Analisi Statistica per le Imprese

Prof. L. Neri

Dip. di Economia Politica e Statistica

4.3. Multiple Regression Model

You should be able to:

- Understand model building using multiple regression analysis
- Apply multiple regression analysis to business decision-making situations
- Analyze and interpret the computer output for a multiple regression model
- Test the significance of the independent variables in a multiple regression model
- Incorporate qualitative variables into the regression model by using dummy variables

The systematic part

- One may need a mathematical model to quantify the existing relationship between a response variable Y and k explicative variables $X_1, \ldots, X_k$

$$y = f(X_1, \ldots X_k)$$

The multiple linear regression model specifies the functional relationship as

$$f(X_1, \ldots X_k) = \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_k X_k \quad (1)$$

Geometrically this corresponds to a hyper-plan in k dimensions. The model is extremely useful because:

1. It has an intuitive geometrical interpretation
2. It has a simple estimation of the model's parameters

The stochastic part

The model always includes a random component, that identifies the stochastic component. This can be expressed as follows:

$$Y = \underbrace{f(X_1, \ldots X_k)}_{\text{systematic}} + \underbrace{\varepsilon}_{\text{stochastic}} \tag{2}$$

The Model specification

Standard notation: for each statistiocal unit *i=1...n*

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \dots \beta_k X_{ik} + \varepsilon_i \quad (3)$$

Matrix notation

$$Y = X\beta + \varepsilon \quad (4)$$

$Y : (n \times 1)$ vector of n dependent variable observations
$X : (n \times k)$ matrix of k regressors with each n observations
$\beta : (k \times 1)$ vector of k parameters
$\varepsilon : (n \times 1)$ vector of n

The Matrices

$$X = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1k} \\ X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix}; \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}$$

$$y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}; \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

The matrix X will have a unitary first column if the model is with intercept. In this case the intercept would be β_1 in the multidimensional notation

Standard Assumptions

Assumptions

- The functional relationship must be linear
- Covariates must have a deterministic nature
- The X matrix has full rank
- The error term has a null expected value $E[\varepsilon_i] = 0$
- The error term is homosckedastic: $Var[\varepsilon_i] = \sigma^2$
- The error terms are not correlated: $Cov[\varepsilon_i \varepsilon_j]_{\forall i \neq j} = 0$

OLS estimator

The OLS estimator in multiple linear regression is the vector $\widehat{\beta}$ that minimize the following function of $\widetilde{\beta}$, where X_i is the i-th row of the X matrix.

$$\min \sum_{i=1}^{n} e_i^2 = min \sum \left(Y_i - X_i \widetilde{\beta} \right)^2 \tag{5}$$

$$e = \left(Y - X \widetilde{\beta} \right) \tag{6}$$

So

$$\widetilde{\beta} = \left(X' X \right)^{-1} X' Y \equiv \widehat{\beta} \tag{7}$$

Notice that the matrix $X'X$ (cross product matrix) must be of full rank in order to be invertible.

OLS estimator properties

$\widehat{\beta}$ is BLUE (Best Linear Unbiased Estimator)

One can notice that $\left(\left(X'X\right)^{-1}X'\right)$ is a matrix of constant elements, therefore $\widehat{\beta}$ is a linear transformation of Y.
One can prove that $\widehat{\beta}$ is a correct estimator as follows

$$\widehat{\beta} = \left(X'X\right)^{-1}X'Y = \left(X'X\right)^{-1}X'\left(X\beta + \varepsilon\right) = \beta + \left(X'X\right)^{-1}X'\varepsilon \tag{8}$$

$$E\left[\widehat{\beta}\right] = \beta + \left(X'X\right)^{-1}X'E\left[\varepsilon\right] = \beta \tag{9}$$

We could also prove that in the class of the linear and unbiased estimator is the one presenting the minimum variance.

