Name _____ Spring, 2006

Applied Statistics Comprehensive Examination

Regression Methods & Linear Models

Calculators are permitted on this part of the examination.

1. A study was done to investigate the role of firearms in homicides in Detroit from 1961 through 1973. The following variables are included in the analysis. The unit of observation is one year (n=13). (Note: all results were rounded to three decimal places in order to fit the output on one page).

   H = Number of homicides per 100,000 people (dependent variable)
   G = Number of government workers (in thousands)
   M = Number of manufacturing workers (in thousands)
   W = Number of White males
   LIC = Number of handgun licenses issued per 100,000 people
   GR = Number of handgun registrations issued per 100,000 people

   Output from <u>all</u> possible models of the main effects are below, including MSE, R^2, and adjusted-R^2 values. P-values for each parameter and variance inflation factors are also included. For example, model 6 shows $H = b_0 - 0.003$ LIC $+ 0.046$ GR. The p-values associated with LIC is 0.886 and for GR is 0.068. The variance inflation factor for LIC is 5.46 and for GR is 5.46.

   (a) Which model would you choose as the final model to predict homicides? Explain the process(es) you used.

(b) What additional models might you consider that are not included here? Explain what relationship you are modeling.

	MSE	Rsq	Adj-Rsq	G	M	W	LIC	GR
Model 1	97.73	0.666	0.636	.	.	.	.	0.043
p-value(s)	.	.	.	.	.	.	.	0.001
VIF(s)	.	.	.	.	.	.	.	1
Model 2	138.39	0.527	0.485	.	.	.	0.038	.
p-value(s)	.	.	.	.	.	.	0.005	.
VIF(s)	.	.	.	.	.	.	1	.
Model 3	30.27	0.897	0.887	.	.	0	.	.
p-value(s)	.	.	.	.	.	0	.	.
VIF(s)	.	.	.	.	.	1	.	.
Model 4	205.44	0.299	0.235	.	0.18	.	.	.
p-value(s)	.	.	.	.	0.053	.	.	.
VIF(s)	.	.	.	.	1	.	.	.
Model 5	24.06	0.918	0.91	0.424	.	.	.	.
p-value(s)	.	.	.	0	.	.	.	.
VIF(s)	.	.	.	1	.	.	.	.
Model 6	107.27	0.667	0.6	.	.	.	-0.003	0.046
p-value(s)	.	.	.	.	.	.	0.886	0.068
VIF(s)	.	.	.	.	.	.	5.46	5.46
Model 7	33.09	0.897	0.877	.	.	0	.	0.003
p-value(s)	.	.	.	.	.	0.001	.	0.807
VIF(s)	.	.	.	.	.	3.556	.	3.556
Model 8	33.14	0.897	0.877	.	.	0	-0.002	.
p-value(s)	.	.	.	.	.	0	0.832	.
VIF(s)	.	.	.	.	.	2.569	2.569	.
Model 9	106.89	0.668	0.602	.	0.018	.	.	0.041
p-value(s)	.	.	.	.	0.815	.	.	0.008
VIF(s)	.	.	.	.	1.65	.	.	1.65
Model 10	151.26	0.531	0.437	.	0.025	.	0.035	.
p-value(s)	.	.	.	.	0.805	.	0.05	.
VIF(s)	.	.	.	.	1.952	.	1.952	.
Model 11	8.61	0.973	0.968	.	-0.141	0	.	.
p-value(s)	.	.	.	.	0	0	.	.
VIF(s)	.	.	.	.	2.413	2.413	.	.
Model 12	26.34	0.918	0.902	0.439	.	.	.	-0.002
p-value(s)	.	.	.	0	.	.	.	0.832
VIF(s)	.	.	.	3.903	.	.	.	3.903
Model 13	24.72	0.923	0.908	0.468	.	.	-0.006	.
p-value(s)	.	.	.	0	.	.	0.421	.
VIF(s)	.	.	.	2.824	.	.	2.824	.
Model 14	26.38	0.918	0.902	0.384	.	0	.	.
p-value(s)	.	.	.	0.137	.	0.868	.	.
VIF(s)	.	.	.	35.145	.	35.145	.	.
Model 15	14.88	0.954	0.945	0.508	-0.089	.	.	.
p-value(s)	.	.	.	0	0.019	.	.	.
VIF(s)	.	.	.	2.015	2.015	.	.	.
Model 16	35.5	0.901	0.868	.	.	0	-0.007	0.009
p-value(s)	.	.	.	.	.	0.001	0.586	0.578
VIF(s)	.	.	.	.	.	3.572	5.485	7.591

