Demand Modeling - 1.202

Problem Set 2: Linear Time Series Regression

Due date: Friday April 5, 2019

1 Case Study 2 (60 points)

1.1 Objective

The objective of this case study set is to make you familiar with linear time series regression. In this case study, you will estimate a linear time series regression model using the dataset of your choice where demand is affected by stocks.

1.2 Software

The reference software is R. However, you can use any other regression software of your choice that has enough functionality to complete the assigned tasks. Please check with your TA if you want to use a software application different than R.

For Durbin-Watson, please do not use the built-in function but compute it yourself, as it is important to understand how to calculate this statistic.

For Cochrane-Orcutt, you can use the built-in function for the complete set of iterations. But please explain clearly how to perform each step and the reasoning behind each step in the report. Then comment on the results like stated in the tasks.

1.3 Datasets

Either *one* of the following datasets can be used for the case study.

Estimating Electricity Demand

In this case, you have to estimate the electricity demand with equipment stocks included explicitly. The data includes 34 annual observations of total kilowatt-hours consumed by all US customers (1951-84) along with some candidate variables affecting the demand. Related data file is *Eler.csv*.

Estimating Transportation Demand (VMT)

In this case, you have to estimate the transportation demand with vehicle stocks included explicitly. You have to develop models to estimate income and price elasticities of vehicular travel in the U.S. for 1950-1994. The variables in this data set include socio-economic attributes and physical infrastructure characteristics. Related data file is *Tran.csv*.

The detailed description of each dataset and the assigned tasks for each dataset are described in Appendices A and B, respectively.

1.4 Report Content

Your final report should include the answers to the questions asked in the Your Tasks sections of your chosen case. Electronic copies of the relevant input files and spreadsheets should be zipped in a single folder (yourName_CS2.zip) and uploaded to the Stellar website.

2 Supplemental Problems (40 points)

1. The following regression equation is estimated as a production function of Q.

$$\ln Q = 1.35 + 0.633 \ln K + 0.453 \ln L$$

where $\hat{\beta}_K = 0.633$, $\hat{\beta}_L = 0.453$, $R^2 = 0.98$, $\operatorname{cov}(\hat{\beta}_K, \hat{\beta}_L) = 0.055$, and the standard errors of the coefficients corresponding to $\ln K$ and $\ln L$ are 0.258 and 0.219, respectively, i.e. $s_{\hat{\beta}_K} = 0.258$ and $s_{\hat{\beta}_L} = 0.219$.

(a) Test the following null hypothesis (assuming a normally distributed error term):

(i) The capital K and labor L elasticities of output are identical.

(ii) There are constant returns to scale (i.e., the coefficients of $\ln K$ and $\ln L$ add to one).

(b) The problem does not give the number of sample observations. Does this omission affect your conclusions?

2. Being concerned with wage rates of urban planners all over the country, you estimate a regression of the following form:

$$\ln(WAGE) = \beta_1 + \beta_2 ED + \beta_3 EX + \beta_4 FE + \beta_5 NONWH + \varepsilon$$

where: WAGE is the hourly wage; ED is years of education; EX is years of experience; FE is a dummy variable, 1 if female, 0 if male; and NONWH is a dummy variable, 1 if non-white (minority), 0 otherwise. Out of a total sample of 650 planners, 207 were females and 148 were non-white.

You would like to test the following hypotheses, each separately. The base regression equation may or may not provide all of the information needed to perform each test. For each hypothesis, state clearly the null and alternative hypothesis, any additional regressions you will need, the equation of the statistic, the test applied, the level of significance, the degrees of freedom, and the conclusion of the test.

(a) Men and women of equal education, experience, and race have the same wages.

(b) Once one controls for education and experience, sex and race have no effect on wages.

(c) The percentage change in wage due to an additional year of experience is not affected by the sex of the worker.

(d) Males and females have the same coefficients in their wage equations.

