Chapter 4

Further development and analysis of the classical linear regression model

Generalising the Simple Model to Multiple Linear Regression

- Before, we have used the model

$$y_t = \alpha + \beta x_t + u_t, \quad t = 1,2,...,T$$

- But what if our dependent variable (y) depends on more than one independent variable?

  For example the **number of cars sold** might plausibly depend on

  1. the price of cars
  2. the price of public transport
  3. the price of petrol
  4. the extent of the public's concern about global warming

- Similarly, **stock returns** might depend on several factors: inflation, the difference in returns on short and long dated bonds, industrial production, default risks.
- Having just one independent variable is no good in this case - we want to have more than one x variable. It is very easy to **generalise the simple model to the one with *k* regressors.**

Multiple Regression and the Constant Term

- Now we write

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \ldots + \beta_k x_{kt} + u_t \quad , t=1,2,\ldots,T$$

- Where is x_1? By convention it is the constant column of length T:

$$x_1 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

- *k*: **the number of regressors** *(including* x_1), which is the number of parameters β_i

Different Ways of Expressing the Multiple Linear Regression Model

- We could write out a separate equation for every value of t:

$$
\begin{aligned}
y_1 &= \beta_1 + \beta_2 x_{21} + \beta_3 x_{31} + \ldots + \beta_k x_{k1} + u_1 \\
y_2 &= \beta_1 + \beta_2 x_{22} + \beta_3 x_{32} + \ldots + \beta_k x_{k2} + u_2 \\
&\vdots \qquad\qquad\quad \vdots \qquad\qquad\quad \vdots \\
y_T &= \beta_1 + \beta_2 x_{2T} + \beta_3 x_{3T} + \ldots + \beta_k x_{kT} + u_T
\end{aligned}
$$

- We can write this in matrix form

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$$

where
y is $T \times 1$
X is $T \times k$
β is $k \times 1$
u is $T \times 1$

Inside the Matrices of the Multiple Linear Regression Model

- Example: if k is 2, we have 2 regressors, one of which is a column of ones:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix} = \begin{bmatrix} 1 & x_{21} \\ 1 & x_{22} \\ \vdots & \vdots \\ 1 & x_{2T} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{bmatrix}$$

$$T \times 1 \qquad T \times 2 \qquad 2 \times 1 \qquad T \times 1$$

- Notice that the matrices written this way are comfortable, ***x_{it} is the element in the t^{th} row and i^{th} column***

How Do We Calculate the Parameters (the β) in this Generalised Case?

- Previously, we took the residual sum of squares, and minimised it w.r.t. α and β.
- Similarly, with the matrix notation, we have

$$\hat{u} = \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_T \end{bmatrix}$$

- The **Residual Sum Square** (**RSS**) would be given by

$$\hat{u}'\hat{u} = \begin{bmatrix} \hat{u}_1 & \hat{u}_2 & \dots & \hat{u}_T \end{bmatrix} \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \\ \hat{u}_T \end{bmatrix} = \hat{u}_1^2 + \hat{u}_2^2 + \ldots + \hat{u}_T^2 = \sum \hat{u}_t^2$$

The OLS Estimator for the Multiple Regression Model

- In order to obtain the parameter estimates, β_1, β_2,..., β_k, we would minimise the RSS with respect to all the βs.

- It can be shown that

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = (X'X)^{-1}X'y$$

Calculating the Standard Errors for the Multiple Regression Model

- Check the dimensions: $\hat{\beta}$ is $k \times 1$ as required.

- But how do we calculate the standard errors of the coefficient estimates?

- Previously, to estimate the variance of the errors, σ^2, we use $s^2 = \frac{\hat{u}'\hat{u}}{T-k}$

  where k = number of parameters.
- Parameter Variance-Covariance matrix: $Var(\hat{\beta}) \approx s^2 (X'X)^{-1}$
- Then $SE(\hat{\beta})$ is obtained by taking square root of the leading diagonal of the Parameter Variance-Covariance matrix

Calculating Parameter and Standard Error Estimates for Multiple Regression Models: An Example

- **Example**: The following model with k=3 is estimated over 15 observations:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

and the following data have been calculated from the original X's.

$$(X'X)^{-1} = \begin{bmatrix} 2.0 & 3.5 & -1.0 \\ 3.5 & 1.0 & 6.5 \\ -1.0 & 6.5 & 4.3 \end{bmatrix}, (X'y) = \begin{bmatrix} -3.0 \\ 2.2 \\ 0.6 \end{bmatrix}, \hat{u}'\hat{u} = 10.96$$

Calculate the coefficient estimates and their standard errors.

