1 A Non-technical Introduction to Regression

- Chapters 1 and Chapter 2 of the textbook are reviews of material you should know from your previous study (e.g. in your second year course). They cover, in a non-technical fashion some basic concepts (e.g. data types, graphs, descriptive statistics, correlation and regression).

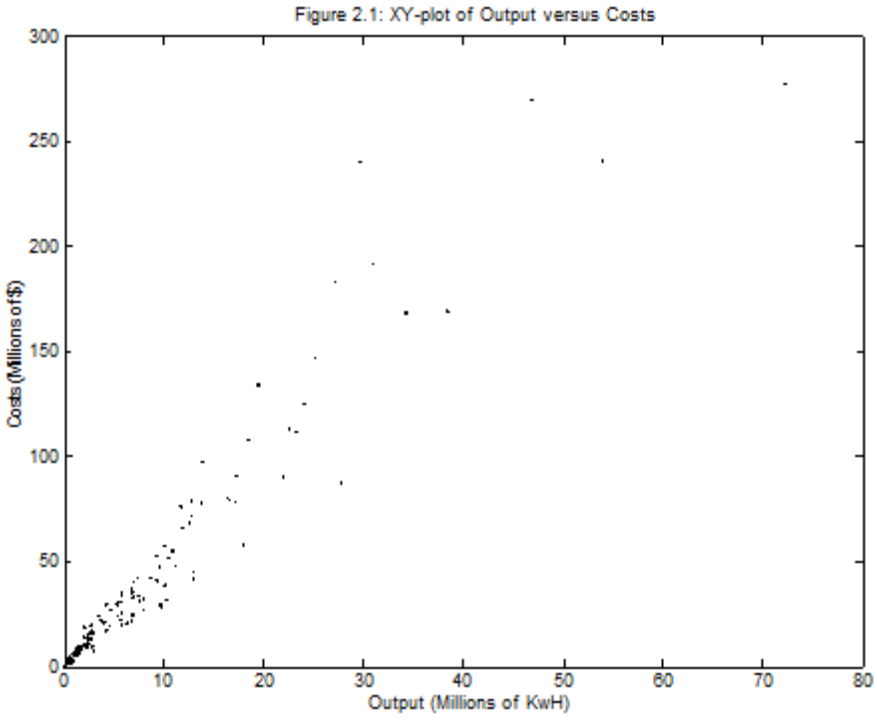
- Since you have covered this material before, I will go through this material quickly, with a focus on the most important tool of the applied economist: regression. But please read through chapters 1 and 2, particularly if you need some review of this material.

- Regression is used to help understand the relationships between many variables.

Regression as a Best Fitting Line

- We begin with simple regression to understand the relationship between two variables, X and Y.

- Example: see Figure 2.1 which is XY-plot of X = output versus Y = costs of production for 123 electric utility companies in the U.S. in 1970.



- The microeconomist will want to understand the relationship between output and costs.

- Regression fits a line through the points in the XY-plot that best captures the relationship between output and costs.

Simple Regression: Some Theory

- Question: What do we mean by "best fitting" line?

- Assume a linear relationship between X = output and Y = costs

$$Y = \alpha + \beta X,$$

where α is the intercept of the line and β its slope.

- Even if straight line relationship were true, we would never get all points on an XY-plot lying precisely on it due to measurement error.

- True relationship probably more complicated, straight line may just be an approximation.

- Important variables which affect Y may be omitted.

- Due to these factors we add an error, ε, which yields the *regression model*:

$$Y = \alpha + \beta X + \varepsilon.$$

- What we know: X and Y.

- What we do not know: α, β and ε.

- Regression analysis uses data (X and Y) to make a guess or **estimate** of what α and β are.

- Notation: $\widehat{\alpha}$ and $\widehat{\beta}$ are the estimates of α and β.

Distinction Between Errors and Residuals

- We have data for $i = 1, ., , N$ individuals (or countries, or companies, etc.).

- Individual observations are denoted using subscripts: Y_i for $i = 1, .., N$ and X_i for $i = 1, .., N$

- **True Regression Line** hold for every observation:

$$Y_i = \alpha + \beta X_i + \varepsilon_i.$$

- Error for i^{th} individual can be written as:

$$\varepsilon_i = Y_i - \alpha - \beta X_i.$$

- If we replace α and β by estimates, we get the **fitted (or estimated) regression line**:

$$\widehat{Y_i} = \widehat{\alpha} + \widehat{\beta} X_i.$$

and **residuals** are given by

$$\widehat{\varepsilon}_i = Y_i - \widehat{\alpha} - \widehat{\beta} X_i.$$

- Residuals measure distance that each observation is from the fitted regression line.

