MACQUARIE University

Business Forecasting

Introduction to Regression Models

Generating Process

For the time series Y_t, a basic representation of the generating process is;

$$Y_t = f(\textbf{systematic}, \textbf{random component})$$

$$Y_t = \mu_t + \varepsilon_t$$

The modeller needs to determine the **relevant functional form** for μ_t. This will depend on the **type of patterns observed in the time series** and the **type of model** (time series/causal)

Regression

In regression modelling the systematic component, μ_t is $f(X_1, X_2, X_3 \ldots\ldots X_k)$ where X_j are explanatory variables

$$Y_t = f(X_1, X_2, X_3 \ldots\ldots X_k) + \varepsilon_t$$

The **exact functional form** and the particular **independent variables (X_j)** to be included in the model is a matter of judgement.

A regression model is a **causal model** since the prediction of the target time series is linked to other time series.

Why Use Regression?

MACQUARIE University

Advantages:

1. Regression allows the forecaster to incorporate **theoretical knowledge** of the time series, independent variables and the functional form

2. Regression provides a **"causal" explanation** of why the prediction may be appropriate.

3. Regression models can be used to provide **strategy and scenario based prediction**. In particular, regression can be used to analyse the **best and worst case** scenarios. This provides some indication of the **range of likely values** for the target time series and helps management identify sources of **risk**

Why use Regression? (cont)

Disadvantages:

1. Regression requires much more data than other forecasts methods. **Data is required on the target time series and the independent variables**. In addition more **theoretical knowledge of the time series generating process** is required

2. Regression analysis requires **more resources (time,money,skill)** to produce forecasts than previously examined time series methods. This time and effort **may not necessarily result in greater predictive accuracy**.

Regression may prove to be an expensive, time consuming way of producing inferior forecasts.

When to use Regression?

Regression is a **relatively time consuming and expensive way** of generating forecasts

Typically the use of regression should be for forecasts of **some importance** for the firm or organisation or when strategic options and/or scenario analysis is required

The **benefits to the firm** (improved understanding, scenarios) must outweigh the **considerable costs**

Since regression is resource hungry it should only be undertaken when there are **sufficient resources (time, money, data etc)** to enable a proper regression analysis.

Regression Forecasting

There are basically three tasks involved

1. **Choose an appropriate model**. This includes independent variable selection and choosing the specific functional form of the model

2. **Use a joint sample of observations** on the dependent and independent variables to derive **estimates of the regression coefficients**

3. Use the estimated model and **predicted values of independent variables** to generate **forecasts of the dependent variable**

Choosing an Appropriate Model

Choosing appropriate independent variables relies on **economic theory, logic, the observed time series and the experience of the modeller**

Typically, the modeller considers the above and selects a **candidate group of variables** which may be independent variables in a final regression model

Functional form is another issue. Once again use logic, theory, experience and the observed time series (**Linear v Non-Linear, Statics v Dynamics, Levels v Changes)**

How the predictive model will be used and **what information and/or forecasts are required** may also influence the functional form.

Estimation

Basic Functional Form (**population** model)

$$Y_t = \beta_o + \beta_1 X_{1t} + \beta_2 X_{2t} + \ldots + \beta_k X_{kt} + \varepsilon_t$$

A **joint** sample on Y_t and $X_{1t,\ 2t,\ kt}$ is collected.

Estimation of regression model coefficients **($b_{o,}$ $b_{1\ ,}$ $b_{2\ldots}$ b_k)** typically via **Ordinary Least Squares (OLS)** in EXCEL or Minitab. This generates **the sample regression model**

$$E(Y_t) = b_o + b_1 X_{1t} + b_2 X_{2t} + \ldots + b_k X_{kt}$$

Why Use OLS?

