

The Basics of Single Variable Linear Regression in Gretl

Sometimes we are interested in predicting the value of a variable of interest (the “dependent” variable) given the value of some other variable (the “independent” variable). Our object is to determine how these variables are quantitatively related. Regression analysis allows us to estimate this relationship from data on the two variables, and tells us how confident we can be in our estimate.

The flow of causation in regression analysis is assumed to run from the independent variable to the dependent variable. In addition *linear* regression constrains the relationship between the two variables to be linear (a straight line). If this assumption seems to be inappropriate for the relationship in question (as determined by prior graphical data exploration – such as a scatter plot), then we should seek a more suitable method (perhaps some non-linear transformation of the data).

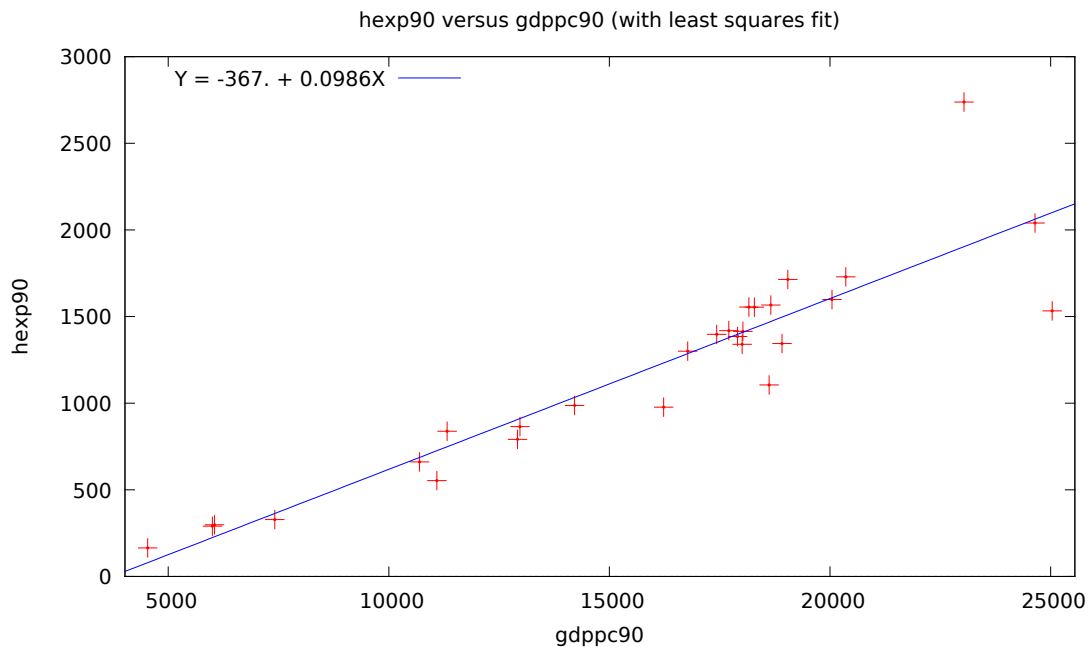
An Example

We have seen from our OECD database that expenditures on health increase with the size of a country’s GDP. The table from the OECD 1990 data set is:

Country	hexp90	gdppc90
Turkey	165	4532
Mexico	290	6001
Poland	298	6048
Korea	328	7416
Portugal	661	10695
Czech Republic	553	11087
Greece	838	11321
Ireland	791	12917
Spain	865	12971
New Zealand	987	14209
UK	977	16228
Australia	1300	16774
Italy	1397	17430
Netherlands	1419	17707
Norway	1385	17905
Belgium	1340	18008
Finland	1414	18025
France	1555	18162
Denmark	1554	18285
Japan	1105	18622
Sweden	1566	18660
Austria	1344	18914
Canada	1714	19044
Iceland	1598	20046

Germany	1729	20359
USA	2738	23038
Switzerland	2040	24648
Luxembourg	1533	25038
Hungary	NA	NA
Slovak Republic	NA	NA