Variance of the OLS estimator

The variance of the OLS estimator is calculated as follows:

$$Var\left(\widehat{\beta}\right) = E\left[\left(\widehat{\beta} - \beta\right)\left(\widehat{\beta} - \beta\right)'\right] \tag{10}$$

$$Var\left(\widehat{\beta}\right) = E\left[\left(X'X\right)^{-1} X'\varepsilon\varepsilon' X \left(X'X\right)^{-1}\right] \tag{11}$$

$$Var\left(\widehat{\beta}\right) = \left(X'X\right)^{-1} X' E\left[\varepsilon\varepsilon'\right] X \left(X'X\right)^{-1} = \sigma^2 \left(X'X\right)^{-1} X'X \left(X'X\right)^{-} \tag{12}$$

$$Var\left(\widehat{\beta}\right) = \sigma^2 \left(X'X\right)^{-1} \tag{13}$$

so for each parameter estimator

$$Var\left(\widehat{\beta}_j\right) = \sigma^2 \left[\left(X'X\right)^{-1}\right]_{jj} \tag{14}$$

Empirical example

A distributor of frozen desert pies wants to evaluate factors that influence demand

- Dependent variable: y= Pie sales (units per week)
- Independent variables: x_1=Price ($) and x_2=Advertising ($100's)
- Data is collected for 15 weeks

The OLS estimates gives the following estimated model:

$$\widehat{y} = 306 - 24x_1 + 74x_2 \tag{15}$$

Interpretation of the estimated coefficient

- each $\widehat{\beta}_j$ estimates the average value of Y changes by $\widehat{\beta}_j$ units for each 1 unit increase in X_j, holding all other independent variables constant
    - example: $\widehat{\beta}_1 = -24$ then sales (y) are expected to decrease, on average, by 24 pies per week for each $1 increase in selling price (x_1), net of the effects due to advertising (x_2)
    - example: $\widehat{\beta}_2 = 74$ then sales (y) is expected to increase, on average, by 74 pies per week for each $100 increase in advertising (x_2), net of the effects due to price (x_1)
- Intercept
    - the estimated average value of y when all X variables are zero.

Using the model to make predictions

We can calculate the predicted sales per week given the selling price (\$5) and advertising expenses (\$350):

$$\widehat{y} = 306 - 24\,(5) + 74\,(3.5) = 445 \tag{16}$$

We predict to sell 445 pies per week.

The σ^2estimator

$$e = Y - X\widehat{\beta} = X\beta + \varepsilon - X\left[\left(X^{'}X\right)^{-1}X^{'}\left(X\beta + \varepsilon\right)\right] \tag{17}$$

$$e = X\beta + \varepsilon - X\beta - X\left(X^{'}X\right)^{-1}X^{'}\varepsilon \tag{18}$$

$$e = \left(1 - X\left(X^{'}X\right)^{-1}X^{'}\right)\varepsilon = M_x\varepsilon \tag{19}$$

M_x is a idempotent symmetric matrix. This implies that:

$$M_x = M_x^{'} = M_x^k; \quad \forall k > 0 \tag{20}$$

The σ^2 estimator

$$Q = e^{'} e = \varepsilon^{'} M_x^{'} M_x \varepsilon = \varepsilon^{'} M_x \varepsilon \tag{21}$$

we can prove that

$$E\left[Q\right] = \sigma^2 \left(n - k\right) \tag{22}$$

so the unbiased estimator of the parameter σ^2 is

$$E\left[\frac{Q}{n-k}\right] = \sigma^2 \Rightarrow \widehat{\sigma}^2 = \frac{Q}{n-k} = \frac{e^{'} e}{n-k} \tag{23}$$

ANOVA

Adding to the standard assumptions the following

- The error term has a normal distribution so:$\varepsilon_i \sim N(0, \sigma^2)$

Reminding that

$$\widehat{\beta} = \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \vdots \\ \widehat{\beta}_k \end{pmatrix} = \left(X^{'}X\right)^{-1} X^{'}Y \tag{24}$$

follows

$$Y_i = N\left(X\beta, \sigma^2\right) \tag{25}$$

$$\widehat{\beta}_i = N\left(\beta_i, \sigma^2 \left\lfloor \left(X^{'}X\right)^{-1} \right\rfloor_{ii}\right) \tag{26}$$

$$\widehat{\beta} = N\left(\beta, \sigma^2 \left(X^{'}X\right)^{-1}\right) \tag{27}$$

ANOVA

In order to test the meaning of the whole model, we need to test the hypothesis system

$$H_0 : \beta_1 = \beta_2 == ....\beta_k = 0$$

$$H_1 : \, almeno\, un\, \beta_j \neq 0$$

the test is

$$F = \frac{ESS/(k-1)}{RSS/(n-k)} = \frac{R^2/(k-1)}{\left(1-R^2\right)/n-k} \sim F_{(k-1,n-k)}$$