- To calculate the coefficients, just multiply the matrix by the vector to obtain

$$\hat{\beta} = \left(X'X\right)^{-1} X'y$$

- To calculate the standard errors, we need an estimate of σ^2.

$$s^2 = \frac{RSS}{T-k} = \frac{10.96}{15-3} = 0.91$$

Calculating Parameter and Standard Error Estimates for Multiple Regression Models: An Example (cont'd)

- The variance-covariance matrix of $\hat{\beta}$ is given by

$$s^2(X'X)^{-1} = 0.91(X'X)^{-1} = \begin{bmatrix} 1.83 & 3.20 & -0.91 \\ 3.20 & 0.91 & 5.94 \\ -0.91 & 5.94 & 3.93 \end{bmatrix}$$

- The variances are on the leading diagonal:

$$\begin{aligned} Var(\hat{\beta}_1) &= 1.83 \qquad & SE(\hat{\beta}_1) &= 1.35 \\ Var(\hat{\beta}_2) &= 0.91 \Leftrightarrow & SE(\hat{\beta}_2) &= 0.96 \\ Var(\hat{\beta}_3) &= 3.93 & SE(\hat{\beta}_3) &= 1.98 \end{aligned}$$

- We write: $\hat{y} = 1.10 - 4.40x_{2t} + 19.88x_{3t}$

$$\quad (1.35) \quad (0.96) \quad (1.98)$$

Testing Multiple Hypotheses: The *F*-test

- We used the *t*-test to test single hypotheses, i.e. hypotheses involving only one coefficient. But what if we want to test more than one coefficient simultaneously?

- We do this using the *F*-test. The *F*-test involves estimating 2 regressions.

- The **unrestricted regression** is the one in which the coefficients are freely determined by the data, as we have done before, with residual sum of squares **URSS**

- The **restricted regression** is the one in which the coefficients are restricted, i.e. the restrictions are imposed on some βs, with residual sum of squares **RRSS**

The *F*-test: Restricted and Unrestricted Regressions

- **Example:**

  The general regression is

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + u_t \qquad (1)$$

- We want to test the restriction that $\beta_3 + \beta_4 = 1$ (we have some hypothesis from theory which suggests that this would be an interesting hypothesis to study). The **unrestricted regression is (1) above**, but what is the restricted regression?

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + u_t \quad s.t.\ \beta_3 + \beta_4 = 1$$

- We substitute the restriction ($\beta_3 + \beta_4 = 1$) into the regression so that it is automatically imposed on the data.

$$\beta_3 + \beta_4 = 1 \Rightarrow \beta_4 = 1 - \beta_3$$

The *F*-test: Forming the Restricted Regression

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + (1-\beta_3)x_{4t} + u_t$$
$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + x_{4t} - \beta_3 x_{4t} + u_t$$

- Gather terms in β's together and rearrange

$$(y_t - x_{4t}) = \beta_1 + \beta_2 x_{2t} + \beta_3 (x_{3t} - x_{4t}) + u_t$$

- This is the restricted regression. We actually estimate it by creating two new variables, call them, say, P_t and Q_t.

$$P_t = y_t - x_{4t}$$
$$Q_t = x_{3t} - x_{4t}$$

so

$\boldsymbol{P_t = \beta_1 + \beta_2 x_{2t} + \beta_3 Q_t + u_t}$ **is the restricted regression** we actually estimate.

Calculating the F-Test Statistic

- The **test statistic** is given by

$$test\ statistic = \frac{RRSS - URSS}{URSS} \times \frac{T-k}{m}$$

where $URSS$ = RSS from unrestricted regression

$RRSS$ = RSS from restricted regression

m = number of restrictions

T = number of observations

k = number of regressors in unrestricted regression including the constant (or the total number of parameters to be estimated).