And the scatter plot is:



If all the points fell exactly in a straight line, we could find the exact relationship using the formula for a straight line:

$$y = a + b \cdot x$$

To find the slope, calculate the change in y per unit change in x :

$$b = \Delta y / \Delta x$$

For example, consider the UK and Finland from the table above, with $x = \text{gdppc90}$ and $y = \text{hexp90}$. We'll take Δy to be the difference in hexp90, Finland minus UK, and similarly for Δx :

$$b = \Delta y / \Delta x = (1414 - 977) / (18025 - 16228) = 0.243$$

This would indicate that when GDP per capita increases by \$1, health spending per capita increases by 24.3 cents.

However, this calculation does not agree with the “best fitting” slope shown on the graph, namely 0.0986. Clearly, all of the data do not lie in a straight line. No estimated

relationship is measured without error, and health spending levels are “caused” by other variables besides GDP.¹

The random error inherent in sampling is called “sampling error.” The further error of omitting relevant variables is called “misspecification”. Since we expect at least the former to occur in any random sampling procedure, the best we can do in estimating relationships is to minimize this error. This is what linear regression as a method, and in the form of a computer routine, does. It finds the straight line that minimizes the “lack of fit” between the actual data points and the predictions given by the line, as explained more fully below. Note, however, that the best fitting straight line may actually be a very poor fit to the data. In that case some other estimation technique may be appropriate. (In gretl you can explore this possibility starting from a scatter plot: click on the plot for a popup menu, select Edit, and in the dialog box that appears you’ll see some more “fitted line” alternatives).

To estimate a linear model in gretl, go to the “Model” menu and choose Ordinary Least Squares; select a dependent variable and an independent variable; and click OK. If we do this for the two variables above, gretl will produce the output shown below. (In the regression output window, select the menu item “Edit/copy/RTF (MS Word)” to paste it into a Word document.)

Model 1: OLS estimates using 28 observations from 1-30
 Missing or incomplete observations dropped: 2
 Dependent variable: hexp90

<i>Variable</i>	<i>Coefficient</i>	<i>Std. Error</i>	<i>t-statistic</i>	<i>p-value</i>	
Const	-367.243	137.904	-2.6630	0.01311	**
gdppc90	0.0985539	0.00823951	11.9611	<0.00001	***

This “basic” output is followed by several lines of additional information; for our purposes we’ve pruned all but the first 5 lines (the deleted portion would be relevant to more knowledgeable analysts. Students who wish to know more, see ECN 215):

Mean of dependent variable = 1195.86
 Standard deviation of dep. var. = 583.204
 Sum of squared residuals = 1.41226e+06
 Standard error of residuals = 233.062
 Unadjusted R² = 0.846217

From this information we can glean the following:

- From the available 28 complete pairs of data, the regression equation, the estimate

¹For a thorough investigation of the many forces affecting health spending across nations, see G. J. Schieber and A. Maeda, “A Curmudgeon’s Guide to Financing Health Care in Developing Countries,” World Bank Discussion Paper no. 365, in *Innovation in Health Care Financing*, ed. G. J. Schieber (Washington: World Bank, 1997).

of the line of best fit is:

$$\text{hexp90} = -367 + .0986 \times \text{gdppc90}$$

This tells us that in the OECD countries in 1990, as GDP/per capita rose by \$1.00 in purchasing power parity units, health expenditures in these countries generally rose by 9.8 cents.

- The mean value of health spending in the OECD in 1990 was \$1196.

As stated earlier, the least squares method finds the “best fitting” straight line in a certain sense. To be more precise, this line minimizes the squares of the differences between observed values of the dependent variable, y_i , and the predicted or fitted values given by the regression equation (\hat{y}_i or “y-hat”). These gaps between actual and fitted y are called *residuals*. So we may also say that Ordinary Least Squares minimizes the sum of squared residuals.

