

Computer Printout for STA302F94 Final Exam

Part 1: The Body Fat data

First we have an analysis of the body fat example you looked at for homework. The command file is followed by part of the list file.

```
options linesize=79 pagesize=50;
title 'Diagnostics from Ch. 11 for Final';
data pudgy;
    infile 'bodyfat.dat';
    input subject triceps thigh midarm bodyfat; /* Data on p. 272 */
    label triceps = 'Triceps Skinfold Thickness'
          thigh   = 'Thigh Circumference'
          midarm  = 'Midarm Circumference'
          bodyfat = 'Percent Body Fat';
proc reg simple corr;
    model bodyfat = triceps thigh midarm
      / clm cli r influence partial tol;
```

Diagnostics from Ch. 11 for Final

1

Descriptive Statistics

Variables	Sum	Mean	
INTERCEP	20	1	Intercept
TRICEPS	506.1	25.305	Triceps Skinfold Thickness
THIGH	1023.4	51.17	Thigh Circumference
MIDARM	552.4	27.62	Midarm Circumference
BODYFAT	403.9	20.195	Percent Body Fat

Variables	Uncorrected SS	Variance	
INTERCEP	20	0	Intercept
TRICEPS	13286.29	25.233131579	Triceps Skinfold Thickness
THIGH	52888	27.401157895	Thigh Circumference
MIDARM	15510.02	13.301684211	Midarm Circumference
BODYFAT	8652.15	26.073131579	Percent Body Fat

Variables	Std Deviation	
INTERCEP	0	Intercept
TRICEPS	5.0232590595	Triceps Skinfold Thickness
THIGH	5.2346115324	Thigh Circumference
MIDARM	3.6471474073	Midarm Circumference
BODYFAT	5.1061856193	Percent Body Fat

Correlation

CORR	TRICEPS	THIGH	
TRICEPS	1.0000	0.9238	Triceps Skinfold Thickness
THIGH	0.9238	1.0000	Thigh Circumference
MIDARM	0.4578	0.0847	Midarm Circumference
BODYFAT	0.8433	0.8781	Percent Body Fat
CORR	MIDARM	BODYFAT	
TRICEPS	0.4578	0.8433	Triceps Skinfold Thickness
THIGH	0.0847	0.8781	Thigh Circumference
MIDARM	1.0000	0.1424	Midarm Circumference
BODYFAT	0.1424	1.0000	Percent Body Fat

Model: MODEL1

Dependent Variable: BODYFAT Percent Body Fat

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	396.98461	132.32820	21.516	0.0001
Error	16	98.40489	6.15031		
C Total	19	495.38950			
Root MSE	2.47998	R-square	0.8014		
Dep Mean	20.19500	Adj R-sq	0.7641		
C.V.	12.28017				

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	117.084695	99.78240295	1.173	0.2578
TRICEPS	1	4.334092	3.01551136	1.437	0.1699
THIGH	1	-2.856848	2.58201527	-1.106	0.2849
MIDARM	1	-2.186060	1.59549900	-1.370	0.1896

Variable	DF	Tolerance	Variable Label
INTERCEP	1	.	Intercept
TRICEPS	1	0.00141075	Triceps Skinfold Thickness
THIGH	1	0.00177197	Thigh Circumference
MIDARM	1	0.00955968	Midarm Circumference

Obs	Dep Var BODYFAT	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean	Lower95% Predict	Upper95% Predict
1	11.9000	14.8550	1.449	11.7840	17.9259	8.7665	20.9435
2	22.8000	20.2188	0.981	18.1388	22.2989	14.5650	25.8727
3	18.7000	20.9867	1.646	17.4977	24.4757	14.6770	27.2964
4	20.1000	23.1273	0.832	21.3645	24.8901	17.5824	28.6723
5	12.9000	11.7576	1.490	8.5984	14.9168	5.6241	17.8911
6	21.7000	22.2437	0.899	20.3372	24.1503	16.6514	27.8360
7	27.1000	25.7143	1.093	23.3967	28.0319	19.9688	31.4598
8	25.4000	22.2706	1.005	20.1404	24.4009	16.5982	27.9431
9	21.3000	19.5948	1.089	17.2865	21.9032	13.8531	25.3366
10	19.3000	20.5484	1.216	17.9701	23.1267	14.6929	26.4039
11	25.4000	24.5956	0.926	22.6330	26.5581	18.9839	30.2072
12	27.2000	24.9923	0.820	23.2543	26.7304	19.4552	30.5295
13	11.7000	15.0094	1.146	12.5798	17.4390	9.2178	20.8010
14	17.8000	13.6723	1.076	11.3923	15.9523	7.9419	19.4027
15	12.8000	11.8119	1.464	8.7092	14.9147	5.7074	17.9165
16	23.9000	23.7275	0.839	21.9494	25.5056	18.1776	29.2773
17	22.6000	22.9736	0.878	21.1124	24.8348	17.3966	28.5506
18	25.4000	26.7859	1.185	24.2740	29.2978	20.9593	32.6125
19	14.8000	18.5263	0.902	16.6136	20.4389	12.9319	24.1207
20	21.1000	20.4879	0.637	19.1375	21.8383	15.0600	25.9159

