

This is a back-to-basics hint sheet for regression. The aim is to highlight a number of points which are basic but essential to good econometric practice, and which might not seem obvious to the beginner. Some of these points are based on questions asked by students undertaking their first solo econometric analysis. This looks long but should be a fairly easy and informative read.

1) **The dependant variable**

For your initial project your dependent variable is likely to be continuous, binomial or maybe ordered. Each provides a number of challenges to be thought about.

- *Continuous (e.g. income in pounds)*

Think about the relationship between the dependant variable and the key explanatory variables. Should the dependant variable be transformed into logs? (Use existing literature, economic theory and your own common sense to guide you in this decision.) If so, remember that any zeros will now become missing variables, and dropped from your regressions automatically by most econometric packages. If your dependant variable is income, and your unit of analysis is the individual rather than the household, you will almost certainly lose some observations. If you have learnt techniques such as Tobit models, consider using this.

Another transformation might be considered. Instead of a continuous dependant variable, you might find it useful to transform it into an ordinal categorical variable. For example, equal to 1 if income is between £0 and £5,000, 2 if income is between £5,001 and £15,000 and so on. This would ensure you keep any zero observations, but brings with it its own challenges.

- *Ordinal*

For example $Y=1$ if income between £0 and £5,000; $Y=2$ if income between £5,001 and £15,000; $Y=3$ if income between £15,001 and £25,000 and so on.

This model can be estimated by OLS, providing all assumptions are fulfilled. For example, the errors must be normally distributed (do a Jarque-Bera test to check for

this). It is always worth running an initial OLS model to get an idea of what is happening in your model.

The preferable model to use here is an “ordered probit”. Your introductory course does not cover this, however you will be rewarded for attempting to use the correct methods and for your analysis and the fact that you have had to research the method yourself. It will involve you reading up on the technique however.

If you are faced with this type of dependant variable, it is always useful to convert it into a binomial and run a probit. For example, a dependant variable taking values, $Y=1,2,3,\dots, 9,10$ could be split into high and low categories so that

$$Y^* = 0 \text{ if } Y = 1,2,3,4,5 \text{ and} \\ = 1 \text{ if } Y = 6,7,8,9,10$$

You don't have to split the values in the middle. Try and think about a sensible cutoff point depending upon what you want to verify and the distribution of Y . Any cutoff point is necessarily arbitrary. Nonetheless, this is always a good method to try.

- ***Binomial***

The dependant variable is equal to 1 or 0. Also known as a categorical, dichotomous or binary variable.

Here, the technique to use should be obvious: use a probit or logit.

2) Explanatory Variables

- How many to use?

There is no correct answer to this. Use as many as required (anything from 1 to 1,000), remembering that excluding relevant variables can bias (make incorrect) the estimates of the coefficients on included variables. There is a lot to be said however, for parsimony, and the fewer observations you have, the fewer independent variables you can use. Thus, variables should not be thrown in at random. Think about what is going into the regression, and justify each inclusion.

Including just one independent variable (plus a constant) is effectively the correlation between the dependant and independent variable.

- **Which independent variables should be included?**

Again, here, there is no correct answer, but there are guidelines. You should be guided by **economic theory**. If economic theory says that X influences Y, you should include X in your regression – even if it is insignificant.

Once you have decided on a topic, you need to look at **previous work** done on the topic. What variables have other people included and which are significant? Include them.

Finally, use your own **common sense**. Justify each variable included. But remember, if you choose to exclude a variable other people include, you might be asked to justify that decision too.

If you don't have a variable in your data set which you feel should be included, try and look for something to **proxy** it. That is, find something else that can be a proxy for the variable you don't have but want to include. For example, you might not have number of years someone has been working for, but you do have their age and you can reasonably argue that these are very strongly related/correlated. So age can be a proxy for experience.

