Name________________________ Spring, 2014

Applied Statistics Comprehensive Examination
Regression & Linear Models

1. (15 points) A forest manager is responsible for the selection and purchase of chainsaws for her field crew. Her primary interest is worker safety. She is provided with data on chainsaw kickback values for 4 models of chainsaws with 5 observations each. Models A and D are residential models whereas Models B and C are industrial grade. The sample mean chainsaw kickback for each model is:

Model	Mean
A	33
B	43
C	49
D	31

The Mean Square Error resulting from the ANOVA model is 101.25.

(a) (5 Points) Construct a complete set of mutually orthogonal contrasts for testing the differences in mean chainsaw kickback between the four models.

(b) (10 Points) Perform a 0.05-level test to determine if the mean chainsaw kickback is larger for industrial grade models than the mean for residential models.

2. (25 points) The following results were obtained from a preliminary study examining whether one of three types of diet combined with aerobic exercise or a control would reduce cholesterol levels. Subjects who participated in the study had elevated cholesterol levels. The cholesterol levels of 16 subjects at completion of the pretrial are given in the table below. Suppose an effects model will be used to analyze the data where the model will be $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$, where $i = 1, 2$, $j = 1, 2, 3$, α_i and β_j are the main effects of Physical Activity and Diet, respectively, γ_{ij} is an interaction effect, and ϵ_{ijk} is a random error term. The following table displays the data from the experiment:

	Diet		
Physical Activity	Control	Diet 1	Diet 2
Control	243	236	219
	229	248	228
	252		213
Aerobic	221	212	206
	237	227	199
		209	211

(a) (10 Points) The solution vector using sum to zero restrictions is $\hat{\beta} = \left(\hat{\mu}, \hat{\alpha_1}, \hat{\beta}_1, \hat{\beta}_2, \hat{\gamma_{11}}, \hat{\gamma_{12}}\right) = (225.6, 8.8, 9.6, 3.4, -2.7, 4.2)$. Use this solution vector to obtain estimates of the parameters which were in the full design matrix but not included in the reparameterized design matrix.

(b) (8 Points) Calculate the means and adjusted means (least squares means) for the three levels of Diet.

(c) (7 Points) Show that the difference in sample means for Physical Activity, $\hat{\mu_{1.}} - \hat{\mu_{2.}}$, is biased for $\alpha_1 - \alpha_2$, which is why adjusted means should be used in the presence of unequal sample sizes. Note: $\hat{\mu_{i.}}$ is the sample mean calculated from all observations at the ith level of Physical Activity or $\bar{Y_{i.}}$.

3. (15 points) Interpreting coefficients.

   a) Consider the following regression equation:

      $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

      Interpret the values of the parameters.

   b) Consider the following regression equation:

      $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$

      Interpret the value of β_3.

   c) A knot/piecewise regression was calculated using the following equation:

      $y = \beta_0 + \beta_1 x_1 + \beta_2 x_{t1} + \beta_3 x_{t2} + \varepsilon$

      $$\text{where} \quad x_{t1} = \begin{cases} x_1 - 20 & \text{if } x_1 > 20 \\ 0 & \textit{otherwise} \end{cases} \quad \text{and} \quad x_{t2} = \begin{cases} x_1 - 40 & \text{if } x_1 > 40 \\ 0 & \textit{otherwise} \end{cases}$$

      Explain what hypothesis should be tested to determine whether a simple linear regression is sufficient for these data (vs. whether the knot/piecewise regression should be used). Describe how you would go about conducting this hypothesis test.