Obs	Residual	Std Err Residual	Student Residual	-2	-1	0	1	2	Cook's D	Rstudent
1	-2.9550	2.013	-1.468		**				0.279	-1.5280
2	2.5812	2.278	1.133				**		0.060	1.1442
3	-2.2867	1.855	-1.233		**				0.299	-1.2545
4	-3.0273	2.336	-1.296		**				0.053	-1.3261
5	1.1424	1.982	0.576				*		0.047	0.5639
6	-0.5437	2.311	-0.235						0.002	-0.2282
7	1.3857	2.226	0.622				*		0.023	0.6102
8	3.1294	2.267	1.380				**		0.094	1.4239
9	1.7052	2.228	0.765				*		0.035	0.7549
10	-1.2484	2.161	-0.578		*				0.026	-0.5652
11	0.8044	2.301	0.350						0.005	0.3399
12	2.2077	2.341	0.943				*		0.027	0.9398
13	-3.3094	2.199	-1.505		***				0.154	-1.5725
14	4.1277	2.235	1.847				***		0.198	2.0164
15	0.9881	2.002	0.494						0.033	0.4815
16	0.1725	2.334	0.074						0.000	0.0716
17	-0.3736	2.319	-0.161						0.001	-0.1561
18	-1.3859	2.179	-0.636		*				0.030	-0.6239
19	-3.7263	2.310	-1.613		***				0.099	-1.7068
20	0.6121	2.397	0.255						0.001	0.2478

	Hat	Diag	Cov		INTERCEP	TRICEPS	THIGH	MIDARM
Obs		H	Ratio	Dffits	Dfbetas	Dfbetas	Dfbetas	Dfbetas
1	0.3412		1.1017	-1.0997	-0.7600	-0.7312	0.7595	0.7049
2	0.1565		1.0982	0.4929	0.3991	0.3963	-0.3993	-0.3893
3	0.4404		1.5521	-1.1130	0.3818	0.3436	-0.3420	-0.4389
4	0.1124		0.9361	-0.4720	0.0460	0.0257	-0.0328	-0.0543
5	0.3611		1.8631	0.4239	-0.2114	-0.2217	0.2092	0.2372
6	0.1315		1.4701	-0.0888	-0.0081	-0.0070	0.0056	0.0132
7	0.1943		1.4567	0.2997	-0.1468	-0.1365	0.1456	0.1339
8	0.1642		0.9328	0.6311	0.4224	0.4361	-0.4341	-0.4058
9	0.1928		1.3815	0.3689	0.2254	0.2133	-0.2143	-0.2349
10	0.2405		1.5667	-0.3181	0.2448	0.2472	-0.2502	-0.2341
11	0.1394		1.4587	0.1368	-0.0522	-0.0461	0.0495	0.0505
12	0.1093		1.1561	0.3292	-0.0056	0.0074	0.0026	-0.0052
13	0.2136		0.8941	-0.8195	0.3401	0.3822	-0.3684	-0.3327
14	0.1881		0.6109	0.9705	-0.3965	-0.4308	0.4009	0.4480
15	0.3483		1.8682	0.3520	-0.0735	-0.0953	0.0866	0.0733
16	0.1144		1.4598	0.0257	0.0107	0.0117	-0.0112	-0.0105
17	0.1253		1.4705	-0.0591	0.0263	0.0258	-0.0271	-0.0234
18	0.2283		1.5141	-0.3393	-0.1051	-0.1114	0.1010	0.1260
19	0.1324		0.7336	-0.6666	-0.4830	-0.4670	0.4764	0.4686
20	0.0660		1.3635	0.0659	0.0327	0.0323	-0.0323	-0.0323
Sum of Residuals				0				
Sum of Squared Residuals				98.4049				
Predicted Resid SS (Press)				160.7366				