3. A researcher **wishes to** estimate the multivariate regression of Y on X **without an intercept term**. Unfortunately, the regression program at hand automatically computes an intercept term, which precludes him from proceeding further. One fine morning, his gray cells go "click" and he concludes that a simple trick would work – estimate the "correct" intercept free regression by entering each data point twice - once in its correct form (Y_i, X_i) and once with the opposite sign ($-Y_i$, $-X_i$). Will this work? Why or why not.

4. Suppose we have a random utility model with a single generic attribute.

The utility of alternative i for individual n has the form:

$$U_{in} = \alpha_i + \beta X_{in} + \varepsilon_{in}, \quad \forall i \in C_n$$

where α_i is an alternative-specific constant for alternative i and β is a scalar coefficient. Consider the following four cases regarding different possible sources of error:

1. There exists an omitted variable Z_{in}.
2. The value X_{in} is an imperfect proxy for some true attribute.
3. The coefficient β varies in some unobserved way across the population.
4. X_{in} is measured with error.

For each of these cases, discuss the reasonableness of the assumption that the ε_{in}'s are independent and identically distributed across the population and across the alternatives. Consider what restrictions are needed for this i.i.d. assertion to be true, and whether such restrictions are reasonable. You need only analyze one source of error at a time.

Appendix A
Econometric Models of Electricity Demand with Equipment Stocks Included Explicitly

In this case study, you will examine the demand for electricity, which is affected by stocks capital goods. These capital goods (appliances) have their own demand schedules. In the short run, capital stocks are fixed, so demand will depend on the utilization of these capital goods, in response to fuel prices. In the long run, capital stocks can be replaced, so changes in fuel prices may lead consumers to purchase goods of better or worse fuel efficiency, so that desired energy consumption patterns may be maintained. This suggests a two-part model, and thus a two-equation framework, in which one stage is a short-run utilization model with electricity demand conditional on the equipment stock, and the second is a long-run model of factors affecting the equipment stock. We begin with the first stage, a short-run utilization model.

One of the first explicit models of the short-run demand for electricity is due to Franklin M. Fisher and Carl Kaysen (1962). Focusing primarily on the residential sector, Fisher-Kaysen called the set of electricity-using appliances and fixtures "white goods", and noted that household demand for electricity use is derived from the demand for the services of the households' various stocks of white goods. In the short run, these stocks are fixed. Since white goods vary in their capacity to consume kilowatts of electricity per hour of normal usage, Fisher-Kaysen proposed to measure the effects of the aggregate equipment stock on electricity consumption in terms of the total kilowatt hours that could be consumed were all appliances employed at their normal use. This was done by obtaining engineering information on kilowatts used per hour of normal use for each type of appliance, and then summing over the various household appliances.

Denote the aggregate equipment stock for the i^{th} household at time t, measured in units of kilowatts per hour, as W_{it}. Actual electricity consumption q_{it} depends on the rate of utilization of the various stocks, denoted u_{it}, which in turn are hypothesized to depend on real per capita income Y_{it} and the real price of electricity, P_{it}:

$$q_{it} = u_{it} \bullet W_{it} = u_{it}(Y_{it},P_{it}) \bullet W_{it} \qquad (1)$$

Fisher-Kaysen specified the functional form for this relationship as

$$q_{it} = P^{\alpha}{}_{it}Y^{\beta}{}_{it}W_{it} \qquad (2)$$

which, after a logarithmic transformation, became

$$\ln q_{it} = \alpha \ln P_{it} + \beta \ln Y_{it} + \ln W_{it}\,. \qquad (3)$$

In implementing this model empirically, Fisher-Kaysen expended considerable efforts in gathering data on stocks of seven major white goods, by state, over the 1944-1957 time period. Stating that the quality of this data ranged "...from somewhat below the sublime to a bit above the ridiculous," Fisher-Kaysen discovered that these seven white goods did not account for enough of total residential electricity consumption, that it would be impossible to estimate stocks for other white goods, and thus concluded that "...to estimate W_{it} by states and years with any kind of reliability is simply out of the question."

