Regression Models for Time Series Situations

November 21, 2013

Time Series Data

In a general regression model we assume that the random error terms ϵ_i are either uncorrelated random variables or independent normal random variables. In business and economics, many regression applications involve time series data. The independency of the error term is unreasonable if we estimate the regression model on time series data. Error terms correlated over time are said to be autocorrelated or serially correlated. There are several sources of autocorrelation. Perhaps the primary cause of the autocorrelation in regression problems involving time series data is failure to include one or more important regressors in the model.
When the error terms are positively autocorrelated

- Ordinary least square regression coefficients are unbiased but they are no longer minimum variance estimates
- MSE may seriously underestimate σ^2
- The confidence intervals and tests of hypotheses based on the t and F distributions are no longer appropriate.

Time Series Analysis

Time series data are vectors of numbers, typically regularly spaced in time. Yearly counts of animals, daily prices of shares, monthly means of temperature, and minute-by-minute details of blood pressure are all examples of time series, but they are measured on different time scales. Sometimes the interest is in the time series itself (e.g. whether or not it is cyclic, or how well the data fit a particular theoretical model), and sometimes the time series is incidental to a designed experiment (e.g. repeated measures). The three key concepts in time series analysis are

- trend,
- serial dependence, and
- stationarity

Most time series analyses assume that the data are untrended. If they do show a consistent upward or downward trend, then they can be detrended before analysis (e.g. by differencing). Stationarity is a technical concept, but it can be thought of simply as meaning that the time series has the same properties wherever you start looking at it.

Example

The Australian ecologist, A.J. Nicholson, collected and studied the data on a laboratory population of Blowflies. The data is in the `gamair` package.

```
> library(gamair)
> data(blowfly)
> attach(blowfly)
> names(blowfly)
> ts.plot(pop)
```

There are two important ideas to understand in time series analysis: **autocorrelation** and **partial autocorrelation**. The first describes how the current population is related to last time's population. This is the autocorrelation at lag 1. The second describes the relationship between this current population and the population at lag t once we have controlled for the correlations between all of the successive weeks between current time and at time t.

Autocorrelation

The autocorrelation function $\rho(k)$ at lag k is

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)}$$

where $\gamma(k)$ is the autocovariance function at lag k of a stationary random function $\{Y(t)\}$ given by

$$\gamma(k) = cov\{Y(t), Y(t-k)\}$$

The most important properties of the autocorrelation coefficient are these

- They are symmetric backwards and forwards, so $\rho(k) = \rho(-k)$
- The limits are $-1 \leq \rho(k) \leq 1$
- When $Y(t)$ and $Y(t-k)$ are independent, then $\rho(k) = 0$.
- The converse of this is not true, so that $\rho(k) = 0$ does not imply that $Y(t)$ and $Y(t-k)$ are independent

First Order Autoregressive Model

We will replace the standard case index "i" by "t" in order to emphasize the fact that the data is a time series and is not cross sectional data. This means we assume that $(y_t, x_{t1}, x_{t2}, \cdots, x_{tp})$ represent the measurements on the response y and the p regressor variables $x_1, x_2, \cdots, x_p$ over time t.
A simple linear regression model with first order autoregressive error (AR(1)), is given by

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t, \qquad \epsilon_t = \rho \epsilon_{t-1} + u_t$$

where, ρ is an autocorrelation parameter such that $\mid \rho \mid < 1$,
u_t are called the disturbance term and they are independent and normally distributed with mean 0 and variance σ^2. This means $u_t \sim N(0, \sigma^2)$.
Observe that by successively substituting for $\epsilon_{t-1}, \epsilon_{t-2}, \ldots$ we have

$$\epsilon_t = \sum_{s=0}^{\infty} \rho^s u_{t-s}$$

Thus , the error term ϵ_t in period t is a linear combination of the current and preceding disturbance terms. Note that

$$\begin{aligned} E(\epsilon_t) &= E\left(\sum_{s=0}^{\infty} \rho^s u_{t-s}\right) = \sum_{s=0}^{\infty} \rho^s E(u_{t-s}) = 0 \\ Var(\epsilon_t) &= \sigma^2/(1-\rho^2) \end{aligned}$$

Durbin-Watson Test for Autocorrelation

The Durbin-Watson test for autocorrelation assumes the first order autoregressive error model with the values of the predictor variable(s) fixed. The test determines whether or not the autoregressive parameter ρ is zero. Note that if $\rho = 0$, then $\epsilon_t = u_t$ so the error terms will be independent and normally distributed. Because most regression problems involving time series data exhibit positive autocorrelation, the hypotheses usually considered in the Durbin-Watson test are

$$H_0 : \rho = 0, \qquad H_1 : \rho > 0$$

The test statistic is

$$D = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

where $e_t = y_t - \hat{y}_t$ are the residuals from an ordinary least squares analysis applied to the (y_t, x_t) data. Exact critical values cannot be obtained but Durbin-Watson show that D lies between two bounds, say d_L and d_U, such that if D is outside these limits, a conclusion regarding the the above hypothesis can be reached. The decision criteria is as follows:

