Please put your name on the back of your answer book.
Do NOT put it on the front. Thanks.

Do not start until I tell you to.

- The exam is closed book, closed notes. Use only the formula sheet and tables I provide today. You may use a calculator.
- Write your answers in your blue book. Ask if you need a second (or third) blue book.
- You have 2:15 hours (135 minutes) to complete the exam.
  Stop working when the end of the exam is announced.
- Points are indicated for each question. There are 130 total points.
- Important reminders:
  - budget your time. Some parts of each question should be easy; others may be hard. Make sure you do all parts you can.
  - notice that some parts do not require any computations.
  - show your work neatly so you can receive partial credit.
- Good luck!

1. 50 points. Health of factory workers. The following data were collected in a study of the health of paint sprayers in an auto assembly plant. Two of the variables that were measured on each of the 103 workers in the study were H, the haemoglobin concentration, and L, the lymphocyte count. These are measures of two different components of the blood.

   The following quantities may help you answer the questions:
   The observed intercept and slope in the regression $H_i = \beta_0 + \beta_1 L_i + \epsilon_i$ are $b_0 = -55.6$, $b_1 = 1.98$
   The estimated s.d. of observations around the line is $s_e = 4.95$
   The error SS for the "intercept only" model $H_i = \mu + \epsilon_i$ is 8050.2
   The error SS for the regression $H_i = \beta_0 + \beta_1 L_i + \epsilon_i$ is 2474.2
   The error SS for the regression $H_i = \beta_0 + \beta_1 L_i + \beta_2 L_i^2 + \epsilon_i$ is 2470.8
   The error SS for the regression $H_i = \beta_0 + \beta_1 L_i + \beta_2 L_i^2 + \beta_3 L_i^3 + \epsilon_i$ is 2396.4
   The error SS for the loess regression $H_i = f(L_i) + \epsilon_i$ is 2391.4 with 97.5 d.f.
   The mean lymphocyte count is 30.9.
   The sum-of-squares of lymphocyte counts, $\sum(x_i - \overline{x})^2$, $= 1428$.
   The correlation coefficient between H and L is 0.838.

   (a) What statistic is the most appropriate to describe the association between haemoglobin concentration and lymphocyte count? You may answer with one of the values I've provided, or some other statistic. Briefly explain why you chose your statistic.

   No matter how you answered the previous question, the investigators want you to fit the regression: $H_i = \beta_0 + \beta_1 L_i + \epsilon_i$.

   (b) Calculate the s.e. of b_1

   (c) Test H0: $\beta_1 = 0$. Report your test statistic and two-sided p-value.
   Note: If you were not able to do the previous question and need a s.e. for your test, use s.e.= 0.49.

