Q-STEP STATA 'HOW TO' GUIDES:

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.

Preparing our Variables

We have a variable – W3WES43a – 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).

First, let's look at how our dependent variable is distributed.

hist likeCon

As we can see, there is a good range of responses to this variable. Next, we can check our main independent variable – housing tenure. The variable (profile_house_tenure) measures the respondents' living situation. Let's use the tab function to explore this variable:

tab profile_house_tenure

```
. tab profile_house_tenure
```

House Tenure	Freq.	Percent	Cum.
Own – outright	1,586	38.10	38.10
Own – with a mortgage	1,126	27.05	65.15
Own (part-own) – through shared ownersh	15	0.36	65.51
Rent – from a private landlord	618	14.85	80.35
Rent – from my local authority	185	4.44	84.79
Rent – from a housing association	205	4.92	89.72
Neither – I live with my parents, famil	169	4.06	93.78
Neither – I live rent-free with my pare	185	4.44	98.22
Other	73	1.75	99.98
Skipped	1	0.02	100.00
Total	4,163	100.00	

There are a lot of responses to this variable. Let's group some of the responses into intuitive categories to simplify our analysis. We can use the codebook function to check how the variable is coded:

codebook profile_house_tenure, tab(100)

We can then recode our variable so that the responses are grouped into "own", "rent", or "live with family or friends", and create a new variable – tenurecat.

recode profile_house_tenure (1/3=1 "Own") (4/6=2 "Rent") (7/8=3 "Live with Family/Friends") (9 98=.), gen(tenurecat)

We can now check the coding of our new variable before we use it for our analysis:

codebook tenurecat, tab(100)

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

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

reg likeCon tenurecat

The output will look like this:

```
. reg W3WES43a tenurecat
```

Source	SS	df	MS	Number of obs	=	2,988
				F(1, 2986)	=	44.38
Model	439.051594	1	439.051594	Prob > F	=	0.0000
Residual	29539.939	2,986	9.8928128	R-squared	=	0.0146
				Adj R-squared	=	0.0143
Total	29978.9906	2,987	10.0364883	Root MSE	=	3.1453

W3WES43a	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tenurecat	-.6326856	.0949708	-6.66	0.000	-.8189004	-.4464708
_cons	5.180391	.1410293	36.73	0.000	4.903867	5.456916

We would interpret the tenurecat coefficient (-0.6) as the change we see in the dependent variable (W3WES43a) when the independent variable (tenurecat) increases by one unit. But: the problem here is that a "one unit increase in the independent variable" doesn't make much sense in this instance… Why? Because there are several, nominal categories – "own", "Rent" and "living with family/friends". It isn't logical to measure a one unit increase between these categories that are clearly not ordered.

Instead, we want to tell STATA that tenurecat is a categorical variable. We instead use the following code, adding "i." before the variable:

reg likeCon i.tenurecat

This then produces the following output:

This then 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

Interpreting the Output:

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. Linear Regression – Age and Liking the Conservative Party

Let's run another regression model – this time using age as our independent variable:

reg likeCon age

The variable "age" is an interval level variable, so there's no need to prefix it with "i."

```
. reg W3WES43a age
```

Source	SS	df	MS	Number of obs	=	3,030
				F(1, 3028)	=	89.75
Model	874.373539	1	874.373539	Prob > F	=	0.0000
Residual	29499.7717	3,028	9.74232882	R-squared	=	0.0288
				Adj R-squared	=	0.0285
Total	30374.1452	3,029	10.0277799	Root MSE	=	3.1213

W3WES43a	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0333064	.0035157	9.47	0.000	.026413	.0401997
_cons	2.55767	.1938216	13.20	0.000	2.177635	2.937705

The 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. In other words, older respondents tend to like the Conservative Party more than younger respondents. Check the p-value associated with this coefficient – at what level is the result statistically significant?

III. Regression Formula:

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

$$Y = a + B X$$

Dep. Variable constant coefficient Ind. Variable

We can use the output from the regression analysis to calculate predictions of Y (e.g. likeCon) given certain values of X (e.g. age). 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. to calculate a likeCon prediction for a 19 year old, we would enter the following figures into the regression equation:

$$Y = 2.6 + 0.03\ (19)$$

$$Y = 3.2$$

Tasks:

1. Run a regression model with gender (profile_gender) as the independent variable and interpret the results (i.e. coefficient and p-value) . (NB: consider how the gender variable is coded and what this means for our interpretation)

2. How does age explain support for the other main political parties? Fit regression models for Labour (likeLab), Lib Dems (likeLD), Plaid Cymru (likePC), and UKIP (likeUKIP), using age as the independent variable. Compare the results.

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

   a. A 67-year-old
   b. A 21-year-old
   c. A 42-year-old