Final Paper Guide for iTV Matching/PS590

Jake Bowers

Fall 2011

Each of you will write a paper which you will turn in three weeks after the end of class. The paper is a methods paper, not a substantive paper. The goal of the paper is to allow you to practice writing a technical report either (1) comparing a linear model (covariance adjustment) to the kind of matching we've practiced (post-stratification adjustment) as methods of adjustment or (2) executing a post-stratified analysis and discussing and justifying the choices you made. You can think of it as an appendix to a paper that justifies your choice of analytic strategy to reviewers who are asking why you made this or that choice.

Most papers for this course will take a specific form that I roughly outline here.

1. Find a linear model in which a comparative claim is assessed: for example, the claim may be that two groups differ on values of their outcome "controlling for" covariates, or the claim may be more about "dose" or about differences among multiple groups. The key feature of such a model is that you should be able to write down what kind of comparison it implies. I would encourage you to use a ideal linear model you've developed in in other work — a (g)lm that is a centerpiece of a dissertation chapter, a conference paper, a seminar paper.
2. Ensure that you can reproduce the linear model output table.
3. Re-specify and execute the analysis using direct comparison (i.e matching) rather than covariance adjustment. What are the potential outcomes? What are the covariates? What would statistical inference mean (i.e. what unobserved quantities are you inferering about when you infer?)? What justifies the statistical inference strategy you choose? Do you need to execute a sensitivity analysis? Will you use a linear model after matching for further covariance adjustment?
4. Explain why the two analyses differ (or not). How do the two analyses reflect differently on your substantive and theoretical concerns that motivated the covariance adjustment in the first place?

If you would rather write a different paper, you'll need to get explicit consent from me. I encourage you to contact me to discuss the shape of your paper as the class goes along.

1 Rubric

Here is the overview of the questions I will ask of your paper You can think of this list as the rubric that I will use to grade your papers.

Writing Can I read your writing? Is your paper written clearly? [I agree with Becker (1986)'s analysis of the evils of passive voice and the literature review.]

Conceptualizing the Comparison Is it clear what you are comparing and why? What kinds of comparisons reflect on theory in which kinds of ways? What are the potential outcomes in this study?

Description Did you provide a useful and clear sense of the provenance of the data, the measurement of the variables, the typical values and variation of important variables? Did this sense link to a useful substantive discussion?

Redo the Regression If you are doing a paper comparing a linear model for adjustment to matching with adjustment, did you actually redo and show the linear model?

Operationalizing the Comparison What values on what variables do you plan to use for the key causal comparison? What variables are covariates? How can you justify that they are covariates and not intermediate outcomes?

Why adjust? What covariates require adjustment? Why?

Justification of the matching setup Did you explain why and how you plan to setup the distance matrices, penalties, calipers, etc.. of your matching? Do propensity scores play a role? Or mahalanobis distances? Why or why not?

Engagment with balance Did you assess balance? Was your balance assessment thoughtful? Were you thoughtful and creative in weighing excluding observations versus giving up some balance?

Engagement with statistical inference about causal inferences Did you show thoughtful engagement with the process of doing statistical inference about your, possibly causal, or at least comparative, inferences? Where you clear about the target of your statistical inferences? Where you clear about what kinds of additional assumptions you were making in order to believe the stastistical inferences? Were you able to check on any of these assumptions?[1]

Sensitivity Analysis Did you execute a sensitivity (or at least a bounds) analysis?

Reproducible Research Did you include a code appendix? Or the source Sweave of your document? Can I run all of the code in your appendix without cutting and pasting to replicate all of your analyses in order? To recreate your figures and tables?

2 More General Points About Graduate Seminar Papers

Data summaries should reflect substantive questions.

First and foremost data summaries (like linear regression coefficients) ought to actually present the comparisons that we claim. Implications of this:

- If your claims are about democracy but your variable records amount of chocolate ice cream eaten per capita you have a lot to do to convince us that democracy is reasonably represented by chocolate ice cream eating. If this statement seems overly self-evident or simple, read Adcock and Collier (2001) and see especially their Figure 1 on measurement.
- Is your hypothesis about the total relationship between X and Y, or the partial relationship [where a partial relationship always is defined relative to some specific other set of variables and functional form]? If it is about the total relationship, then why are there control variables in your model? If it is about a partial relationship, then how can you guard against (a) interpolation and extrapolation of the kind discussed in Gelman and Hill (2007); (b) multi-dimensional influential points (i.e. points which would not be influential were it not for the loss of degrees of freedom occuring from the addition of variables to your model); (c) other micronumerosity and overfitting problems that occur as models become more complex [i.e. it does not take a lot of data to tell us something precise about as single difference of means, but it takes ever more data to tell us something precise about a difference of means conditioning on the kitchen sink.] [2]
- If your data summary mostly reflects the influence of just one observation, then you don't really have a very good summary, do you? That is, your summary does not reflect the substantive comparison you care about but rather it tells us a lot about Sweden. So here you must be alert to issues of influential points and overfitting, too.

