Notes 17: Scaled variables and misspecification

ECO 231W - Undergraduate Econometrics

Prof. Carolina Caetano

1 Scaled variables

What happens if you add a variable in a scaled format? I bet you don't know the answer. However, I also bet that once I give you a specific example you will know the answer right away. For example, what happens if instead of using the temperature in Fahrenheit, I write it in Celsius? What happens if instead of using weigh in grams, I write it in ounces? The answer is _____

So, I can tell you the takeaway right now: you should never worry about the units of measurement in which you add a specific variable. It does not affect the model in any important way. However, it does affect the coefficients, in the sense that they have to adapt to make sense. The easiest way to see what will happen when we use scaled variables is to think of the unit of measurement. Fahrenheit and Celsius, ounces and grams, all are units of measurement. This can be extended to anything. For example, cigarettes are measured in average per day. However, the questionnaire could have asked: how many packs do you smoke per week? If you want to know the effect of cigarettes per day, then you could transform the data you have on cigarettes per day. Say that cigarettes per day is x_1, and the data set information (packs per week) is x_1'. One pack has 20 cigarettes. Over a week that is 20/7=2.86 cigarettes per day. So, $x_1' = 1$ implies $x_1 = 2.86$. The conversion is thus

You could go to the data and transform it into the unit you want (cigarettes per day) using the simple operation above. Alternatively, you could throw x_1' into the model instead of x_1, and in that case, the effect of one extra pack per week β_1' is the same as the effect of 2.86. So, the model is

If you estimate β_1', but you want β_1, then just do $\beta_1' = \beta_1 \cdot 2.86$. In the example, it means that the effect of each extra pack per week is the same as 2.86 times the effect of one extra daily cigarette.

This kind of operation is easy to do. The rule, if you are one to memorize, is that if $x_1' = a + bx_1$, and you use x_1' in the regression, then $\beta_1 = b\beta_1'$. I am actually confused by this, so usually I just think about it in a case by case basis. Try it, it is very easy!

Observe that the standard errors change as well. The rule is

It seems as if these changes were important, but nothing fundamentally important is happening. Models are hard because of unobservables, linearity, sample selection, etc. Scaled variables are just a matter of paying attention and adjusting a thing or two.

2 Misspecification

Misspecification is the failure of the linearity condition. Specifically, the model is misspecified if

The word misspecification is usually employed to denote a failure of the linearity assumption. However, sometimes some researchers refer to the problem of omitted variables also as misspecification (and it makes sense), so you should always read carefully. If you want to be specific, you can always say: "misspecification of the functional form."

There are four types of misspecification of the functional form:

1. Model is linear, but you added extra terms which should not be there. For example, the true model is

   $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u,$$

   but you run

or

or

What happens in those cases? _____

This is exactly the same as a case with superfluous variables. The fact is that $x_1 x_2$ in the first equation, x_1^2 in the second, and $\log(x_2)$ in the third are all superfluous. Why? _____

Hence, there are no consequences for $\hat{\beta}_1$ and $\hat{\beta}_2$, but the variance can (and likely will) increase. Since it is only one superfluous variable, this is not a worrisome issue. In practice this is of no concern. The following cases are actually troublesome.

2. The model is linear, but you did not add extra terms which should have been there. For example, the true model is

but you run

In this example, you should have included x_2^2, but you did not. In other words, you omitted x_2^2. It doesn't matter that you do observe x_2, the fact is that you omitted the variable x_2^2. Hence, you will incur an omitted variable bias problem. Remember how I said that many problems could be understood as omitted variable bias? This is one of those. Unfortunately we cannot apply the formulas directly. Let's check this out. Remember the bias formula:

$$\hat{\beta}_1 \approx \beta_1 + \hat{\beta}_w \hat{\theta}_1$$

where

- $\hat{\theta}_1 \approx \mathbb{E}[w|x_1 = 1, x_2, \ldots, x_k] - \mathbb{E}[w|x_1 = 0, x_2, \ldots, x_k]$

  (Formally, it is the slope coefficient of x_1 in a regression of w onto $x_1, \ldots, x_k$. This is mathematically equivalent to the following: define $\eta = w - d_0 - d_1 x_2 - \cdots - d_k x_k$, the residual of a regression of w onto the controls $x_2, \ldots, x_k$. Then θ_1 is the coefficient of a regression of η onto x_1).