- A "good fitting" regression line will have observations lying near the regression line and, thus, residuals will be small.

Derivation of OLS Estimator

- How do we choose and $\widehat{\alpha}$ and $\widehat{\beta}$?

- A regression line which fits well will make residuals as small as possible.

- Usual way of measuring size of the residuals is the *sum of squared residuals* (SSR), which can be written in the following (equivalent) ways:

$$\begin{aligned} SSR &= \sum_{i=1}^{N} \widehat{\varepsilon}_i^2 \\ &= \sum_{i=1}^{N} \left(Y_i - \widehat{\alpha} - \widehat{\beta} X_i\right)^2 \\ &= \sum_{i=1}^{N} \left(Y_i - \widehat{Y_i}\right)^2. \end{aligned}$$

- The ordinary least squares (OLS) estimator finds values of $\widehat{\alpha}$ and $\widehat{\beta}$ which minimize SSR

- The formula for the OLS estimator will be discussed later. For now, note that standard econometrics software packages (e.g. PC-Give, E-views, Stata or Microfit) will calculate $\widehat{\alpha}$ and $\widehat{\beta}$.

Jargon of Regression

- $Y =$ dependent variable.
- $X =$ explanatory (or independent) variable.
- α and β are coefficients.
- $\widehat{\alpha}$ and $\widehat{\beta}$ and are OLS estimates of coefficients
- "Run a regression of Y on X"

Interpreting OLS Estimates

- Remember fitted regression line is

$$\widehat{Y_i} = \widehat{\alpha} + \widehat{\beta} X_i.$$

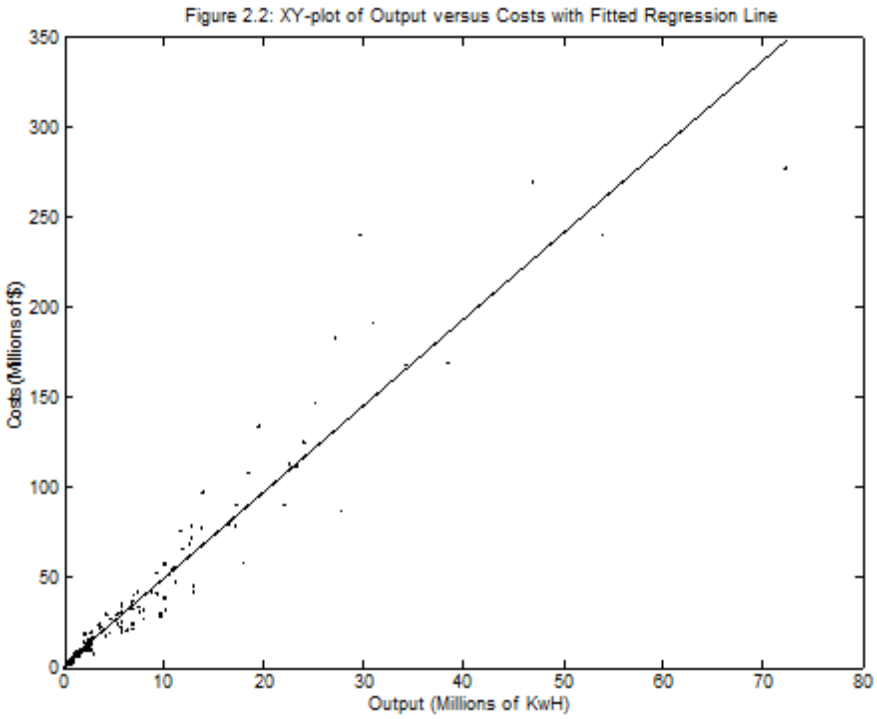
- Interpretation of $\widehat{\alpha}$ is estimated value of Y if $X = 0$. This is often not of interest.

- Example: $X =$ lot size, $Y =$ house price. $\widehat{\alpha} =$ estimated value of a house with lot size $= 0$ (not of interest since houses with lot size equal zero do not exist).

- $\widehat{\beta}$ is usually (but not always) the coefficient of most interest.

- The following are a few different ways of interpreting $\widehat{\beta}$.