ANOVA

	SS	d.f.	Mean Square (MS)
Model	ESS $= \widehat{\beta}' X' y = y' y R^2$	k-1	$\widehat{\beta}' X' y / (k-1)$
Residual	RSS $= e' e = y' y \left(1 - R^2\right)$	n-k	$e' e / (n-k)$
Total	TSS $= y' y = \sum y_i^2$	n-1	

Table 1:

- Construct the F statistic $F = \frac{ESS/(k-1)}{RSS/(n-k)}$
- Find the 95^{th} or the 99^{th} quantile of the distribution $F_{(k-1),(n-k)}$
- If $F > F_{(1-\alpha);(k-1),(n-k)}$ one rejects

R^2

$$R^2 = \frac{ESS}{TSS} \geq 0 \tag{28}$$

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum Y_i^2} \leq 1 \tag{29}$$

$$0 \leq R^2 \leq 1 \tag{30}$$

Adjusted R^2

- R^2 never decreases when a new X variable is added to the model
  - this can be a disadvantage when comparing different models
- What is the net effect of adding a new variable?
  - we lose a degree of freedom when a new X variable is added
  - did the new X variable add enough explanatory power to offset the loss of one degree of freedom?
- The adjusted R^2 adjusts for the number of variables (k).

$$R^2_{adj} = 1 - \frac{\sum e_i^2/(n-k)}{\sum Y_i^2/(n-1)} = 1 - \frac{(n-1)}{(n-k)}\left(1 + R^2\right) \qquad (31)$$

Empirical example: ANOVA

	SS	d.f.	MS
Model	29460	2	14730
Residual	27033	12	2252
Total	56493	14	

Table 2:

$$R^2 = 29460/56493 = 0.52 \tag{32}$$

52% of the variation in pie sales is explained by the variation in price and advertising

$$R^2_{adj} = 0.44 \tag{33}$$

44% of the variation in pie sales is explained by the variation in price and advertising, taking into account the sample size and the number of independent variables

Empirical example: is the model significant?

- F-test for the overall significance of the model
- Shows if there is a linear relationship between all of the X variables considered together and Y
- use F test statistic
  - in the estimated model the empirical value of F is equal to 6.54
  - the critical value of $F_{0.05;2,12}$ is equal to 3.88
  - $6.54 > 3.88$ therefore the regression model explains a significant portion of the variation in pie sales (there is evidence that at least one independent variable affects y)

Single parameter t-test

$$\widehat{\beta} \sim N\left(\beta, \sigma^2 \left\lfloor \left(X'X\right)^{-1} \right\rfloor\right) \tag{34}$$

To test the hypothesis if the individual variable X_i has a significant effect on Y we have to test
H_0:β_i=0 (no linear relationship)
H_1:$\beta_i \neq$0 (linear relationship does exist between X_i and Y)
if σ^2 is known, under H_0 :

$$\frac{\widehat{\beta}_i - 0}{\sqrt{\sigma^2 \left\lfloor (X'X)^{-1} \right\rfloor_{ii}}} \sim N(0,1) \tag{35}$$

Single parameter t-test

Generally σ^2 is unknown and we have to use its estimator

$$\widehat{\sigma}^2 = \frac{e^{'}e}{n-k} \tag{36}$$

Therefore the Standard Errors of β is

$$se_{\widehat{\beta}} = \sqrt{\widehat{\sigma}^2 \left\lfloor (X'X)^{-1} \right\rfloor} \tag{37}$$

Single parameter t-test

under H_0 we have:

$$\frac{\widehat{\beta}_i - 0}{\frac{\sqrt{\sigma^2\left\lfloor\left(X'X\right)^{-1}\right\rfloor_{ii}}}{\sqrt{\frac{e'e}{\sigma^2}/(n-k)}}} = \frac{\widehat{\beta}_i}{se_{\widehat{\beta}}} \sim t_{n-k} \qquad (38)$$

- Find the 95^{th} or the 99^{th} quantile of the distribution $t_{(n-k)}$
- if $|t| > t_{(1-\alpha/2);(n-k)}$ one rejects H_0