4. (45 points) A random sample of 45 college students were asked whether they favor criminal penalties for students who purchase or download papers from the internet and turn them in as their own work. A score was created from 0 to 25 where 0 indicated the student was totally opposed to criminal penalties and 25 meant that he or she fully support criminal penalties. Data were also gathered on student age, income of parents (in $1000s) and gender (coded as “male”, where 1=male and 0=female). Selected output from a regression analysis is included below. The model analyzed was:

   $\text{Score} = \beta_0 + \beta_1 x_{\text{age}} + \beta_2 x_{\text{income}} + \beta_3 x_{\text{male}} + \varepsilon$

a) Calculate and interpret the value of R^2 for this model.

b) Conduct the global hypothesis test for this model at $\alpha=0.05$.

c) State all the assumptions for this multiple regression. For each one, provide evidence of whether the assumption is met. If you cannot verify it from the given output, state what additional output you would need in order to verify it.

d) Consider a backwards elimination stepwise procedure with the significance level set at $\alpha=0.05$. What is the next step after viewing these results?

e) The researchers suggested adding a quadratic term for age by squaring the age variable.

   i. Why do you think this was suggested?
   ii. What problem may adding this variable create?
   iii. How might you solve the problem introduced by adding the square of the age variable?

f) The last page of the output presents a table for the Lack of Fit test for these data.

i. State the null and alternative hypotheses for the Lack of Fit test.

ii. What decision should be made about the hypotheses based on the results in the table?

iii. Comment on the appropriateness of this test for these data.

The REG Procedure
Model: MODEL1
Dependent Variable: CRIME

Number of Observations Read	45
Number of Observations Used	45

Source	DF	Sum of Squares	Mean Square	F Value
Model	3	1315.21799	438.40600	69.91
Error	41	257.09312	6.27056	
Corrected Total	44	1572.31111		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	**1**	-11.10839	2.29597	-4.84	<.0001
AGE	**1**	0.44652	0.07621	5.86	<.0001
INCOME	**1**	0.28621	0.02574	11.12	<.0001
male	**1**	2.79259	0.84490	3.31	0.0020

