Lecture 6: Specifying the Regression Model

Thomas Chadefaux

Contents

Tests are usually performed under the assumption that the model has been correctly specified. For instance, the computed standard errors of estimated coefficients and their P-values depend on this assumption. In practice, in initial rounds of the empirical cycle we may work with first-guess models that are not appropriately specified. This may lead, for instance, to underestimate the standard errors

1 How many variables should I include?

Assume that a set of explanatory variables has been selected as possible determinants of the variable y. Even if one is interested in the effect of only one of these explanatory variables—say, x_2—it is important not to exclude the other variables a priori. The reason is that variation in the other variables may cause variations in the variable y, and, if these variables are excluded from the model, then all the variations in y will be attributed to the variable x_2 alone.

On the other hand, the list of possibly influential variables may be very long. If all these variables are included, it may be impossible to estimate the model (if the number of parameters becomes larger than the number of observations) or the estimates may become very inefficient (owing to a lack of degrees of freedom if there are insufficient observations available). The question then is how many variables to include in the model.

1.1 Direct, indirect and total effects

Suppose the true model is

$$\text{Unrestricted: } y = X_1\beta_1 + X_2\beta_2 + \varepsilon, \tag{1}$$

but we estimate instead a restricted version of it, namely:

$$\text{Restricted: } y = X_1\beta_1 + \varepsilon. \tag{2}$$

I.e., we've omitted the variable(s) X_2. Now let us compare the estimate we get from the unrestricted model, b_1, and from the restricted model, b_R. We know that $b_1 = (X'X)^{-1}X'y$, but what is b_R?

$$\begin{aligned} b_R &= (X_1'X_1)^{-1}X_1'y \\ b_R &= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + e) \\ &= (X_1'X_1)^{-1}X_1'X_1\beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 \quad + (X_1'X_1)^{-1}X_1'X_1e \\ &\Rightarrow E[\beta_R] = \beta_1 + P\beta_2, \quad \text{since } Xe = 0 \end{aligned} \tag{3}$$

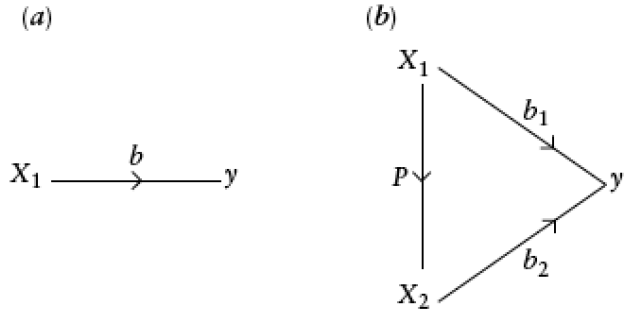

Figure 1: Direct and indirect effects

where $P = (X_1'X_1)^{-1}X_1'X_2$. So b_R is equal to b_1 plus some function of b_2. What is the intuition for P? P is the projection of X_2 onto the X_1 space. Intuitively, this means that P returns the effect of regressing X_1 on X_2 (see fig. 1). Thinking in terms of the Venn diagrams, b_R returns the blue portion AND red portions, as opposed to the blue portion only returned by b_1.

There are many reasons why P may be different from 0. I.e., why X_1 may affect X_2 or vice-versa. [From Heij , de Boer, Franse]: For instance, X_1 may 'cause' X_2 or X_2 may 'cause' X_1, or there could exist a third 'cause' in the background that influences both X_1 and X_2. It may be useful to keep this in mind when interpreting the restricted estimate b_R and the unrestricted estimate b_1. Consider the second element of b_R (the first element is the intercept). Traditionally, in a linear relationship this measures the partial derivative $\partial y/\partial z$ (where z now denotes the second explanatory variable?that is, the second column of X1). It answers the question how Y will react on a change in Z ceteris paribus?that is, if all other things remain equal. Now the question is: which 'other things'? In the restricted model, the 'other things' clearly are the remaining columns of the matrix X_1 and the residual e_R, and in the unrestricted model the ?other things? are the same columns of X_1 and in addition the columns of X_2 and the residual e.

Take the particular case that X_1 'causes' X_2. Then a change of X_1 may have two effects on y, a direct effect measured by b_1 and an indirect effect measured by Pb_2. Under these circumstances it may be hard to keep X2 constant if X1 changes. So in this case it may be more natural to look at the restricted model. That is, bR gives a better idea of the total effect on y of changes in X_1 than b_1, as it is unnatural to assume that X_2 remains fixed.