At the end of the day, if you don't have a variable, you can't include it. You have to make do with what you have. But do ask yourself whether it is worth going ahead with your analysis or not. If the variable is a key one, then you should forget it. For example if you want to test the impact of religion on happiness, but don't know what religion people are, you simple cannot test this.

Explanatory variables can be any mix of continuous, binomial (dummies), or ordered categorical variables (1,2,3,4,5,...). However, you should take care not to include unordered categorical variables. For example

Profession = 1 if respondent a manager
=2 if student
=3 if teacher
=4 if other

What would the coefficient on this variable mean? Nothing, in fact.

Instead, you can create a series of dummy variables indicating whether or not the respondent belongs to each profession. (Remember, one dummy would need dropping!!!).

As an example, imagine you are estimating INCOME using:

$$\text{INCOME} = a + b_1(\text{EDUC}) + b_2(\text{AGE}) + b_3(\text{PROFESSION}) + \text{error}.$$

The coefficient, b_3 is meaningless as described above. Instead we should create a series of dummy variables indicating whether someone belongs to a profession or not. So that PROFESSION become

MANAGER=1 if respondent a manager, 0 otherwise

STUDENT=1 if a student, 0 otherwise

TEACHER=1 if a teacher, 0 otherwise

OTHER=1 if other, 0 otherwise (i.e. 0 if a manager, student or teacher in our example)

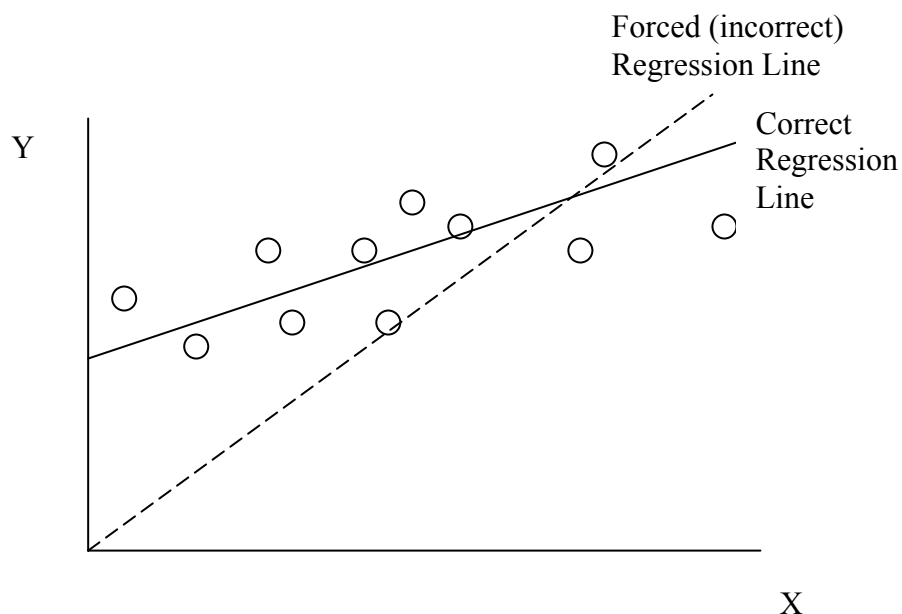
The regression now becomes:

$$\text{INCOME} = a + b_1(\text{EDUC}) + b_2(\text{AGE}) + b_3(\text{STUDENT}) + b_4(\text{TEACHER}) + b_5(\text{OTHER}) + \text{error}.$$

Note that we have dropped MANAGER. This now becomes the baseline, and the coefficients on the other dummies, b_3 , b_4 and b_5 indicate the effect of belonging to each of these professions *in relation to* being a manager. So, a positive coefficient on TEACHER means that, on average, a teacher will earn more than a manager (the excluded variable).

Which dummy to drop is largely a matter of personal choice, however, it is good practice to include any OTHER category you may have since comparing coefficients to an OTHER category is less informative than comparing differences amongst known categories.