The REG Procedure
Model: MODEL1
Dependent Variable: CRIME

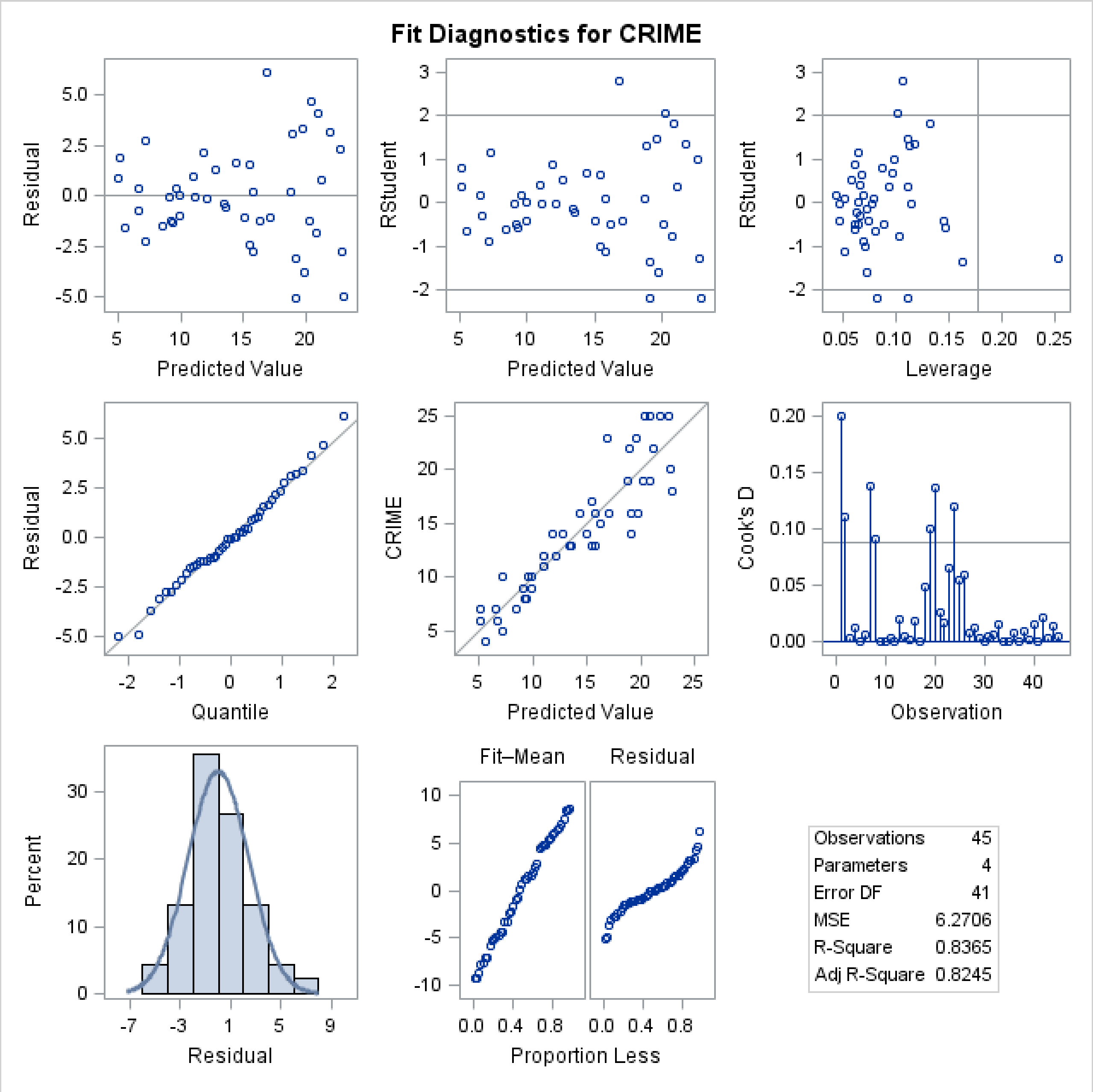
Output Statistics

Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS			
						Intercept	AGE	INCOME	male
1	6.1404	2.8025	0.1064	0.6030	0.9672	0.4321	-0.6169	0.1303	0.4012
2	4.6715	2.0442	0.1026	0.8266	0.6914	0.0896	-0.2973	0.3079	0.2200
3	0.8128	0.3408	0.1124	1.2292	0.1213	0.0087	-0.0487	0.0628	0.0324
4	1.6123	0.6733	0.0977	1.1694	0.2216	0.0675	-0.1200	0.1165	-0.1349
5	0.2284	0.0939	0.0794	1.1979	0.0276	0.0055	-0.0112	0.0072	0.0130
6	-1.2026	-0.4987	0.0896	1.1828	-0.1565	-0.0142	0.0572	-0.0655	-0.0585
7	-4.9388	-2.1871	0.1118	0.7909	-0.7758	0.1050	0.1638	-0.4887	-0.1878
8	-3.1273	-1.3807	0.1638	1.0957	-0.6110	0.0270	0.1858	-0.4857	0.3776
9	-0.0980	-0.0403	0.0774	1.1963	-0.0117	-0.0059	0.0074	-0.0034	0.0066
10	-0.4032	-0.1653	0.0738	1.1885	-0.0466	-0.0169	0.0247	-0.0199	0.0288
11	-1.0544	-0.4332	0.0741	1.1700	-0.1225	-0.0411	0.0545	-0.0055	-0.0707
12	-0.5635	-0.2298	0.0636	1.1725	-0.0599	-0.0186	0.0272	-0.0259	0.0385
13	-2.4410	-1.0120	0.0717	1.0747	-0.2812	-0.0245	0.0830	-0.1716	0.1899
14	1.2951	0.5281	0.0578	1.1394	0.1308	0.0521	-0.0641	0.0425	-0.0815
15	-1.0634	-0.4309	0.0481	1.1383	-0.0968	0.0086	-0.0026	-0.0490	0.0666
16	-2.7961	-1.1513	0.0518	1.0218	-0.2691	0.0607	-0.0443	-0.1471	0.1810
17	0.2039	0.0826	0.0518	1.1633	0.0193	-0.0044	0.0032	0.0106	-0.0130
18	-3.7679	-1.5919	0.0731	0.9314	-0.4470	0.1633	-0.1973	-0.0103	-0.3040
19	-5.0696	-2.2112	0.0822	0.7583	-0.6618	0.2341	-0.3360	0.0734	-0.4664
20	-2.7457	-1.2787	0.2533	1.2593	-0.7447	0.6092	-0.6471	-0.3049	0.2162
21	2.3356	0.9818	0.0983	1.1129	0.3242	-0.1978	0.1939	0.0873	0.1641
22	-1.8213	-0.7640	0.1030	1.1613	-0.2589	0.1378	-0.1691	-0.0043	-0.1560
23	3.3799	1.4518	0.1123	1.0126	0.5164	-0.3547	0.3456	0.2839	-0.2445
24	4.1443	1.8268	0.1325	0.9234	0.7139	-0.4024	0.5242	-0.0379	0.4104