Instead, they postulated that the stock of white goods in the ith state grew at a constant rate of γ_i percent per year, i.e.,

$$W_{i,t}/W_{i,t-1} = \exp(\gamma_i), \text{ or } \ln W_{i,t} - \ln W_{i,t-l} = \gamma_i . \quad (4)$$

Equation (3) was then lagged one time period, this was subtracted from Equation (3) and then Equation (4) was substituted in, thereby yielding the first-difference equation

$$\ln q_{it} - \ln q_{i,t-l} = \gamma_i + \alpha_i(\ln P_{it} - \ln P_{i,t-1}) + \beta_i(\ln Y_{it} - \ln Y_{i,t-1}). \quad (5)$$

A random disturbance term was added to Equation (5), reflecting the effects of inherent stochastic elements and omitted variables (assumed to be uncorrelated with the included regressors); this random disturbance term was assumed to be independently and identically normally distributed. Fisher-Kaysen then estimated the parameters α_i, ,β_i and γ_i in Equation (5) for each of the states in the US, using 1946-1957 annual data and ordinary least squares estimation.

Data

The data file contains 34 annual observations, 1951-84, on the following variables: the year of the observation (YEAR), total kilowatt hours consumed by all US customers, in millions (KWH), the average price of electricity per kwh in constant 1972 cents per kwh (PELEC), US gross national product in billions of 1972 dollars (GNP), and the NERC Summary Forecast average annual electricity demand growth rate forecast, in percentage points, using data up through and including the current year (NSF). The NSF variable has non-zero values only for 1973 to 1984. Further, the PELEC variable should be interpreted as an ex post average price; it is computed as total revenue from sales of utilities to all sectors (residential, commercial, industrial, etc.) in 1972 dollars, divided by the total kwh sold. Data sources for the variables in the file NERC are listed in Nelson-Peck (1985).

Your Tasks: The Fisher-Kaysen Specification with US Time Series Data

The purpose of this exercise is to have you assess the Fisher-Kaysen specification using aggregate US time series data from the Nelson-Peck study, both in estimation and forecasting.

In the background section, electricity demand specifications were considered with equipment stocks included explicitly. Due to difficulties in obtaining data on appliance stocks, Fisher-Kaysen made the assumption that appliance stocks grew at a constant annual rate γ. This resulted in Equation (5), in which first differences in the logarithms of electricity consumption were related to a constant term γ, as well as to first differences in the logarithms of real electricity price and real income.

Fisher-Kaysen's data file contains 34 observations on variables named YEAR, KWH, PELEC, GNP and NSF; definitions of these variables were given above.

(a) The first task is to generate variables required to estimate parameters in Equation (5). At the aggregate national level, GNP is often used as a proxy measure of real income; hence, let real GNP replace the income measure in Equation (5). A common procedure for reducing collinearity with time series data is to first difference the data, and then to work with the first differenced data. Logarithmic first differencing is particularly attractive, for $\ln Y_t - \ln Y_{t-1}$ equals $\ln (Y_t/Y_{t-1})$, which for small changes can be interpreted as the percentage change in Y from period t-1 to period t. This way of computing percentage change is also attractive in that $\ln (Y_t/Y_{t-1})$ always yields a value in between $(Y_t - Y_{t-1})/ Y_{t-1}$ and $(Y_t - Y_{t-1})/ Y_t$. Using the annual data from 1952 to 1984, compute logarithmic first differences for PELEC, GNP and KWH, and call these new variables LNP1, LNG1 and LNK1, respectively.

(b) Using OLS and the 1952-84 annual observations, estimate and interpret parameters in Equation (5). In particular, do these estimates represent short- or long-run elasticity measures? Are they statistically significant?

(c) Suppose that the real price of electricity was forecast to fall at an annual rate of 2%, but that real GNP growth was forecast to grow at a 4% rate annually. Based on your estimated equation in part (b), what would be the growth rates forecasted for electricity demand?