If $D < d_L$	Reject H_0
If $D > d_U$	Fail to reject H_0
If $d_L < D < d_U$	Test is inconclusive

Example

A software beverage company wishes to predict annual sales for a particular product and want to expresses the sales as a function of the annual advertising cost for the product. Table below displays the data for 20 years.

year	t	sales(y)	Ad.Cost(x)	year	t	sales(y)	Ad.Cost(x)
1980	1	3083	75	1981	2	3149	78
1982	3	3218	80	1983	4	3239	82
1984	5	3295	84	1985	6	3374	88
1986	7	3475	93	1987	8	3569	97
1988	9	3597	99	1989	10	3725	104
1990	11	3794	109	1991	12	3959	115
1992	13	4043	120	1993	14	4194	127
1994	15	4318	135	1995	16	4493	144
1996	17	4683	153	1997	18	4850	161
1998	19	5005	170	1999	20	5236	182

Example

For the give data we have

```
> model=lm(y~x)
> summary(model)
Call:
lm(formula = y ~ x)
Residuals:
    Min      1Q  Median      3Q     Max
-32.330 -10.696  -1.558   8.053  40.032
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1608.5078    17.0223   94.49   <2e-16 ***
x             20.0910     0.1428  140.71   <2e-16 ***
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 20.53 on 18 degrees of freedom
Multiple R-squared: 0.9991,     Adjusted R-squared: 0.999
F-statistic: 1.98e+04 on 1 and 18 DF,  p-value: < 2.2e-16
```

Hence the resulting model is $\hat{y} = 1608.51 + 20.09x$ with $R^2 = 0.9991$.

Example

Observe that there is a pattern in the residuals. There is definite upward and downward drift in the residuals. Autocorrelation could be a problem of such pattern in the residuals.

```
> library(lmtest)
> dwtest(y~x)
        Durbin-Watson test
data:  y ~ x
DW = 1.08, p-value = 0.006108
alternative hypothesis: true autocorrelation is greater than 0
```

Note that the p-value of the DW test is 0.006108 which conclude that errors are positively autocorrelated.

Remedial Measures for Autocorrelation

A significance value of the Durbin-Watson statistic or a suspicious residual plot indicates a model specification error. The model misspecification could be because of

a) Omission of important regressor variable

b) Lack of fit

The misspecification of the model can be removed by adding a new regressor variable or by taking an appropriate transformation.

Now consider adding a new regressor variable Population (z) in the above data so the new data is as below

year	t	sales(y)	AC(x)	pop(z)	year	t	sales(y)	AC(x)	pop(z)
1980	1	3083	75	825,000	1981	2	3149	78	830,445
1982	3	3218	80	838,750	1983	4	3239	82	842,940
1984	5	3295	84	846,315	1985	6	3374	88	852,240
1986	7	3475	93	860,760	1987	8	3569	97	865,925
1988	9	3597	99	871,640	1989	10	3725	104	877,745
1990	11	3794	109	886,520	1991	12	3959	115	894,500
1992	13	4043	120	900,400	1993	14	4194	127	904,005
1994	15	4318	135	908,525	1995	16	4493	144	912,160
1996	17	4683	153	917,630	1997	18	4850	161	922,220
1998	19	5005	170	925,910	1999	20	5236	182	929,610

Example

```
> model2=lm(y~x+z)
> summary(model2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.203e+02  2.173e+02   1.474    0.159
x           1.843e+01  2.915e-01  63.232  < 2e-16 ***
z           1.679e-03  2.829e-04   5.934 1.63e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 12.06 on 17 degrees of freedom
Multiple R-squared: 0.9997,     Adjusted R-squared: 0.9997
F-statistic: 2.873e+04 on 2 and 17 DF,  p-value: < 2.2e-16

> dwtest(y~x+z)
        Durbin-Watson test
data:  y ~ x + z
DW = 3.0593, p-value = 0.9822
alternative hypothesis: true autocorrelation is greater than 0
```

Hence the resulting model is $\hat{y} = 320.33956 + 18.43421x + 0.00168z$ with and

Eliminating the Autocorrelation

When the use of additional predictor variables is not helpful in eliminating the problem of autocorrelated errors we should use transformation methods. Three different methods are available

- Cochrane-Orcutt Procedure
- Hildreth-Lu Procedure
- First Difference Procedure

All of these method are based on the property of the first order autoregressive error term. Suppose we transform the response variable so that

$$y_t^{'} = y_t - \rho y_{t-1}$$

Now substituting for y_t and y_{t-1} we have

$$\begin{aligned} y_t^{'} &= y_t - \rho y_{t-1} \\ &= \beta_0 + \beta_1 x_t + \epsilon_t - \rho(\beta_0 + \beta_1 x_{t-1} + \epsilon_{t-1}) \\ &= \beta_0(1-\rho) + \beta_1(x_t - \rho x_{t-1}) + \epsilon_t - \rho\epsilon_{t-1} \\ &= \beta_0^{'} + \beta_1^{'} x_t^{'} + u(t) \end{aligned}$$

Hence, by the use of the transformed variables $y_t^{'}$ and $x_t^{'}$ we obtained a standard simple linear regression model with independent error terms.

