MACQUARIE University

Business Forecasting

Multiple Regression Models

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?

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**

Statistical Testing

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

Testing Overall Model Significance

Overall Significance test (F Test):

One joint test in particular is useful; we test the **null** hypothesis **all of the slope coefficients** in the population are jointly zero which is a test of the **explanatory power of the model**

Non-rejection of the null (all coefficients jointly zero) indicates that as a **group** the variables selected and the precise model chosen has **no significant explanatory power**

Rejection of the null indicates **some explanatory power of the model** and importantly potentially some predictive ability

F test - check p-value **(<0.05 then Reject H_0)**

Is the Model Significant?

F Test for Overall Significance of the Model

Shows if there is a relationship between all the X variables considered together and Y

Use F-test statistic

Hypotheses:

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ (no relationship)

H_1: at least one $\beta_i \neq 0$ (at least one independent variable affects Y)

F Test for Overall Significance

Test statistic: Given in **ANOVA** table in output

$$F = \frac{MSR}{MSE} = \frac{\dfrac{SSR}{k}}{\dfrac{SSE}{n-k-1}}$$

where F has (numerator) = k & (denominator) = (n – k - 1) degrees of freedom

k =(number of variables in model)

n = (number of observations)

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)

$$H_o: \beta_j = 0 \quad vs \quad H_1: \beta_j \neq 0$$

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)

MACQUARIE University

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 – F test from 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%	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

Sig-value (α) is estimated probability of a similar or more extreme sample F value given the null hypothesis is true

Excel: Example 1 - t tests from 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	[illegible]			
Regression	1	185.658	185.658	30.980	[illegible]			
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

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

Forecasting with Regression

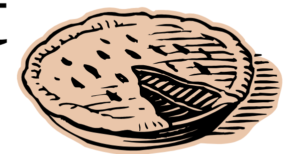
The diagnostic tests are used as a tool to check specified models and to suggest **potential improvements to model specifications**

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 $\mathbf{Y_f}$

Example 2: 2 Independent Variables



A distributor of frozen desert pies wants to evaluate factors thought to influence demand (Y)

Dependent variable: Y - Pie sales (units per week)

Independent variables:
Price (X_1) (in \$),
Advertising(X_2) (\$100's)

$$\text{Sales} = \beta_0 + \beta_1 (\text{Price}) + \beta_2 (\text{Advertising}) + \varepsilon_t$$

Week	Pie Sales	Price (\$)	Advertising (\$100s)
1	350	5.50	3.3
2	460	7.50	3.3
3	350	8.00	3.0
4	430	8.00	4.5
5	350	6.80	3.0
6	380	7.50	4.0
7	430	4.50	3.0
8	470	6.40	3.7
9	450	7.00	3.5
10	490	5.00	4.0
11	340	7.20	3.5
12	300	7.90	3.2
13	440	5.90	4.0
14	450	5.00	3.5
15	300	7.00	2.7

Multiple Regression Output

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$\text{Sales} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$$

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

The Multiple Regression Equation

$$\widehat{\text{Sales}} = 306.526 - 24.975 * (\text{Price}) + 74.131 * (\text{Advertising})$$

where
Sales is in number of pies per week
Price is in $
Advertising is in $100's.

$b_1 = -24.975$: sales will decrease, on average, by 24.975 pies per week for each $1 increase in selling price, net of the effects of changes due to advertising

$b_2 = 74.131$: sales will increase, on average, by 74.131 pies per week for each $100 increase in advertising, net of the effects of changes due to price

Using The Equation to Make Predictions/Forecasts

Predict sales for a week in which the selling price is \$5.50 and advertising is \$350:

$$\widehat{\text{Sales}} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$$

$$= 306.526 - 24.975\,(5.50) + 74.131\,(3.5)$$

$$= 428.62$$

Predicted sales is 428.62 (429) pies

Note that Advertising is in \$100's, so \$350 means that $X_2 = 3.5$

Coefficient of Determination R^2

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$R^2 = \frac{SSR}{SST} = \frac{29460.0}{56493.3} = .52148$$

52.1% of the variation in pie sales is explained by the variation in price and advertising

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Adjusted R²

MACQUARIE University

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$r^2_{adj} = .44172$$

44.2% of the variation in pie sales is explained by the variation in price and advertising, taking into account the sample size and number of independent variables

ANOVA	df	SS	MS	F	Significance F
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

F Test for Overall Significance

MACQUARIE University

(continued)