Statistical inferences about data summaries should have clear targets of inference

Why are you doing statistical inference at all? What is unobserved and stochastic that demands your guessing? Are you asking about an unobserved population? About an unobserved series of repeated experiments? About your own (or the field's own) prior beliefs? If you can answer this question without saying, "Journal editors make me do it." then you have a clear target of statistical inference.

On the assumptions behind canned lm/glm tables

If you want a list of the standard assumptions of iid+CLT least squares and/or mle based statistical inference consult Fox and Achen, Kennedy, or many other available sources. Yet, remember, Normality of outcomes or errors is **not** a crucial assumption unless you are using the likelihood based approach (remember Achen (1982) for how and why). For the MLE based justification, Normality of outcomes does matter. But for the large-sample, iid+CLT based story of regression (as well articulated by Achen), Normality arises in the course of the iid+CLT.

If you are not using a simple linear regression but some glm model (logit, probit, poisson, negative binomal, . . .)

Then it is not adequate justification to say, "I have counts. . . " or "My regression could, in theory, predict values lower than zero or higher than 1." Nor is "heteroskedasticity" alone a reason to leave the simple linear regression model. The bootstrap,

[1] See Berk (2004) and Achen (2004, 2002, 1986, 1982)) for lots of good material on the issues and difficulties of statistical inference.
[2] This is a partial unpacking of some of the reasons behind Achen's arguments against garbage can regressions Achen (2004, 2002).

for example, makes no assumptions about homoskedasticity. And, in many cases, a linear fit to a binary outcome will not predict below 0 or 1 (or outside the range of the counts).

This is not to say that you can't or shouldn't do logit/poisson/negative-binomial models, but that you should be aware of what are strong ustifications and what are weak ones. For this paper, it is enough to say, "I found this negative binomial model" or "Others are using this." — putting the burden of justification on someone else. It is fine to start there in order to deeply engage with / support / criticize such decisions. And if you want to justify said model, then that is fine. But I encourage you to spend a moment thinking about how such a move improves upon the basic smoothed difference of means of the simple regression. In the end, you may desire a more flexible functional form for **Xb** rather than assuming that a particular link function is useful for you. For example, you might ask about the substantive reason for the logit versus the cloglog link versus probit link in the binomial glm.[3]

On assumptions:

It is not enough to claim that some model involves assumptions, and that said assumptions are not satisfied, in order to justify use of another model. The other model will have often as many if not more assumptions as the model left behind; it too would be easily discarded under the rhetoric of "makes assumptions."

A very common move in this regard when using binary outcomes is to say, "Linear regression assumes linearity and allows predicted values beyond 0 or 1. Therefore I use the logistic regression/logit model." Notice, that the logit model and the linear regression both assume the same **Xb**, but the logit model *adds* the assumption that one should mash **Xb** into the [0,1] interval via the logit link function g(Xb) where $g()$ is the logit function. So, (1) both require linearity in how **X** relates to **b** (i.e. neither estimate b in x^b) and (2) logit adds another assumption on top of linearity and now requires a correct likelihood function whereas Achen shows that the OLS model does not require a correct likelihood function when n is large. In addition, if the linear regression does not predict beyond 0 or 1 in your particular data with your particular regression, then using the logit transform solely for the purpose of limiting $\hat{\mathbf{y}}$ to [0,1] is no longer necessary. In this case, the homoskedasticity requirement for the standard errors of the linear model is assumed indirectly in the logit model in which all of the observations are assumed to come from the same binomial/Bernoulli data generating process with p (the probability of a 1) parameterized by g(**Xb**). In fact concerns about the variance of the binomial outcomes are verycommon

Of course, we don't care about homoskedasticity if we are using design-respecting permutations/shuffling or bootstrap resampling. And, even if we are using iid+CLT justified statistical inference, as Fox noted in his textbook, heteroskedasticity is usually the very least of our worries.

There *are* good reasons to use logit. The best reasons to do something are positive: logit may offer you something that would help you rather than merely be a way to not do simple linear regression.

R Code should be an appendix to the paper.