MACQUARIE University

OLS estimates have good forecast properties

Under the conditions the model is correctly specified and the random error at any observation is independently derived with zero mean and constant variance the OLS estimates and forecasts will be

1. **Unbiased** - **on average** the OLS estimates and forecasts will be equal to the true values

2. **Efficient**- the OLS estimators and forecasts will be the most precise of any **linear unbiased estimators or forecasts.**

Estimation (cont)

Before the model can be used for forecasts it needs to be checked for **adequacy** and **violations of assumptions underpinning OLS**

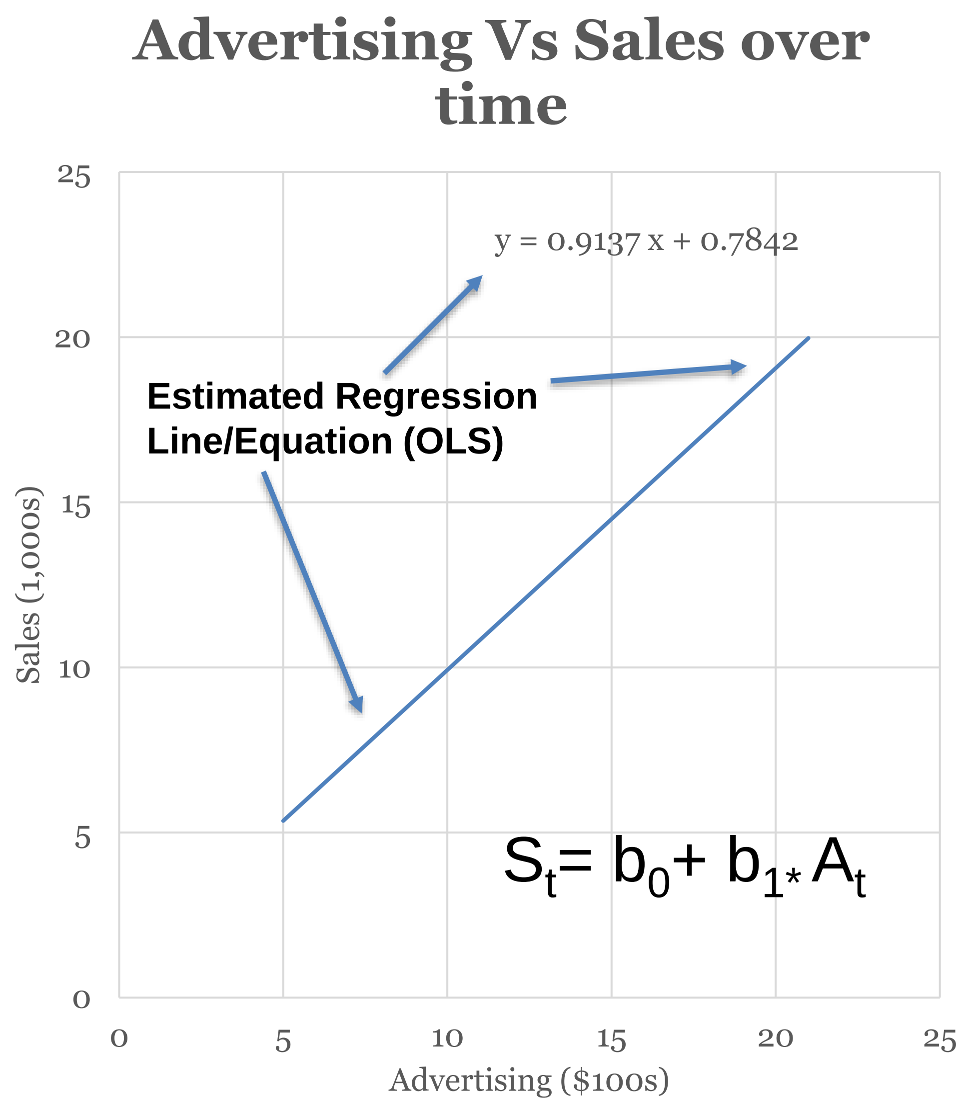
Assumptions of OLS include **correctly specified functional form** and **error term (ε_t) behaviour**

Residuals, other **diagnostics** and associated **relevant statistical tests** used to determine the adequacy of model