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$F = \frac{MSR}{MSE} = \frac{14730.0}{2252.8} = 6.5386$$

With 2 and 12 degrees of freedom

P-value for the F Test

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

F Test for Overall Significance (2)

MACQUARIE University

H_0: $\beta_1 = \beta_2 = 0$

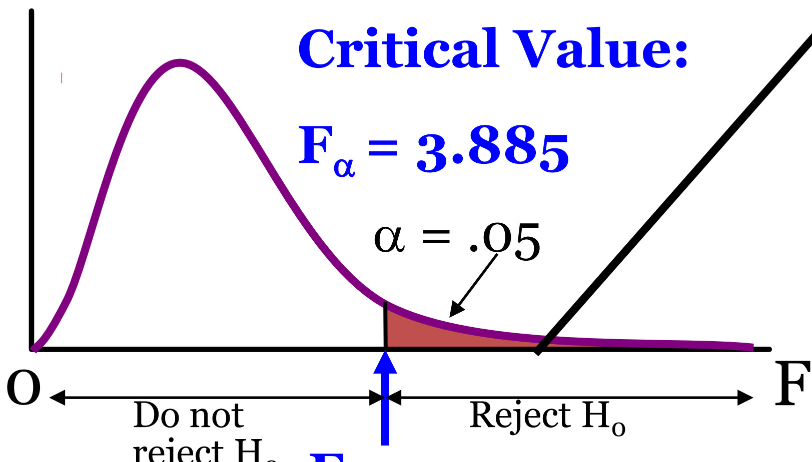
H_1: β_1 and β_2 not both zero

$\alpha = .05$

$df_1 = 2 \quad df_2 = 12$

Critical Value:

$F_\alpha = 3.885$



$F_{.05} = 3.885$

Test Statistic:

$$F = \frac{MSR}{MSE} = 6.5386$$

Decision:

Since F test statistic is in the rejection region (p-value < .05), reject H_0

Conclusion:

There is evidence that at least one independent variable affects Y

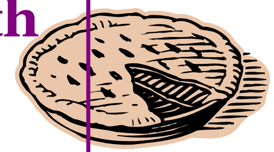
Are Individual Variables Significant?

(continued)

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

t-value for Price is t = -2.306, with p-value .0398 (Significant)

t-value for Advertising is t = 2.855, with p-value .0145 (Significant)



ANOVA	df	SS	MS	F	Significance F
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Inferences about the Slope: t test

$H_0: \beta_i = 0$

$H_1: \beta_i \neq 0$

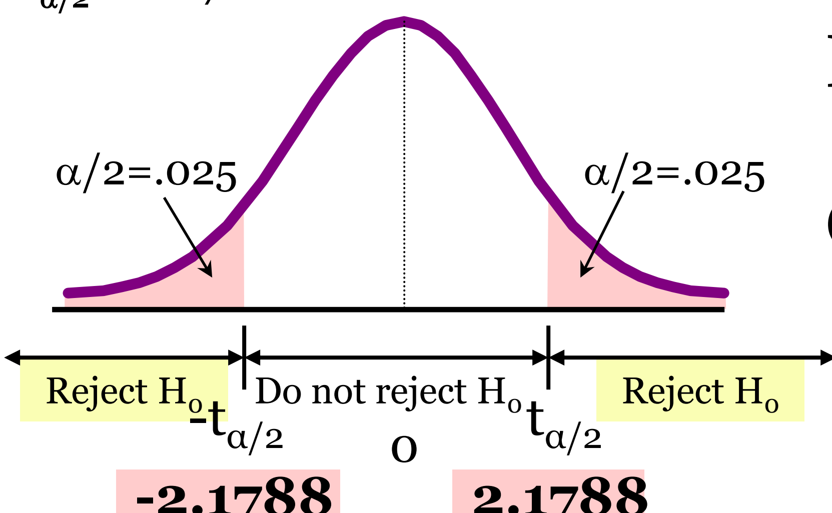
	Coefficients	Standard Error	t Stat	P-value
Price	-24.97509	10.83213	-2.30565	0.03979
Advertising	74.13096	25.96732	2.85478	0.01449

d.f. = 15-2-1 = 12

α = **.05**

$t_{\alpha/2}$ **= 2.1788**

The test statistic for each variable falls in the rejection region (p-values < .05)



Decision:

Reject H_0 for each variable

Conclusion:

There is evidence that both Price and Advertising affect Pie sales at α = .05