Introduction to Econometrics

Arthur Campbell

MIT

16th February 2007
What is a Regression?
Regression Equation
Regression Coefficients, Standard Errors, T-statistics, Level of Significance, $R^2$ values
Interaction terms
What is a Regression?

It is a statistical tool for understanding the relationship between different variables

- Usually we want to know the causal effect of one variable on another
- For instance we might ask the question how much extra income do people receive if they have had one more year of education all other things equal?
- When $I$ represents income and $E$ education this is equivalent to asking what is $\frac{\partial I}{\partial E}$?
- To answer this question the econometrician collects data on income and education, and uses it to run a regression equation
What is a Regression?

The most simple regression is a regression with a single explanatory variable. In the case of income and education this could be

\[ I = \beta_0 + \beta_1 E + \varepsilon \]

\( I \) is called the dependent (endogenous) variable and \( E \) is known as the explanatory (exogenous)

\( \beta_0 \) and \( \beta_1 \) are the regression co-efficients

\( \varepsilon \) is the noise term

This regression equation will put a straight line through the data
Fitting the regression equation

Consider the following set of data on income and education

Figure by MIT OCW and adapted from:

Fitting the regression equation

The regression will typically fit the line which minimizes the sum of the squared distances of the data points to the line.

Figure by MIT OCW and adapted from:

Fitting the regression equation

The criteria we have used here is

\[ \min_{\beta_0, \beta_1} \sum (y_i - \beta_0 - \beta_1 X_i)^2 \]

This determines the values of \( \beta_0 \) and \( \beta_1 \) and hence the position of the line.

There are many potential criteria we could use such as

\[ \min_{\beta_0, \beta_1} \sum |y_i - \beta_0 - \beta_1 X_i| \]

However provided the noise term from earlier \( \epsilon \) satisfies certain assumptions the sum of squared distances is optimal.
\( \beta_0 \) is the intercept of the line
\( \beta_1 \) is the slope of the line or in other words is \( \frac{\partial I}{\partial E} \)
If for instance \( \beta_1 = \frac{\partial I}{\partial E} = 15,000 \) this would imply that for every additional year of schooling an individual would on average earn $15,000 more
For a given level of income and education we could now work out the elasticity of income wrt education
Interpreting the coefficients in the log-log regression model

Consider now an isoelastic demand curve

\[ Q_D = \beta_0 P^{\beta_1} \]

Now take the logarithm of both sides

\[ \ln Q_D = \ln \beta_0 + \beta_1 \ln P \]

We can estimate the following regression relationship

\[ \ln Q_D = \ln \beta_0 + \beta_1 \ln P + \epsilon \]

to determine \( \beta_0 \) and \( \beta_1 \)

Here each data point would be \((\ln Q_D, \ln P)\) and the value of the intercept is \( \ln \beta_0 \) and the slope is \( \beta_1 \)
Interpreting the coefficients in the log-log regression model

In this log-log specification $\beta_1$ is again the derivative of the dependent variable wrt the explanatory variable $\frac{\partial \ln Q_D}{\partial \ln P} = \frac{\partial Q_D}{\partial P} \frac{P}{Q}$ and has the natural interpretation of the elasticity of demand with respect to price. In Problem Set 2 you will be asked to calculate elasticities from the regression results.
The regression may in fact contain more than one explanatory variable. For instance we might think that a person’s income is influenced by both the number of years of education and the number of years experience in the labour force.

In this case we might run the following multi-variable regression

\[ I = \beta_0 + \beta_1 E + \beta_2 L \]

Here we can find the effect education and labour force experience on income separately.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\beta}_0 )</td>
<td>-0.615 (0.929)</td>
<td>-1.697*** (0.587)</td>
</tr>
<tr>
<td>In(( P ))</td>
<td>-0.335*** (0.024)</td>
<td>-0.042*** (0.009)</td>
</tr>
<tr>
<td>In(( Y ))</td>
<td>0.467*** (0.096)</td>
<td>0.530*** (0.058)</td>
</tr>
<tr>
<td>Jan</td>
<td>-0.079*** (0.010)</td>
<td>-0.044*** (0.006)</td>
</tr>
<tr>
<td>Feb</td>
<td>-0.129*** (0.019)</td>
<td>-0.122*** (0.010)</td>
</tr>
<tr>
<td>Mar</td>
<td>-0.019*** (0.006)</td>
<td>-0.008 (0.005)</td>
</tr>
<tr>
<td>Apr</td>
<td>-0.021 (0.016)</td>
<td>-0.024*** (-0.005)</td>
</tr>
<tr>
<td>May</td>
<td>0.013 (0.011)</td>
<td>0.026*** (0.004)</td>
</tr>
<tr>
<td>Jun</td>
<td>0.020 (0.010)</td>
<td>0.000 (0.004)</td>
</tr>
<tr>
<td>Jul</td>
<td>0.031*** (0.010)</td>
<td>0.040*** (0.005)</td>
</tr>
<tr>
<td>Aug</td>
<td>0.042*** (0.010)</td>
<td>0.046*** (0.004)</td>
</tr>
<tr>
<td>Sep</td>
<td>-0.028*** (0.006)</td>
<td>-0.039*** (0.005)</td>
</tr>
<tr>
<td>Oct</td>
<td>0.002 (0.010)</td>
<td>0.008 (0.005)</td>
</tr>
<tr>
<td>Nov</td>
<td>-0.058*** (0.012)</td>
<td>-0.032*** (0.004)</td>
</tr>
</tbody>
</table>