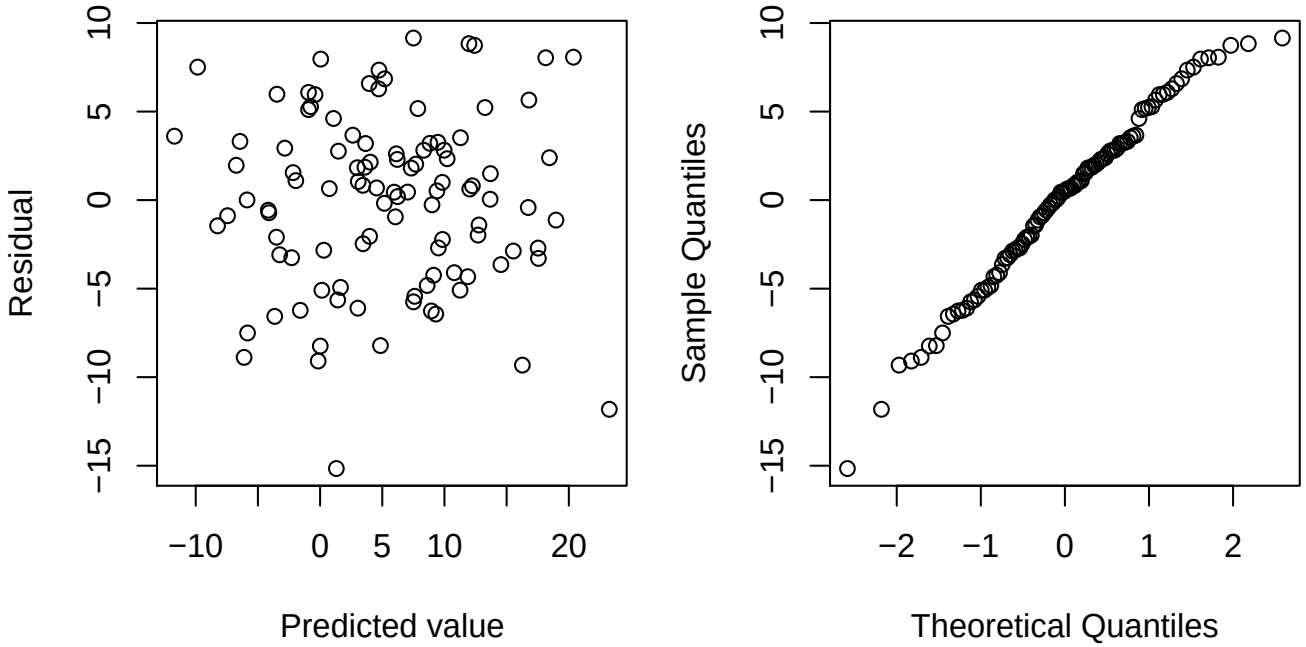
   (d) The usual ANOVA table for this regression has rows and columns labelled:

Source	d.f.	SS	MS	F
Model	??	??	??	??
Error	??	??	??	
Total	??	??		

   Calculate as many of the missing entries as you can from the available data and what you know about the study.

   (e) The investigators use the fitted regression to predict average haemoglobin concentration at three possible lymphocyte counts: $L_i = 26$, $L_i = 32$, and $L_i = 35$. Which prediction is the most precise? Explain your choice.

(f) Here are a residual plot and a normal quantile-quantile plot for the fitted regression. List the assumptions made in the regression, then assess each using the information in the plots.



(g) The investigators want to know whether the relationship between haemoglobin concentration and lymphocyte count follows a straight line. Test this, if possible from the available information. Provide a test statistic, p-value and short conclusions. If not possible, say what additional information you need.

(h) There are a few additional workers for whom the lymphocyte count was not measured. One of those workers had a haemoglobin concentration of 5.4.

If it is possible given the available data, estimate the lympocyte count for that individual.

(i) The investigators are concerned about collecting redundant data. If the correlation between haemoglobin and lymphocyte exceeds 0.8, they will consider collecting only one variable, instead of two. Test whether the correlation is larger than 0.8, i.e. test H0: $\rho \leq 0.8$ vs Ha: $\rho > 0.8$. Provide your test statistic, p-value and a short conclusion.

2. 55 points. Estimating demand for gasoline. Econometricians are often interested in estimating demand curves. A demand curves describes the relationship between price and consumption of a specific product (demand). If the relationship is linear, the deman curve is described by the slope coefficient for price..

   Estimating such curves correctly involves many issues well beyond what we have considered in this class. You are to use the regression tools we have considered in this class.

   The following data are from a study to estimate the relationship between gasoline price and consumption. The data are the annual data on the US economy from 1960 to 1986. There are twenty seven (27) observations in the data set, one for each year. An economic detail: all variables are adjusted for the effects of inflation. The prices and income look small, but that is because they are expressed in 1967 dollars. The variables in the data set are:

Variable	explanation	units
gascons	total gasoline consumption	million barrels
gasprice	price index for gasoline	dollars
income	per capita disposable income	dollars
newcar	price index for new cars	dollars
usedcar	price index for used cars	dollars
bus	price index for public transportation	dollars
durable	price index for durable goods	dollars
nondur	price index for non-durable goods	dollars
service	price index for services	dollars
year	calendar year	
year2	year squared	

   The goal of the study is to estimate the relationship between gasoline price and gasoline consumption. Is an increase in gas price associated with a change in consumption? If so, how large is the effect?