Output Statistics

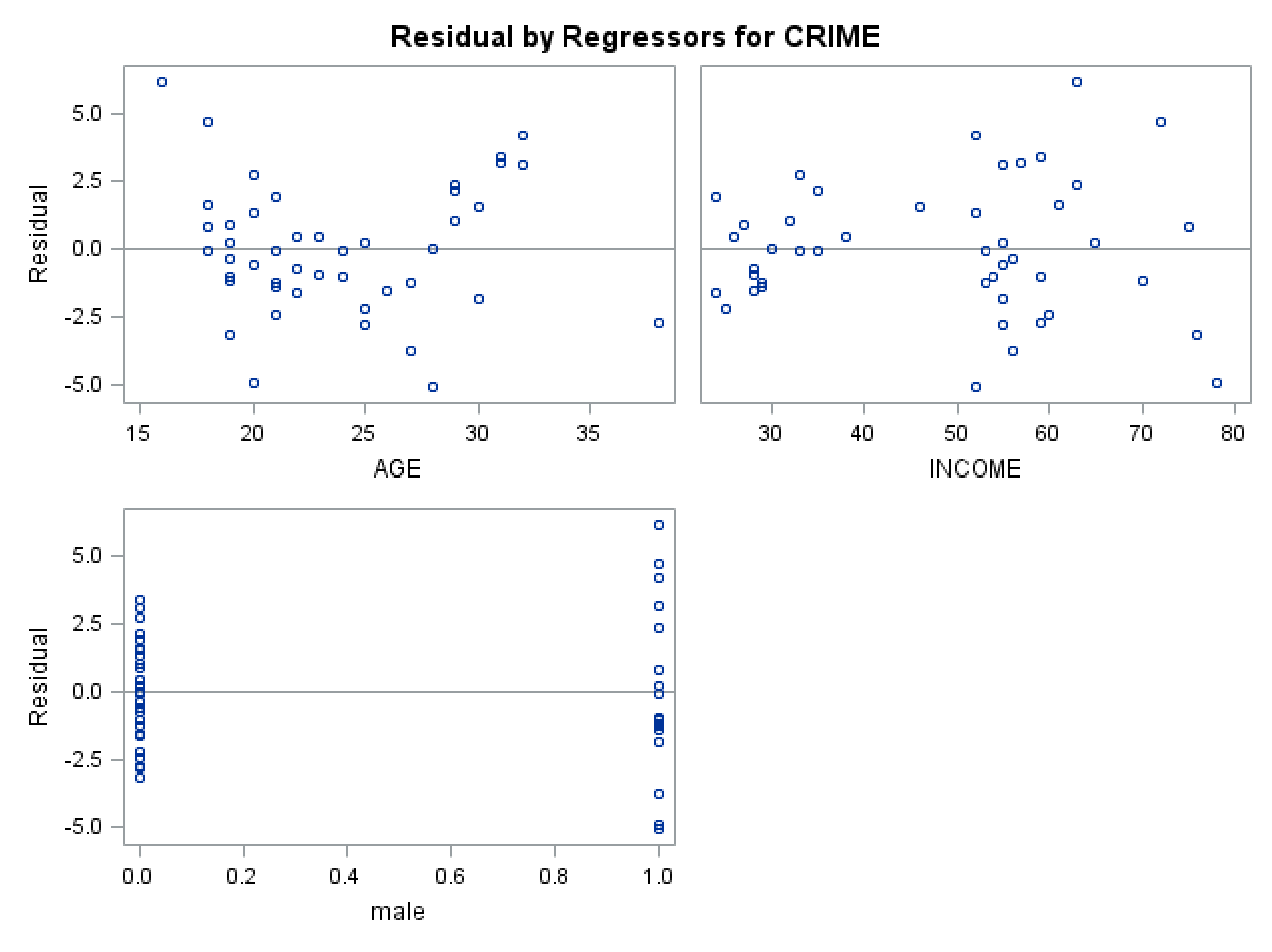
Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS			
						Intercept	AGE	INCOME	male
25	3.0782	1.3170	0.1132	1.0503	0.4705	-0.3256	0.3503	0.2045	-0.1964
26	3.1597	1.3569	0.1175	1.0447	0.4950	-0.3009	0.3497	0.0418	0.2700
27	1.5471	0.6357	0.0690	1.1389	0.1731	-0.0847	0.1187	0.0313	-0.0706
28	2.1419	0.8806	0.0617	1.0895	0.2259	-0.0352	0.1256	-0.0625	-0.0580
29	1.0006	0.4094	0.0669	1.1634	0.1097	-0.0090	0.0572	-0.0422	-0.0218
30	0.0195	0.007952	0.0643	1.1796	0.0021	0.0001	0.0008	-0.0010	-0.0004
31	-1.2478	-0.5097	0.0617	1.1463	-0.1307	-0.0253	-0.0374	0.0719	0.0225
32	-1.5150	-0.6197	0.0613	1.1317	-0.1583	-0.0510	-0.0253	0.0960	0.0258
33	-2.2099	-0.9133	0.0702	1.0930	-0.2509	-0.1167	-0.0056	0.1742	0.0283
34	-0.0530	-0.0214	0.0467	1.1579	-0.0047	-0.0023	0.0004	0.0023	0.0014
35	0.3970	0.1624	0.0694	1.1830	0.0443	0.0285	-0.0091	-0.0303	-0.0063
36	-0.9681	-0.4139	0.1453	1.2696	-0.1707	-0.0751	0.0143	0.1317	-0.1306
