Thinking About Issues in the Methodology

Sara Stoudt

March 2, 2016

1 Find significantly different counties.

$v_{p,t}$ defined as CropFailure$_p, t$ / DroughtIndex$_p, t$

p over counties, t over time

$$v_{p,t} = \frac{c_{p,t}}{d_{p,t}}$$

Equation 1 (ANOVA)

$$v_{p,t} = \alpha + \beta_p Dummy_p + \epsilon_{p,t}$$

Equation 2 (equivalent to equation 2 after moving denominator to other side)

$$c_{p,t} = \alpha d_{p,t} + \beta_p d_{p,t} Dummy_p + \epsilon^*_{p,t}$$

1. Equation 1 is getting a different β_p for each county (answering the question: does resiliency differ by county?).

2. The interaction between the continuous variable and the categorical dummy variable in Equation 2 allows for different intercepts (see Figure 1). It's interpretation is the effect of $d_{p,t}$ on $c_{p,t}$ depends on the level of place (which exactly gets at the question of whether the effects of drought differ on crops [in this case] by place as before).

3. We could have specified Equation 1 without an intercept. The intercept acts as a base for county which all β's are comparing with respect to. Without an intercept we would just have an extra β value. The results using software such as R will be the same with or without the intercept in this case (see second answer here `http://stats.stackexchange.com/questions/7948/when-is-it-ok-to-remove-the-intercept-in-lm`).

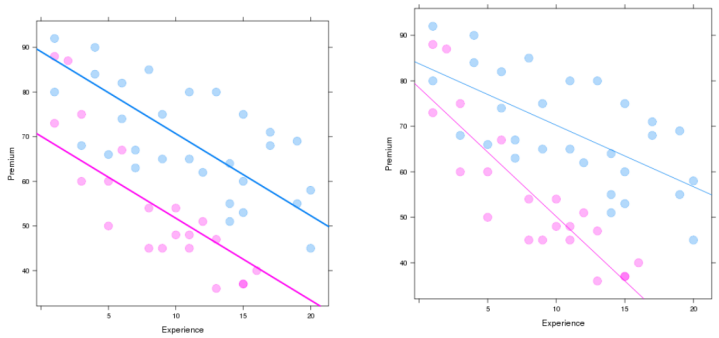

Figure 1: Parallel Slopes Model v. Separate Slope Model

2 Find significant trends over time in certain places.

Equation 1

$$v_{p,t} = \alpha_p + \beta_t Dummy_p * (timeVariable) + \epsilon_{p,t}$$

Equation 2

$$c_{p,t} = \alpha_p d_{p,t} + \beta_t d_{p,t} Dummy_p * (timeVariable) + \epsilon^*_{p,t}$$

1. In Equation 1 we have a separate intercept per place and we answering the question whether or not resilience differs over time in different places.

2. In Equation 2 we now have an interaction between two continuous variables (timeVariable and $d_{p,t}$). This is interpreted as the effect of drought on crops differs over time. This gets at the question whether effects of drought on crops differs over time per place as before.

3. We also have an interaction with the dummy variable for place again allowing a different slope for each county. $d_{p,t}$ is absorbed into each α_p.

In both cases above, moving the denominator to the other side just adds another variable to assess in an ANOVA approach. Both equations answer the same questions.

3 Still Not Convinced about the Zero Intercept?

Mathematically, a regression with a zero intercept means that we assume that the expected value of y given $x = 0$ is 0.

1. x will never be zero in our case since x is a ratio. The denominator would have to be huge to even approach zero.

2. $index = \frac{y}{x}$ goes to infinity as x goes to zero which breaks are assumption.

3. **However** $x = 0$ means that the average value of drought is much (much, much) less/greater than the average. Similarly $y = 0$ means that the predicted value of crop failure is much (much, much) less/greater than predicted by the detrending. What we are saying by the assumption that $E[Y|X = 0] = 0$ is that you can't be resilient to crazy (way way different than we would ever expect given past data) behavior which is reasonable in context.

4 Multiple Testing

Because we are essentially getting a coefficient for every county, we do have to worry about multiple testing which isn't really addressed in the paper. The easiest way to deal with this is to use the Bonferroni correction (divide the p-value at which we reject the null by the number of tests we perform). This is a conservative measure.

A less conservative measure that is still relatively easy to perform can be found here `https://en.wikipedia.org/wiki/Holm%E2%80%93Bonferroni_method`. The main idea is that you test only the most extreme p-value against the strictest criterion and the others against progressively less strict criteria. I can look into more methods if we want to be a little less naive about correcting for multiple testing (but by addressing it at all we are already doing better than the paper...).

Another way to assess this issue is to hope for robustness in the results. If we can find trends using a variety of measures that are related in context, then that gives us more credibility than if we found a certain county to be significant under one specific metric only.

5 Use of Spearman Correlation Coefficient in the Second Stage

This is a non-parametric way to measure of statistical dependence between two variables. It assesses how well the relationship between two variables can be described using a monotonic function (rather than a line as in a regression approach). The benefit of this is that we are interested in relationships between socioeconomic variables and our index. We don't have any reason to think that this relationship will be linear, and we don't want to mess around with trying to guess the functional form.

6 What if a weird scenario happens that breaks this measure?

If we can identify what type of scenario might break this measure we can formalize it and simulate data from a model where the scenario occurs. We can see what kinds of index values we would get and compare to what we get in real life and assess whether or not it is reasonable to worry about the bad scenario.

7 Paper Claims Causality

We just can't claim causality.

In the paper Ian posted about health outcomes and drought (The Effect of Drought on Health Outcomes and Health Expenditures in Rural Vietnam), I am pretty convinced that their instrumental variable is doing its job, and their causality claim seems fair. However, they do this at the individual level. We can't use this approach because we are working at an aggregated level and we would run into the ecological fallacy about reasoning about individuals given group data since the instrumental variable really needs to be at the individual level to be convincing.