   Summary statistics for each of the 11 variables (10 X variables and *gascons*, the response) are:

Variable	N	Mean	Std Dev	Minimum	Maximum
year	27	73.0000000	7.9372539	60.0000000	86.0000000
gascons	27	207.0333333	43.7989287	129.7000000	269.4000000
gasprice	27	1.9021111	1.1679056	0.9140000	4.1090000
income	27	8513.52	1455.63	6036.00	10780.00
newcar	27	1.3813333	0.4279747	0.9910000	2.2400000
usedcare	27	1.7122963	0.9747435	0.8360000	3.7970000
bus	27	1.8623333	1.1083126	0.8100000	4.2640000
durable	27	0.6748889	0.2231702	0.4440000	1.0530000
nondur	27	0.6112593	0.2731639	0.3310000	1.0750000
service	27	0.6008148	0.3026135	0.3020000	1.2240000
year2	27	5389.67	1160.15	3600.00	7396.00

   A scatterplot matrix of the data is on page 7.

The investigators considered 5 models. The error SS are included for each model

SS_{error}		model equation	
127.29	$gascons =$	$\beta_0 + \beta_1 gasprice + \beta_2 year + \beta_4 income + \beta_5 newcar + \beta_6 usedcar + \beta_7 bus + \beta_8 durable + \beta_9 nondur + \beta_{10} service + \epsilon$	(1)
207.64	$gascons =$	$\beta_0 + \beta_1 gasprice + \beta_2 year + \beta_3 year^2 + \beta_4 income + \beta_5 newcar + \beta_6 usedcar + \beta_7 bus + \beta_8 durable + \beta_9 nondur + \beta_{10} service + \epsilon$	(2)
800.96	$gascons =$	$\beta_0 + \beta_1 gasprice + \beta_4 income + \beta_6 usedcar + \epsilon$	(3)
841.25	$gascons =$	$\beta_0 + \beta_1 gasprice + \beta_4 income + \beta_7 bus + \epsilon$	(4)
607.42	$gascons =$	$\beta_0 + \beta_1 gasprice + \beta_2 year + \beta_3 year^2 + \beta_{10} service + \epsilon$	(5)

Please use these results to answer the following:

(a) The estimate $\hat{\beta}_1$ in model 2 is -23.65. Please give a careful interpretation of this value.

(b) Construct a test of whether model 2 fits significantly better than model 3, if this is possible from the available information. If not, explain what other information is needed.

(c) Explain, in a way understandable by a non-statistician, the hypothesis being tested by the comparison of models 2 and 3.

(d) Construct a test of whether model 5 fits significantly better than model 4, if this is possible from the available information. If not, explain what other information is needed.

The remaining seven questions concern model 3:

$$gascons = \beta_0 + \beta_1 gasprice + \beta_4 income + \beta_6 usedcar + \epsilon.$$

SAS output with additional information for this model is included at the end of the exam questions. This information includes parameter estimates, numeric values of diagnostics, and results from Breusch-Pagan and Durbin-Watson tests.

(e) Use model 3 to predict gas consumption when GASPRICE = 1.10, USEDCAR=0.90, and INCOME = 7000.

(f) Estimate the standard deviation of a predicted **observation** at GASPRICE = 1.10, USEDCAR=0.90, and INCOME = 7000. The appropriate value of $x_i(X'X)^{-1}x_i'$ is 0.0830.

