Q-STEP STATA 'HOW TO' GUIDES:

MULTIVARIATE LINEAR REGRESSION IN STATA

Creator: Joshua Townsley

One of the most widely-used methods in the quantitative analysis toolbox is regression. There are many types of regression, but this guide will focus on Ordinary Least Squares (or 'Linear Regression').

Linear regression would be used when our dependent variable is an interval level variable. This guide will start with a simple bivariate linear regression.

Remember to clean your variables first – i.e. check that missing values are indeed set as missing values. This can be done by examining the coding of the variable first (e.g. **codebook W3WES43a, tab(100)**) then setting desired values as missing (e.g. **mvdecode W3WES43a, mv(12)**).

Q-Step
A step-change in quantitative social science skills
Funded by the Nuffield Foundation, ESRC and HEFCE
Nuffield Foundation
Sheffield Methods Institute.
The University Of Sheffield.

I. Linear Regression – Housing Tenure and Liking the Conservative Party

We have a variable – likeCon – that measures how much a respondent likes or dislikes the Conservative Party (on a scale of 0-10). This is our dependent variable – we want to know what other variables help explain it (i.e. our independent variables).

Let's run our regression model with our housing tenure variable – tenurecat. To do this, we use the regress (or "reg") function:

reg likeCon i.tenurecat

This produces the following output:

```
. reg W3WES43a i.tenurecat
```

Source	SS	df	MS	Number of obs	=	2,988
				F(2, 2985)	=	34.28
Model	673.033711	2	336.516855	Prob > F	=	0.0000
Residual	29305.9569	2,985	9.81774101	R-squared	=	0.0225
				Adj R-squared	=	0.0218
Total	29978.9906	2,987	10.0364883	Root MSE	=	3.1333

W3WES43a	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tenurecat						
Rent	-1.138544	.1403144	-8.11	0.000	-1.413667	-.8634217
Live with Family/Friends	-.6390067	.2286177	-2.80	0.005	-1.087271	-.1907425
_cons	4.614735	.0678757	67.99	0.000	4.481647	4.747823

There are several important pieces of information we want to pick out from the output:

- **Number of obs** = the number of observations (in our case, respondents) that are included in the analysis
- **R-squared** = the model 'fit'
- **Coef.** = the coefficient associated with each variable/category
- **_cons** = the constant (what the DV would be if the IV was 0)
- **P>|t|** = the p value associated with each coefficient
- **[95% Conf. Interval]** = the 95% confident interval around the coefficient

The model provides a coefficient for two of the three responses in our tenurecat variable. This is because we specified to STATA that the variable is categorical (using the "i." term). Each coefficient tells us the difference between that value and the value that is left out (i.e. "Own").

So, the average response from those who "Rent" is 1.1 lower on the dependent variable – W3WES43a – than those who "Own". In other words, renters like the Conservatives less than those who own their homes.

How confident are we in this result of -1.1? The associated p-value is "0.000" (or, <.001), meaning we can 99.9% confident that this coefficient exists in the population.

II. Multivariate Linear Regression

But we know that other factors influence how much one likes the Conservative Party besides their housing situation. For instance, we suspect age might also be an important factor. So let's run another regression model – this time adding age to our model:

reg likeCon age i.tenurecat

```
. reg likeCon age i.tenurecat
```

Source	SS	df	MS	Number of obs	=	2,988
				F(3, 2984)	=	40.33
Model	1168.04663	3	389.348875	Prob > F	=	0.0000
Residual	28810.944	2,984	9.65514209	R-squared	=	0.0390
				Adj R-squared	=	0.0380
Total	29978.9906	2,987	10.0364883	Root MSE	=	3.1073

likeCon	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0293473	.0040986	7.16	0.000	.0213108	.0373837
tenurecat						
Rent	-.7260503	.1506015	-4.82	0.000	-1.021344	-.4307569
Live with Family/Friends	.1283385	.2507693	0.51	0.609	-.3633597	.6200367
_cons	2.922096	.2457898	11.89	0.000	2.440161	3.404031

How do we interpret the coefficients in a multivariate model? Simple: each coefficient tells us the expected change to the likeCon score when that variable increases by 1 unit controlling for the other variables in the model (i.e. holding them constant).

The age coefficient in this model is 0.03 – meaning that a one unit increase in age is associated with a 0.03 increase in our dependent variable, when controlling for one's living situation.

In other words, if we had two individuals that had the same housing/living situation, but one was a year older, we would expect the older one to like the Conservatives 0.03 more than their counterpart who is a year younger (based on the information we have in this model).

III. R-Squared

So, both variables are associated with our dependent variable, but how good is the model as a whole at explaining our dependent variable? To judge this, we can consider the R-Squared, which is presented in the top right of the output.

The R-squared is a measure of how close each data point in our model is to the regression line. It ranges from 0-1, and tells us how good our model is (i.e. all of our independent variables combined) at explaining our dependent variable.

In this case, our R-squared is 0.039, meaning our model (comprised of housing tenure status and age) accounts for 3.9% of the variation in likeCon scores.

IV. Regression Formula:

Another way to look at the output is by way of the regression formula.

$$Y = a + B X + B X$$

Dep. Variable constant coefficient Ind. Variable coefficient Ind. Variable

We can use the output from the regression analysis to calculate predictions of Y (e.g. likeCon) given certain values of X1 (age) and X2 (housing tenure). We do this by 'plugging' the coefficient and constant in to the formula, along with the value of X we want to use, and calculating the formula.

E.g.: what is predicted likeCon (Y) score for 19-year-old (X1 = 19) renter (X2 = 1 & X3 = 0)?

The coefficients are: Age = 0.03, Rent = -0.7, Fam/Friends = 0.1, and the constant = 2.9.

We can plug these figures in to calculate our prediction:

$$Y = a + B\,X1 + B\,X2 + B\,X3$$
$$Y = 2.9 + 0.03\,(19) - 0.7\,(1) + 0.1\,(0)$$
$$Y = 2.8$$

Tasks:

1. Run a regression model with likeLab as the dependent variable and gender (profile_gender) and age (age) as the independent variables. Interpret the results (i.e. coefficients and p-values).

2. Use the R-squared to comment on the model overall.

3. Use the regression formula to calculate likeLab predictions for the following respondents:

   a. A 28-year-old man and renter
   b. A 58-year-old woman and home-owner
   c. An 18-year-old woman living with parents