8.31 Correlated regression errors. What are the consequences of running least squares regression when the errors are correlated?

Ans: Inflated *t*-statistics for testing and model parameters

8.34 Forecasting car sales. Forecasts of automotive vehicle sales in the United States provide the basis for financial and strategic planning of large automotive corporations. The following forecasting model was developed for *y*, total monthly passenger car and light truck sales (in thousands):

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

where

x_1 = Average monthly retail price of regular gasoline
x_2 = Annual percentage change in GNP per quarter
x_3 = Monthly consumer confidence index
x_4 = Total number of vehicles scrapped (millions) per month
x_5 = Vehicle seasonality

The model was fitted to monthly data collected over a 12-year period (i.e., $n = 144$ months) with the following results:

$\hat{y} = -676.42 - 19.3x_1 + 6.54x_2 + 2.02x_3 + .08x_4 + 9.82x_5$
$R^2 = .856$
Durbin–Watson $d = 1.01$

(a) Is there sufficient evidence to indicate that the model contributes information for the prediction of *y*? Test using $\alpha = .05$.

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0, H_a:$ At least one $\beta_i \neq 0$

$$F = \frac{R^2/k}{(1-R^2)/[n-(k+1)]} = \frac{0.856/5}{(1-0.856)/[144-(5+1)]} = \frac{0.1712}{0.00104} = 164.067,$$

$F_{.05,5,\infty} = 2.57 < 164.067$. We reject null hypothesis and conclude that the model contributes information for the prediction of *y*.

Ans: There is sufficient evidence

(b) Is there sufficient evidence to indicate that the regression errors are positively correlated? Test using $\alpha = .05$.

$H_0:$ No residual correlation; $H_a:$ Positive residual correlation,

$n = 144, k = 5, d = 1.01, d_{L,\alpha} = 1.57$

Because $d < d_{L,\alpha}$, we reject null hypothesis and conclude that the residuals of the straight-line model for sales are positively correlated.

Ans: positively correlated.

(c) Comment on the validity of the inference concerning model adequacy in light of the result of part (b).

Ans: The residual correlation can be taken into account in a time series model, thereby improving both the fit of the model and the reliability of model inferences.

6.1 Selecting the best one-variable predictor. There are six independent variables, x_1, x_2, x_3, x_4, x_5, and x_6, that might be useful in predicting a response y. A total of $n = 50$ observations are available, and it is decided to employ stepwise regression to help in selecting the independent variables that appear to be useful. The computer fits all possible one-variable models of the form $E(y) = \beta_0 + \beta_1 x_1$, where x_i is the ith independent variable, $i = 1, 2, \ldots, 6$. The information in the table is provided from the computer printout.

INDEPENDENT VARIABLE	$\hat{\beta}_i$	$s_{\hat{\beta}_i}$
x_1	1.6	.42
x_2	−.9	.01
x_3	3.4	1.14
x_4	2.5	2.06
x_5	−4.4	.73
x_6	.3	.35

(a) Which independent variable is declared the best one-variable predictor of y? Explain.

$$t_{x_1} = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{1.6}{.42} = 3.8095, t_{x_2} = \frac{\hat{\beta}_2}{s_{\hat{\beta}_2}} = \frac{-.9}{.01} = -90, t_{x_3} = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{3.4}{1.14} = 2.9825$$

$$t_{x_4} = \frac{\hat{\beta}_4}{s_{\hat{\beta}_4}} = \frac{2.5}{2.06} = 1.2136, t_{x_5} = \frac{\hat{\beta}_5}{s_{\hat{\beta}_5}} = \frac{-4.4}{.73} = -6.0274, t_{x_6} = \frac{\hat{\beta}_6}{s_{\hat{\beta}_6}} = \frac{.3}{.35} = 0.8571$$

Ans: x_2 because it produces the highest absolute *t*-value

(b) Would this variable be included in the model at this stage? Explain.

Ans: Yes. It would be included until adding other terms makes the *t*-value insignificant

(c) Describe the next phase that a stepwise procedure would execute.

Fitting all two-variable models containing x_2 and each of the other options for the second variable (i.e., x_1, x_3, x_4, x_5, and x_6) in the form of $E(y) = \beta_0 + \beta_1 x_2 + \beta_2 x_j$. The t-values for the test $H_0: \beta_2 = 0$ are computed for each of the five models, and the variable having the largest t is retained.

Ans: fit all possible 2-variable models, $E(y) = \beta_0 + \beta_1 x_2 + \beta_2 x_j$

6.5 Modeling marine life in the gulf. A marine biologist was hired by the EPA to determine whether the hot-water runoff from a particular power plant located near a large gulf is having an adverse effect on the marine life in the area. The biologist's goal is to acquire a prediction equation for the number of marine animals located at certain designed areas, or stations, in the gulf. Based on past experience, the EPA considered the following environmental factors as predictors for the number of animals at a particular station:

x_1 = Temperature of water (TEMP)
x_2 = Salinity of water (SAL)
x_3 = Dissolved oxygen content of water (DO)
x_4 = Turbidity index, a measure of the turbidity of the water (TI)
x_5 = Depth of the water at the station (ST_DEPTH)
x_6 = Total weight of sea grasses in sampled area (TGRSWT)

As a preliminary step in the construction of this model, the biologist used a stepwise regression procedure to identity the most important of these six variables. A total of 716 samples were taken at different stations in the gulf, producing the SAS printout shown at the top of the next page. (The response measured was *y*, the logarithm of the number of marine animals found in the sampled area.)

SAS Output for Exercise 6.5

The REG Procedure
Model: MODEL1
Dependent Variable: LOGNUM

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	ST_DEPTH		1	0.1223	0.1223	51.57	99.47	<.0001
2	TGRSWT		2	0.0924	0.1821	1.52	79.38	<.0001
3	TI		3	0.0368	0.1870	3.51	54.59	<.0001
4	DO		4	0.0134	0.1889	1.03	41.40	<.0001
5		DO	3	0.0368	0.1870	3.51	54.59	<.0001

(a) According to the SAS printout, which of the six independent variables should be used in the model?

Ans: x_5 (ST_DEPTH), x_6 (TGRSWT), and x_4 (TI)

(b) Are we able to assume that the marine biologist has identified all the important independent variables for the prediction of *y*? Why?

Ans: No, because x_1 and x_2 were not tested in stepwise selection

(c) Using the variables identified in part (a), write the first-order model with interaction that may be used to predict *y*.

Ans: $E(y) = \beta_0 + \beta_1 x_4 + \beta_2 x_5 + \beta_3 x_6 + \beta_4 x_4 x_5 + \beta_5 x_4 x_6 + \beta_6 x_5 x_6$

(d) How would the marine biologist determine whether the model specified in part (c) is better than the first-order model?

Ans: nested *F*-test of $H_0: \beta_4 = \beta_5 = 6_6 = 0$

(e) Note the small value of R^2. What action might the biologist take to improve the model?

Ans: consider interaction and higher-order terms