SYSTEMS 302

LECTURE 24

- **AUTOCORRELATION PROBLEM**
  - **Durbin-Watson Test**
  - **Two-Stage Regression Approach**
  - **Auxiliary-Variable Approach**
- **REGRESSION OUTLIERS**
  - **Cook's Distance**
- **For next time:**
  - **Logistic Regression Notes**

SALES FORCASTING PROBLEM

One of the oldest 'economic laws' is that *increased income leads to increased expenditures*. This time-honored relation can be tested for the U.S. by regressing retail sales against per capita income for a number of years.

DATA: In a study of $T = 15$ years, data was collected from 1965 to 1980 on per capita retail sales ($SALES_t : t = 1,..,T$), per capita income ($PCI_t : t = 1,..,T$), and the unemployment rate ($UR_t : t = 1,..,T$) the U.S.

ANALYSIS: Consider the two regression models

$$SALES_t = \beta_0 + \beta_1 PCI_t + \varepsilon_t \ , \quad t = 1,..,T$$

$$SALES_t = \beta_0 + \beta_1 PCI_t + \beta_2 UR_t + \varepsilon_t \ , \quad t = 1,..,T$$

GENERAL TWO-STAGE ESTIMATION

MULTIVARIATE LINEAR MODEL

(1) $$y_t = \beta_0 + \sum_{i=1}^{k} \beta_i x_{it} + \varepsilon_t , \quad t = 1,..,T$$

(2) $$\varepsilon_t = \rho \varepsilon_{t-1} + u_t , \quad t = 2,..,T$$

STEP 1. **Do a *multiple regression* to estimate (1). If *P-value* in *Durbin-Watson test* is small (say < .05), continue.**

STEP 2. Set $\hat{\rho}$ equal to the *Autocorrelation value* in JMPIN

STEP 3. Estimate the *transformed variables*:

(3) $$\hat{z}_t = y_t - \hat{\rho} y_{t-1} , \quad t = 2,..,T$$

(4) $$\hat{w}_{it} = x_{it} - \hat{\rho} x_{i,t-1} , \quad t = 2,..,T , \; i = 1,..,k$$

STEP 4. Do multiple regression to estimate *new linear model*:

(5) $$\hat{z}_t = \alpha_0 + \sum_{i=1}^{k} \beta_i \hat{w}_{it} + u_t , \quad t = 1,..,T$$

STEP 5. Use $(\hat{\beta}_1,..,\hat{\beta}_k)$ from Step 4 to estimate $(\beta_1,..,\beta_k)$ in (1), and use $\hat{\alpha}_0$ from Step 4 plus $\hat{\rho}$ in Step 2 to estimate β_0 in (1) by

(6) $$\hat{\beta}_0 = \hat{\alpha}_0 / (1 - \hat{\rho})$$

COOK'S D-MEASURE FOR OUTLIERS

Given a regression of Y on $(x_1,..,x_k)$ using data set $(y_j, x_{1j},..,x_{kj}),\ j=1,..,n$, if s is the root mean square error, and if

$\hat{Y}_j =$ **regression prediction of** $E(Y_j \mid x_{1j},..,x_{kj})$

$\hat{Y}_j(i) =$ **regression prediction of** $E(Y_j \mid x_{1j},..,x_{kj})$ **with the** i^{th} **data point** $(y_i, x_{1i},..,x_{ki})$ **removed.**

then Cook's Distance Measure for point i is defined by

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{Y}_j - \hat{Y}_j(i)\right)^2}{(k+1)s^2},\ i=1,..,n$$

Data point i is then considered to be a statistical outlier whenever the following "Rule of Thumb" holds:

$$D_i \geq \frac{4}{n-(k+1)}$$