\[ \varepsilon_j^t s \]
\[ \overline{R}^2 \]
\[ \hat{\sigma} \]
0.85  0.94
0.027  0.011

***(p < 0.01)

In the previous slide the regression included 11 dummy variables for the months Jan-Nov.

These variables take a value of 1 if the data point was observed during that month and 0 otherwise.

They are included to remove any seasonality in the data, a positive value means that there was more (gasoline) consumed during that month compared to the month without a dummy variable (December).
When the error terms $\varepsilon$ are normally distributed it is possible to show that our estimates from the regression of the $\beta'$s are also normally distributed.

Standard errors represent how accurately we have estimated a coefficient.

A very small standard error means it is a very accurate estimate.

In the regression results from earlier these standard errors are typically reported in parantheses beneath the coefficient’s value.
A t-statistic is used to measure how confident we are given the results of the regression that the true $\beta$ is different from 0. For instance if we measured a very high value for $\beta$ with a very small standard error we would be very confident. On the other hand if we found a small value of $\beta$ with a high standard error we would be far less confident. The t-statistic is calculated as:

$$\frac{\beta}{s}$$

The magnitude of this term not the sign is what is important since $\beta$ can be positive or negative.
Associated with a t-statistic is a level of significance
The level of significance is the probability we attach to the real value of $\beta$ being 0 given the evidence we have found through our regression
As the magnitude of $\frac{\beta}{\sigma}$ increases the level of significance decreases
The significance of an estimate is often indicated with a *, **, or ***
the meaning of these is usually indicated below the regression results
Goodness of fit ($R^2$)

The goodness of fit measure $R^2$ is a measure of the extent to which the variation of the dependent variable is explained by the explanatory variable(s).

The formula for it is

$$R^2 = 1 - \frac{\text{sum of squared errors}}{\text{sum of deviations from mean}}$$

$$R^2 = 1 - \frac{\sum_i (y_i - \beta_0 - \beta_1 x_i)^2}{\sum_i (y_i - \bar{y})^2}$$

where $\bar{y}$ is the average value of $y$

$$\frac{\text{sum of squared errors}}{\text{sum of deviations from mean}}$$

is the amount of the total variation of $y$ that is unexplained by the regression, so $1-\frac{\text{sum of squared errors}}{\text{sum of deviations from mean}}$ is the amount which is explained by the regression.

Clearly $R^2$ will be between 0 and 1, values close to 1 indicate good explanatory power.
An obvious way to increase the $R^2$ of a regression is to simply increase the number of explanatory variables since including additional variables cannot decrease its explanatory power.

The adjusted $R^2$ is a measure of explanatory power which is adjusted for the number of explanatory variables included in the regression. The formula for the adjusted $R^2$ is

$$R_{Adjusted}^2 = 1 - (1 - R^2) \frac{n - 1}{n - m - 1}$$

where $n$ is the number of data points and $m$ is the number of explanatory variables.

The adjusted $R^2$ increases when a new variable is added if the new term improves the model more than would be expected by chance. It is always less than the actual $R^2$. 

---

Arthur Campbell  
Introduction to Econometrics  
02/16/07  18 / 19
An interaction term is where we construct a new explanatory variable from 2 or more underlying variables.

For instance, we could multiply two variables together, say Price and Income.

The regression equation we would estimate would then be:

$$Q_D = \beta_0 + \beta_1 P + \beta_2 Y + \beta_3 PY$$

We do this if we think that the effect of $P$ on $Q_D$ is different when $Y$ is high or low, and similarly, the effect of $Y$ on $Q_D$ is different when $P$ is high or low.

Consider the demand elasticity wrt price:

$$E_D = \frac{\partial Q_D}{\partial P} \frac{P}{Q} = (\beta_1 + \beta_3 Y) \frac{P}{Q_D}$$

We see here that holding everything else constant increasing $Y$ by 1 unit will increase $E_D$ by $\beta_3 \frac{P}{Q_D}$. 