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

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

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

**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:

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

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

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

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

**STEP 5.** Use \( (\hat{\beta}_1, \ldots, \hat{\beta}_k) \) from Step 4 to estimate \( (\beta_1, \ldots, \beta_k) \) in (1), and use \( \hat{\alpha}_0 \) from Step 4 plus \( \hat{\rho} \) in Step 2 to estimate \( \beta_0 \) in (1) by

\( \hat{\beta}_0 = \hat{\alpha}_0 / (1 - \hat{\rho}) \)
COOK’S D-MEASURE FOR OUTLIERS

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

\[
\hat{Y}_j = \text{regression prediction of } E(Y_j \mid x_{1j}, \ldots, x_{kj})
\]

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

then Cook’s Distance Measure for point \( i \) is defined by

\[
D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_j(i))^2}{(k + 1)s^2}, \ i = 1, \ldots, n
\]

Data point \( i \) is then considered to be a statistical outlier whenever the following “Rule of Thumb” holds:

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