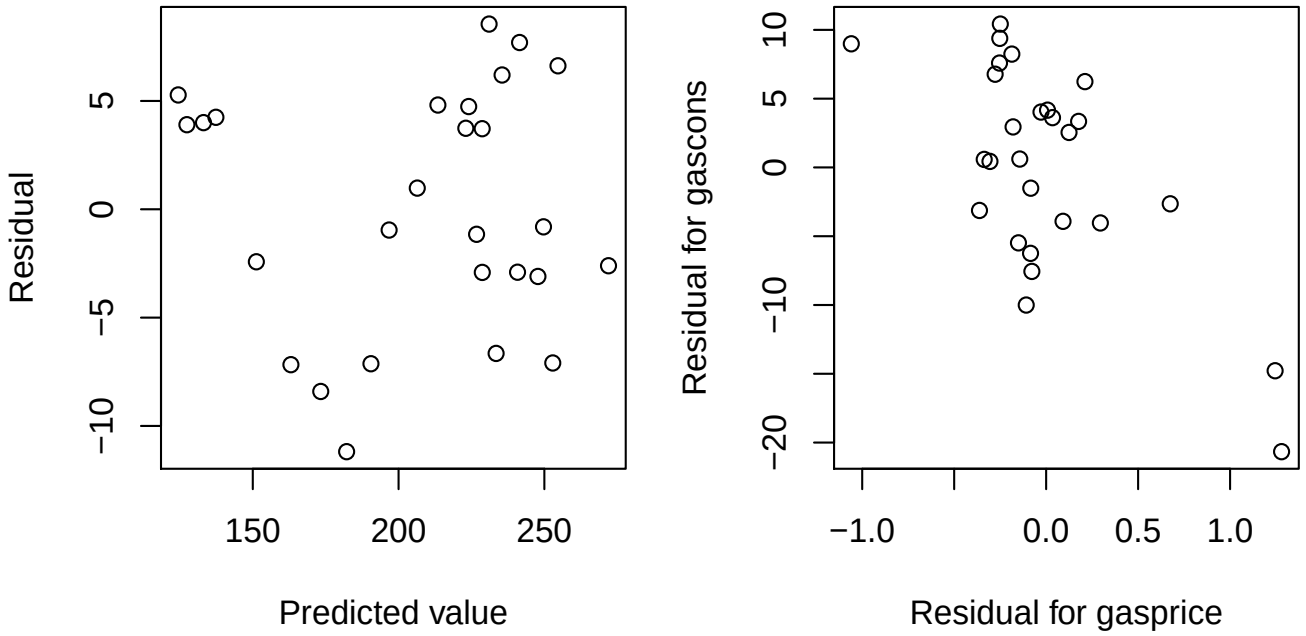
(g) The plot of residuals vs. predicted values (below part 2j) suggests a problem with lack of fit. We have ignored violations of some assumptions because they have little effect on the desired answers. If there is lack of fit, do you have any concerns about the prediction and standard deviation you calculated in previous two parts? If so, briefly describe your concerns.

(h) Do you have any concerns with multicollinearity? Explain why or why not.

(i) The primary goal of the study is to estimate β_1, the partial regression coefficient for gas price. Do you have any concerns about the effects of outlying observations? Briefly explain your concerns. If you don't have any, just say 'none'.

For reference, $F_{4,23,0.2} = 0.41$ and $F_{4,23,0.5} = 0.86$.

(j) Below is the partial regression residual plot (sometimes called the added variable plot) for *gasprice*. Do you have any concerns about the linearity of the relationship with *gasprice*? Explain why or why not.



(k) If appropriate, test for first-order autocorrelation in the errors. Report your test statistic and a p-value. If not appropriate, say explain why a test of autocorrelation is not appropriate for these data.

3. 12 points. An adjuvant is a substance added to a vaccine to improve the immune response and reduce the amount of vaccine needed to provide protection. Imagine a study of Y= the vaccine RESPONSE at five DOSEs of vaccine with and without the ADJUVANT.

(a) You are told that:
the relationship between DOSE and RESPONSE is assumed to be linear.
The effect of the adjuvant is the same at all doses.

Write down an appropriate model for the relationship between RESPONSE and both DOSE and ADJUVANT. Describe briefly the variables in your model.

(b) Using the parameters in your model, what parameter or function of parameters describes the increase in vaccine response caused by adding the adjuvant?

(c) You now told that the biologists believe that the effect of adjuvant increases with the vaccine dose. Write down an appropriate model that that allows this. Describe any new variables used in this model.

(d) Describe how you could test whether the effect of the adjuvant is constant over doses or increases with dose.