If one wants to estimate only the direct effect of an explanatory variable? that is, under the assumption that all other explanatory variables remain fixed?then one should estimate the unrestricted model that includes all explanatory variables. On the other hand, if one wants to estimate the total effect of an explanatory variable?that is, the direct effect and all the indirect effects that run via the other explanatory variables?then one should estimate the restricted model where all the other explanatory variables are deleted.

An Example:

Two variables (education and starting salary) influence the current salary, and education also influences the starting salary. The total effect of education on salary consists of two parts, a

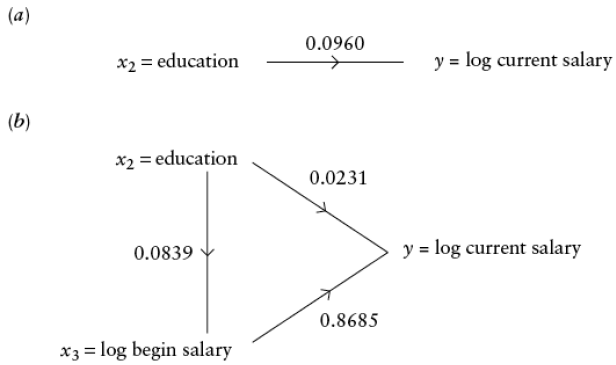

Figure 3: Bank Wages

direct effect and an indirect effect that runs via the begin salary. If salary is regressed on education alone, the estimated effect is 0.0960, and if salary is regressed on education and begin salary together, then the estimated effects are respectively 0.0231 and 0.8685. If begin salary is regressed on education, the estimated effect is 0.0839. In this case the direct effect is 0.0231, the indirect effect is $0.0839 \times 0.8685 = 0.0729$, and the total effect is $0.0231 + 0.0729 = 0.0960$

1.2 Omitting Relevant Variables: Omitted Variable Bias

We just noted that

$$E[b_R] = \beta_1 + P\beta_2. \tag{4}$$

In other words, b_R is biased and the last term in (4) is often called the omitted variable bias. This means that we should be cautious not to forget important variables—to control for them—or else our estimates will be biased.

On the other hand, however, the variance of b_R will be smaller than the variance of b_1. Why? Intuitively, we have more data to estimate b_R. Looking at the venn diagrams, we are using both the blue and the red portions to estimate b, whereas if we include the control variable W, we'd only have the blue portion left. Note that omitted variable bias is not a problem for our estimate of b_1 if (it always leads to bias for our constant estimate):

- only X_1 and X_2 are not correlated. In that case, $P = 0$ and so $E[b_R] = \beta_1$
- $\beta_2 = 0$. In other words, there will be no bias if the omitted X_2 has no effect on Y anyway. In effect, there is no specification bias here. This tells us that omitting an irrelevant variable does not induce bias

Note that sometimes we know that we have omitted a variable but have no way of including it, possibly because it is difficult to measure or observe. In these circumstances, we should at least think about the likely direction of the bias. So there is a trade-off between bias and efficiency.

1.3 Adding Irrelevant Variables Leads to Inefficiency

Adding variables that do not explain y leads to inefficiency, but the estimates will remain unbiased. Why is it inefficient? Because it eats up a degree of freedom. Most of the time, moreover, there will be some collinearity with other Xs, and hence the variance of the bs will increase.

1.4 trade-off between bias and efficiency

Which estimator of b1 should be preferred, b_1 or b_R? If $\beta_2 = 0$, then clearly we need to use b_R to avoid the loss of efficiency. BUT usually $\beta_2 \neq 0$. If we remove the estimator b_2, we obtain a more efficient b_1, but also one that is biased. The fact that restrictions improve the efficiency is one of the main motivations for modelling, but of course the restrictions should not introduce too much bias. So how to choose between the two models?