(d) From (b), test for the absence of first order autocorrelation by employing the Durbin-Watson test statistic. If the null hypothesis of no autocorrelation is rejected, proceed by estimating the same equation but allowing for first order autocorrelation (use the estimation procedure discussed in the Appendix C- iterative Cochrane-Orcutt). With first order autocorrelation permitted, are your elasticity estimates substantially changed? Which results are more plausible and credible, those in part (b) or here? Why?

(e) Now repeat parts (b), (c) and (d), employing instead data only through 1973. Peck and Nelson report that estimates based on data through 1973 are very similar to those based on data through 1984. Do you agree or disagree? Why? Over the 1951-73 time period, average annual growth rates for GNP and PELEC were 3.5% and -2.4%, respectively. If in 1973 electric utility forecasters predicted these same growth rates to continue into the future, and had they estimated Equation (5) as you just have using data through 1973, what would their electricity demand growth forecasts have been on the basis of these assumptions and the OLS (or GLS) estimated equation? How would their 1974-83 forecast compare to that published by the NERC, reproduced in Table 1? How would their forecasts have changed, had they known that from 1973 to 1984 GNP would only grow at an annual rate of 2.5%, and that the real price of electricity would increase at 4.2% per annum? If they had known this, might they have adjusted their estimate of γ downward? Why?

(f) Briefly assess the empirical performance of the Fisher-Kaysen specification, in terms of implied estimated elasticities and growth rates of appliance stocks, and in terms of its forecasting properties.

Table 1
North American Electric Reliability Council Ten Year Forecasts
and Actual Experience, 1973 – 1984

Using data through year	**For ten-year time period**	**NERC 10-year Avg% growth forecast**	**Actual 10-year Avg% growth realized**
1973	1974-83	7.5	2.3
1974	1975-84	6.7	3.0
1975	1976-85	6.3	2.9
1976	1977-86	5.8	2.4
1977	1978-87	5.3	2.3
1978	1979-88	4.8	2.4
1979	1980-89	4.1	
1980	1981-90	3.7	
1981	1982-91	3.3	
1982	1983-92	3.2	
1983	1984-93	2.8	
1984	1985-94	2.4	

Sources: Charles R. Nelson and Stephen C. Peck (1985) and Charles R. Nelson, Stephen C. Peck and Robert G. Uhler (1987). Updates by E. R. Berndt.

References

Ernst R. Brendt (1991), *The Practice of Econometrics: Classic and Contemporary*, New York, Addison-Wisley.

Franklin Fisher and Carl Kaysen (1962), *A Study in Econometrics: The Demand for Electricity in the United States*, Amsterdam, North-Holland.

Charles Nelson and Stephen Peck (1985), "The NERC Fan: A Retrospective Analysis of NERC Summary Forecasts", *Journal of Business and Economic Statistics*, 3:3, July, pp. 179-187.

Appendix B

Econometric Models of Transportation Demand (VMT) with Vehicle Stocks Included Explicitly

Transportation demand can be measured by vehicle-miles of travel (VMT). In the short run, transportation demand is constrained by vehicle stocks. Travelers may use their vehicle more or less in response to changes in fuel prices or their own incomes, but generally, people hold automobiles for several years at a time. In the long run, of course, households may vary the number of autos they own based on their travel needs, income, fuel prices and the fuel efficiency of available models.

Pickrell developed a data set to estimate income and price elasticities of vehicular travel in the U.S. for 1950-1994. The variables in this data set are described in Table 2. Pickrell's model used VMT per vehicle as the explanatory variable, but otherwise, it's similar to the Fisher-Kaysen specification. The specification Pickrell used is

$$\text{VMT/Vehicle}_t = k_0(\text{GDP/capita}_t)^{\beta 1}(P_{\text{fuel/gal},t})^{\beta 2}(\text{MPG}_t)^{\beta 3}(\text{Vehicles/driver}_t)^{\beta 4}(\text{psuburbs}_t)^{\beta 5}(\text{VMT/Vehicle}_{t-1})^{\beta 6}\eta_t \qquad (6)$$

