Robustness Tests: What, Why, and How

In your econometrics class you learn all sorts of analytic tools: ordinary least squares, fixed effects, autoregressive processes, and many more. The purpose of these tools is to be able to use data to answer questions. In fact, they promise something pretty spectacular: if you have the appropriate data and the tool is used correctly, you can uncover hidden truths about the world. Does the minimum wage harm employment? What was the impact of quantitative easing on investment? Does free trade reduce or increase inequality?

At the same time, you also learn about a bevy of tests and additional analyses that you can run, called "robustness tests." These are things like the White test, the Hausman test, the overidentification test, the Breusch-Pagan test, or just running your model again with an additional control variable. These are often presented as things you will want to do alongside your main analysis to check whether the results are "robust."

But what does that mean? What do these tests do, why are we running them, and how should we use them?

This page won't teach you how to run any specific test. But it will tell you what the tests are *for*, and how you should think about them when you're using them. That sort of thinking will apply no matter *what* robustness test you're thinking about. I will also address several common misconceptions regarding robustness tests. You might find this page handy if you are in an econometrics class, or if you are working on a term paper or capstone project that uses econometrics.

Why We Run Robustness Tests

It's easy to feel like robustness tests are a thing you *just do*. Running fixed effects? Do a Hausman. That's the thing you do when running fixed effects. But this is not a good way to think about robustness tests! It can lead to running tests that aren't necessary, or not running ones that are. Or, even if you do the right test, you probably won't write about the findings properly in your paper.

So then what are robustness tests for?

Robustness tests are all about *assumptions*. Do you remember the list of assumptions you had to learn every time your class went into a new method, like the Gauss-Markov assumptions for ordinary least squares? Or do you at least remember that there *was* such a list (good luck on that midterm)? That's because every empirical analysis that you could ever possibly run depends on *assumptions* in order to make sense of its results. It's impossible to avoid assumptions, even if those assumptions are pretty obviously true. Sure, you may have observed that the sun has risen in the East every day for several billion days in a row. But if you want to predict that it will also rise in the East tomorrow, you must *assume* that nothing will prevent it from occurring - perhaps today is the day that it turns out Superman exists and he decides to reverse the Earth's rotation so the sun rises in the West.

These assumptions are pretty important. Without any assumptions, we can't even predict with confidence that the sun will rise in the East tomorrow, much less determine how quantitative easing affected investment. So we have to make assumptions. But then, what if, to our shock and horror, *those assumptions aren't true?* In that case, our analysis would be wrong.

So that's what robustness tests are for. We are worried *whether our assumptions are true*, and we've devised a test that is capable of checking either (1) whether that assumption is true, or (2) whether our results would change if the assumption WASN'T true.[1]

Often, robustness tests test hypotheses of the format:
H0: The assumption made in the analysis is true.
H1: The assumption made in the analysis is false.

This tells us what "robustness test" actually means - we're checking if our results are *robust* to the possibility that one of our assumptions might not be true.

Thinking about robustness tests in this way - as ways of evaluating our assumptions - gives us a clear way of thinking about using them.

Every time you do a robustness test, you should be able to fill in the letters in the following list:

1. My analysis assumes A.
2. If A is not true, then my results might be wrong in way B [estimate too high/estimate too low/standard errors too small/etc...].
3. I suspect that A might not be true in my analysis because of C.
4. EITHER: D is a test of whether or not A is true,
   OR: D is an alternate analysis I can run that does not assume A, allowing me to see how big of a problem B is.
5. If it turns out that [A is not true OR B is a big problem], then I

will do E instead of my original analysis.

If you can't fill in that list, don't run the test! Filling in the list includes filling in C, even if your answer for C is just "because A is not true in lots of analyses," although you can hopefully do better than that.[2] As a bonus, once you've filled in the list you've basically already written a paragraph of your paper.

Keep in mind, sometimes filling in this list might be pretty scary! If the D you come up with can't be run with your data, or if you can't think of a D, then you have no way of checking that assumption - that might be fine, but in that case you'll definitely want to discuss your A, B, and C in the paper so the reader is aware of the potential problem. Also, sometimes, there's not a good E to fix the problem if you fail the robustness test. Sometimes, the only available E is "don't run the analysis and pick a different project." Regardless, we have to make the list!