1. F-test
2. AIC and BIC
3. Out-of-sample predictions. The idea here is to compare the predictive performance of two models. The data is split in two parts, an 'estimation' sample and a 'prediction' sample. So the models are estimated using only the data in the first subsample, and the estimated models are then used to predict the y-values in the prediction sample. Possible evaluation criteria are the root mean squared error (RMSE) and the mean absolute error (MAE). These are defined by:

$$RMSE = \sqrt{\frac{1}{n}\sum_{1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

$$MAE = \frac{1}{n}\sum_{1}^{n}|y_i - \hat{y}_i|, \tag{6}$$

   where n is the number of observations in the prediction sample
4. Iterative variable selection methods: Brute force, bottom-up or top-bottom.

2 Non-linearities

2.1 Diagnostics

Residual plots

2.2 Data Transformations

First note that the scaling of variables is of no intrinsic importance, as you saw in HW1 (if you multiply x by 3, its coefficient will simply also be multiplied by 3). However, an important transformation is to take the logarithm of a variable.

Reminder: the logarithm of a number is the exponent to which another fixed value, the base, must be raised to produce that number. For example, the logarithm of 1000 to base 10 is 3, because 10 to the power 3 is 1000: $1000 = 10 \times 10 \times 10 = 10^3$.

Logarithmically transforming variables in a regression model is a very common way to handle situations where a non-linear relationship exists between the independent and dependent variables. Using the logarithm of one or more variables instead of the un-logged form makes the effective relationship non-linear, while still preserving the linear model. Logarithmic transformations are also a convenient means of transforming a highly skewed variable into one that is more approximately normal. (In fact, there is a distribution called the log-normal distribution defined as a distribution whose logarithm is normally distributed — but whose untransformed scale is skewed.)

Interpretation of models with log transformations.

- Linear model: $Y_i = \alpha + \beta X_i + \varepsilon_i$. b gives us the change in expected value of Y for a one-unit change in x
- Linear-log model: $Y_i = \alpha + \beta log(X_i) + \varepsilon_i$. Here, adding one to $log(X_i)$ will produce an expected increase in Y of β units. What does it mean to add 1 to $log(X_i)$? It means that X increases goes from $e^{log(X_i)}$ to $e^{1+log(X_i)}$, which means that X itself is multiplied by $e \approx 2.72$. But that's not very convenient. The problem is that we cannot make a general statement saying "if you increase x by 1 unit, y will increase by so many units", because the increase depends on the value of X (to see this, note that moving from $log(1)$ to $log(2)$ means increasing y by $\beta 0.3$. However, from $log(2)$ to $log(3)$ increases y by 0.17 only... However, we can talk in terms of percentages. We might want the effect on Y of an increase in X by, say, 10%. Then we can say that if X increases by 10%, Y increases by $\beta log(1.1) = .095$. In other words, $0.095b$ is the expected change in Y when X is multiplied by 1.1, i.e. increases by 10%.
- Log-linear model: $log(Y_i) = \alpha + \beta X_i + \varepsilon_i$. Here one unit increase in X_i leads to an expected increase in $log(Y_i)$ of β units. This means that the expected value of Y is multiplied by e^{β} for a one unit change in X.[1] For a v change in X, the expected value of Y is multiplied by e^{cb}
- Log-log model: $log(Y_i) = \alpha + \beta log(X_i) + \varepsilon_i$. The interpretation is that a 1% change in X leads to a 1% change in E[Y].

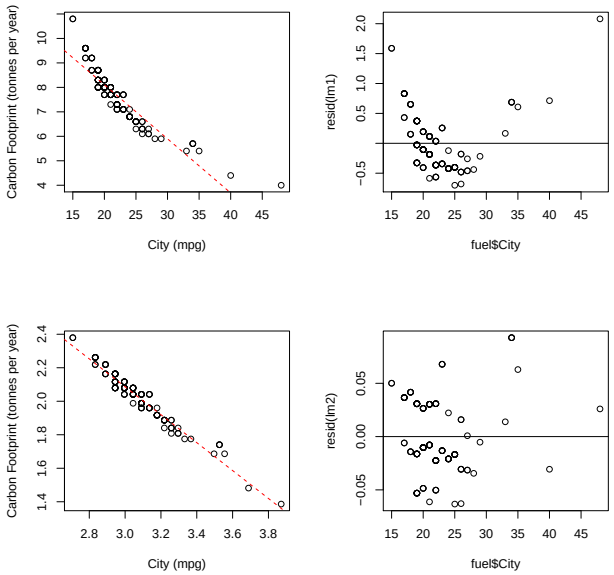

Figure 4: log log transformation. Top 2 plots show xy plot of untransformed variables (left), and corresponding residuals (right). Bottom row shows the same plots but where the log of each variable is used instead. Source: loglog.R

[1] Why?

$$\frac{\partial log(Y)}{\partial X} = \beta \tag{7}$$

$$\Rightarrow \frac{\partial e^{log(Y)}}{\partial X} = e^{\beta} \Rightarrow \frac{\partial Y}{\partial X} = e^{\beta} \tag{8}$$

3 Dummy Variables

Your variables may not be quantitative, but rather qualitative. In that case, you will need dummy variables.