At any given data point, fitted y is found by substituting the corresponding x value into the regression equation. For example, the UK had an hexp90 value of 997 (y) but using the UK’s GDP per capita ($x = 16228$) we get a fitted value of

$$-367 + .0986 \times 16228 = 1233$$

The UK’s health spending fell short of what the regression predicts: the residual for the UK is negative ($997 - 1233 = -236$). If we were to repeat this exercise for each country, square the residuals in each case, and add them up, we’d get the number reported by `gretl` for the sum of squared residuals, namely $1.41226\text{e}+06$ (in scientific notation), or 1412260. This may seem a large number, but we can be confident any other straight line will produce a larger sum.

The sum of squared residuals can be processed to give a measure of how closely the data cluster around the line. The Standard Error of Estimate (Se) is one such measure. It is calculated according to the following formula, where Σ denotes summation:

$$Se = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}}$$

For example, the Standard Error of Estimate for the preceding model is:

$$Se = \sqrt{\frac{1412260}{26}} = 233$$

This number is reported under the name “Standard error of residuals” in the `gretl` output shown above.

Another measure of the goodness of fit, which is easier to interpret, is the coefficient of determination, R^2 . The calculation for R^2 is based on a comparison of two sums of

squares: the sum of squared residuals and the “total sum of squares” for y , that is, the sum of squared deviations of y from \bar{y} (mean y). The formula is:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

R^2 equals 1 if the data all lie exactly on a straight line (the sum of squared residuals equals zero), and it equals 0 if the data are unrelated. In the latter case the sum of squared residuals equals the total sum of squares for y , which is to say that the best linear predictor for y is just mean y : x is no help. Values of R^2 between 1 and 0 indicate that the data have some linear relationship, but also some scatter. What’s a “good” R^2 ? It depends on the nature of the data. As a very rough rule of thumb, a value of 0.15 or greater might be considered strong evidence of a real relationship for cross sectional data (as in the example above), while for time series data with a trend the bar is higher, maybe 0.8 or better.

All data measurement is subject to random sampling error. This means the estimated slope coefficient for the relationship between two variables might turn out different given a different data sample. Our estimates of the slope and intercept are just what we say they are: *estimates*. They are themselves random variables that are distributed around the true population slope and intercept (which we call parameters).

Thus our data, and our estimates drawn from that data, are random variables. How close are our estimates to the true population values of the slope and intercept? We can never know for sure. But we can construct a range of values such that we are confident, at some definite level of probability, that the true value will fall within this range. This is called a “confidence interval”.

A “confidence interval” always has a certain probability attached to it. A 90 percent interval is one for which we can be 90 percent confident that the interval brackets the true value. A 99 percent interval gives us 99 percent confidence that we’ve included the true value in the range. The higher the confidence level, the greater the chance that the true value of the parameter falls within the interval constructed. This higher level of confidence will be associated with a wider range of values within which the parameter may fall.

The most common choice is the 95% level of confidence. This choice means that there is a 95% chance that the true parameter falls within the interval constructed. Correspondingly, there is a 5% chance that the true parameter lies outside of this range.

The *gretl* output shown above also gives a standard error, a t -statistic and a p -value for each of the estimated coefficients. The standard error is a measure of our uncertainty regarding the true value. It is used in constructing confidence intervals: the rule of thumb is that you get a 95 percent interval by taking the estimated value plus or minus two standard errors. The t -statistic is just the ratio of the estimate to its own standard error. The rule of thumb here is that if the t -statistic is greater than 2 in absolute value, the estimate is “statistically significant at the 5 percent level”, meaning that the true value is unlikely to be zero. This corresponds to a 95 percent confidence interval that does not include zero.

The p-value is tricky to interpret at first but quite useful: it answers the following question: Suppose the true parameter value were zero, what then would be the probability of getting, in a random sample, an estimate as far away from zero as the one we actually got, or further? In the example above, the p-value for the slope coefficient is shown as “less than 0.00001”. In other words, if there really were no relationship between GDP and health spending, the chances of coming up with a slope estimate of 0.0986 or greater would be minuscule. Since we did come up with such a value, we can be pretty confident that a relationship really exists.