Only after examination of the above diagnostics and determination of adequacy **should the estimated model be used for forecasts**

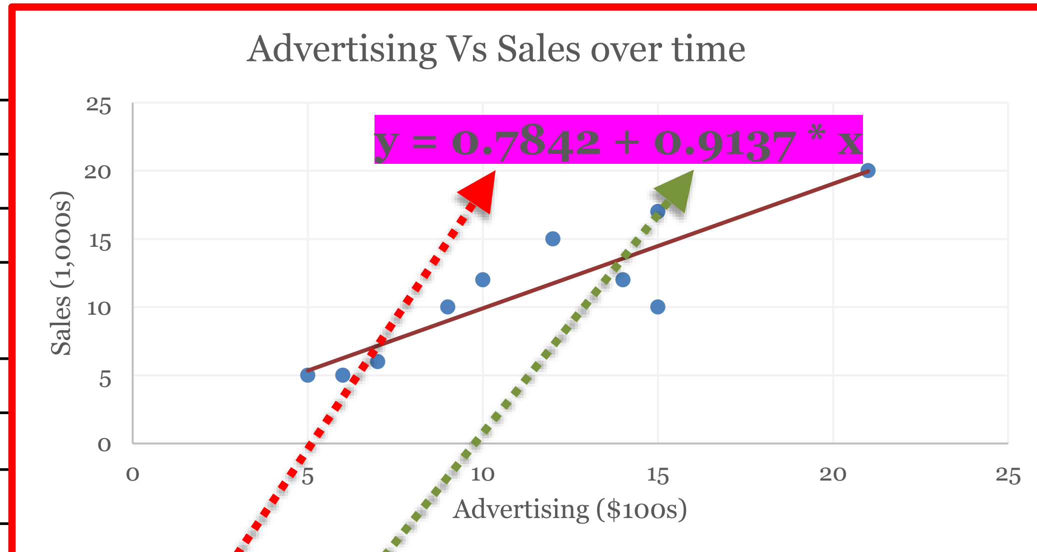
Regression: Example 1

Week	Sales (Y) (1,000s)	Advertising (X) ($100s)
1	10	9
2	6	7
3	5	5
4	12	14
5	10	15
6	15	12
7	5	6
8	12	10
9	17	15
10	20	21



Excel: Regression Output

SUMMARY OUTPUT



Regression Statistics						
Multiple R	0.891					
R Square	0.795					
Adjusted R Square	0.769					
Standard Error	2.448					
Observations	10					
ANOVA						
	df	**SS**	**MS**	**F**	**Sig F**	
Regression	1	185.658	185.658	30.980	0.001	
Residual	8	47.942	5.993			
Total	9	233.6				
	Coefficients	**Standard Error**	**t Stat**	**P-value**	**Lower 95%**	**Upper 95%**
Intercept	0.784	2.025	0.387	0.709	-3.886	5.454
Advertising	0.914	0.164	5.566	0.001	0.535	1.292

Regression: Example 1 (cont.)

The sample estimated equation is

$$Sales = 0.7842 + 0.9137 * A$$

(Estimated through Excel or MINITAB)

Intercept: **0.7842** – Estimated Sales when A = 0

Slope: **0.9137** – Estimated constant increase in Sales (000's) when Advertising increases by 1 unit ($100)

Forecasts:

When A = 10: $S = 0.7842 + 0.9137 * 10 = 9.921$ (000's)

When A = 15: $S = 0.7842 + 0.9137 * 15 = 14.490$ (000's)

Measures of Estimated Model Performance

R² - Coefficient of Determination:

- R^2 is the % of dependent variable variation (sample) explained by the estimated regression

- **R^2 is between 0 and 1** and the closer the R^2 to 1 the better the estimated model fits the sample data.

- EXCEL calculates R^2 as part of the standard regression estimation routine

- In practice, many unskilled modellers place too much emphasis on R^2 or a close counterpart **R^2 adjusted**.