4. 13 points. A study evaluated the relationship between N_i, the number of CPU's in a parallel computing environment, and Y_i, the time to solve a specific hard problem. A reasonable model for the relationship is:

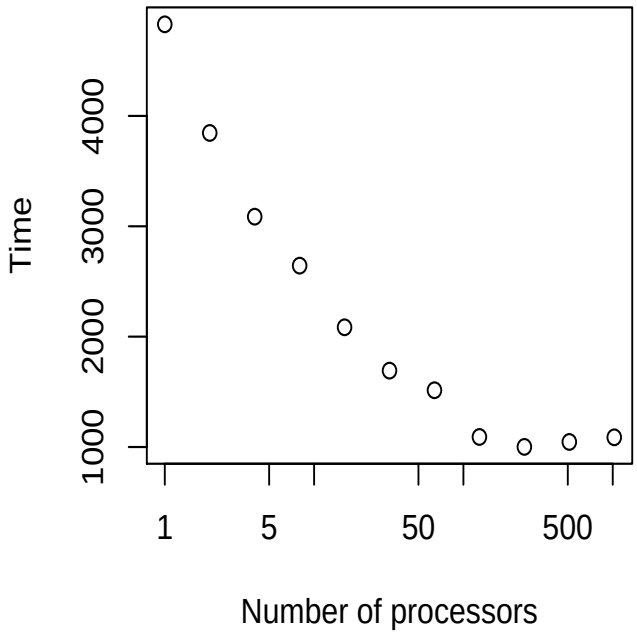
$$Y_i = \beta_0 + \beta_1 \log N_i + \beta_2 (\log N_i)^2 + \varepsilon_i.$$

Data were collected for 11 numbers of processors (1, 2, 4, 8, ... 512, 1024). The model was fit to centered X's, i.e.:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i,$$

where $X_i = \log N_i - \log 32$. Note: all logarithms are base e, i.e. natural logarithms. After centering, the mean $X_i = 0$ and Cov $(b_1, b_2) = 0$.

The data are plotted below:



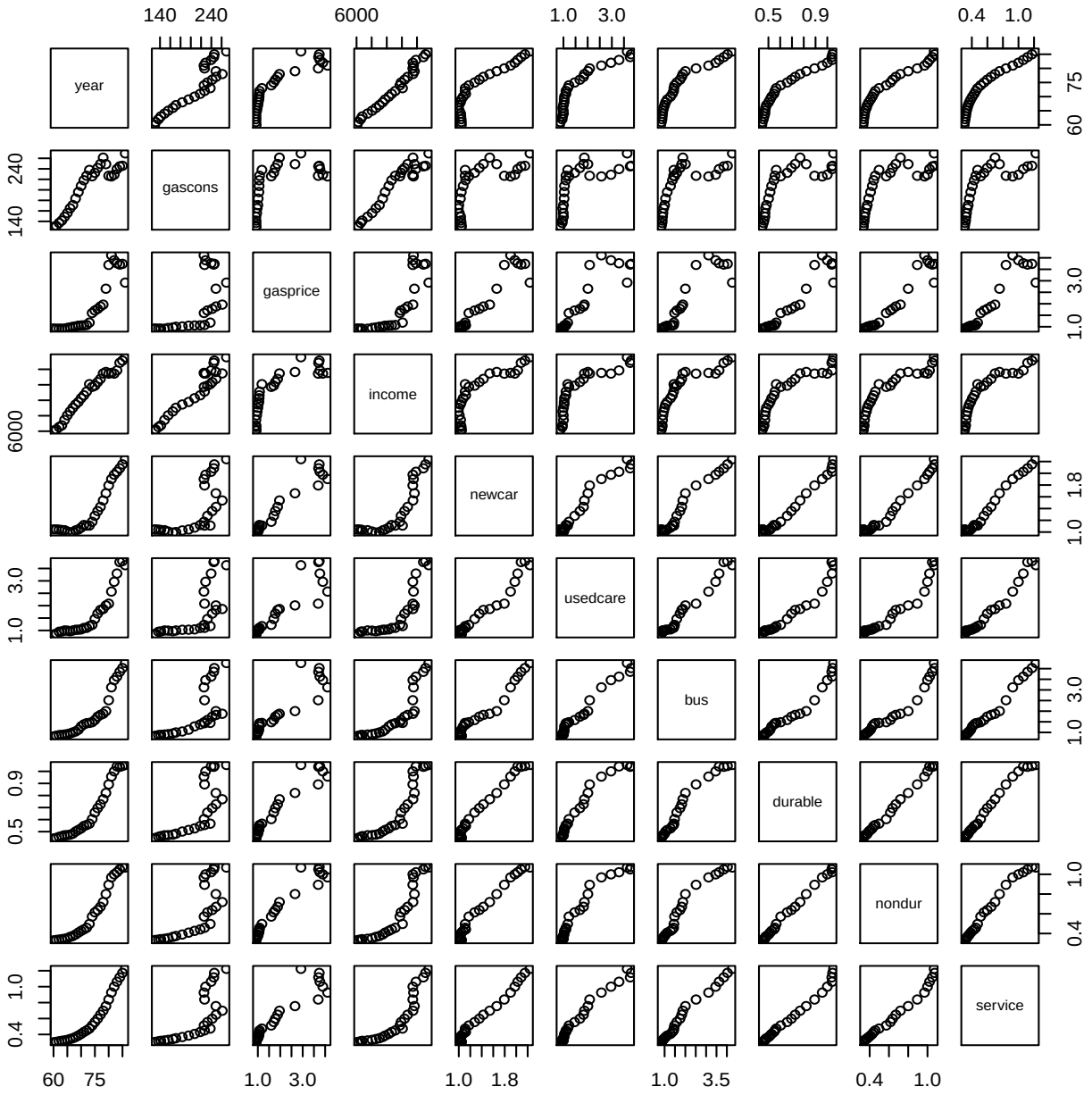
Here is some summary information about the regression:

Coefficient	Estimate	s.e.
b0	1681.97	39.87
b1	-522.57	12.04
b2	102.64	6.22

(a) Estimate the value of X at which the time is minimized.

(b) Estimate the number of processors at which the time is minimized.

(c) Prior to collecting the data, the investigators expected 350 processors to minimize the time to solve the problem. Test whether the data are consistent with the minimum being at 350 processors, if this is possible with the data available. Report your test statistic and a p-value. If not possible, state what additional information you need.


year
gascons
gasprice
income
newcar
usedcare
bus
durable
nondur
service
140 240
6000
1.0 3.0
0.5 0.9
0.4 1.0
60 75
1.0 3.0
1.0 1.8
1.0 3.5
0.4 1.0

The REG Procedure
Model: MODEL3
Dependent Variable: gascons

Number of Observations Read 27
Number of Observations Used 27

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	49076	16359	469.75	<.0001
Error	23	800.96554	34.82459		
Corrected Total	26	49877			

Root MSE	5.90124	R-Square	0.9839
Dependent Mean	207.03333	Adj R-Sq	0.9818
Coeff Var	2.85038		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-103.50606	9.68365	-10.69	<.0001	0
gasprice	1	-10.94542	2.47158	-4.43	0.0002	6.22086
income	1	0.04058	0.00147	27.52	<.0001	3.44073
usedcar	1	-8.25978	3.11278	-2.65	0.0142	6.87331

Durbin-Watson D	0.738
Pr < DW	<.0001
Pr > DW	1.0000
Number of Observations	27
1st Order Autocorrelation	0.610

NOTE: Pr<DW is the p-value for testing positive autocorrelation, and
Pr>DW is the p-value for testing negative autocorrelation.