The *F*-Distribution

- The test statistic follows the *F*-distribution, which has 2 d.f. parameters.

- The value of the **degrees of freedom parameters are (m, T-k)** respectively (the order of the d.f. parameters is important).

- The appropriate critical value will be in column m, row (T-k).

- The *F*-distribution has only positive values and is not symmetrical. **We therefore only reject the null if the test statistic > critical *F*-value**

Determining the Number of Restrictions m in an *F*-test

- Examples:

H_0: hypothesis	No. of restrictions, m
$\beta_1 + \beta_2 = 2$	1
$\beta_2 = 1$ and $\beta_3 = -1$	2
$\beta_2 = 0$, $\beta_3 = 0$ and $\beta_4 = 0$	3

- If the model is $y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + u_t$,
  then the null hypothesis
  H_0: $\beta_2 = 0$, and $\beta_3 = 0$ and $\beta_4 = 0$ is tested by "**THE regression *F*-statistic", or "junk regression"**. It tests the null hypothesis that all of the coefficients except the intercept coefficient are zero.

- Note the form of the alternative hypothesis for all tests when more than one restriction is involved: **H_1**: $\beta_2 \neq 0$, or $\beta_3 \neq 0$ or $\beta_4 \neq 0$

What we Cannot Test with Either an *F* or a *t*-test

- We cannot test using this framework hypotheses which are not linear or which are multiplicative, e.g.

  $H_0: \beta_2 \beta_3 = 2$ or $H_0: \beta_2^{\,2} = 1$

  cannot be tested.

The Relationship between the *t* and the *F*-Distributions

- Any hypothesis which could be tested with a *t*-test could have been tested using an *F*-test, but not the other way around.

  For example, consider the hypothesis

  H_0: $\beta_2 = 0.5$

  H_1: $\beta_2 \neq 0.5$

  We could have tested this using the usual *t*-test: $test\ stat = \dfrac{\hat{\beta}_2 - 0.5}{SE(\hat{\beta}_2)}$

  or it could be tested in the framework above for the *F*-test.
- Note that the two tests always give the same result since the *t*-distribution is just a special case of the *F*-distribution.
- **If we have some random variable $Z \sim t\,(T\text{-}k)$ then $Z^2 \sim F(1,T\text{-}k)$**

F-test Example

- **Question**: Suppose a researcher wants to test whether the returns on a company stock (y) show unit sensitivity to two factors (factor x_2 and factor x_3) among three considered. The regression is carried out on 144 monthly observations. The regression is $y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + u_t$

  - What are the restricted and unrestricted regressions?

  - If the two RSS are 436.1 and 397.2 respectively, perform the test.

- **Solution**:

  Unit sensitivity implies $H_0: \beta_2=1$ and $\beta_3=1$. The unrestricted regression is the one in the question. The restricted regression is $(y_t - x_{2t} - x_{3t}) = \beta_1 + \beta_4 x_{4t} + u_t$ or letting $z_t = y_t - x_{2t} - x_{3t}$, the restricted regression is $z_t = \beta_1 + \beta_4 x_{4t} + u_t$

  In the *F*-test formula, $T=144$, $k=4$, $m=2$, $RRSS=436.1$, $URSS=397.2$

  F-test statistic = 6.68. Critical value is an $F(2,140) = 3.07$ (5%) and 4.79 (1%).

  Conclusion: Reject H_0.

Goodness of Fit Statistics

- We would like some measure of how well our regression model actually fits the data.
- We have goodness of fit statistics to test this: i.e. how well the sample regression function (SRF) fits the data.
- The most common goodness of fit statistic is known as R^2. One way to define R^2 is to say that it is the square of the correlation coefficient between y and $\hat{y}$.
- For another explanation, recall that what we are interested in doing is explaining the variability of y about its unconditional mean $\bar{y}$, i.e. the **total sum of squares *TSS***:

$$TSS = \sum_t \left(y_t - \bar{y}\right)^2$$

- We can split the *TSS* into two parts: the part explained by the model **(explained sum of squares *ESS*)**, and the part not explained by the model **(residual sum of squares *RSS*)**.