Empirical example: are individual variables significant?

	coefficients	Standard Error	t
Intercept	306	114.25	2.67
Price	-24	10.83	-2.21
Advertising	74	25.96	2.85

Table 3:

At α =0.05 significant level, the t-value for price is
$|$-2.21$|$>t$_{(\alpha/2;12)} = 2.1788$ so we refuse H_0
At α =0.05 significant level, the t-value for advertising is
$|$2.85$|$>t$_{(\alpha/2;12)}$=2.1788 so we refuse H_0
There is evidence that both Price and Advertising affect pie sales at α =0.05

Dummy Variables

Until now we have assumed that in $Y = X\beta + u$, X are cardinal variables.
One can also use categorial explanatory (i.e. "dummy") variables that identify specific factors depending on categories:

- Temporal effects
- Spacial effects
- Qualitative variables

We suppose that the Dummies influence just the model intercept (not the slopes).
There will be different regression interecepts corresponding to the different groups/situations, if the dummy variable is significant.

Dummy Variables

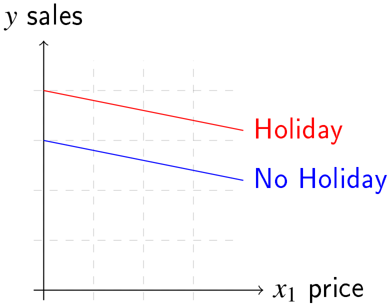
Lets assume the following generic model:

$$\widehat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{39}$$

Where Y is the pie sale, X_1 is the price and X_2 is a holiday indicator function, that will be equal to 1 when a holiday has occurred during the week, and equal to 0 when there is no holiday during the week.

Dummy Variables

$$\begin{array}{lccccc}
\widehat{Y} = \beta_0 + \beta_1 X_1 + \beta_2(1) = & (\beta_0 + \beta_2) & + & \beta_1 & X_1 \\
\widehat{Y} = \beta_0 + \beta_1 X_1 + \beta_2(0) = & \underbrace{(\beta_0)}_{\substack{\textit{different}\\ \textit{intercept}}} & + & \underbrace{\beta_1}_{\substack{\textit{same}\\ \textit{slope}}} & X_1
\end{array} \tag{40}$$



If $H_0 : \beta_2 = 0$
is rejected
Holiday has
a significant effect
on pie sales

Dummy Variables

Example:

$$Sales = 300 - 30\,(Price) + 15\,(Holiday) \tag{41}$$

- Sales = number of pies sold per week
- Price = pie price in \$
- Holiday = $\begin{cases} 1 & \text{if holiday has occurred} \\ 0 & \text{if holiday has not occurred} \end{cases}$

As we see the dummy coefficient $\beta_2 = 15$. This implies that on average 15 more pies were sold in weeks with holidays then in weeks without holidays, given the same price.

Dummy Variables

The number of dummy variables must always be one less then the number of discriminated levels. Imagine that we are analyzing the house market have three different housing levels {ranch, split level, condo}.
Example:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \tag{42}$$

Y = house price
X_1 = square meters

$$X_2 = \begin{cases} 1 & \text{if ranch} \\ 0 & \text{if not} \end{cases} \Rightarrow \beta_2 \text{ impact of ranch vs. condo}$$

$$X_3 = \begin{cases} 1 & \text{if split level} \\ 0 & \text{if not} \end{cases} \Rightarrow \beta_3 \text{ split level vs. condo}$$

Dummy Variables

Suppose the estimated equation is:

$$\hat{Y} = 20 + 0.05X_1 + 24X_2 + 15X_3 \tag{43}$$

The we will have as follows:

For a condo $(x_2 = x_3 = 0)$:

$$\hat{y} = 20 + 0.05X_1 \tag{44}$$

For a ranch $(x_3 = 0)$:

$$\widehat{y} = 44 + 0.05X_1 \tag{45}$$

For a split level $(x_2 = 0)$:

$$\widehat{y} = 35 + 0.05X_1 \tag{46}$$

- Pindyck R. and Rubinfeld D., 1998 “Econometrics Models and Economic Forecasts” (4-th Ed.)
- De Luca Amedeo,1996 “Marketing Bancario e Metodi Statistici Applicati”, vol. 1, Franco Angeli.