37	0.4090	0.1650	0.0436	1.1510	0.0352	0.0212	-0.0121	-0.0101	-0.0150
38	-1.5841	-0.6555	0.0815	1.1515	-0.1953	-0.1414	0.0591	0.1392	0.0213
39	-0.7289	-0.2979	0.0665	1.1720	-0.0795	-0.0566	0.0254	0.0501	0.0147
40	-1.3612	-0.5838	0.1468	1.2506	-0.2422	-0.1352	0.0579	0.1833	-0.1794
41	-0.0785	-0.0329	0.1149	1.2469	-0.0118	-0.0063	0.0030	0.0080	-0.0091
42	2.7331	1.1329	0.0655	1.0410	0.2999	0.2418	-0.1615	-0.1374	-0.0841
43	0.8968	0.3722	0.0934	1.2009	0.1195	0.1044	-0.0682	-0.0712	-0.0191
44	1.8624	0.7749	0.0877	1.1399	0.2402	0.1888	-0.0948	-0.1677	-0.0268
45	-1.2302	-0.5033	0.0646	1.1506	-0.1322	-0.0386	0.0346	0.0225	-0.0965

The REG Procedure
Model: MODEL1
Dependent Variable: CRIME



Fit Diagnostics for CRIME

Observations	45
Parameters	4
Error DF	41
MSE	6.2706
R-Square	0.8365
Adj R-Square	0.8245



Residual by Regressors for CRIME

Residual	DF	Sum of Squares	Mean Square	F Value	Pr > F
Lack of Fit	40	252.593125	6.314828	1.40	0.5964
Pure Error	1	4.500000	4.500000		
Total Error	41	257.093125	6.270564		