- $\widehat{\beta}$ is slope of the best fitting straight line through an XY-plot such as Figure 2.1:



- $$\widehat{\beta} = \frac{d\widehat{Y}_i}{dX_i}.$$

- $\widehat{\beta}$ is the marginal effect of X on Y. It is a measure of how much the explanatory variable influences the dependent variable.

- $\widehat{\beta}$ is measure of how much Y tends to change when X is changed by one unit.

- The definition of "unit" depends on the particular data set being studied.

Example: Costs of production in the electric utility industry data set

- Using data set in Figures 2.1 and 2.2 we find $\widehat{\beta} = 4.79$.

- This is a measure of how much costs tend to change when output changes by a small amount.

- Costs are measured in terms of millions of dollars and output is measured as millions of kilowatt hours of electricity produced.

- Thus: if output is increased by one million kilowatt hours (i.e. a change of one unit in the explanatory variable), costs will tend to increase by $\$4,790,000$.

Measuring the Fit of a Regression Model

- The most common measure of fit is referred to as the R^2.

- Intuition: "Variability" $=$ (e.g.) how costs vary across companies

- Total variability in dependent variable $Y =$

Variability explained by the explanatory variable (X) in the regression

$+$

Variability that cannot be explained and is left as an error.

- R^2 measures the proportion of the variability in Y that can be explained by X.

Formalizing the Definition of R^2

- Remember (or see Chapter 1 or Appendix B) that variance is a measure of dispersion or variability.

- Variance of any variable can be estimated by:

$$var\left(Y\right) = \frac{\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2}{N-1},$$

where $\overline{Y} = \frac{\sum_{i=1}^{N} Y_i}{N}$ is the mean, or average value, of the variable.

- Total sum of squares (TSS) is proportional to variance of dependent variable:

$$TSS = \sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2.$$

- The following is not hard to prove:

$$TSS = RSS + SSR$$

- RSS is regression sum of squares, a measure of the explanation provided by the regression model:

$$RSS = \sum_{i=1}^{N} \left(\widehat{Y}_i - \overline{Y}\right)^2.$$

- SSR is the sum of squared residuals.

- This formalizes the idea that "variability in Y can be broken into explained and unexplained parts"

- We can now define our measure of fit:

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

or, equivalently,

$$R^2 = 1 - \frac{SSR}{TSS}.$$

- Note that TSS, RSS and SSR are all sums of squared numbers and, hence, are all non-negative. This implies $TSS \geq RSS$ and $TSS \geq SSR$. Using these facts, it can be seen that $0 \leq R^2 \leq 1$.

- Intuition: small values of SSR indicate that the residuals are small and, hence, that the regression model is fitting well. Thus, values of R^2 near 1 imply a good fit and that $R^2 = 1$ implies a perfect fit.

- Intuition: RSS measures how much of the variation in Y the explanatory variables explain. If RSS is near zero, then we have little explanatory power (a bad fit) and R^2 near zero.

- Example: In the regression of $Y =$ cost of production on $X =$ output for the 123 electric utility companies, $R^2 = .92$. The fit of the regression line is quite good.

- 92% of the variation in costs across companies can be explained by the variation in output.

- In simple regression (but not multiple regression), R^2 is the correlation between Y and X squared.

Basic Statistical Concepts in the Regression Model

- $\widehat{\alpha}$ and $\widehat{\beta}$ or only estimates of α and β. How accurate are the estimates?

- This can be investigated through *confidence intervals*.

- Closely related to the confidence interval is the concept of a *hypothesis test*.

- Intuition relating to confidence intervals and hypothesis tests given here, formal derivation provided in next chapter.

Confidence Intervals

- Example: $\widehat{\beta} = 4.79$ is the point estimate of β in the regression of costs of production on output using our electric utility industry data set

- Point estimate is best guess of what β is.

- Confidence intervals provide interval estimates which give a range in which you are highly confident that β must lie.

- Example: If confidence interval is $[4.53, 5.05]$ "We are confident that β is greater than 4.53 and less than 5.05"

- We can obtain different confidence intervals corresponding to different levels of confidence.

- 95% confidence interval: "we are 95% confident that β lies in the interval"

- 90% confidence interval we can say that "we are 90% confident that β lies in the interval", etc..

- The degree of confidence (e.g. 95%) is referred to as the *confidence level*.

- Example: for the electric utility data set, the 95% confidence interval for β is $[4.53, 5.05]$.

- "We are 95% confident that the marginal effect of output on costs is at least 4.53 and at most 5.05".

Hypothesis Testing

- Hypothesis testing involves specifying a hypothesis to test. This is referred to as the *null hypothesis*, H_0.

