ECON 423 - Multiple Regression Forecasting Lab

Introduction

Regression methods are useful tools to forecasters. They are more sophisticated than naive methods because regression models use more information, in the form of explanatory variables, to forecasting applications. Regression models also allow for the incorporation of economic theories into the forecasting process. They also require more data, additional computational burden, and the judgement of the forecaster.

This lab focuses on using regression methods to forecast demand for tickets at Major League Baseball games. You will estimate a simple demand model for tickets using OLS and use the estimated parameters from this to generate *ex post* and *ex ante* forecasts of ticket sales.

Goals

1. Estimate a demand function for Major League Baseball ticket sales
2. Understand how to use multiple regression models to generate forecasts
3. Estimate a univariate regression model for attendance at baseball games.
4. Gain more experience using the Excel Regression Wizard

Data

The Excel file `baseball_data.xls` contains data on attendance, ticket prices and other factors for two professional baseball teams - the Kansas City Royals and the Philadelphia Phillies - for the 1990 through 2001 baseball seasons. The file contains the following variables:

Variable	Description
`year`	Calendar year, 1990-2001
`teamname`	Name of baseball team
`avg_attend`	Average attendance per game for season
`price`	Average ticket price
`playoff`	Dummy variable, equal 1 if team made the post season in that season
`strike`	Dummy variable, equal 1 if a baseball strike took place in that season
`wins`	Total number of games won in season
`pct`	Percent of games won in season

The Excel file `attendance_data.xls` contains data on total and average attendance for the National Footabll League (NFL) and Major League Baseball (MLB) for much of the 20th century. The file contains the following variables:

Variable	Description
`year`	Calendar year
`GP`	Total games played
`Total`	Total attendance for season
`Avg`	Average attendance per game for season

Methods

Recall the general form of a simple regression model

$$Y_i = \beta_0 + \beta_1 X_i + e_i \tag{1}$$

where Y_i is the dependent variable, X_i is the independent or explanatory variable, e_i is the unobservable equation error that captures all factors except X_i that affect Y_i, and β_0 and β_1 are unknown parameters to be estimated. e_i is a random variable and by assumption

1. $E(e_i) = 0$
2. $var(e_i) = \sigma^2$
3. $cov(e_i, e_j) = 0 \; for \; i \neq j$
4. $e_i \sim N(0, \sigma^2)$

Also recall the general form of a multiple regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + e_i. \tag{2}$$

In this lab, the economic theory underlying the simple regression model is the theory of consumer behavior. In particular, the demand function that emerges from the model of consumer behavior from microeconomics. According to this model, the quantity of any good demanded by consumers varies inversely with the price of that good, and demand for any good at any price also changes in response to changes in factors like income, the price of substitute and complementary goods, tastes and preferences, and other factors. Demand curves slope down and things like income and other prices shift demand curves to the left or right.

In this context, Y_i in equation (1) is demand for attendance at baseball games in season i and X_i is the average price of a ticket to games in season i.

Procedures

1. Using data for the Kansas City Royals, estimate the demand function

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$

where Y_i is average attendance, X_{1i} is average ticket price and X_{2i} is the number of wins in each season. Note that X_{1i} and X_{2i} must be in contiguous columns for the Excel regression wizard to work.

The results are

```
SUMMARY OUTPUT

Regression Statistics
Multiple R          0.73
R Square            0.53
Adjusted R Square   0.43
Standard Error   2715.37
Observations          12

ANOVA
            df          SS             MS    F
Regression   2    75824042    37912020.88 5.14
Residual     9    66359295     7373254.99
Total       11   142183337

               Coefficients    Standard Error  t Stat  P-value
Intercept         32225.82          14518.85    2.22    0.05
X Variable 1      -1490.96            596.19   -2.50    0.03
X Variable 2         55.30            144.13    0.38    0.71
```

2. Interpret the coefficients of the regression model.
3. Perform a hypothesis test on the coefficient on X_{2i}. What does this tell you?
4. Does X_{2i} belong in the regression model?

   recall that the F-satistic tests the overall significance of all the explanatory variables (Xs) in the model. The null hypothesis on this test is $H_o : \beta_1 = \beta_2 = 0$. If rejected, all variables belong in the model.
5. Plot the average annual attendance for the National League from the file `attendance_data.xls`
6. Using data for the National League from the file `attendance_data.xls`, estimate the parameters of the AR(4) autoregressive regression model

$$X_t = \beta_0 + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \beta_3 X_{t-3} + \beta_4 X_{t-4} + e_t$$

   - Start on a new worksheet
   - Label to four columns to the right of average attendance `L1`, `L2`, `L3`,`L4`. These are the "lagged" variables that make up the explanatory variables in the regression model.
   - What is the first year you can estimate this model for?
   - Should you use the entire series? Why or why not?
   - The results for the entire series are

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.98
R Square	0.96
Adjusted R Square	0.96
Standard Error	1646.29
Observations	98

ANOVA

	df	SS	MS	F
Regression	4	6353429256	1588357314	586
Residual	93	252056029	2710280	
Total	97	6605485285		

	Coefficients	Standard Error	t Stat	P-value
Intercept	342.01	330.38	1.035	0.30
X Variable 1	0.99	0.10	9.593	0.00
X Variable 2	-0.28	0.15	-1.954	0.05
X Variable 3	0.17	0.15	1.178	0.24
X Variable 4	0.12	0.11	1.141	0.26

- Discuss the significance of the parameters.
- Plot the actual and predicted values for the AR(4) model. How well does this track the data?
- Re-estimate for the period 1970 on. How do the results differ?