11**	3	x2 x3 x4	2.662816e-01	0.22042417	3.122484
11	12	3	x1 x2 x3	2.483994e-01	0.20142436	4.270952
13	13	3	x1 x3 x4	2.389245e-01	0.19135733	4.879466
12	14	3	x1 x2 x4	9.128705e-02	0.03449250	14.361367
15	**15**	4	x1 x2 x3 x4	2.681887e-01	0.20590688	5.000000