	MSE	Rsq	Adj-Rsq	G	M	W	LIC	GR
Model 17	117.95	0.671	0.561	.	0.027	.	-0.006	0.046
p-value(s)	.	.	.	.	0.765	.	0.808	0.082
VIF(s)	.	.	.	.	1.952	.	6.457	5.46
Model 18	9.54	0.973	0.964	.	-0.141	0	.	0.001
p-value(s)	.	.	.	.	0.001	0	.	0.881
VIF(s)	.	.	.	.	2.422	5.219	.	3.57
Model 19	8.72	0.976	0.968	.	-0.148	0	0.004	.
p-value(s)	.	.	.	.	0	0	0.375	.
VIF(s)	.	.	.	.	2.573	3.388	2.74	.
Model 20	26.77	0.925	0.9	0.447	.	.	-0.01	0.006
p-value(s)	.	.	.	0	.	.	0.384	0.638
VIF(s)	.	.	.	3.952	.	.	5.529	7.641
Model 21	29.18	0.918	0.891	0.4	.	0	.	-0.002
p-value(s)	.	.	.	0.161	.	0.879	.	0.844
VIF(s)	.	.	.	38.598	.	35.169	.	3.906
Model 22	27.45	0.923	0.898	0.447	.	0	-0.006	.
p-value(s)	.	.	.	0.114	.	0.936	0.454	.
VIF(s)	.	.	.	39.07	.	35.544	2.856	.
Model 23	16.48	0.954	0.939	0.517	-0.088	.	.	-0.001
p-value(s)	.	.	.	0	0.027	.	.	0.877
VIF(s)	.	.	.	4.775	2.019	.	.	3.91
Model 24	16.43	0.954	0.939	0.516	-0.086	.	-0.001	.
p-value(s)	.	.	.	0	0.036	.	0.824	.
VIF(s)	.	.	.	3.215	2.222	.	3.114	.
Model 25	9.1	0.975	0.966	0.104	-0.133	0	.	.
p-value(s)	.	.	.	0.513	0.002	0.024	.	.
VIF(s)	.	.	.	42.234	2.9	50.569	.	.
Model 26	8.98	0.978	0.967	.	-0.157	0	0.009	-0.007
p-value(s)	.	.	.	.	0.001	0	0.247	0.414
VIF(s)	.	.	.	.	2.975	5.444	6.736	8.775
Model 27	30.1	0.925	0.888	0.435	.	0	-0.01	0.006
p-value(s)	.	.	.	0.145	.	0.963	0.419	0.662
VIF(s)	.	.	.	39.503	.	35.706	5.613	7.675
Model 28	18.49	0.954	0.931	0.516	-0.086	.	-0.002	0
p-value(s)	.	.	.	0	0.055	.	0.88	0.994
VIF(s)	.	.	.	4.796	2.369	.	6.487	8.145
Model 29	10.23	0.975	0.962	0.107	-0.133	0	.	0
p-value(s)	.	.	.	0.548	0.003	0.034	.	0.964
VIF(s)	.	.	.	46.436	2.914	50.756	.	3.925
Model 30	9.72	0.976	0.964	0.049	-0.143	0	0.003	.
p-value(s)	.	.	.	0.792	0.003	0.028	0.535	.
VIF(s)	.	.	.	54.511	3.59	57.433	3.537	.
Model 31	10.22	0.978	0.962	0.032	-0.154	0	0.008	-0.007
p-value(s)	.	.	.	0.867	0.005	0.029	0.348	0.46
VIF(s)	.	.	.	55.245	4.16	62.707	8.014	8.894

2.Short Answers.

(a) Name three ways of dealing with multicollinearity.

(b) A categorical variable with k categories is often analyzed in the context of regression with $k-1$ indicator variables. As an alternative, one could use k indicator variables instead. Explain the different interpretations of the model coefficients and hypothesis tests for these two choices ($k-1$ vs. k indicator variables).

(c) If $y = X\beta + \varepsilon$ and $\hat{\beta} = (X'X)^{-1}X'y$, then show that the sums of squares due to error (or residual) is $y'(I - H)y$.

(d) If $y = X\beta + \varepsilon$ and $\hat{\beta} = (X'X)^{-1}X'y$, show that $\mathrm{E}[\hat{\beta}] = \beta$ and find Var($\hat{\beta}$).

3. An experiment was conducted to determine the effect of three different pesticides and two different fertilizers on the yield of fruit from a citrus tree. Twelve trees were randomly selected from an orchard. Each of the six pesticide by fertilizer combinations were then randomly assigned to two of the trees. The yield of fruit, in bushels per tree, was obtained for each tree after the test period. The mean yield for each pesticide by fertilizer combination is given in the following table:

	Fertilizer	
Pesticide	1	2
1	44	48
2	52	62
3	40	46

A partially completed ANOVA table is given below:

Soucre	df	SS	MS
Pesticide			217.3
Fertilizer		133.3	
Interaction		18.7	
Error			
Total		852.7	

(a) For this design, write out the mathematical model for both the effects model and the cell means model. Be sure to explain all of the terms for both of the models and state the model assumptions.

(b) Use this design with only one observation per cell to write out the design matrix using the sum-to-zero restrictions.

(c) Create an interaction plot and discuss the presence/type of interaction. Discuss the effect of this type of interaction on the ability of the researcher to test the main effects individually.

(d) Fill out the partially completed ANOVA table and test the significance of the interaction.

(e) For the effects model, write out the general form of the estimable functions and provide a basis set of estimable functions. Is $\gamma_{11} - \gamma_{12} - \gamma_{21} - \gamma_{22}$ estimable? Explain.