Time Series Regression

Multiple Linear Regression Analysis

on Time Series Data

Contents

Linear Regression Model for Time Series Data

$$Y_t = b_0 + b_1X_{1t} + b_2X_{2t} + \ \dots + b_pX_{pt} + e_t$$

Where,

Y_t	:	Dependent Variable at time t
$X_{1t}, X_{2t}, \dots, X_{pt}$	:	Independent Variables at time t
$b_0, b_1, \dots, b_p$	:	Parameters of Model
e_t	:	Random Error Component at time t

Variables can either be Continuous or Categorical

The Concept of Autocorrelation

- Autocorrelation is a **correlation between a time series (Y_t) and another time series representing lagged values of the same time series(Y_{t-k}).**

- Autocorrelation refers to the **correlation of a time series with its own past values**

t	Y_t
1	Y_1
2	Y_2
"	"
"	"
10	Y_{10}

t	Y_t	Y_{t-1}
2	Y_2	Y_1
3	Y_3	Y_2
"	"	"
"	"	"
10	Y_{10}	Y_9

- Correlation between Y_t and Y_{t-1} is **"first order autocorrelation"**

Autocorrelation in Multiple Linear Regression

- Autocorrelation is observed only when the data is a time series or panel data
- In regression model, absence **of autocorrelation among errors is one of the key assumptions**

Autocorrelation → **Incorrect Standard Errors of Parameters**

If we obtain the OLS estimators of the regression model without correcting error autocorrelation, the model may yield

- **Underestimation of True Error Variance**
- **Overestimation of R^2**
- **Misleading Results of t & F Tests**

Case Study – Predicting Sales

Background

- A retail store undertakes two-fold marketing campaign, one for print media and the other for digital. The company also collects information on yearly increments (price adjusted) in sales. The company wishes to check if the marketing expenses have any bearing on the sales

Objective

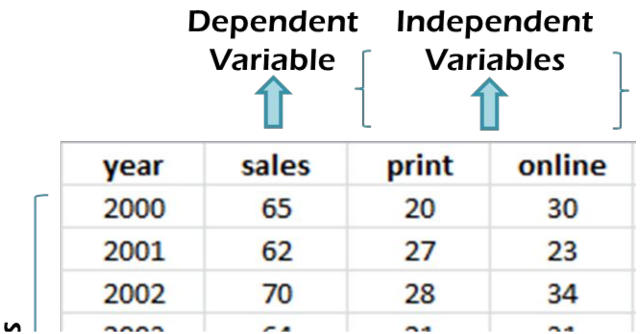
- To predict price adjusted incremental sales based on expenses on marketing campaigns

Available Information

- Yearly time series data, of 11 years
- Independent Variables: Marketing expenses for – Print Media and Online
- Dependent Variable: Price Adjusted Incremental Sales

Data Snapshot

SALES VS MARKETING COSTS



year	sales	print	online
2000	65	20	30
2001	62	27	23
2002	70	28	34

Columns0	Description	Type	Measurement	Possible values
Year	Year	Numeric	2000 to 2010	11
Sales	Price Adjusted Incremental Sales	Numeric	in Euro Million	Positive values
Print	Expenses on Print Marketing	Numeric	in Euro Million	Positive values
online	Expenses on Online Marketing	Numeric	in Euro Million	Positive values

Detecting Presence of Autocorrelation – Durbin Watson Test

Objective	**To check the assumption of 'No Autocorrelation'**

Null Hypothesis H_0: Autocorrelation is not present among errors

Alternate Hypothesis H_1: Not H0

Test Statistic	$d \cong 2(1-r)$ Where r is sample autocorrelation based on residuals obtained in regression model. Ideal Value of d = 2 (taking r = 0)
Decision Criteria	Reject the null hypothesis if p-value < 0.05

Durbin Watson Test in R

```
# Importing the Data and Fitting Linear Model
```

```
sales<-read.csv("SALES VS MARKETING COSTS.csv",header=TRUE)
salesmodel<-lm(sales~print+online,data=sales)

summary(salesmodel)
```

```
# Output
```

```
Call:
lm(formula = sales ~ print + online, data = sales)

Residuals:
     Min       1Q   Median       3Q      Max
-2.16109 -1.00608  0.07342  0.85923  2.32128

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.05815    3.78520   7.941 4.61e-05 ***
print        0.48154    0.14275   3.373  0.00974 **
online       0.76633    0.06828  11.224 3.56e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.486 on 8 degrees of freedom
Multiple R-squared:  0.952,     Adjusted R-squared:   0.94
F-statistic:  79.3 on 2 and 8 DF,  p-value: 5.317e-06
```

Durbin Watson Test in R

```
# Install & Load package "car"
# Durbin Watson Test
```

```
install.packages("car")
library(car)

durbinWatsonTest(salesmodel)
```

durbinWatsonTest() in package car carries out the Durbin Watson test. Output gives the d-w statistic as well as p-value.

```
# Output
```

```
> durbinWatsonTest(salesmodel)
 lag Autocorrelation D-W Statistic p-value
   1       0.4804249     0.7259509   0.004
 Alternative hypothesis: rho != 0
```

Interpretation :

- Reject H0, p-value<0.05.
- Errors have significant autocorrelation.