- It is compared to an *alternative hypothesis*, H_1.

- E.g. $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$ is common (and software packages will print out results for this hypothesis test)

- Many economic questions of interest have form: "Does the explanatory variable have an effect on the dependent variable?" or, equivalently, "Does $\beta = 0$ in the regression of Y on X?"

Aside on Confidence Intervals and Hypothesis Testing

- Hypothesis testing and confidence intervals are closely related.

- Can test whether $\beta = 0$ by looking at the confidence interval for β and see whether it contains zero.

- If it does not then we can "reject the hypothesis that $\beta = 0$" or conclude "X has significant explanatory power for Y" or "β is significantly different from zero" or "β is statistically significant".

- If confidence interval does include zero then we change the word "reject" to "accept" and "has significant explanatory power" with "does not have significant explanatory power", and so on.

- Confidence interval approach to hypothesis testing is equivalent to approach to hypothesis testing discussed next

- Just as confidence intervals came with various levels of confidence (e.g. 95%), hypothesis tests come with various *levels of significance*.

- Level of significance is 100% minus the confidence level.

- E.g. if a 95% confidence interval does not include zero, then you may say "I reject the hypothesis that $\beta = 0$ at the 5% level of significance" (i.e. 100%-95%=5%).

Hypothesis Testing (continued)

- First step: specify a hypothesis to test and choosing a significance level.

- E.g. H_0: $\beta = 0$ and the **5%** level of significance.

- Second step: calculate a test statistic and compare it to a *critical value* (a concept we will define in Chapter 3).

- E.g. For H_0: $\beta = 0$, the test statistic is known as a *t-statistic* (or t-ratio or t-stat):

$$t = \frac{\widehat{\beta}}{s_b},$$

where we will explain s_b later.

- Idea underlying hypothesis testing is that we accept H_0 if the value of the test statistic is consistent with what could plausibly happen if H_0 is true.

- If H_0 is true, then we would expect $\widehat{\beta}$ to be small (i.e. if $\beta = 0$ then expect $\widehat{\beta}$ near zero).

- But if $\widehat{\beta}$ is large this is evidence against H_0.

- Formally test statistic is large or small relative to "critical value taken from statistical tables of the Student-t distribution" (define later).

- For empirical practice, do not need critical value since *P-value* for this and other tests produced by computer packages.

- P-value is level of significance at which you can reject H_0.

- E.g. with 5% level of significance and software package gives P-value of 0.05 then reject H_0.

- If the P-value is less than 0.05 then you can also reject H_0.

- Students often want to interpret the P-value as measuring the probability that $\beta = 0$.

- E.g. if P-value less than 0.05 one wants to say "There is less than a 5% probability that $\beta = 0$ and, since this is very small, I can reject the hypothesis that $\beta = 0$."

- This is not formally correct. But, it does provide some informal intuition to motivate why small P-values lead you to reject H_0.

Hypothesis Testing involving R^2: The F-statistic

- Another popular hypothesis to test is H_0: $R^2 = 0$.
- If $R^2 = 0$ then X does not have any explanatory power for Y.
- Note: for simple regression, this test is equivalent to a test of $\beta = 0$. However, for multiple regression (which we will discuss shortly), the test of $R^2 = 0$ will be different than tests of whether regression coefficients equal zero.
- Same strategy: calculate a test statistic and compare to a critical value.
- Or most software will also calculate a P-value which directly gives a measure of the plausibility of H_0 : $R^2 = 0$

- Test statistic is called the F-statistic:

$$F = \frac{(N-2)\,R^2}{\left(1-R^2\right)}.$$

- The appropriate statistical table for obtaining the critical value is F-distribution (to be explained later)

- Or if the P-value for the F-test is less than 5% (i.e. 0.05), we conclude $R^2 \neq 0$.

- If the P-value for the F-test is greater than or equal to 5% , we conclude $R^2 = 0$.

- Can use levels of significance other than 5%.

Computer packages typically provide the following:

- $\widehat{\beta}$, the OLS estimate of β.
- The 95% confidence interval, which gives an interval where we are 95% confident β will lie.
- Standard deviation (or standard error) of $\widehat{\beta}$, s_b, which is a measure of how accurate $\widehat{\beta}$.
- The test statistic, t, for testing H_0: $\beta = 0$.
- The P-value for testing H_0: $\beta = 0$.
- R^2 which measures the proportion of the variability in Y explained by X
- The F-statistic and P-value for testing $H_0 : R^2 = 0$.