Defining R^2

- That is, $TSS = ESS + RSS$

$$\sum_t (y_t - \bar{y})^2 = \sum_t (\hat{y}_t - \bar{y})^2 + \sum_t \hat{u}_t^2$$

- Our **goodness of fit statistic** is

$$R^2 = \frac{ESS}{TSS}$$

- But since $TSS = ESS + RSS$, we can also write

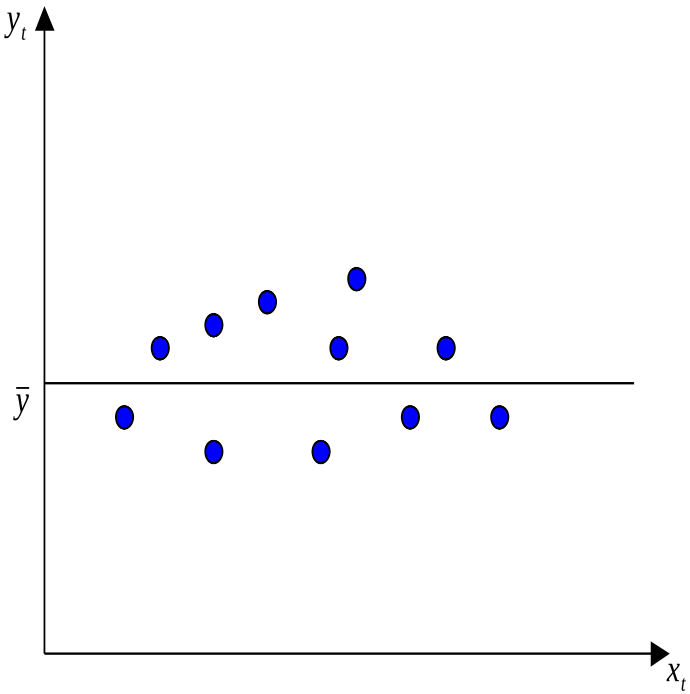
$$R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

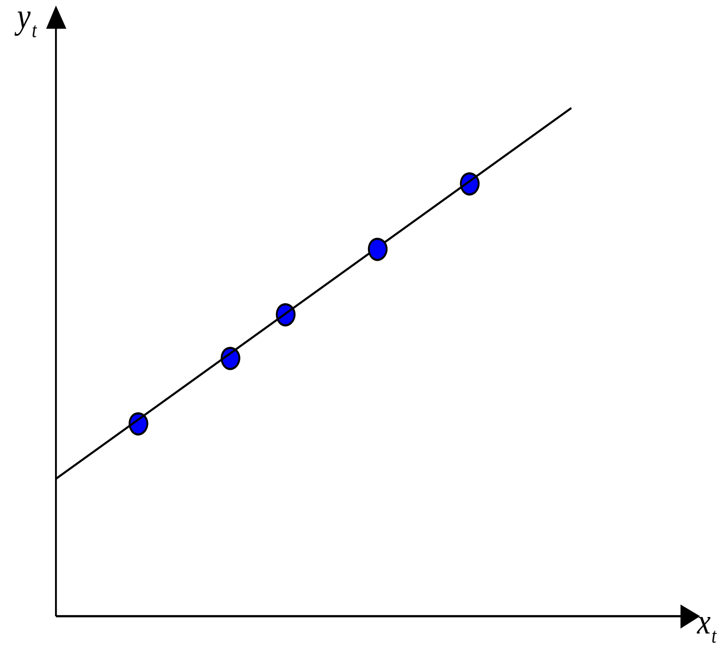
- R^2 must always lie between zero and one. To understand this, consider two extremes

  $RSS = TSS$ i.e. $ESS = 0$ so $R^2 = ESS/TSS = 0$

  $ESS = TSS$ i.e. $RSS = 0$ so $R^2 = ESS/TSS = 1$

The Limit Cases: $R^2 = 0$ and $R^2 = 1$





Problems with R^2 as a Goodness of Fit Measure

- There are a number of them:

1. R^2 is defined in terms of variation about the mean of y so that if a model is reparameterised (rearranged) and the dependent variable changes, R^2 will change.

2. R^2 never falls if more regressors are added to the regression, e.g. consider:

Regression 1: $y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t$

Regression 2: $y = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + u_t$

R^2 will always be at least as high for regression 2 relative to regression 1.