Let's put this list to the test with two common robustness tests to see how we might fill them in. First, let's look at the White test. The White test is one way (of many) of testing for the presence of heteroskedasticity in your regression. Heteroskedasticity is when the variance of the error term is related to one of the predictors in the model. Let's say that we are interested in the effect of your parents' income on your own income, so we regress your own income on your parents' income when you were 18, and some controls.

1. My analysis assumes that the variance of the error term is constant and unrelated to the predictors (homoskedasticity).
2. If my error term is heteroskedastic, then my results might have incorrect standard errors.
3. I suspect that the error term might be heteroskedastic in my analysis because among groups with higher incomes, income will be more variable, since there will be some very high earners. So if parental income does increase your income, it will also likely increase the variance of your income in ways my control variables won't account for, and so be correlated with the variance of the error term.
4. The White test is a test of whether or not the error term is homoskedastic.
5. If it turns out that the error term is heteroskedastic, then I will use heteroskedasticity-robust standard errors instead of my original analysis.

Second, let's look at the common practice of running a model, then running it again with some additional controls to see if our coefficient of interest changes.[3] Why do we do that? Let's fill in our list. Let's imagine that we're interested in the effect of regime change on economic growth in a country. So we are running a regression of GDP growth on several lags of GDP growth, and a

variable indicating a regime change in that country that year.

1. My analysis assumes that my variables are unrelated to the error term (no omitted variable bias).
2. If there is omitted variable bias, then the coefficient on regime change might be biased up or down, depending on which variables are omitted.
3. I suspect that there might be omitted variable bias in my analysis because regime change often follows heightened levels of violence, and violence affects economic growth, so violence will be related to GDP growth and will be in the error term if not controlled for.
4. Adding recent violence as a control is an alternate analysis I can run that does not assume that violence is not in the error term, allowing me to see how big of a problem omitted variable bias due to violence is.
5. If it turns out that the coefficient on regime change is very different with the new control, then I will include violence as a control instead of my original analysis.

Notice that in both of these examples, we had to think about the robustness tests *in context*. We didn't run a White test just-because we could. We ran it because, in the context of the income analysis, homoskedasticity was unlikely to hold. We didn't just add an additional control just-because we had a variable on hand we could add. We added it because, in the context of the regime change analysis, that additional variable might reasonably cause omitted variable bias.

Why bother with this list? A few reasons! First, it will make sure that you actually understand what a given robustness test means. No more running a test and then thinking "okay... it's significant... what now?" Second, the list will encourage you to think hard about your actual setting - econometrics is all about picking appropriate assumptions and analyses for the setting and question you're working with. Thinking about robustness tests in that light will help your whole analysis. Third, it will help you understand what robustness tests actually *are* - they're not just a list of post-regression Stata or R commands you hammer out, they're ways of checking assumptions. *Any* analysis that checks an assumption can be a robustness test, it doesn't have to have a big red "robustness test" sticker on it. Heck, sometimes you might even do them *before* doing your analysis.

Fourth, it will organize your thinking about your analysis, and help you avoid several common misconceptions about robustness tests...

Common Misconceptions

I should "do all the robustness tests."

There are lots of robustness tests out there to apply to any given analysis. You can test for heteroskedasticity, serial correlation, linearity, multicollinearity, any number of additional controls, different specifications for your model, and so on and so on. In most cases there are actually multiple different tests you can run for any given assumption. Since you have tests at your fingertips you can run for these, seems like you should run them all, right?

No! Why not? The reason has to do with *multiple hypothesis testing*, especially when discussing robustness tests that take the form of statistical significance tests. Roughly, if you have 20 null hypotheses that are *true*, and you run statistical significance tests on all of them at the 95% level, then you will on average reject one of those true nulls just by chance.[4] We commonly think of this problem in terms of looking for results - if you are disappointed with an insignificant result in your analysis and so keep changing your model until you find a significant effect, then that significant effect is likely just an illusion, and not really significant. You just found a significant coefficient by random chance, even though the true effect is likely zero. The same problem applies in the opposite direction with robustness tests. If you just run a whole bunch of robustness tests for no good reason, some of them will fail just by random chance, even if your analysis is totally fine!