Part 2: The SMSA data

Now we have a model selection analysis (and other stuff) on the SMSA data set. **Here is the beginning of the raw data file.** Notice I have added a few lines at the beginning.

```
.      .      .      .      .      1778.15      .      54.54      .      .      .      1
.      .      .      .      .      1778.15      .      54.54      .      .      .      2
.      .      .      .      .      1778.15      .      54.54      .      .      .      3
.      .      .      .      .      1778.15      .      54.54      .      .      .      4
1      1384      9387      78.1      12.3      25627      69678      50.1      4083.9      72100      709234      1
2      4069      7031      44.0      10.0      15389      39699      62.0      3353.6      52737      499813      4
3      3719      7017      43.9      9.4      13326      43292      53.9      3305.9      54542      393162      2
4      3553      4794      37.4      10.7      9724      33731      50.6      2066.3      33216      198102      1
5      3916      4370      29.9      8.8      6402      24167      52.2      1966.7      32906      294466      2
```

etc.

Now the command file ...

```

options linesize=79 pagesize=500;
title 'STA 302f94: Final SAS Job -- Model Selection for SMSA Data';

proc format; /* Used to label values of the categorical variables */
  value regfmt      1 = 'Northeast'
                   2 = 'North central'
                   3 = 'South'
                   4 = 'West';

data census;
  infile 'smsa.raw';
  input id landarea totpop urban oldfolks doctors hospbeds hsgrads
        labforce income crimes region;
  label
    id          = 'Identification number'
    landarea    = 'Land area in square miles'
    totpop      = 'Estimated 1977 population in thousands'
    urban       = 'Percent 1976 population in cities'
    oldfolks    = 'Percent 1976 population 65 & older'
    doctors     = 'Number active nonfederal physicians 1977'
    hospbeds    = 'Number hospital beds & cribs 1977'
    hsgrads     = 'Percent of pop 25+ with 12+ yrs schooling 1970'
    labforce    = 'Number emp & unemp in civilian labor force 1977'
    income      = 'Total personal income before taxes 1976'
    crimes      = 'Total serious crimes 1977'
    region      = 'Geographic region' ;
  format region regfmt.;
  if region=2 then r1=1; else r1=0;
  if region=3 then r2=1; else r2=0;
  if region=4 then r3=1; else r3=0;
  crimrate = crimes/totpop;
  label crimrate = 'Crime Rate: Serious crimes per thousand pop';
  popdens = totpop/landarea;
  label popdens = 'Population Density';
proc reg simple corr;
  model crimrate = {r1 r2 r3} landarea totpop urban oldfolks doctors
    hospbeds hsgrads labforce income
    / selection = stepwise slentry=.05 slstay=.05
    groupnames = 'region' 'landarea' 'totpop' 'urban' 'oldfolks'
    'doctors' 'hospbeds' 'hsgrads' 'labforce' 'income';
  /* Specifying all names again is a pain, but otherwise group6 etc.
    It's worth it to be able to keep the dummy vars together. */
  model crimrate = r1 r2 r3 landarea totpop urban oldfolks doctors
    hospbeds hsgrads labforce income
    / selection = rsquare adjrsq include=3; /* Can't group r1-r3 */
  /* Look at the selected model */
  model crimrate = r1 r2 r3 doctors hsgrads/ r clm;
proc means;
  var crimrate;
  class region;

```

=== Now part of the 1st file. First descriptive stats & then stepwise ===

Descriptive Statistics

Variables	Sum	Mean	Uncorrected SS
INTERCEP	141	1	141
R1	35	0.2482269504	35
R2	51	0.3617021277	51
R3	28	0.1985815603	28
LANDAREA	363721	2579.5815603	2024868829
TOTPOP	132838	942.11347518	357271388
URBAN	5891.6	41.784397163	289671.56
OLDFOLKS	1384.7	9.8205673759	14489.35
DOCTORS	250719	1778.1489362	1715566309
HOSPBEDS	864913	6134.1347518	15583556605
HSGRADES	7689.7	54.536879433	428124.21
LABFORCE	60872.2	431.7177305	75116314.72
INCOME	914794	6487.9007092	19631090146
CRIMRATE	7877.2520489	55.867035808	469244.66354