Adjusted R^2

- In order to get around these problems, a modification is often made which takes into account the loss of degrees of freedom associated with adding extra variables. This is known as $\bar{R}^2$, or **adjusted R^2**:

$$\bar{R}^2 = 1 - \left[\frac{T-1}{T-k}(1-R^2)\right]$$

- So if we add an extra regressor, k increases and unless R^2 increases by a more than offsetting amount, $\bar{R}^2$ will actually fall.

- Example: If **adjusted R^2** is 0.65 then 65% of the total variability of the dependent variable about the mean is explained by the model.

A Regression Example:
Hedonic House Pricing Models

- **Hedonic models** are used to value real assets, especially housing, and view the asset as representing a bundle of characteristics.
- Des Rosiers and Thérialt (1996) consider the effect of various amenities on rental values for buildings and apartments 5 sub-markets in the Quebec area of Canada.
- The **rental value** in Canadian Dollars per month (the dependent variable) **is a function of 9 to 14 variables** (depending on the area under consideration). The paper employs 1990 data, and for the Quebec City region, there are 13,378 observations, and the 12 explanatory variables are:

LnAGE - log of the apparent age of the property

NBROOMS - number of bedrooms

AREABYRM - area per room (in square metres)

ELEVATOR - a dummy variable = 1 if the building has an elevator; 0 otherwise

BASEMENT - a dummy variable = 1 if the unit is located in a basement; 0 otherwise

Hedonic House Pricing Models: Variable Definitions

OUTPARK - number of outdoor parking spaces

INDPARK - number of indoor parking spaces

NOLEASE - a dummy variable = 1 if the unit has no lease attached to it; 0 otherwise

LnDISTCBD - log of the distance in kilometres to the central business district

SINGLPAR - percentage of single parent families in the area where the building stands

DSHOPCNTR- distance in kilometres to the nearest shopping centre

VACDIFF1 - vacancy difference between the building and the census figure

- Examine the signs and sizes of the coefficients.
  - The coefficient estimates themselves show the Canadian dollar rental price per month of each feature of the dwelling.

Hedonic House Price Results

Dependent Variable: Canadian Dollars per Month

Variable	Coefficient	*t*-ratio	*A priori* sign expected
Intercept	282.21	56.09	+
LnAGE	-53.10	-59.71	-
NBROOMS	48.47	104.81	+
AREABYRM	3.97	29.99	+
ELEVATOR	88.51	45.04	+
BASEMENT	-15.90	-11.32	-
OUTPARK	7.17	7.07	+
INDPARK	73.76	31.25	+
NOLEASE	-16.99	-7.62	-
LnDISTCBD	5.84	4.60	-
SINGLPAR	-4.27	-38.88	-
DSHOPCNTR	-10.04	-5.97	-
VACDIFF1	0.29	5.98	-

Notes: Adjusted R^2 = 0.65l; regression *F*-statistic = 2082.27. Source: Des Rosiers and Thérialt
(1996). Reprinted with permission of the American Real Estate Society.

Tests of Non-nested Hypotheses

- All of the hypothesis tests concluded thus far have been in the context of "**nested**" models.

- But what if we wanted to compare between the following models?

$$\text{Model 1}: y_t = \alpha_1 + \alpha_2 x_{2t} + u_t$$

$$\text{Model 2}: y_t = \beta_1 + \beta_2 x_{3t} + v_t$$

- We could use R^2 or adjusted R^2, but what if the number of explanatory variables were different across the 2 models?

- An alternative approach is an encompassing test, based on examination of the **hybrid model**: $\text{Model 3}: y_t = \gamma_1 + \gamma_2 x_{2t} + \gamma_3 x_{3t} + w_t$

Tests of Non-nested Hypotheses (cont'd)

- There are 4 possible outcomes when Model 3 is estimated:
  - γ_2 is significant but γ_3 is not
  - γ_3 is significant but γ_2 is not
  - γ_2 and γ_3 are both statistically significant
  - Neither γ_2 nor γ_3 are significant

- Problems with encompassing approach
  - Hybrid model may be meaningless
  - Possible high correlation between x_2 and x_3.