A Few Notes on Analysis Writeups

Dale W.R. Rosenthal

Last updated: November 26, 2018

Having seen your first two homeworks, I thought it would be a good time to give some guidance on the specifics which make for a good writeup of a modeling exercise of analysis.

1 Narrate

First and foremost: always comment on numbers and results. I once saw a TA (at UChicago) throw a homework at a student because the student had merely printed out 50 pages of output from *R*. He also yelled at the student: "You are making *me* do the analysis. That is *your* job! *You* show me what is important and tell me why."[1]

While harsh, the TA was not incorrect about the need to point out the results and note both what they mean and why they are important. Perhaps a better way to think of this is to imagine yourself like a tour guide: explain to the reader what they should be noticing — what is interesting and surprising. This creates a *narrative*: a story.

You might think other people should know what the analysis means. They might, but they also might be so busy that they forget or miss an important detail. Good narration makes it easy for them to see the value in the work you did. Poor or non-existent narration implies the work has little value and lets useful results hide.

2 Setup

While I do not ask for it in the homeworks, in your work life a summary table with statistics for all variables is crucial. Showing plots of all variables is also useful; however, that is not as critical as a summary table.

[1] To be fair, the student was from the Financial Mathematics program and the TA was from the business school — which has a *strong* dislike of the FinMath program.

3 Models

When we work with models, we need to be clear on a few details:

1. What is the dependent variable (aka response, output, LHS)?
2. What are the independent variables (aka covariates, inputs, RHS)?
3. If there are multiple stages of a model, we need to see this for each stage.

Linear models generally have the form of

$$Y = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon \tag{1}$$

or, in linear algebra notation:

$$Y = X\beta + \epsilon. \tag{2}$$

Implicit in that is that X has a column (X_0) of all 1's which estimates the y-intercept. When we work with returns, the y-intercept is very useful because it soaks up the general trend in prices. Often, we cannot depend on any trend to continue, so this is useful.

When we estimate these models, there are usually two details we must know and one that is mildly interesting.

The must-know details of a model are (1) coefficient estimates, and (2) t-stats or standard errors (or a confidence interval for the coefficient estimate). Typically, these are reported to 3–4 significant digits for estimates and 2–3 significant figures for standard errors or 1–2 decimals for t-stats. Sometimes, people denote standard errors by putting them in parentheses. Often, standard errors or t-stats are in grey, italicized, or smaller to make it easier to distinguish them from coefficient estimates.

These help us assess what information in the model is reliably related to the response. The coefficient estimates are often interesting for both sign and magnitude. Sometimes, theory tells us we should expect certain signs or magnitudes for the coefficients. If the coefficients differ from what we expect, it may be due to a trend in the data (check the y-intercept) or because the model is down-weighting some inputs and pushing the result toward an overall average (which is in the y-intercept).

We would hope that any variables added to a model are *statistically significant* (*i.e.* their coefficients are statistically significant). We would also hope that any variables are economically significant: we want them to have a non-trivial benefit to being in the model. If a variable is statistically

significant but only predicts 0.001% better for annual returns... well, then that variable is not very useful.

To look at models, we typically present these data in a table for multiple models. That allows us to compare models and see if certain coefficient estimates (or statistical significance) holds up across different models. If so, that gives us stronger evidence that the estimated effects are because we have isolated an important variable.

Do *not* use stars to denote statistical significance since (1) the meanings vary between different analysis applications, (2) it adds bulk to the table, and (3) some analysts *hate* stars since they imply a lack of thinking about statistical, economic, and theoretical significance.

An example is given in Table 1.

Dep. Variable:	T3M	T1Y	T2Y	T5Y	T10Y	T20Y	T30Y
Intercept	-0.01	-0.007	-0.004	0.009	0.001	0.015	0.002
$t =$	*-20.5*	*-17.0*	*-11.5*	*36.1*	*66.0*	*71.0*	*73.2*
TIPS	0.70	0.72	0.71	0.80	0.81	1.04	0.87
$t =$	*83.0*	*94.4*	*106.5*	*155.6*	*159.1*	*182.1*	*112.7*
E(Inflation)	1.27	1.20	1.18	0.65	0.45	0.31	0.28
$t =$	*57.0*	*62.5*	*70.1*	*51.8*	*40.1*	*28.3*	*19.1*
Var(TIPS)	-2.02	-2.03	-1.59	-21.0	-10.9	-4.7	-12.8
$t =$	*-17.2*	*-19.0*	*-17.1*	*-24.8*	*-11.2*	*-3.7*	*-7.4*
Var(Infl. Surp.)	-0.91	-0.82	-0.68	-0.22	-0.07	-0.07	-0.10
$t =$	*-26.1*	*-26.0*	*-24.7*	*-12.7*	*-4.3*	*-4.3*	*-4.6*
R^2	83.4%	86.3%	88.8%	94.1%	93.5%	93.4%	85.1%

Table 1: Estimated descriptive linear models of nominal interest rates with windowed variance estimators. Note that (1) real rates are not fully priced into nominal rates, (2) expected inflation may be over- or under-priced into nominal rates with less effect at longer tenors, and (3) uncertainty in real rates and inflation implies lower rates.

Finally, what of R^2? While R^2 is the proportion of variance explained by the model; thus it is a metric based on the data we analyzed, *i.e.* an in-sample metric. We can always increase the R^2 by merely stuffing more inputs into our model. That need not yield a better model in any sense.

That said, we often expect R^2 to be low — since we are often making predictions. Market efficiency suggests that we should not expect high R^2 values if we are predicting returns. In that case, most of future returns are unexplained (*i.e.* noise) and so a 3% R^2 might be a nice result and explaining

an additional 0.05% of the future data might be OK.

Thus while we may look at or show (in-sample) R^2, it is *at most* used to check if something is grossly wrong: if you are predicting returns and your model has a 30% or 80% R^2, you almost surely messed up and should not show those results to anyone without much more verification. You should never choose a model based on in-sample R^2.

As mentioned before, commentary is crucial. Looking at different model forms, it might appear that one is clearly better or disproves theoretically-suggested models. This may not be the case, however, and your commentary should make that clear. For example, a model for returns with an insignificant y-intercept (alpha) and positive coefficients from a market index and a size factor might just be confirming that a more broad-based index ("market" plus some fraction of SMB) yields a better CAPM with smaller (maybe even insignificant) alpha.

4 Polish

Finally, there are a few things to do that make your writeups more polished. These are not needed for your writeups in this class; however, doing these is critical for any writeups you do in industry.

Note that some firms use Jupyter, R Notebooks/RMarkdown, or Beaker notebooks for storing analyses. These guidelines apply to those environments as well.

First, put tabular data in real tables. If you are using a word processor, create a centered table with some rows and columns. If it requires its own page, so be it. If you are using LaTeX, your writeup will likely look better. In that case, use `center` and `table` + `tabular` environments. Give each table a number and a caption (`\caption` in LaTeX).

Second, do likewise for figures: they should all be numbered, centered, and captioned.

Captions should draw readers in. Sadly, many people (yes, your future boss!) will only look at the figures (and maybe tables) and read their captions — so the captions should not require the reader to have read the text to understand. If you have a successful caption, it may pull the reader in to actually reading the text.

And finally, always summarize: repeat your findings at the end. You want to make it easy for people to take your advice.