You will need to consider how travel is related to population, drivers' licenses, employment or other factors. In log form, the model is

$$\ln \text{VMT/Vehicle}_t = \beta_0 + \beta_1 \ln \text{GDP/capita}_t + \beta_2 \ln P_{\text{fuel/gal},t} + \beta_3 \ln \text{MPG}_t + \beta_4 \ln \text{Vehicles/driver}_t + \beta_5 \ln \text{psuburbs}_t + \beta_6 \ln \text{VMT/Vehicle}_{t-1} + \varepsilon_t \qquad (7)$$

Where $\beta_0 = \ln k_0$, and $\varepsilon_t = \ln \eta_t$ is a stochastic disturbance representing variables omitted from the model and other noise in the data.

The vehicles-per-driver term reflects the size and rate of utilization of the vehicle stock. Since the vehicle stock may change in the long run, Pickrell modeled it as

$$\ln \text{Vehicles/driver}_t = \gamma_0 + \gamma_1 \ln \text{GDP/capita}_t + \gamma_2 \ln P_{\text{fuel/gal},t} + \gamma_3 \ln P_{\text{veh},t} + \gamma_4 \ln \text{LFP}_t + \gamma_5 \ln \text{psuburbs}_t + \gamma_6 \ln \text{Vehicles/driver}_{t-1} + \xi_t \qquad (8)$$

Your tasks:

(a) Create the variables you need to estimate Pickrell's VMT and vehicle stock models by taking logarithms, lagging variables, or performing other mathematical manipulations to create variables you think will help explain VMT growth. Explain your intuition.

(b) Replicate each of Pickrell's models (equations 7 and 8) using OLS. Interpret your estimates of the parameters. Are they statistically significant? What are the short- and long-run estimates of the elasticity of VMT with respect to gasoline price and income (GDP per capita)?

(c) From (b), use one of the procedures described in Appendix D to test for the absence of first order autocorrelation. If autocorrelation is present, can you use a lagged endogenous variable as an explanatory variable? Why or why not?

(d) For each model where you conclude autocorrelation exists (based on (c)), reestimate the Pickrell's model *without* the lagged endogenous variable. Examine specifications of your own. Can you come up with a model that has more explanatory power than Pickrell's? Use the Durbin-Watson statistic to test for first order autocorrelation in your best model. What are the properties (e.g., consistency, efficiency) of your OLS estimates when autocorrelation is present and when autocorrelation is not present?

(e) Use the Iterative Cochrane-Orcutt procedure described in Appendix C to correct for first order autocorrelation and to estimate the correlation between successive stochastic error terms. With first order autocorrelation permitted, are your elasticity estimates substantially changed? Which results are more plausible and credible, those in part (b) or here? Why?

(f) Now reestimate your best model specification from parts (b), (d) and (e) using only the data through 1970. It may be argued that the oil shocks of the 1970s and growing environmental consciousness may have led to a structural change in travel demand patterns. Do you agree with this assertion? Describe how you might test this assertion statistically.

(g) Many urban planners and transportation professionals believe the hypothesis of "if you build it, they will come," that is, that any increase in road capacity is immediately consumed by additional travel at congestion levels similar to before the road construction. Mark Kiefer and Shomik Mehndiratta (1998) present a counter-argument that while VMT and congestion have risen dramatically from the '50s to the '90s, this growth in VMT should not be expected to continue in the future for two reasons: (1) the growth in labor-force participation should be tapering off now that labor-force participation among women is approaching that of men, and (2) the growth in household auto ownership appears to slowing to a rate of one vehicle per driver.