- $\hat{\beta}_w \approx E[y|x_1, \ldots, x_k, w = 1] - E[y|x_1, \ldots, x_k, w = 0]$

  (Formally, it is the coefficient of w if we could run a regression of y onto $x_1, \ldots, x_k, w$. This is mathematically equivalent to the slope coefficient of a regression of y on η, which is the average change in y when we increase η in one unit.)

The problem here is that we always have to condition on x_2, and so x_2^2 is always pre-determined. So, $\hat{\theta}_1 \approx$ _____

and $\hat{\beta}_{x_2^2} \approx E[y|x_1, x_2, x_2^2 = 1] - E[y|x_1, x_2, x_2^2 = 0]$ doesn't even make sense. How can we keep x_2 fixed and vary x_2^2? This is why we need to look at the formal definition (the "$\approx$" sign is there for a reason).

The variable η is the residual of a regression of x_2^2 onto x_2, $x_3, \ldots, x_k$. We know that x_2 cannot perfectly predict x_2^2. Now regress η onto x_1, there is no reason why $\hat{\theta}_1$ should be zero.

The term $\hat{\beta}_{x_2^2}$ is the slope of a regression of y onto η, which is mathematically equivalent to the coefficient of x_2^2 on a regression of y onto x_1, x_2 and x_2^2, which should be similar to β_3. (So, for approximation purposes, we can say $\hat{\beta}_{x_2^2} = \beta_3$.

Thus, the OLS estimator $\hat{\beta}_1$ will be biased. To figure our the signal depends on the specific variables.

3. The model is linear, but you added terms in the wrong functional form. For example, the true model is

but you run

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

or, suppose that the true model is

but you run

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u.$$

The above cases are not different from the second type of misspecification. The only thing you must do is to understand what is the omitted variable. To do this, you must start with the true model, and transform it into the model you run. The remainder is the omitted variable. Take the first example. The true model is

Hence, the omitted variable is _____

What is the omitted variable in the second case? Try it at home.

4. The model is not linear, but you think it is. For example, the true model is

but you run

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u.$$

This is a very troublesome case. To figure out the bias, you can proceed in the same way as above:

2.1 Addressing misspecification

Misspecification is not less serious than endogeneity. It causes the same kind of trouble. However, it is much easier to solve. The econometric theory to solve misspecification problems is very advanced, although most techniques are beyond the level of this course. What should you do?

- It is easy to test whether there is trouble in your model. At the very least (unless you really don't have much data) you should attempt to test whether your specification is correct. How do you do it? Throw in some extra terms, some squared terms (of course, don't square dummies), some interactions, some logs. Then, do an F test to check that the coefficients of the extra terms are all zero.

- If your test does not reject, don't breathe as if you are done. Try other tests, throw different variables. Also, make sure that the reason you are not rejecting is simply because your standard errors are blown up.

- If your test rejects, beware. It could be a sign of misspecification, but it could also be a sign of endogeneity. You need to investigate some more. Think about your problem, try to figure out what the right functional form is using economic theory. You can also plot (especially graphs of averages) to better understand the shapes of the relations between the variables. If you investigated enough, and your test still rejects, it could be endogeneity. If this is the case, you have to face that problem. Either look for proxies, or look for something called instrumental variables (search instrumental variables, IV regression, 2SLS regression), or panel data (search for panel data models, fixed effects). There is a lot of material about how to solve endogeneity, but have no doubt, it is one of the hardest problems to solve in applied research.

- If you want to be serious about the possibility of misspecification, you can learn more advanced techniques. If your model is non-linear (in parameters), but you think you know its shape, you need to study GMM estimators. If you don't know the shape at all, you need to study semi-parametric and non-parametric models. This is pretty advanced, but if you need it, you should know to look for them (for example, look for partially linear models, nonparametric additively-separable models, non-parametric estimators, kernel regression, local polynomial regression, series regression, sieves, and ask for help.) Solving misspecification may be fancy, but it is doable.

- Whatever you do, do not try to solve misspecification by throwing in every interaction, squares, and cubes of the variables in your model without thinking about it. If you get to this point, you must be dealing with a highly non-linear model. You are misguidedly doing something called a power series regression (read the previous item). It is a fair approach, but you need to understand what you are doing, and the issues associated with this specific technique. Moreover, if you do get to this point, there are better strategies.