3.1 Dummy variables

Consider for example the effect of education on income. You think the effect is the same for men and women, but that men tend to get more on average for the same level of education. i.e. the slope is the same, but the intercept is not. One way of formulating this argument is :

$$Y_i = \alpha + \beta_1 X_1 + \gamma D_i + \varepsilon_i,$$

where D_i is what is called a dummy variable, which takes value 0 for women and 1 for men. Thus for women, the model becomes:

$$Y_i = \alpha + \beta_1 X_1 + \gamma(0) + \varepsilon_i = \alpha + \beta_1 X_1 + \varepsilon_i$$

whereas for men

$$Y_i = \alpha + \beta_1 X_1 + \gamma(1) + \varepsilon_i = (\alpha + \gamma) + \beta_1 X_1 + \varepsilon_i$$

Interpretation: γ gives the difference in intercepts for the two regression lines—that is, the expected income advantage accruing to men when education is held constant. If men were disadvantaged relative to women, then λ would be negative.

Note that it doesn't really matter which group is coded 1 or 0 (just make sure you interpret correctly...).

Of course, we can add more dummy variable, for example for ethnicity, etc.

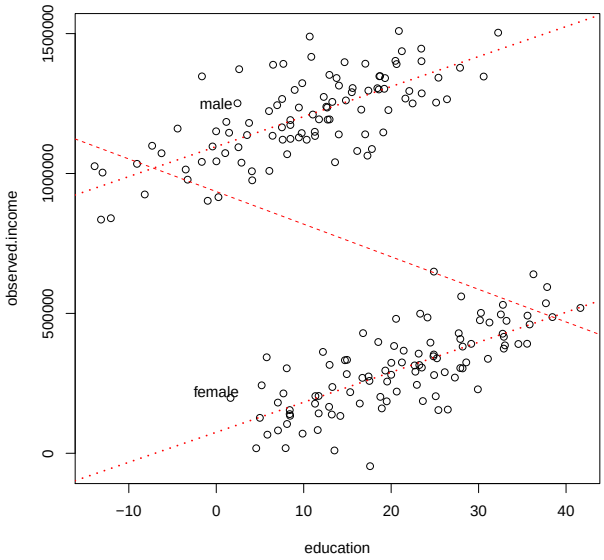

Figure 5: Effect of education on income for men and women. Source: dummies.R

3.2 Polytomous variables

You 'dummies' need not be binary. Suppose, for example, that you are interested in the effect of GDP on a country's level of democracy but think that its geographic location matters. In particular, you want to control for the region a country belongs to. Say, Europe, America, Middle East, Asia, Australia. That's 5 possible values. Then you would create 4 (yes, 4, not 5. Why?) dummy variables:

$$Democracy_i = \alpha + \beta_1 GDP_i + \beta_2 Europe_i + \beta_3 America_i + \beta_4 MiddleEast_i + \beta_5 Asia_i + \varepsilon_i$$

What this will tell you in that $E[Democracy_{Ger}]$ for Germany is

$$E[Democracy_{Ger}] = \alpha + \beta_1 GDP_{Ger} + \beta_2$$

. Note that for Australia,

$$E[Democracy_{Aus}] = \alpha + \beta_1 GDP_{Aus}$$

.