Forecast the VMT/Vehicle for the year 2016 (that is, 1994, the last year of the data, plus a 22-year planning horizon) assuming LFP continues to grow at 0.3% per year and auto ownership (vehicles per drive) continues to grow at 0.9% per year — these are the average growth rates for 1950-1994. (Other average growth rates are GDP per capita, 1.8% per year; gasoline price, -0.8% per year; vehicle cost, 0.2% per year; fuel efficiency, 0.7% per year; and proportion of suburbs, 1.7% per year. Do final forecast variables look reasonable?)

Rerun your forecast of VMT/Vehicle for 2016 assuming that LFP stays constant at 66.3 percent, and auto ownership remains at 1.089 vehicles per driver. Is the difference in forecasts large enough that Kiefer and Mehndiratta's hypothesis should be given more analysis?

Table 2

Variable Definitions for FHWA VMT Data Set

Column	Name	Description of Variable
1	Year	1950-1994
2	vmt	annual vehicle-miles traveled in automobiles and two-axle, four-tire trucks (millions)
3	Pop	U.S. resident population (millions)
4	pop16	U.S. resident population 16 years of age and older (millions)
5	Vmtcap	annual VMT per resident person (=vmt/pop)
6	vmtcap16	annual VMT per resident person 16 years of age and older (=vmt/pop16)
7	gdp	total U.S. Gross Domestic Product in 1987 dollars (billions)
8	gdpcap	U.S. Gross Domestic Product per resident person in 1987 dollars (=gdp/pop)
9	drivers	licensed drivers in the U.S. (millions)
10	lirate	proportion of persons 16 years of age and older holding drivers' licenses (=drivers/pop16)
11	vehicles	number of automobiles and two-axle, four-tire trucks registered in the U.S. (thousands)
12	vmtveh	average annual miles driven per vehicle (=vmt/vehicles)
13	vehcap	registered vehicles per resident person (=vehicles/pop)
14	vehdriv	registered vehicles per licensed driver (=vehicles/drivers)
15	vehp16	registered vehicles per resident person 16 years of age and older (=vehicles/pop16)
16	fuel	gasoline consumption by automobiles and two-axle, four-tire trucks (millions gallons)
17	fuelcap	fuel consumption per resident person (gallons)
18	pgas	price of gasoline per gallon in 1987 U.S. dollars
19	mpg	average fuel efficiency of automobiles and two-axle, four-tire trucks (miles per gallon)
20	pmile	gasoline cost per mile driven in 1987 U.S. dollars (=pgas/mpg)
21	roads	road and highway mileage in the continental U.S. at midyear
22	roadcap	road mileage per thousand resident person at midyear (=roads/pop)
23	lfp	percent of working-age U.S. population employed or actively seeking work
24	pvehicle	average price of a new automobile in 1987 U.S. dollars
25	psuburbs	proportion of U.S. population living in urbanized areas who live in suburban portions of urban areas (using Census Bureau definitions)

Notes: Some variables are not available for 1991-1994.
Labor Force Participation, lfp, is presented as a percentage, that is, it takes values from zero to 100. Other variables are presented as proportions, that is, ranging from zero to one.

References

Don Pickrell (1998) Personal Communication.

Mark Kiefer and Shomik Mehndiratta (1998) "If We Build It, Will They Really Keep Coming? Critical Analysis of the Induced Demand Hypothesis." Presented at the 77th Annual Meeting of Transportation Research Board, 11-15 January 1998. Washington, DC.

Appendix C
Iterative Cochrane-Orcutt Procedure

Serial Correlation

The assumption that errors corresponding to different observations are uncorrelated often breaks down in time-series studies but can be a problem in some cross-section work as well. Recall that when the error terms from different (usually adjacent) time periods (or cross-section observations) are correlated, we say that the error term is *autocorrelated* or *serially correlated*. Serial correlation occurs in time-series studies when the errors associated with observations in a given time period carry over into future time periods. For example, if we are predicting the growth of stock dividends, an overestimate in one year is likely to lead to overestimates in succeeding years.