Example: Cost of Production in the Electric Utility Industry

- Regression of $Y =$ the costs of production and $X =$ output of electricity by 123 electric utility companies.

- Table 2.1 presents regression results in the form they would be produced by most software packages.

Table 2.1: Regression Results Using Electric Utility Data Set

Variable	Coeff	Stand Error	t-stat	P-value	95% conf. interval
Intercept	2.19	1.88	1.16	0.25	$[-1.53, 5.91$
Output	4.79	0.13	36.36	5×10^{-67}	$[4.53, 5.05]$

$R^2 = 0.92$ and the P-value for testing $H_0 : R^2 = 0$ is 5.4×10^{-67}.

Multiple Regression

- Multiple regression same as simple regression except many explanatory variables.

- Intuition and derivation of multiple and simple regression very similar.

- We will emphasise only the few differences between simple and multiple regression.

Example: Explaining House Prices

- Data on $N = 546$ houses sold in Windsor, Canada.
- Dependent variable, Y, is the sales price of the house in Canadian dollars.
- Four explanatory variables:
- X_1= the lot size of the property (in square feet)
- X_2 = the number of bedrooms
- X_3 = the number of bathrooms
- X_4 = the number of storeys (excluding the basement).

OLS Estimation of the Multiple Regression Model

- With k explanatory variables model is:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + .. + \beta_k X_{ki} + \varepsilon_i,$$

- i subscripts to denote observations, $i = 1, .., N$.

- With multiple regression have to estimate α and $\beta_1, .., \beta_k$.

- OLS estimates are found by choosing the values of $\widehat{\alpha}$ and $\widehat{\beta}_1, \widehat{\beta}_2, .., \widehat{\beta}_k$ that minimize the SSR:

$$SSR = \sum_{i=1}^{N} \left(Y_i - \widehat{\alpha} - \widehat{\beta}_1 X_{1i} - \widehat{\beta}_2 X_{2i} - .. - \widehat{\beta}_k X_{ki} \right)^2.$$

- Computer packages will calculate OLS estimates.

Statistical Aspects of Multiple Regression

- Largely the same as for simple regression.

- Formulae for confidence intervals, test statistics, etc. have only minor modifications.

- R^2 is still a measure of fit.

- Can test $R^2 = 0$ in same manner as for simple regression.

- If you find $R^2 \neq 0$ then you conclude that the explanatory variables *together* provide significant explanatory power (Note: this does not necessarily mean each individual explanatory variable is significant).

- Confidence intervals can be calculated for each *individual* coefficient as before.

- Can test $\beta_j = 0$ for each individual coefficient ($j = 1, 2, .., k$) as before.

- Emphasize: now we have a confidence interval and a test statistic for each coefficient.

Interpreting OLS Estimates in the Multiple Regression Model

- *Mathematical Intuition:* Total vs. partial derivative

Simple regression:

$$\beta = \frac{dY}{dX}$$

Multiple Regression:

$$\beta_j = \frac{\partial Y}{\partial X_j}$$

for the j^{th} coefficient $j = 1, .., k$.

Interpreting OLS Estimates in the Multiple Regression Model

- *Verbal intuition:* with simple regression β is the marginal effect of X on Y

Multiple regression: β_j is the marginal effect of X_j on Y, *ceteris paribus*

β_j is the effect of a small change in the j^{th} explanatory variable on the dependent variable, *holding all the other explanatory variables constant.*

Example: Explaining House Prices (continued)

Multiple regression results using the house price data set:

Table 2.2: Multiple Regression Using House Price Data Set

Variable	Coefficient	t-stat	P-value	95% conf. interval
Intercept	-4009.55	-1.11	0.27	$[-11087, 3068]$
Lot Size	5.43	14.70	2×10^{-41}	$[4.70, 6.15]$
# bedrm	2824.61	2.33	0.02	$[439, 5211]$
# bathrm	17105.17	9.86	3×10^{-21}	$[13698, 20512]$
# storeys	7634.90	7.57	1×10^{-13}	$[5655, 9615]$

Furthermore, $R^2 = 0.54$ and the P-value for testing H_0 : $R^2 = 0$ is 1.2×10^{-88}.

Example: Explaining House Prices (continued)

- How can we interpret the fact that $\widehat{\beta}_1 = 5.43$?

- An extra square foot of lot size will tend to add \$5.43 onto the price of a house, ceteris paribus.