And that might leave you in a pickle - do you stick with the original analysis because your failed test was probably just random chance, or do you adjust your analysis because of the failed test, possibly ending up with the wrong analysis?

We can minimize this problem by sticking to testing assumptions you think might actually be dubious in your analysis, or assumptions that, if they fail, would be really bad for the analysis. A good rule of thumb for econometrics in general: don't do anything unless you have a reason for it.

If my analysis passes the robustness tests I do, then it's correct.

This page is pretty heavy on not just doing robustness tests because they're *there*. One of the reasons I warn against that approach to robustness tests so much is that I think it promotes a false amount of confidence in results.

After all, if you are doing a fixed effects analysis, for example, and you did the fixed effects tests you learned about in class, and you passed, then your analysis is good, right?

No! Why not? Because your analysis depends on *all* the assumptions that go into your analysis, not just the ones you have neat and quick tests for. If you really want to do an analysis super-correctly, you shouldn't be doing one of those fill-in lists above for every

robustness check you run - you should be trying to do a fill-in list for every *assumption your analysis makes*. Of course, for some of those assumptions you won't find good reasons to be concerned about them and so won't end up doing a robustness test. But you should think carefully about the A, B, C in the fill-in list for each assumption. This conveniently corresponds to a mnemonic: Ask what each (A)ssumption is, how (B)ad it would be if it were wrong, and whether that assumption is likely to be (C)orrect or not for you.

There's another reason, too - sometimes the test is just weak! Sometimes, even if your assumption is wrong, the test you're using won't be able to pick up the problem and will tell you you're fine, just by chance. Type I error, in other words. There's not much you can do about that. But do keep in mind that passing a test about assumption A is some *evidence* that A is likely to be true, but it doesn't ever really *confirm* that A is true. So you can never really be sure. Just try to be as sure as you reasonably can be, and exercise common sense!

Robustness tests are always specialized *tests*.

Many of the things that exist under the banner of "robustness test" are specialized hypothesis tests that *only* exist to be robustness tests, like White, Hausman, Breusch-Pagan, overidentification, etc. etc.. It's tempting, then, to think that this is what a robustness test *is*. So is it?

No! Why not? Because a robustness test is anything that lets you evaluate the importance of one of your assumptions for your analysis. We've already gone over the robustness test of adding additional controls to your model to see what changes - that's not a specialized robustness test. These kinds of robustness tests can include lots of things, from simply looking at a graph of your data to see if your functional form assumption looks reasonable, to checking if your treatment and control groups appear to have been changing in similar ways in the "before" period of a difference-in-difference (i.e. parallel trends). Don't be fooled by the fancy stuff - getting to know your data and context well is the best way of figuring out what assumptions are likely to be true.

Endnotes

[1] If you want to get formal about it, assumptions made in statistics or econometrics are very rarely strictly *true*. After all, they're usually idealized assumptions that cleanly describe statistical relationships or distributions, or economic theory. Often they assume that two variables are completely unrelated. But the real world is messy, and in social science everything is related to everything else. So the real question isn't really whether the assumptions are *literally true* (they aren't), but rather whether the assumptions are *close enough to true* that we can work with them.

[2] In some cases you might want to run a robustness test even if you have no reason to believe A might be wrong. But this is generally limited to assumptions that are both *super duper important* to your analysis (B is really bad), and might fail just by bad luck. For example, it's generally a good idea in an instrumental variables analysis to test whether your instrument strongly predicts your endogenous variable, even if you have no reason to believe that it won't. That's because the whole analysis falls apart if you're wrong, and even if your analysis is planned out perfectly, in some samples your instrument just doesn't work that well.

[3] Despite being very common practice in economics this isn't really the best way to pick control variables or test for the stability of a coefficient. But that's something for another time...

[4] Technically this is true for the same hypothesis tested in multiple samples, not for multiple different hypotheses in the same sample, etc., etc.. C'mon, statisticians, it's illustrative and I did say "roughly," let me off the hook, I beg you. I have a family.