In this section we deal with the problem of *first-order serial correlation*, in which errors in one time period are correlated directly with errors in the ensuing time period. While it is certainly possible that serial correlation can be negative as well as positive, we concern ourselves primarily with the case of positive serial correlation, in which errors in one time period are positively correlated with errors in the next time period. Positive serial correlation frequently occurs in time-series studies, either because of correlation in the measurement error component of the error term or more likely, because of the high degree of correlation over time present in the cumulative effects of the omitted variables in the regression model.

Corrections for Serial Correlation

We assume that each of the error terms in a linear regression model is drawn from a normal population with zero expected value and constant variance but that the errors are not independent over time. Since serial correlation is usually present in time-series data, we use a subscript of t (in place of i or n) and assume that the total number of observations is T. The model is

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \cdots + \beta_K X_{Kt} + \varepsilon_t \tag{9}$$
$$\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t \quad 0 \le \rho < 1 \tag{10}$$

where ν_t is distributed as $N(0, \sigma_\nu^2)$ and is independent of other errors over time, as well as being independent of ε; and ε_t is distributed as $N(0, \sigma_\varepsilon^2)$ but is not independent of other errors over time. The error process as described in Eq. (10) is generated by a rule which says that the error in time period t is determined by diminishing the value of the error in the previous period (multiplying by ρ) and then adding the effect of a random variable with zero expected value. It is the most elementary form of an *autoregressive* error process.

If ρ were known, it would be easy to adjust the ordinary least squares regression procedure to obtain efficient parameter estimates. The procedure involves the used of *generalized differencing* to alter the linear model into one in which the errors are independent. To describe this procedure, we use the fact that the linear model in Eq. (9) holds *for all time periods*. In particular,

$$Y_{t-1} = \beta_1 + \beta_2 X_{2t-1} + \beta_3 X_{3t-1} + \cdots + \beta_K X_{Kt-1} + \varepsilon_{t-1} \tag{11}$$

Multiplying (11) by ρ and subtracting from Eq. (9), we obtain the desired transformation:

$$Y_t^* = \beta_1(1-\rho) + \beta_2 X_{2t}^* + \beta_3 X_{3t}^* + \cdots + \beta_K X_{Kt}^* + \nu_t \qquad (12)$$

where

$$\begin{aligned} Y_t^* &= Y_t - \rho Y_{t-1} \\ X_{kt}^* &= X_{kt} - \rho X_{kt-1} \\ \nu_t &= \varepsilon_t - \rho \varepsilon_{t-1} \end{aligned}$$

are the *generalized differences* of Y_t, X_{kt} and ε_t. By construction the transformed equation has an error process which is independently distributed with zero mean and constant variance (see Eq. (10)). Thus, ordinary least-squares regression applied to Eq. (12) will yield efficient estimates of all the regression parameters. Of course, the intercept of the original model must be calculated from the estimated intercept associated with Eq. (12).

The Cochrane-Orcutt Procedure

This procedure involves a series of iterations, each of which produces a better estimate of ρ than the previous one. It uses the notion that ρ is a correlation coefficient associated with the errors of adjacent time periods.

Step 1. Ordinary least squares is used to estimate the original model, Eq. (9).

Step 2. The estimates, $\hat{\beta}$, are used to construct residuals, $\hat{\varepsilon}$:

$$\hat{\varepsilon}_t = Y_t - \hat{\beta}_1 - \hat{\beta}_2 X_{2t} - \cdots - \hat{\beta}_K X_{Kt}.$$

These residuals are then used to perform the regression

$$\hat{\varepsilon}_t = \rho \hat{\varepsilon}_{t-1} + \nu_t \qquad (13)$$

to get an estimate for $\hat{\rho}$.