Stepwise Procedure for Dependent Variable CRIMRATE

Step 1 Group region Entered R-square = 0.38251098 C(p) = 19.88763573

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	11156.29269099	3718.76423033	28.29	0.0001
Error	137	18009.64856591	131.45728880		
Total	140	29165.94125690			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	42.20579312	2.20653313	48095.88227487	365.87	0.0001
--- Group region ---			11156.29269099	28.29	0.0001
R1	10.57265646	2.93678583	1703.75977561	12.96	0.0004
R2	15.20634325	2.72880603	4082.14959995	31.05	0.0001
R3	27.88102632	3.09252189	10685.05147400	81.28	0.0001

Bounds on condition number: 1.843972, 15.60835

Step 2 Group doctors Entered R-square = 0.44043767 C(p) = 7.54523499

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	4	12845.77919373	3211.44479843	26.76	0.0001
Error	136	16320.16206317	120.00119164		
Total	140	29165.94125690			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	39.32895012	2.24328643	36884.23710890	307.37	0.0001
--- Group region ---			11506.12670020	31.96	0.0001
R1	11.49231021	2.81658784	1997.80749518	16.65	0.0001
R2	16.57355642	2.63253189	4756.30503712	39.64	0.0001
R3	28.28192558	2.95663002	10980.18212984	91.50	0.0001
--- Group doctors ---			1689.48650274	14.08	0.0003
DOCTORS	0.00116662	0.00031092	1689.48650274	14.08	0.0003

Bounds on condition number: 1.879989, 25.10794

Step 3 Group hsgrads Entered R-square = 0.46380739 C(p) = 3.75899127

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	5	13527.37921525	2705.47584305	23.36	0.0001
Error	135	15638.56204165	115.84120031		
Total	140	29165.94125690			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	21.03850311	7.85586659	830.81536998	7.17	0.0083
--- Group region ---			8468.11094478	24.37	0.0001
R1	9.85744369	2.84822878	1387.52851458	11.98	0.0007
R2	17.10263868	2.59567994	5029.06219977	43.41	0.0001
R3	24.25213403	3.34642406	6084.16552423	52.52	0.0001
--- Group doctors ---			1533.87160813	13.24	0.0004
DOCTORS	0.00111436	0.00030624	1533.87160813	13.24	0.0004
--- Group hsgrads ---			681.60002152	5.88	0.0166
HSGRADS	0.35568719	0.14663409	681.60002152	5.88	0.0166

Bounds on condition number: 2.169283, 42.78862

All groups of variables left in the model are significant at the 0.0500 level. No other group of variables met the 0.0500 significance level for entry into the model.

Summary of Stepwise Procedure for Dependent Variable CRIMRATE

Step	Group Entered	Group Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	region		3	0.3825	0.3825	19.8876	28.2888	0.0000
2	doctors		4	0.0579	0.4404	7.5452	14.0789	0.0003
3	hsgrads		5	0.0234	0.4638	3.7590	5.8839	0.0166

===== Now R-Squared output =====

N = 141 Regression Models for Dependent Variable: CRIMRATE

In	R-square	Adj Rsq	Variables in Model
3	0.382511	0.368989	R1 R2 R3

NOTE: The above variables are included in all models to follow.

4	0.440438	0.423980	DOCTORS
4	0.434824	0.418201	INCOME
4	0.434316	0.417679	HOSPBEDS
4	0.434258	0.417618	TOTPOP
4	0.433118	0.416445	LABFORCE
4	0.411216	0.393899	HSGRA DS
4	0.396925	0.379187	LANDAREA
4	0.384960	0.366871	OLDFOLKS
4	0.383894	0.365773	URBAN

5	0.463807	0.443948	DOCTORS HSGRA DS
5	0.460042	0.440043	TOTPOP HSGRA DS
5	0.459349	0.439325	HSGRA DS INCOME
5	0.459179	0.439149	HOSPBEDS HSGRA DS
5	0.457896	0.437818	HSGRA DS LABFORCE
5	0.449397	0.429005	LANDAREA DOCTORS
5	0.442938	0.422307	LANDAREA INCOME
5	0.442262	0.421605	LANDAREA HOSPBEDS
5	0.442182	0.421522	OLDFOLKS DOCTORS
5	0.441824	0.421151	DOCTORS LABFORCE
5	0.441702	0.421024	TOTPOP DOCTORS
5	0.441652	0.420972	DOCTORS INCOME

