EMPIRICAL RESEARCH:
Where should we be heading?

James Ohlson

(June 20 2016)

The Big Picture: Before you get going on any RQ

- Fancy nouns and fancy econometrics breed attempts at validating fantasies.

- The world we live in is messy. The fact that a scenario/story is plausible does not mean the data will provide support.

- If nobody has heard of your story, the data is *exceedingly* unlikely to co-operate.

The Big Picture: Before you get going on any RQ

Continue …

- Research success is always difficult. But it is much easier if you try to deal with pre-existing issues in an interesting fashion rather than coming up with stories that nobody has heard of.

- Yes, it is true: Theoretical papers do *not* generally yield interesting "new" RQ. Such papers may, however, allow you to frame the (presumably pre-existing) RQ in a more interesting way.

- Sherlock Holmes: " It is a capita mistake to theorize before having examined the facts"

DATA ANALYSIS: BASIC OBSERVATIONS

- **“Make it as simple as possible, but no more.” (Einstein.)**

  What would you do if you never heard of p-values? Focus on educating the reader about the data rather than “arriving at the correct/desired conclusion via a p-value.”

- **To understand data, nothing beats counting (binomial) exercises and two by two matrices.**

  Exercise wrt accruals: What is the incidence of negative accruals? What does the matrix uptick/downtick in sales and positive/negative accruals look like? What percentage of times is a negative discretionary accrual followed by a positive discretionary accrual?

DATA ANALYSIS: BASIC OBSERVATIONS

- **The main research uncertainty pertains to how you frame the question, not p-values.**

  Suppose you want to study the extent to which certain measures of discretionary accruals “reverse”, how would you go about studying it? There may be more than one approach. Then recognize that conclusions can be ambiguous.

  Ambiguous conclusions are entirely consistent with well executed research. By contrast, unqualified claims tend to occur when,

  (i), the author believes it increases the probability of getting a publication or,

  (ii), the author is pre-disposed emotionally to dismiss real world complexities (aka wishful thinking).

DATA ANALYSIS: BASIC OBSERVATIONS

- **Descriptive correlations should be looked at carefully and compared to other studies. Memorize key numbers. Use medians (rather than means) and rank correlations.**

As to rank correlations:

< 0.1	“effectively zero”
0.1 to 0.2	“pretend it is good enough for an A –journal; but do not write home about it”
0.2 to 0.3	“ visible to the eye – it is there”
>0.3	“ you have something quite real”

With sufficiently large N all correlations are significant

What keeps Empiricists so busy?

- **Answer: Regress x on y controlling for z!**

- Y is old and familiar (like cost of equity, or tax-rate, or management earnings forecast, or listing on some exchange, or capital expenditures,.. Or, …

- X is the VDJ (variable-de-jour) or the paper's "substantive" focus. Like IFRS, or gender, or the analyst following, or real earnings management, or discretionary accruals, or earnings smoothing, or management incentives, AND ,nowadays, throw in an interactive effect for good measure …

- Z is , supposedly, inherited from prior research when Y (like, ROA, size, MTB, leverage…) AND, nowadays, FFE and FTE.

What keeps Empiricists so busy?

- The name of the game: X is relevant because (i) the estimated coefficient has the correct sign and (ii) (***) ! Victory can be declared by the author if and only if both conditions are met.

  The proposed story supposedly describes how the real world works!

- The next paper comes around with same Y same Z (more or less), another X say X(2) as opposed X(1). Has Z been modified to include X(1)? Nope. Why?

All too often common sense tell us that X(1) cannot be relevant in a substantive sense. Nor will X(2) be.

So what problems have not been confronted?

Regressing x on y controlling for z: So what are the problems?

- **You end up staring at the signs of estimated coefficients related to the VDJ and p-value—kind of boring, is it not?**

  Why? Because you try to validate rather than find out; uncertainties are viewed as an unfortunate nuisance and fought tooth and nail.

- **OLS leads to "OLS PLUS" – you have to get things "right" (especially the estimated sign) and that takes some trial and error. OLS plus defined: winsorization, trimming, scaling of variables, dummy variable controls, and interactive effects. Yes, they do produce an endless stream of screens, and, yes, one of them will undoubtedly be "right".**

  OLS PLUS tends to challenge your sense of ethics. (And not being alone is of only limited consolation.)

Regressing x on y controlling for z: So what are the problems?

- DISCOMFORT:

Anxiety I: If the VDJ is deleted, then the R^2 does not decline (at best, the third digit). What to do? Do not disclose the offending regression in the paper? What about deleting the two R^2? (The gambit may work in oral presentations – but do not bet on it.) Deny the relevance of R^2?

Anxiety II: The t-statistic results in three stars, but N runs into 20,000 plus. No way around it: the reader that cares will notice that the VDJ could have introduced NOISE in the regression! And the researcher's gut suggests the same. Basic quantitative rule: a t-statistic is less than SQRT(0.4% of N) suggests no "real" significance.