The REG Procedure
Model: MODEL3
Dependent Variable: gascons

Output Statistics

Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS
1	5.2783	0.9918	0.1873	1.2339	0.4761
2	3.9056	0.7216	0.1764	1.3209	0.3340
3	4.0007	0.7319	0.1594	1.2908	0.3188
4	4.2466	0.7713	0.1449	1.2557	0.3176
5	-2.4219	-0.4265	0.1068	1.2941	-0.1475
6	-7.1715	-1.2848	0.0800	0.9722	-0.3788
7	-8.4073	-1.5152	0.0661	0.8600	-0.4032
8	-11.1869	-2.0948	0.0604	0.6142	-0.5312
9	-7.1277	-1.2619	0.0602	0.9612	-0.3194
10	-0.9567	-0.1639	0.0624	1.2679	-0.0423
11	0.9793	0.1687	0.0729	1.2819	0.0473
12	4.8137	0.8462	0.0823	1.1452	0.2534
13	3.7409	0.6612	0.1034	1.2316	0.2246
14	-2.9031	-0.5249	0.1491	1.3358	-0.2197
15	-2.9130	-0.5125	0.1021	1.2688	-0.1728
16	3.7210	0.6462	0.0718	1.1938	0.1798
17	6.1994	1.0954	0.0723	1.0412	0.3058
18	7.6974	1.3864	0.0793	0.9281	0.4070
19	6.6233	1.2024	0.1119	1.0427	0.4268
20	-0.8086	-0.1420	0.1086	1.3352	-0.0496
21	-6.6513	-1.4453	0.3631	1.3049	-1.0913
22	-1.1527	-0.2356	0.3407	1.7938	-0.1693
23	4.7423	0.8873	0.1872	1.2770	0.4259
24	8.5504	1.6426	0.1644	0.9001	0.7287
25	-3.1001	-0.5980	0.2497	1.4928	-0.3450
26	-7.0931	-1.4214	0.2532	1.1256	-0.8277
27	-2.6049	-0.5538	0.3839	1.8345	-0.4372

Output Statistics

	DFBETAS			
Obs	Intercept	gasprice	income	usedcar
1	0.4341	0.0813	-0.3796	0.0790
2	0.3022	0.0414	-0.2635	0.0677
3	0.2831	0.0116	-0.2461	0.0880
4	0.2756	0.0026	-0.2368	0.0890
5	-0.1136	0.0158	0.0923	-0.0434
6	-0.2428	0.0476	0.1763	-0.0712
7	-0.1851	0.0511	0.1030	-0.0156
8	-0.1493	0.0981	0.0356	0.0122
9	-0.0295	0.0820	-0.0393	0.0204
10	0.0023	0.0102	-0.0113	0.0067
11	-0.0122	-0.0132	0.0218	-0.0104
12	-0.0930	-0.0938	0.1403	-0.0427
13	-0.1130	-0.0811	0.1515	-0.0514
14	0.1464	0.0728	-0.1776	0.0649
15	0.0908	-0.0208	-0.1172	0.1071
16	-0.0932	-0.0079	0.1187	-0.0758
17	-0.1758	-0.0889	0.2096	-0.0493
18	-0.2479	-0.1508	0.2824	-0.0227
19	-0.3114	-0.1348	0.3425	-0.0698
20	0.0270	-0.0186	-0.0291	0.0298
21	0.0826	-0.9709	-0.0894	0.8426
22	-0.0170	-0.1513	0.0229	0.0985
23	0.1097	0.2781	-0.1458	-0.0606
24	0.1831	0.1587	-0.2716	0.2409
25	-0.0344	0.0974	0.0798	-0.2396
26	-0.0403	0.2497	0.1500	-0.5678
27	0.0920	0.3130	-0.0548	-0.3429

Sum of Residuals	0
Sum of Squared Residuals	800.96554
Predicted Residual SS (PRESS)	1092.78795

Information for Breusch-Pagan test

The REG Procedure
Model: MODEL3
Dependent Variable: esq

Number of Observations Read	27
Number of Observations Used	27

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.01491	0.01491	0.01	0.9042
Error	25	25.22218	1.00889		
Corrected Total	26	25.23709			

Root MSE	1.00443	R-Square	0.0006
Dependent Mean	1.00001	Adj R-Sq	-0.0394
Coeff Var	100.44268		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.03901	0.37456	2.77	0.0103
gasprice	1	-0.02050	0.16867	-0.12	0.9042