- For houses with the same number of bedrooms, bathrooms and storeys, an extra square foot of lots size will tend to add \$5.43 onto the price of a house.

- If we compare houses with the same number of bedrooms, bathrooms and storeys, those with larger lots tend to be worth more. In particular, an extra square foot in lot size is associated with an increased price of \$5.43.

- Confidence interval for β_1: "I am 95% confident that the marginal effect of lot size on house price (holding other explanatory variables constant) is at least \$4.70 and at most \$6.15"

- Hypothesis testing: "Since the P-value for testing $H_0 : \beta_1 = 0$ is less than 0.05, we can conclude that β_1 is significant at the 5% level of significance"

- Can make similar statements for the other coefficients.

- Since $R^2 = 0.54$ can say: "54% of the variability in house prices can be explained by the four explanatory variables"

- Since the P-value for testing $H_0 : R^2 = 0$ is less than 0.05, we can conclude that the explanatory variables (jointly) have significant explanatory power at the 5% level of significance

Which Explanatory Variables to Choose in a Multiple Regression Model?

- We will relate this question to topics of *omitted variables bias* and *multicollinearity*.

- First note that there are two important considerations which pull in opposite directions.

- It is good to include all variables which help explain the dependent variable (include as many explanatory variables as possible).

- Including irrelevant variables (i.e. ones with no explanatory power) will lead to less precise estimates (include as few explanatory variables as possible).

- Playing off these two competing considerations is an important aspect of any empirical exercise. Hypothesis testing procedures can help with this.

Omitted Variables Bias

- To illustrate this problem we use the house price data set.

- A simple regression of $Y =$ house price on $X =$ number of bedrooms yields a coefficient estimate of $13,269.98$.

- But in multiple regression (see Table 2.2), coefficient on number of bedrooms was $2,824.61$.

- Why are these two coefficients on the same explanatory variable so different? i.e. $13,269.98$ is much bigger than $2,824.61$.

Answer 1: They just come from two different regressions which control for different explanatory variables (different ceteris paribus conditions).

Answer 2:

- Imagine a friend asked: "I have 2 bedrooms and I am thinking of building a third, how much will it raise the price of my house?"

- Simple regression: "Houses with 3 bedrooms tend to cost $13,269.98 more than houses with 2 bedrooms"

- Does this mean adding a 3rd bedroom will tend to raise price of house by $13,269.98? Not necessarily, other factors influence house prices.

- Houses with three bedrooms also tend to be desirable in other ways (e.g. bigger, with larger lots, more bathrooms, more storeys, etc.). Call these "good houses".

- Simple regression notes "good houses" tend to be worth more than others.

- Number of bedrooms is acting as a proxy for all these "good house" characteristics and hence its coefficient becomes very big (13,269.98) in simple regression.

- Multiple regression can estimate separate effects due to lot size, number of bedroom, bathrooms and storeys.

- Tell your friend: "Adding a third bedroom will tend to raise your house price by $2,824.61".

- Multiple regressions which contains all (or most) of house characteristics will tend to be more reliable than simple regression which only uses one characteristic.

- Take a look at the correlation matrix for this data set:

Table 2.3: Correlations Matrix for House Price Data Set					
	Price	Lot Size	# bed	# bath	# storey
Price	1				
Lot Size	0.54	1			
# bed	0.37	0.15	1		
# bath	0.52	0.19	0.37	1	
# storey	0.42	0.08	0.41	0.32	1

- Positive correlations between explanatory variables indicate that houses with more bedrooms also tend to have larger lot size, more bathrooms and more storeys.

Omitted Variable Bias

"Omitted variable bias" is a statistical term for these issues.

IF

1. We exclude explanatory variables that should be present in the regression,

AND

2. these omitted variables are correlated with the included explanatory variables,

THEN

3. the OLS estimates of the coefficients on the included explanatory variables will be biased.

Example: Explaining House Prices (continued)

- Simple regression used Y = house prices and X = number of bedrooms.

- Many important determinants of house prices omitted.

- Omitted variables were correlated with number of bedrooms. Hence, the OLS estimate from the simple regression of $13,269.98$ was biased.

Practical Advice for Selecting Explanatory Variables

- Include (insofar as possible) all explanatory variables which you think might possibly explain your dependent variable. This will reduce the risk of omitted variable bias.

- However, including irrelevant explanatory variables reduces accuracy of estimation and increases confidence intervals.

- So do t-tests (or other hypothesis tests) to decide whether variables are significant. Run a new regression omitting the explanatory variables which are not significant.