Error Process in the Presence of Autocorrelation

Presence of autocorrelation implies that, errors at time t are related to errors at previous time points

$$\boxed{e_t = \rho_1 e_{t-1} + \rho_2 e_{t-2} + \ldots + \rho_k e_{t-k} + u_t}$$

ρ_k is coefficient of autocorrelation of lag k between e_t & e_{t-k}

u_t is such that,

$E(u_t)=0$

$V(u_t)=$ Constant

$Cov(u_t, u_s)=0 \quad s\neq0$

* Note: Typically in time series regression we consider autocorrelation of order 1 only

Consequences of Ignoring Autocorrelation

If we obtain the OLS estimators of the model, without correcting error autocorrelation, the model may yield

- Underestimation of True Error Variance
- Overestimation of R^2
- Misleading Results of t & F Tests

Autocorrelation – Remedial Measures

Cochrane-Orcutt Method

Praise-Winsten Method

Maximum Likelihood Estimation Method

Maximum Likelihood Estimation Method

- Maximum likelihood estimator (MLE) can be obtained by maximizing the log likelihood function with respect to b, σ^2_u, ρ

- Log likelihood function is given as follows:

$$\ln L = -\frac{\sum_{t=1}^{T} u_t^{\,2}}{2\sigma_u^2} + \frac{1}{2}\ln(1-\rho^2) - \frac{T}{2}(\ln 2\pi + \ln \sigma_u^2)$$

Parameter Estimation – Maximum Likelihood Estimation Method

```
# Parameter Estimation (Maximum Likelihood Estimation)
salests<-ts(sales$sales,start=2000,end=2010)
```

ts() converts a column from a data frame to a simple time series object.

```
xvar<-subset(sales,select=c("online","print"))
```

subset() creates a subset of all the independent variables

```
salesmodel<-arima(salests,order=c(1,0,0),xreg=xvar)
```

arima() estimates model parameters. Subset of independent variables (which is a vector or matrix of external regressors) is used in **xreg=**

```
coef(salesmodel)
 ar1   intercept     online      print
 0.7774653 32.0288203  0.8117825  0.3621399
```

Parameter Estimation – Maximum Likelihood Estimation Method

#Parameter Estimation (Maximum Likelihood Estimation)

```
install.packages("lmtest")
library(lmtest)

coeftest(salesmodel)
```

arima() does not give p-values, hence **coeftest()** from package **lmtest()** is used. It performs z and (quasi-)t tests of estimated coefficients.

Output

```
z test of coefficients:

           Estimate Std. Error z value  Pr(>|z|)
ar1        0.777465   0.171698  4.5281 5.952e-06 ***
intercept 32.028820   2.137725 14.9827 < 2.2e-16 ***
online     0.811782   0.033897 23.9486 < 2.2e-16 ***
print      0.362140   0.065133  5.5600 2.697e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation :

- Both print and online are significant variables in this output as well

Predictions

```
# Predictions (For Next Two Years)
online<-c(40,45)

print<-c(50,55)

xvarnew<-data.frame(online,print)


predict(salesmodel, n.ahead=2, newxreg=xvarnew)
```

Expected Values of online and print marketing expenses in next 2 years.

data.frame() combines the future expenses into a single dataframe object, to be used for predictions.

predict() uses the MLE regression model, **n.ahead=2** ensures next two forecasts are generated, **newxreg=** specifies the new values of independent variables.

```
# Output
```

```
$pred
Time Series:
Start = 2011
End = 2012
Frequency = 1
[1] 82.80867 88.63343

$se
Time Series:
Start = 2011
End = 2012
Frequency = 1
[1] 0.8509076 1.0778190
```

Quick Recap

Statistical Model	• $Y_t = b_0 + b_1 x_{1t} + b_2 x_{2t} + \text{- - - - -} + b_p x_{pt} + e_t$ • Here, t – the time element is added to each term
What is Autocorrelation	• Autocorrelation refers to the correlation of a time series with its own past values
Consequences	• Presence of Autocorrelation results in incorrect standard errors of model parameters
Test	• The Durbin Watson test is used to check autocorrelation • **`durbinwatsonTest()`** in package **car** performs the D-W Test in R
Maximum Likelihood Estimation Method	• Method of correcting autocorrelation problem