Anxiety III: Introducing the VDJ in fact reduced the goodness-of-fit. Can be tested using a relative accuracy (RAS) test. Compare a model including VDJ vs. a ceteris paribus model without VDJ. Check which of the two models' implied value of y is the closest to y's actual value and count the number of times the "with VDJ" model wins. Be prepared; at best a trifle above 50% but, more likely than not less than 50%. (Yes, your gut has been sending an accurate signal all along.)

Regressing x on y controlling for z: So what are the problems?

- IN EFFECT:

BY DISALLOWING A FALSE NEGATIVE, YOU HAVE SET YOURSELF UP TO MAXIMIZE THE PROBABILITY OF A FALSE POSITIVE.

AS "N" INCREASES THE PROBABILITY OF A FALSE POSITIVE INCREASES (The Jeffry & Lindley paradox)

NO SOLID APPROACH TO "CONTROLLING " VARIABLES. REFERENCE TO THE PRIOR LITERATURE OFTEN PLAIN NONSENSE.

WHAT TO DO: ANY PRESCRIPTIONS?

- Do NOT pose questions that elevate your emotions by "wanting" certain conclusions or findings. Recognize that a false acceptance of the null may happen.
- Keep in mind that y and z deserve at least as much attention as the VDJ.
- If X, the VDJ, does not help, then so be it. If you picked an interesting RQ – a plausible PRE-EXISTING scenario -- the reader will still learn a lot. Your problems starts when you have a farfetched scenario.(For example, highly paid CEOs are more likely to engage in EM).
- Given any Y (yes, almost any Y) never expect that more than 4 or 5 RHS variable can contribute to an explanation of Y. (Models with 25 or more variables on the RHS can be viewed as a form of academic entertainment.). Step-wise regressions can act as very useful tools to provide a first cut reading as to which variables are likely to be the most relevant. That aside, use RAS tests to convince the reader what variables are (ir-) relevant.

WHAT TO DO: ANY PRESCRIPTIONS?

- If N is large, split the data into subsets to educate the reader about the underlying robustness of estimated models. IF you do not do that, tell the reader why.
- Never take regression t-values literally. The educated person knows that t-values are ALWAYS understated because OLS presumes (x, z) are non-stochastic. When you report on a t-value, in the same breath always inform the reader about N too. In the good old days, the reporting tuple (t=xxx, N=yyyy) was commonly used.
- Avoid playing OLS PLUS games. PLUS is fine – provide you use a hold out sample (evaluate the model's (relative) ability to explain y on data not used to train the model)

WHAT TO DO: ANY PRESCRIPTIONS?

- OLS IS NOT LONG RUN VIABLE AS A STANDARD PARADIGM. IT HAS NO KNOWN VIRTUES (“EXCEPTIONS”: OVERSTATED T-SATISTICS AND SCREEN-PICKING)

  REPLACE IT BY USING THEIL-SEN (TS)

  (i) SIMPLER THAN OLS

  (ii) OUTLIER PROBLEM ABSENT

  (iii) SCALING PROBLEM ABSENT

  (iv) MORE EFFICIENT THAN OLS:

  OLS = TS + NOISE

. BACK TO BASICS. FOCUS ON A DEPENDENT VARIABLE AND FIND THE LHS VARIABLES THAT EXPLAIN IT. MAYBE SOMETHING HAS BEEN MISSING ON THE RHS?

. YES, BOOTSTRAPPING IS A FINE TECHNIQUE . BUT YOU HAVE TO START TO LEARN HOW TO LIVE WITH MUCH LARGER STANDARD ERRORS

A text-book quote on bootstrapping

"21.6 Concluding Remarks

If the bootstrap is so simple and of such broad application, why isn't it used more in the social sciences? Beyond the problem of lack of familiarity (which surely can be remedied), there are, I believe, three serious obstacles to increased use of the bootstrap: 1. Common practice—such as relying on asymptotic results in small samples or treating dependent data as if they were independent—usually understates sampling variation and makes results look stronger than they really are. Researchers are understandably reluctant to report honest standard errors when the usual calculations indicate greater precision. **It is best, however, not to fool yourself, regardless of what you think about fooling others."**

- **A CONVERSION TO SERIOUS RESEARCH INVOLVES TWO STEPS: Internal *and* external deception must be avoided at all costs**

Summary

- Goodness-of-fit considerations are important: Does X **REALLY** help to explain Y?
- Avoid OLS PLUS unless you set aside a hold out sample.
- Serious research allows for ambiguous conclusions.
- Focus on **pre-exiting** issues; in the long run they prevail
- Research uncertainty deals with RQ framing – not p-values
- Report on N whenever you report on a t

AND

Stay away from "emotional story validation"

Make the data analysis as simple as possible