

TEST 2 STAT 572

Spring 1999

1. A study was conducted to investigate the factors most associated with gasoline consumption. Data were obtained for the 48 contiguous states. After some preliminary analysis, the model included the following variables:

GAS: Total gasoline and auto diesel consumption in 10^{12} BTU
 AREA: Of state in 1000 miles
 POP: 1970 population in millions
 MV: Estimated number of motor vehicles in millions
 reg: Region (0 for East and 1 for West of Mississippi river)

SAS results appear below.

The REG Procedure

Model: MODEL1
 Dependent Variable: GAS

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	3177687	794422	1856.17	<.0001
Error	43	18404	427.99001		
Corrected Total	47	3196091			

Root MSE	20.68792	R-Square	0.9942
Dependent Mean	277.23896	Adj R-Sq	0.9937
Coeff Var	7.46213		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
Intercept	1	-9.10118	5.54276	-1.64	0.1079	0
AREA	1	0.46334	0.11038	4.20	0.0001	2.93211
POP	1	1.36976	3.36654	0.41	0.6861	23.67026
MV	1	108.75460	6.73510	16.15	<.0001	25.44253
reg	1	-24.57555	9.53711	-2.58	0.0135	2.51039

Collinearity Diagnostics(intercept adjusted)

Number	Eigenvalue	Condition Index	-----Proportion of Variation-----			
			AREA	POP	MV	reg
1	2.08534	1.00000	0.01273	0.00917	0.00882	0.00032083
2	1.67859	1.11459	0.08896	0.00053149	0.00016212	0.12675
3	0.21560	3.11004	0.75226	0.01004	0.00180	0.85605
4	0.02047	10.09215	0.14605	0.98026	0.98922	0.01688

a) Comment on the characteristics of the model (excluding collinearity).

b) Comment on the collinearity of the data.

c) What seems the most logical action to improve the collinearity level in this situation?

d) Estimate the amount of variability in MV that is explained by the other regressors.

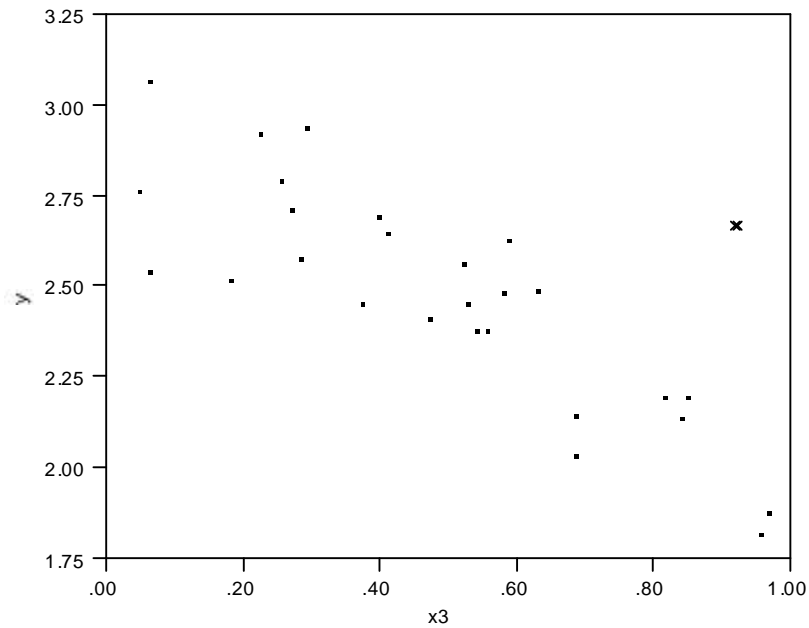
The following are some characteristics of the three highest leverage points

State	Residual	Leverage	Dfbetas			
			AREA	POP	MV	Reg
California	19.18	.4372	-.20	-.20	.41	.39
New York	-14.32	.8135	-.54	-2.14	2.78	.04
Texas	20.25	.5521	1.25	-.20	.17	-.74

e) For each state interpret the largest Dfbetas.

f) The values for the regressor reg are 0 or 1. Why is it that not all the eastern (or western) states have the same effect on this coefficient according to Dfbetas?

2. The following is a **partial residual plot** in a model that contains four regressors x_1 , x_2 , x_3 , and x_4 .



a) Explain the effect of the point marked with an X on the value of the corresponding coefficient and its p-value.

b) How can this point be discovered using influence diagnostics? Will the value of the diagnostic be positive or negative?

3. Determine the truthfulness or falseness of the following statements. Carefully justify your answers and explain any assumptions made.

T F The model found by stepwise methods depends on the procedure used (backward, forward, stepwise)

T F Autocorrelation causes the OLS estimators of the coefficients to be biased

T F A collinearity-influential observation usually has a large DFFITS

T F $SSE(\text{full model}) > SSE(\text{restricted model})$

T F Collinearity is harmful to the fit of the model (R^2)

T F Collinearity is harmful to prediction.

T F Collinearity causes the estimators to be "unstable."

T F If there are interactions among the regressors, some of the correlations among them will be significant