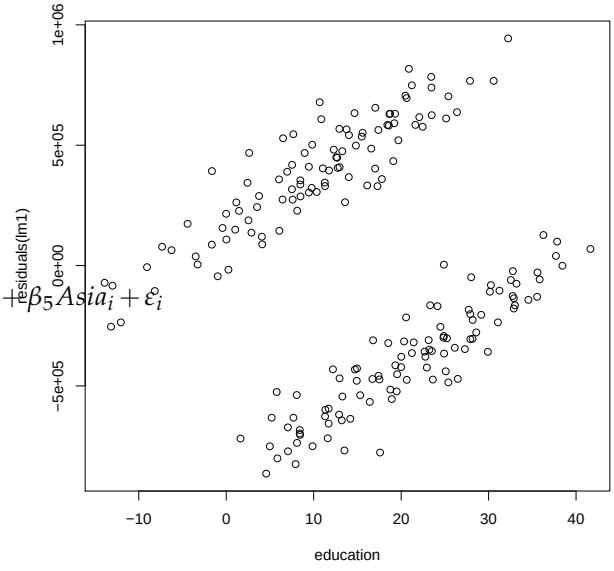

Figure 6: Plot of residuals for Fig. 5. Source: dummies.R

4 Interaction Effects

Now you might not think that the effect of gender on income is simply a shift in intercept. Instead, gender might affect the slope. In other words, one more year of education for a man might have a different effect on income than an additional year of education for a woman.

To model this type of 'interaction', we need a model with different slopes for men and women. Consider the following model:

$$Y_i = \alpha + \beta_1 education_i + \beta_2 gender_i + \beta_3 (education_i \times gender_i) + \varepsilon_i \tag{9}$$

So for males, the model is:

$$Y_i = (\alpha + \beta_2) + (\beta_1 + \beta_3) education_i + \varepsilon_i \tag{10}$$

and for females:

$$Y_i = \alpha + \beta_1 education_i + \varepsilon_i \tag{11}$$

Note that the slope for education for males $(\beta_1 + \beta_3)$ is different than the slope for females (β_1), AND that the intercepts are different, just like it was for dummies.

VERY IMPORTANT: You MUST be careful with the interpretation of interaction terms. Suppose your model is

$$Y = \alpha + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon \tag{12}$$

Then the effect of X on Y holding Z constant is:

$$\frac{\partial Y}{\partial X} = \beta_1 + \beta_3 Z$$

In other words, the effect of X on Y depends on the value of Z! You *can no longer* interpret β_1 as the effect of X_1 on Y.

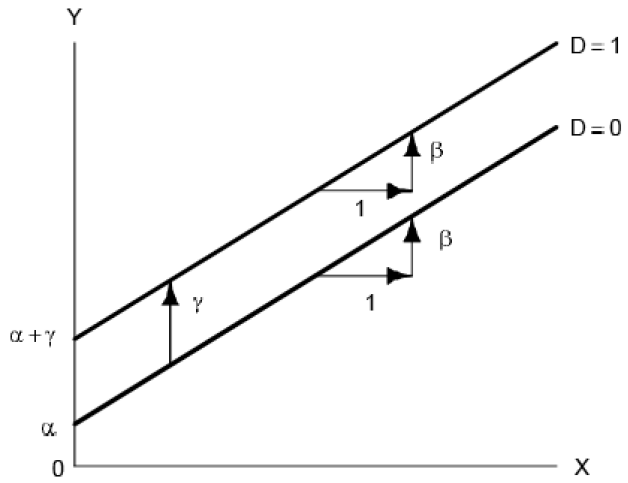

Figure 7: Regression line when Dummy = 1 and when =0

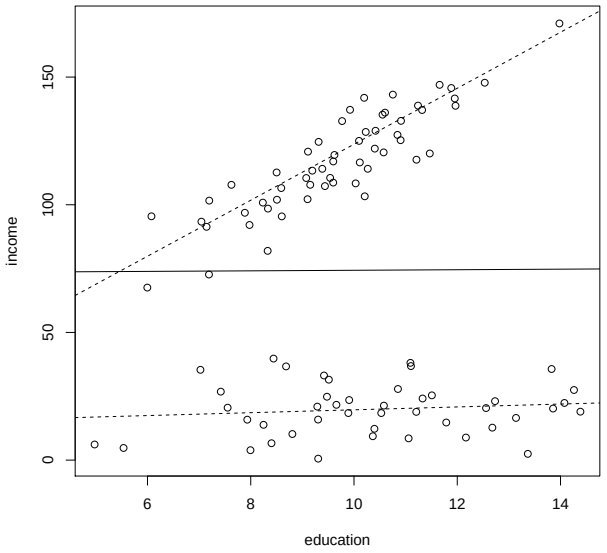

Figure 8: Source: interaction.R