- R^2 has many flaws and can be easily manipulated. Don't place too much reliance on it but use it as one tool of many in deciding the suitability of models

Performance Measures (cont.)

Standard Error

The standard error is approximately the "average" residual of the regression.

It is similar although not identical to **RMSE**

Provided in **standard regression output** in **EXCEL and Minitab**

Standard error can be used as comparison between competing regression models

More on Performance Measures

Out of sample forecast performance:

R^2 and standard error are indicators of the **in- sample** predictive ability of the model.

In-sample prediction is easier since both **dependent and independent variable data are available** and used to obtain the "best" model (most accurate)

For out- of-sample prediction, the dependent variable is not available (that's why we are predicting it!). The values of the independent variables may also need to be estimated

The predictive ability of the model will be different out of sample. The forecaster should test the out-of-sample predictive ability by leaving aside a **portion of the most recent observations as a test set.** Usual error criteria can be used.

Excel: Regression Output

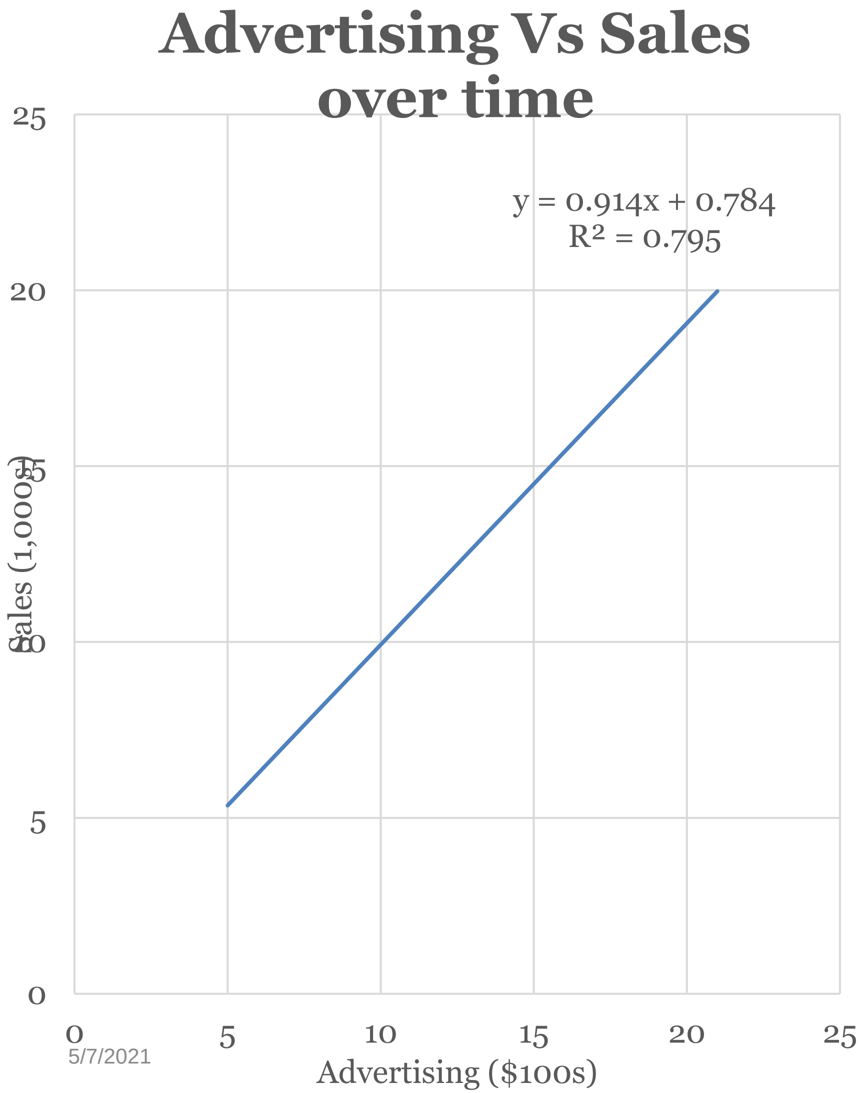
SUMMARY OUTPUT

Standard Error = 2.448

R-sq = SSR/SST =185.7/233.6 =0.795

Regression Statistics								
Multiple R	0.891							
R Square	0.795							
Adjusted R Square	0.769		MSE = Mea					
Standard Error	2.448							
Observations	10							
ANOVA								
	df	SS	MS	F	Sig F			
Regression	1	185.658	185.658	30.980	0.001			
Residual	8	47.942	5.993					
Total	9	233.6						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	0.784	2.025	0.387	0.709	-3.886	5.454	-3.886	5.454
Advertising	0.914	0.164	5.566	0.001	0.535	1.292	0.535	1.292

Coefficient of Determination (R^2)



$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

79.5% of the sample variation in Sales is explained by the variation in Advertising expenditure

Statistical Testing

- Test for **overall model significance** – **F test**
- Test for **individual variable significance** – **t tests**

Testing Individual Variables

Testing Individual coefficients:

Separate tests of **population slope coefficients (β_j) being zero (null hypothesis)**

If the slope coefficient is zero it suggests the **independent variable being examined** does **not influence the dependent variable**

Further, the independent variable being examined **may** be an **irrelevant variable** and could possibly be **dropped from the model specification**

t test - check p-value **(<0.05 then Reject H_o)**

