

INTERPRETING REGRESSION OUTPUT

Simon Davies, University of Bath
sd245@bath.ac.uk

This is a (mostly) non-technical check list of things you need to look out for and consider *after* running a cross sectional OLS regression. Firstly a couple of things to think about *before* running a regression.

Before running a regression:

- 1) **Think carefully about what you are trying to show**, or the hypothesis you want to test. Use economic theory, previous work and your own common sense to decide what your independent variable should be, which dependant variables you might need to include, and about the functional form of the regression. Try to specify a model before running a regression.
- 2) **Know the data.**
 - a. Know what the variables mean. Which are continuous and which are categorical? Are the categorical variables dummies? Do the values they take indicate a logical ordering (e.g. a wealth ranking: 1=poor to 5=stinking rich) or no ordering at all (e.g. 1=teacher; 2=student; 3=unemployed; ...)
 - b. Create new variables if necessary. E.g. log income; a dummy indicating whether or not someone is a student or not.
 - c. Know the means and standard deviations. Know maximums and minimums. Are there any **outliers**? Should they be deleted?

After running a regression:

- 1) ***Look at number of observations***
- 2) ***Look at the r^2***
- 3) ***Look at the F-test***
- 4) ***Interpret the signs of the coefficients***
- 5) ***Interpret the size of the coefficients where relevant***
- 6) ***Look at the significance of the coefficients (most important?)***

- 1) **Look at the number of observations.** Is it what you expect? If not, you should find out why. Remember, any observations with missing values will be

dropped from the regression. For example, if you log income, any observations reporting zero income will be dropped from your sample when you run a regression which includes the log of income.

2) Look at the r^2 .

- a. **Explaining the r^2 :** The r^2 tells you the percentage of the variation in the dependant variable that your model “explains”. The rest is in the error term. E.g. Think about individual incomes in the UK. Different individuals have different income. Our model might want to explain why this variation exists, so our dependant variable might be (the log of) income. Our independent variables could include individual characteristics such as education, age and distance from London (why include this?) so that we run the model:

$$\ln \text{INCOME} = a + b_1(\text{EDUCATION}) + b_2(\text{AGE}) + b_3(\text{MILES FROM LONDON})$$

and we find an r^2 of 0.46, this means that 46% of the variation incomes is explained by our dependent variables.

- b. **Problems with r^2 :** → If you have a “*very low*” r^2 , have a think about whether you might have omitted some important variables. Do other people doing similar work also have low r^2 ? Are you missing variables that many other people have included? However, be careful not to include unnecessary variables only to increase your r^2 . → A “*very high*” r^2 could indicate several problems. **Firstly**, if a high r^2 is combined with many insignificant variables, your independent variables might be highly correlated amongst themselves (multicollinearity). You might consider dropping some in the interests of parsimony. **Secondly**, it might be an indication that you have mis-specified your model. I once read an example (but can’t remember where) that said that if you have an r^2 approaching 1, you have run $\text{LEFT_SHOE} = a + b_1(\text{RIGHT_SHOE}) + \text{error}$. That is, you are trying to explain the number of left shoe people own, so include household characteristics such as wealth, age and job, and, thinking it might be significant, you include the number of right shoes they own. Are you really adding any interesting information here? Is this really what you want to explain, or does your model need re-specifying?

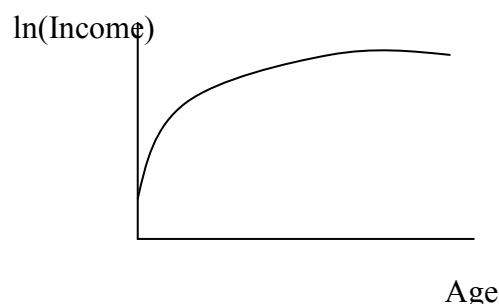
c. **The adjusted r2:** Adjusts the r2 to penalise the inclusion of more variables. Include as many variables as you need but keep your model as parsimonious (www.dictionary.com ☺) as possible.

3) **Look at the F test.** The F test aims to test the “global significance” of the model. More formally it is a test of whether all your coefficients are jointly equal to zero. If they are, effectively your model is not really explaining anything. **Hint:** ideally you want a **high F-value**, and a **low corresponding p-value**.

4) **Interpret the signs of the coefficients.** Which are positive and which are negative? Interpret this! A positive coefficient means that variable has a positive impact on your dependent variable, and a negative one has a negative impact. (Not rocket science!). So, if I estimate:
 $\ln\text{INCOME}=40.4+7.5*(\text{EDUCATION})+3.4*(\text{AGE})+(-2.1)*(MILES FROM LONDON)$ then I can interpret my results as saying that as education increases, so does income, and as some one gets older, their income increases. However, as a person lives further from London, their income declines.
 → **Question:** How about if I add age squared and find the following results:

$$\ln\text{INCOME}=40.4+7.5*(\text{EDUCATION})+3.4*(\text{AGE})-2.2*(\text{AGE_SQ})-2.1*(MILES FRM LNDN)$$

Answer: We have log of income increasing with age, but at a decreasing rate. That is, we have decreasing returns (illustrated below):

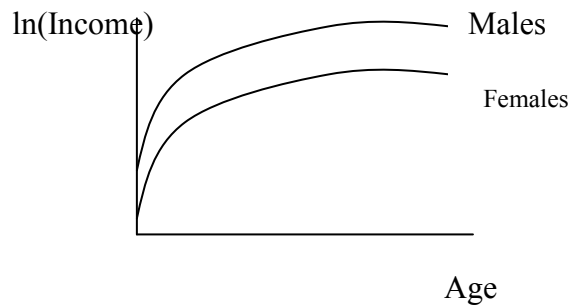


→ **Question:** How do we interpret a dummy variable?