Always include a constant! Actually, this is a slight exaggeration, but not much of one. There are cases in which the constant should be dropped, but these are rare, and you are unlikely to be dealing with such a case. Excluding the constant can severely bias your coefficients as you are, in effect, imposing that the intercept be at $X=0$. The below graph illustrates this.



The dots represent the observations. The solid line is the correct estimate of the regression line with a positive constant. The dashed regression line is the regression line when the constant is excluded (or, equivalently, set equal to 0).

Remembering that the coefficient on the independent variable is the slope of the line, we can see that by excluding the constant, our estimate becomes biased.

3) How can excluded variables bias the coefficients?

Let's take an example to see how this can happen. Imagine you would like to understand what contributes to happiness. This is a growing area of economic

literature, and is important too in psychology. Remember, happiness, or utility, is the basic building block or currency of economics. Increasingly surveys (including the British Household Panel Survey, BHPS) are asking people to rate how happy they are. For example, “On a scale of 1 to 6, where 6 is the most happy and 1 the least happy, how happy are you with your life in general?”. Similar questions can be asked about satisfaction with work, health or any other area of life and is known as Subjective Wellbeing (SWB).

With this background, we want to estimate the following regression (using OLS, or perhaps an ordered probit, or a simple probit – see earlier for how to do this).

$$\text{Happiness} = a + b_1(\text{INCOME}) + b_2(\text{MARRIED_DUMMY}) + b_3(\text{RELIGIOSITY}) + b_4(\text{JOB_DUMMIES}) + b_i(\text{OTHER VARIABLES}) + \text{error}.$$

Let’s suppose we estimate $b_1 = 0.8$. However, we have reason to believe that the number of children (KIDS) affects both Happiness on its own (say a positive impact), and affects INCOME. For example, the more children you have, the more you work to support your family so the higher your income is. Therefore, children have a positive effect on income or $(\uparrow \text{KIDS} \rightarrow \uparrow \text{INCOME})^1$. We also said $\uparrow \text{KIDS} \rightarrow \uparrow \text{Happiness}$.

By excluding KIDS, the coefficient on INCOME, b_1 , captures not only the effect of INCOME on Happiness, but also the effect of KIDS. The coefficient b_1 is biased (upwards, in our case) since people with more children tend to have (c.p.) higher income, and higher Happiness.

What we actually want here is the effect of INCOME “cleansed” of the effect of KIDS. There are two ways of doing this.

One involves “cleaning” income of KIDS by regressing $\text{INCOME} = a + b\text{KIDS} + \text{error}$. The error term is the part of INCOME not explained by KIDS. The INCOME

¹ Arguments could obviously be constructed that lead to the opposite relationship between KIDS and INCOME, so that the overall effect is ambiguous, but let’s go with this example to help illustrate the problem.

component of our regression, can be replaced by this error term. We do not discuss this further here, and interested readers can refer to Basic Econometrics by Gujarati (pages 205 to 207 and 213 to 215 in the 4th Edition (International)) for an example of the relationship between child mortality (dependant variable) and GDP and Female Literacy Rate (FLR, explanatory variables) in which GDP and FLR are related.

The second, and easiest method is simply to include KIDS as an explanatory variable. So that we run:

$$\text{Happiness} = a + b_1(\text{INCOME}) + b_2(\text{MARRIED_DUMMY}) + b_3(\text{RELIGIOSITY}) + b_4(\text{JOB_DUMMIES}) + b_5(\text{KIDS}) + b_i(\text{OTHER VARIABLES}) + \text{error}.$$

In this way, b_1 is cleansed of the influence of KIDS, and this is captured by b_5 (equal to, say, 0.2). We now estimate the true value of b_1 (assuming the model is now correctly specified), which might be, say, 0.6 (lower than the original estimated 0.8).

Finally

Hopefully some of these pointers will be useful to you. Take your time and go slowly, but the single most important thing, that cannot be emphasised enough, is to combine what you learn with **common sense**. Good luck!