6	0.475151	0.451650	LANDAREA DOCTORS HSGRA DS
6	0.469804	0.446064	LANDAREA HSGRA DS INCOME
6	0.469448	0.445692	LANDAREA HOSPBEDS HSGRA DS
6	0.469359	0.445599	LANDAREA TOTPOP HSGRA DS
6	0.467936	0.444112	OLDFOLKS DOCTORS HSGRA DS
6	0.467574	0.443734	LANDAREA HSGRA DS LABFORCE
6	0.464609	0.440636	DOCTORS HSGRA DS LABFORCE
6	0.464481	0.440503	DOCTORS HSGRA DS INCOME
6	0.464341	0.440356	DOCTORS HOSPBEDS HSGRA DS
6	0.464335	0.440350	TOTPOP OLDFOLKS HSGRA DS
6	0.464025	0.440026	TOTPOP DOCTORS HSGRA DS
6	0.463928	0.439925	URBAN DOCTORS HSGRA DS

7	0.478063	0.450593	LANDAREA OLDFOLKS DOCTORS HSGRA DS
7	0.477892	0.450413	LANDAREA DOCTORS HSGRA DS LABFORCE
7	0.477090	0.449568	LANDAREA TOTPOP DOCTORS HSGRA DS
7	0.476866	0.449332	LANDAREA DOCTORS HSGRA DS INCOME
7	0.476767	0.449228	LANDAREA DOCTORS HOSPBEDS HSGRA DS

```

7  0.475692  0.448097  LANDAREA URBAN DOCTORS HSGRA
7  0.473851  0.446159  LANDAREA HSGRA
7  0.473056  0.445322  LANDAREA OLDFOLKS HSGRA
7  0.472529  0.444768  LANDAREA TOTPOP OLDFOLKS HSGRA
7  0.472412  0.444644  LANDAREA TOTPOP HSGRA
7  0.472380  0.444610  LANDAREA OLDFOLKS HOSPBEDS HSGRA
7  0.471032  0.443192  LANDAREA OLDFOLKS HSGRA

```

... (Results for models with more than 7 IVs have been deleted) ...

===== Now a look at the model that was selected automatically =====

Model: MODEL3

Dependent Variable: CRIMRATE Crime Rate: Serious crimes per thousand

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	5	13527.37922	2705.47584	23.355	0.0001
Error	135	15638.56204	115.84120		
C Total	140	29165.94126			

Root MSE	10.76295	R-square	0.4638
Dep Mean	55.86704	Adj R-sq	0.4439
C.V.	19.26531		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	21.038503	7.85586659	2.678	0.0083
R1	1	9.857444	2.84822878	3.461	0.0007
R2	1	17.102639	2.59567994	6.589	0.0001
R3	1	24.252134	3.34642406	7.247	0.0001
DOCTORS	1	0.001114	0.00030624	3.639	0.0004
HSGRA	1	0.355687	0.14663409	2.426	0.0166

Variable	DF	Variable Label
INTERCEP	1	Intercept
R1	1	
R2	1	
R3	1	
DOCTORS	1	Number active nonfederal physicians 1977
HSGRA	1	Percent of pop 25+ with 12+ yrs schoolin

Obs	Dep Var CRIMRATE	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean	Residual	Std Err Residual
1	.	42.4192	2.124	38.2192	46.6192	.	.
2	.	52.2766	1.837	48.6429	55.9103	.	.
3	.	59.5218	1.643	56.2728	62.7709	.	.
4	.	66.6713	2.386	61.9529	71.3898	.	.
5	75.5549	67.4161	7.410	52.7615	82.0706	8.1389	7.806
6	71.0870	84.4921	4.556	75.4817	93.5025	-13.4050	9.751
7	56.0299	64.9174	4.041	56.9256	72.9092	-8.8875	9.976
8	41.3229	49.8723	3.052	43.8361	55.9085	-8.5494	10.321
9	67.3835	56.5969	2.425	51.8014	61.3925	10.7866	10.486
10	80.1892	78.2758	2.834	72.6716	83.8801	1.9134	10.383

... (Rest of the list file omitted up to proc means output) ...

REGION	N Obs	N	Mean	Std Dev	Minimum
Northeast	28	27	42.2057931	13.6756200	15.6902985
North central	36	35	52.7784496	8.7692392	31.8106061
South	52	51	57.4121364	12.0644156	25.5514706
West	29	28	70.0868194	10.9797293	53.7309091