Multicollinearity

- Intuition: If explanatory variables are highly correlated with one another then regression model has trouble telling which individual variable is explaining Y.

- Symptom: Individual coefficients may look insignificant, but regression as a whole may look significant (e.g. R^2 big, F-stat big, but t-stats on individual coefficients small).

- Looking at a correlation matrix for explanatory variables can often be helpful in revealing extent and source of multicollinearity problem.

Example of Multicollinearity

- $Y =$ exchange rate
- Explanatory variable(s) = interest rate
- $X_1 =$ bank prime rate
- $X_2 =$ Treasury bill rate
- Using both X_1 and X_2 will probably cause multicollinearity problem
- Solution: Include either X_1 or X_2 but not both.
- In some cases this "solution" will be unsatisfactory if it causes you to drop out explanatory variables which economic theory says should be there.

Multiple Regression with Dummy Variables

- Dummy variable is either 0 or 1.
- Use to turn qualitative (Yes/No) data into 1/0.
- Example: Explaining House Prices (continued)

- Data set has 5 potential dummy explanatory variables

- $D_1 = 1$ if the house has a driveway ($= 0$ if it does not)

- $D_2 = 1$ if the house has a recreation room ($= 0$ if not)

- $D_3 = 1$ if the house has a basement ($= 0$ if not)

- $D_4 = 1$ if the house has gas central heating ($= 0$ if not)

- $D_5 = 1$ if the house has air conditioning ($= 0$ if not)

Simple Regression with a Dummy Variable

- One dummy explanatory variable, D:

$$Y_i = \alpha + \beta D_i + \varepsilon_i$$

for $i = 1, .., N$ observations.

- OLS estimation produces $\widehat{\alpha}$ and $\widehat{\beta}$, and fitted regression line is:

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta} D_i.$$

- Since D_i is either 0 or 1, we either have $\widehat{Y}_i = \widehat{\alpha}$ or $\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}$.

Example: Explaining House Prices (continued)

- Regress Y = house price on D = dummy for air conditioning (=1 if house has air conditioning, = 0 otherwise).

- Fitted regression line is:

$$\widehat{Y}_i = 59884.85 + 25995.74 D_i.$$

- Average price of house with air conditioning is $\$85,881$

- Average price of house without air conditioning is $\$59,885$

- Remember, however, omitted variables bias (this simple regression no doubt suffers from it)

Multiple Regression with Dummy Variables

$$Y_i = \alpha + \beta_1 D_{1i} + .. + \beta_k D_{ki} + \varepsilon_i$$

- Example: Explaining House Prices (continued)

- Regress Y = house price onD_1 = driveway dummy and D_2 = rec room dummy.

- Fitted regression line:

$$\widehat{Y}_i = 47099.08 + 21159.91 D_{1i} + 16023.69 D_{2i}.$$

- Putting in either 0 or 1 values for the dummy variables, we obtain the fitted values for Y for the four categories of houses:

1. Houses with a driveway and recreation room ($D_1 = 1$ and $D_2 = 1$) have $\widehat{Y}_i = 47099+21160+16024 = \$84,283$.

2. Houses with a driveway but no recreation room ($D_1 = 1$ and $D_2 = 0$) have $\widehat{Y}_i = 47099 + 21160 = \$68,259$.

3. Houses with a recreation room but no driveway ($D_1 = 0$ and $D_2 = 1$) have $\widehat{Y}_i = 47099 + 16024 = \$63,123$.

4. Houses with no driveway and no recreation room ($D_1 = 0$ and $D_2 = 0$) have $\widehat{Y}_i = \$47,099$.

- Multiple regression with dummy variables may be used to classify the houses into different groups and to find average house prices for each group.

Multiple Regression with Dummy and non-Dummy Explanatory Variables

- E.g. one dummy variable (D) and one regular non-dummy explanatory variable (X):

$$Y_i = \alpha + \beta_1 D_i + \beta_2 X_i + \varepsilon_i.$$

- Example: Explaining House Prices (continued)

- Regress Y = house price on D = air conditioning dummy and X = lot size.

- Obtain $\widehat{\alpha} = 32,693$, $\widehat{\beta}_1 = 20175$ and $\widehat{\beta}_2 = 5.638$.

- Get two different fitted regression lines

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}_1 + \widehat{\beta}_2 X_i = 52868 + 5.638 X_i$$

if $D_i = 1$ (i.e. the ith house has an air conditioner) and

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}_2 X_i = 32693 + 5.638 X_i$$

if $D_i = 0$ (i.e. the house does not have an air conditioner).