Individual Coefficient Tests

Single Co-efficient Tests:

Hypothesis tests can be applied to the co-efficients of all variables separately. For a model given by

$$Y = \beta_o + \beta_1 * X_1 + \beta_2 * X_2 + \ldots . \beta_k * X_k + \varepsilon$$

The relevant test (each co-efficient separately)

$$H_o: \beta_j = 0 \quad \text{vs} \quad H_1: \beta_j \neq 0$$

The test statistic has a **t distribution** with p-values indicating support for H_o or H_1.

Are Individual Variables Significant?

Use t tests of individual variable slopes

Shows if there is a relationship between the variable X_j and Y

Hypotheses:

H_0: $\beta_j = 0$ (no linear relationship exists between X_j and Y)

H_1: $\beta_j \neq 0$ (linear relationship does exist between X_j and Y)

Are Individual Variables Significant? - (2)

H_0: $\beta_j = 0$ (no relationship)

H_1: $\beta_j \neq 0$ (relationship does exist between X_j and Y)

Test Statistic:

$$\boxed{t = \frac{b_j - 0}{S_{b_j}}}$$

(df = n – k – 1)

Check p-value in output (<0.05 Reject H_0)

Excel: Example 1 - t tests from Regression Output

MACQUARIE University

SUMMARY OUTPUT

Sig-value (α) is the probability of a similar or more extreme sample t value given β is zero.

Regression Statistics								
Multiple R	0.891							
R Square	0.795							
Adjusted R Square	0.769							
Standard Error	2.448							
Observations	10							
ANOVA								
	df	SS	MS	F	S[illegible]			
Regression	1	185.658	185.658	30.980	[illegible]01			
Residual	8	47.942	5.993					
Total	9	233.6						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	0.784	2.025	0.387	0.709	-3.886	5.454	-3.886	5.454
Advertising	0.914	0.164	5.566	0.001	0.535	1.292	0.535	1.292

Check the Residuals

As in time series models, a necessary condition for adequacy of a forecast model are **non-systematic errors**

Check residuals for **randomness- visual inspection of residual plots** (**vs. time and vs. all explanatory variables separately**)

Examine **ACF and PACF** of residuals

Systematic residuals may indicate **violation of regression assumptions**

Other objective tests can be used (more on this next week)

Forecasting with Regression

The diagnostic tests are used as a tool to check specified models and to suggest **potential improvements to model specifications**

Models may be modified (according to diagnostic information) and the **process of estimation and diagnostic testing is repeated**

Once a **final** model is determined (with acceptable diagnostics) it is used for **forecasts**

Forecasts will use the **estimated equation** and **estimates of future X_j** values to forecast $\mathbf{Y_f}$