Step 3. The estimated value of ρ is used to perform the generalized differencing transformation process, and a new regression is run. The transformed equation is

$$Y_t^* = \beta_1\left(1-\hat{\rho}\right)+\beta_2 X_{2t}^* + \beta_3 X_{3t}^* + \cdots + \beta_K X_{Kt}^* + \nu_t \tag{14}$$

where

$$\begin{aligned} Y_t^* &= Y_t - \hat{\rho} Y_{t-1} \\ X_{kt}^* &= X_{kt} - \hat{\rho} X_{kt-1} \, . \\ \nu_t &= \varepsilon_t - \hat{\rho}\varepsilon_{t-1} \end{aligned}$$

Step 4. The estimated transformed equation yields new estimates for the parameter vector β. These revised parameter estimates are substituted into the *original* equation, and new regression residuals are obtained. That is, the new estimate of β is used upon returning to **Step 2** to obtain a new estimate of ρ.

The iterative process can be carried on for as many steps as the researcher desires. Standard procedure is to stop the iterations when the new estimates of ρ differ from the old ones by less than 0.01 or 0.005, or after 10 or 20 estimates of ρ have been obtained. The specific choice made depends upon the computational costs involved. The primary difficulty with the Cochrane-Orcutt procedure is that there is no guarantee that the final estimate of ρ will be the optimal estimate, in the sense of minimizing the sum of squared residuals. The difficulty arises because the iterative technique may lead to a local rather than a global minimum.

Appendix D
Testing for First Order Autocorrelation in Models with Lagged Endogenous Variables

In models with lagged endogenous variables (such as Pickrell's), the Durbin-Watson statistic may not provide a valid test of serial correlation. Durbin (1970) devised a Lagrange Multiplier Test that scales up the Durbin-Watson statistic based on the significance of the lagged endogenous variable. A modification of the Breusch (1978) - Godfrey (1978) test examines possible serial correlation by regressing residuals on all explanatory variables and lagged residuals.

Durbin's Lagrange Multiplier Test

Under the null hypothesis of no serial correlation, for a regression containing a lagged endogenous variable, Durbin showed that the statistic

$$h = \left(1 - \frac{d}{2}\right)\sqrt{\frac{T}{1 - Ts_c^2}}$$

where d = the Durbin-Watson statistic,
T = the number of observations (or time periods), and
s_c^2 = the estimated variance (standard error squared) of the coefficient on the lagged endogenous variable,

follows a standard normal distribution. Large values of h suggest rejecting the null hypothesis of no serial correlation. Of course, this test cannot be used if $s_c^2 > 1/T$.

The Modified Breusch-Godfrey Test

The Modified Breusch-Godfrey Test requires an auxiliary regression of the fitted residuals, $\hat{\varepsilon}_t$ on all explanatory variables (x_t's and y_{t-1}) and the lagged residuals, $\hat{\varepsilon}_{t-1}$. Missing lagged residuals (e.g., the first one) should be filled in with zeros. Additional lagged residuals can be added to the auxiliary regression to test for higher orders of autocorrelation.

The purpose of this auxiliary regression is similar to estimating ρ during the Cochrane-Orcutt procedure. The presence of the explanatory variables, which under the null hypothesis aren't correlated with the residuals, insures that the coefficient on $\hat{\varepsilon}_{t-1}$ reflects only the influence of correlation that is unexplained by the x's or y's.

The Modified Breusch-Godfrey Test uses familiar significance tests on the coefficient(s) of the lagged residuals. That is, a t-test can be used when only $\hat{\varepsilon}_{t-1}$ is included, and an F test is used to test the joint significance when more lags of residuals are included.

This appendix is a summary of Greene (1997) pages 595-597.

References

T. Breusch (1978) "Testing for Autocorrelation in Dynamic Linear Models." *Australian Economic Papers*, **17**, pages 334-355.

J. Durbin (1970) "Testing for Serial Correlation in Least Squares Regression When Some of the Regressors Are Lagged Dependent Variables." *Econometrica*, **38**, pages 410-421.

L. Godfrey (1978) "Testing Against General Autoregressive and Moving Average Error Models When the Regressors Include Lagged Dependent Variables." *Econometrica*, **46**, pages 1293-1302.

William H. Greene (1997) *Econometric Analysis*, Macmillan Publishing Company, 3rd ed.