→ **Answer:** It shifts the **intercept** for **all** observations in the category. The slope does not change. So, if we now add a dummy for females in our hypothetical regression and find it is negative, for example:

$$\ln \text{INCOME} = 40.4 + 7.5 * (\text{EDUCATION}) + 3.4 * (\text{AGE}) - 2.2 * (\text{AGE_SQ}) - 2.1 * (\text{MILES FRM LNDN}) - 2.1 * \text{FEMALE}$$

where FEMALE=1 if the individual is female, and 0 if he is a male. Then the below graph shows the impact. (Try and interpret meaning of graph/regression results in words).



5) **Interpret the size of the coefficients where relevant.** You've got a statistically significant coefficient – great! So maybe you've found eating more bananas increases worker productivity. But by how much? Is it worth a firm supplying free bananas to their employees or are the productivity benefits too small to make it worth while?

6) **Lastly, but perhaps the MOST IMPORTANT, look at the statistical significance of each variable.** This should in fact become the first thing that your eyes drift towards when you get regression output. You should feel a little hint of excitement as you are waiting to find out whether your model works and whether your theory has been proved correct or not!

→ The test of significance is designed to test whether a coefficient is significantly different from zero. If it is not then you must conclude that your explanatory variable does not, in fact, explain at all your dependent variable.

→ We use a t-test (just like you learnt in first year statistics) to test this, so that we compare a t value taken from a table (at a given significance level, α , with n-k degrees of freedom) with a calculated t, where n= number of observations and k=number of parameters estimated / independent variables; n-k = degrees of freedom.

$$t = \frac{\beta - \beta^*}{se(\beta)} = \frac{\beta - 0}{se(\beta)} = \frac{\beta}{se(\beta)} \quad \text{where } \beta = \text{coefficient value, } \beta^* = \text{hypothesised value}$$

(almost always 0), $se(\beta)$ = standard error of β .

All econometrics packages give the coefficient value, the standard error, and the t value as standard. They also automatically compare the t value with the value taken from a table in a book to find the significance level, α . This is the p-value you are presented with.

Hint: You want a **low standard error**, a **high t-value**, and a low **p-value**.

In general, in economics, we are prepared to accept a p-value of under 0.10 (=10%) as indicating a variable is significant. In other disciplines this value is different. For example, in medical tests, the lowest acceptable level taken to indicate the variable is significant is 0.001. (why?)

Why do we need significance tests? After all, we know the value of our coefficients for our sample. The problem is that it is correct only for our sample. But our sample is one of many that could have been drawn from the whole population. To see this imagine a country with a population of around 60 million, say the United Kingdom. I want to know how education affects the income of all 60 million people, but I don't have the resources to ask them all. So I take a sample of 10,000 and I estimate the equation above for my 10,000. I have found the correct relationship for those 10,000 people, but can I say anything about the other 59,990,000 people, that is, the whole population? After all, there is some chance that I might have picked out a lot of really luck people in my 10,000 sample who happen to have really high salaries but very little formal education (a lot of Richard Bransons, if you will). [Note this will bias the coefficient on education towards zero.]

We could pick out many different samples of 10,000 people (millions of different samples in fact), and we could estimate the same equation for each sample, and each time, find a slightly different β . Thus, β has a variance. The *estimate* of this variance is the standard error you are given. We want to know whether the sample we have picked is a freak or whether it is likely that the population as a whole has similar

characteristics to the sample (a similar coefficient), or whether the coefficient would be zero for the whole population.

Our t test therefore tests whether the population coefficient is likely to be different from zero. The probability of this is indicated by the p-value, or, doing it manually, by your chosen α level. For example, if we have a p-value of 0.01, that indicates that we can be 99% sure that our population parameter differs from zero.

A similar but slightly different way of thinking about this is in terms of **confidence intervals**. A confidence interval is usually centred around β and gives us a range within which we believe the population parameter lies. Supposing our software returns an *estimate* of $\beta=7.5$, and a 95% confidence interval of [6.2 - 8.8] then we can be 95% based on our *sample*, then we can be 95% sure that the population lies between 6.2 and 8.8. Put another way, in 95 out of 100 samples from the population, we can expect the coefficient, β , to lie within the corresponding confidence interval.

Note that the confidence interval gives us a minimum value and a maximum value within which you are $(1-\alpha)\%$ sure the population parameter lies. The upper and lower boundaries are given by $\beta \pm t_{\alpha/2} * se(\beta)$.

Other

Other tests follow, such as testing for normality of error terms, checking for existence of heteroskedasticity, performing specification and robustness tests. But these exciting topics are to be covered later in the course.