- Note that the two regression lines have the same slope and only differ in their intercepts.

Interacting Dummy with non-Dummy Explanatory Variables

- Consider the following regression model:

$$Y_i = \alpha + \beta_1 D_i + \beta_2 X_i + \beta_3 Z_i + \varepsilon_i,$$

where D and X are dummy and non-dummy explanatory variables and $Z = DX$.

- How do we interpret results from a regression of Y on D, X and Z?

- Note that Z_i is either 0 (for observations with $D_i = 0$) or X_i (for observations with $D_i = 1$).

- Fitted regression lines for individuals with $D_i = 0$ and $D_i = 1$ are:

- If $D_i = 0$ then $\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}_2 X_i$

- If $D_i = 1$ then $\widehat{Y}_i = \left(\widehat{\alpha} + \widehat{\beta}_1\right) + \left(\widehat{\beta}_2 + \widehat{\beta}_3\right) X_i$

- Two different regression lines corresponding to $D = 0$ and $D = 1$ exist and have different intercepts and different slopes.

- Marginal effect of X on Y is different for observations with $D_i = 0$ than with $D_i = 1$.

Example: Explaining House Prices (continued)

- Regress Y = house price on D = air conditioner dummy, X = lot size and $Z = D \times X$

- $\widehat{\alpha} = 35684$, $\widehat{\beta}_1 = 7613$, $\widehat{\beta}_2 = 5.02$ and $\widehat{\beta}_3 = 2.25$.

- Marginal effect of lot size on housing is 7.27 (i.e. $\widehat{\beta}_2 + \widehat{\beta}_3$) for houses with air conditioners and only 5.02 for houses without.

Working with Dummy Dependent Variables

- Example: Dependent variable is a transport choice.
- 1 = "Yes I take my car to work"
- 0 = "No I do not take my car to work"
- We will not discuss this case in this course.
- Note only the following points:
- There are some problems with OLS estimation. But OLS estimation might be adequate in many cases.
- Better estimation methods are "Logit" and "Probit" available in many software packages.

Chapter Summary

This non-technical introduction to regression, you should be able to get started in actually doing some empirical work (at least with cross-sectional data).The major points covered in this chapter include:

1. Simple regression quantifies effect of an explanatory variable, X, on a dependent variable, Y, through a regression line $Y = \alpha + \beta X$.

2. Estimation of α and β involves choosing estimates which produces the "best fitting" line through an XY graph. These are called ordinary least squares (OLS) estimates, are labelled $\widehat{\alpha}$ and $\widehat{\beta}$ and are obtained by minimizing the sum of squared residuals (SSR).

3. Regression coefficients should be interpreted as marginal effects (i.e. as measures of the effect on Y of a small change in X).

4. R^2 is a measure of how well the regression line fits the data.

5. The confidence interval provides an interval estimate of any coefficient (e.g. an interval for β in which you can be confident β lies).

6. A hypothesis test of whether $\beta = 0$ can be used to find out whether the explanatory variable belongs in the regression. A hypothesis test can either be done by comparing a test statistic (i.e. the t-stat) to a critical value taken from statistical tables or by examining P-value. If the P-value for the hypothesis test of whether $\beta = 0$ is less than 0.05 then you can reject the hypothesis at the 5% level of significance.

7. The multiple regression model has more than one explanatory variable. The basic intuition (e.g. OLS estimates, confidence intervals, etc.) is the same as for the simple regression model. However, with multiple regression the interpretation of regression coefficients is subject to *ceteris paribus* conditions.

8. If important explanatory variables are omitted from the regression and are correlated with included explanatory variables, omitted variables bias occurs.

9. If explanatory variables are highly correlated with one another, coefficient estimates and statistical tests may be misleading. This is referred to as the multicollinearity problem.

10. The statistical techniques associated with the use of dummy explanatory variables are exactly the same as with non-dummy explanatory variables.

11. A regression involving only dummy explanatory variables classifies the observations into various groups (e.g. houses with air conditioners and houses without). Interpretation of results is aided by careful consideration of what the groups are.

12. A regression involving dummy and non-dummy explanatory variables classifies the observations into groups and says that each group will have a regression line with a different intercept. All these regression lines have the same slope.

13. Regression involving dummy, non-dummy and interaction (i.e. dummy times non-dummy variables) explanatory variables classifies the observations into groups and says that each group will have a different regression line with different intercept and slope.