Eliminating the Autocorrelation

By the use of the transformed variables $y_t^{'}$ and $x_t^{'}$ we obtained a standard simple linear regression model with independent error terms. This means that we produce a model that satisfies the usual regression assumptions and ordinary least squares can be used but we need to estimate the autocorrelation parameter ρ.
Estimation of ρ

- Cochrane-Orcutt proposed that

$$\hat{\rho} = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_{t-1}^2}$$

- Hildreth-Lu use the value of ρ that minimizes SSE of the transformed regression model.
- First difference Procedure suggest to take $\rho = 1$ in the transformed model.

Remark: There exists an approximate relation between the Durbin-Watson test statistic and the estimated autocorrelation parameter $\hat{\rho}$ from the Cochrane-Orcutt method which is given by

$$D \approx 2(1 - \hat{\rho})$$

Example

A staff analyst for a manufacturer of microcomputer components has compiled monthly data for the past 16 months on the value of industry production of processing units that use these components (X, in millions dollars) and the value of the firm's components used (y, in thousand dollars).

t	Y_t	X_t	t	Y_t	X_t
1	102.9	2.052	2	101.5	2.026
3	100.8	2.002	4	98.0	1.949
5	97.3	1.942	6	93.5	1.887
7	97.5	1.986	8	102.2	2.053
9	105.0	2.102	10	107.2	2.113
11	105.1	2.058	12	103.9	2.060
13	103.0	2.035	14	104.8	2.080
15	105.0	2.102	16	107.2	2.150

Using the ordinary least square we have the following model for the data

$$\hat{y} = -7.739 + 53.953x$$

Example

```
>  x= c(2.052, 2.026, 2.002,1.949, 1.942, 1.887, 1.986,2.053,
     2.102, 2.113, 2.058, 2.060, 2.035, 2.080, 2.102, 2.150)
>  y=c(102.9, 101.5, 100.8,  98.0,  97.3,  93.5,  97.5, 102.2,
      105.0, 107.2, 105.1, 103.9, 103.0, 104.8, 105.0, 107.2)
>  model1=lm(y~x)
Coefficients:
(Intercept)            x
     -7.739       53.953
> library(lmtest)
> dwtest(y~x)
        Durbin-Watson test
data:  y ~ x
DW = 0.8566, p-value = 0.002502
alternative hypothesis: true autocorrelation is greater than 0
```

In the First difference procedure we take $\rho = 1$ for the transformed model $y_t^{'} = \beta_0^{'} + \beta_1^{'} x_t^{'} + u(t)$, where $y_t^{'} = y_t - y_{t-1}$ and $x_t^{'} = x_t - x_{t-1}$ and the regression model is based on "regression through Origin".
The resulting transformed model is

$$y^{'} = 49.81 x^{'}.$$

Example

We can use the R code below to calculate the first difference and fit the model

```
>  x= c(2.052, 2.026, 2.002,1.949, 1.942, 1.887, 1.986,2.053,
     2.102, 2.113, 2.058, 2.060, 2.035, 2.080, 2.102, 2.150)
>  y=c(102.9, 101.5, 100.8,  98.0,  97.3,  93.5,  97.5, 102.2,
      105.0, 107.2, 105.1, 103.9, 103.0, 104.8, 105.0, 107.2)
> y1=diff(y)
> x1=diff(x)
> model2=lm(y1~-1+x1)
> model2
Call:
lm(formula = y1 ~ -1 + x1)
Coefficients:
   x1
49.81

> dwtest(y1~x1)
        Durbin-Watson test
data:  y1 ~ x1
DW = 1.7538, p-value = 0.273
alternative hypothesis: true autocorrelation is greater than 0
```

Example

In order to check whether any positive autocorrelation remains after the first iteration. Hence we would like to check

$$H_0 : \rho = 0 \quad VS. \quad \rho > 0$$

We observe that the Durbin-Watson test statistic $D = 1.754$.
Since $p - value = 0.273$ we fail to reject H_0 and conclude that there is no positive autocorrelation in the transformed model.
Similarly we can perform the transformation based on Cochrane-Orcutt method and Hildreth-Lu method