If you are using Sweave to produce your final papers there are two options for making this appendix:

You can collect the code using the Rtangle(mypaper.Rnw) function from within R. Edit the resulting mypaper.R file adding `\begin{Verbatim}` at the top and `\end{Verbatim}` at the bottom. Then you can just do something like:

Listing 1: "Simple R Code Appendix"

```
\section{R Code Appendix}
\input{mypaper.R}
```

A schnazzier method is to ensure that each Sweave code chunk has a label, like the following:

Listing 2: "Schnazzy R Code Appendix"

```
  The regression of $y$ on $x$ is simple and shows no relationship, as
  you can see in Table~\ref{tab:lm1}.

<<mymainregression,eval=TRUE,echo=FALSE,results=hide>>=
## Fit the model y~x on the data.frame mydata
lm1<-lm(y~x,data=mydata)
@

<<tableformainregression,echo=FALSE,results=tex>>=
## Make a table of the regression results
xtable(summary(lm1)$coef,label="tab:lm1")
```

[3] There are good answers to this question, by the way, especially in the economics literature on latent choices and in the epidemiology literature. And there is a large literature on nonparametric smoothing and other ways to fit functions to conditional moments of outcomes without making the strong assumptions of (generalized) linear models.

```
@

Then you can repeat <<mymainregression>>= by just referring to it in your appendix.

\appendix
\section{The Appendix}
Here is the code appendix.

<<appendix, eval=FALSE, echo=TRUE, results=verbatim>>=
<<mymainregression>>
<<tableformainregression>>
@
```

At least, here is a quick and dirty approach. For example, if I were to do it for real, I'd put better R comments at the front of each code chunk which would serve to make the Appendix more intelligible. Perhaps also fo interest for playing would be the condordance option to the Rtangle() function [which itself is just a call to the Stangle() function].

If you are not using Sweave and/or LaTeX, then you can just add the R command file to your Word/OpenOffice/Pages document. I am assuming that you will be able to rerun all of the analyses in your papers without cutting and pasting by just doing "source(mycode.R)" in R or R CMD BATCH mycode.R from the unix command line (I'm not sure what the Windows command line is like these days).

More on literate programming.

- OpenOffice has it's own kind of Sweave possible called odfWeave.
- Of course you can use LyX with Sweave.

In case you are wondering about editors, I just switched back to plain Emacs from Aquamacs as my editor although I used to use Vi a lot (and I think MacVim is pretty cool) and others really like Textmate. I don't know the state of editors for Windows (although Emacs and Vi will run there as they do on Linux too). Once social science computing environments left the VAX, I used Emacs on Windows and Emacs and vi/vim on different flavors of Linux and Solaris.[4]

About writing:

If you don't already have it, get and read Becker's Writing for Social Scientists Becker (1986). Do whatever he says to do.

Here is a list of some of what I think are best practices for writing scholarly papers. Not everything on this list applies directly to the final paper for this class. So, in no particular order, here are my thoughts:

Dump all sections entitled "Literature Review" Read Becker (1986) for his chapter on using literature. The point of using past literature is to *motivate* your paper. Demonstrating knowledge of the literature occurs as you *situate* and *frame* your paper — as you show how your ideas are great and important and novel and different from and building upon but in conversation with past ideas. Talking about a literature without linking each and every sentence to your own particular project is like putting an undergraduate mid-term exam into the middle of a scholarly paper. It is boring, distracting, and inappropriate. Don't do it. If you find yourself labeling a section, "Literature Review" rather than "Why my ideas are so great and interesting" stop and ask yourself why you are doing this.

Motivate, motivate, motivate Readers should not have to ask "So?" or "So what?" or "Who cares?" I call paper that generates a lot of "So what?" questions *unmotivated*.

Even if you think that you are answering the "So what?" question, often you are not: people often misunderstand what it means to adequately answer the "So what?" question.

Two examples of common misguided attempts at motivation illustrate: (1) Motivation attempt #1 is to say "The literature has a hole." and even to point to the hole. Of course, the existence of something unstudied does not, in fact, mean it ought to be studied — or if it is studied that it will yield great and deep insights into the causes of war or voting or violence or democratic transitions. To stop at noticing the hole in the literature is to leave the paper unmotivated. To explain why the hole ought to be filled is to motivate the paper. (2) Motivation attempt #2 consists of claiming that some variation in (presumably causal) effects across some subgroup has not been studied before. This is a form of the first failed motivation attempt. The hole consists of a lack of knowledge about subgroup specific effects or about

[4] Jeff Gill and I periodically maintain a list of stuff to help folks who use Macs make them useful scientific workstations: https://github.com/jwbowers/SocialScienceMacConfig. Feel free to fork and contribute.

extent of variation across subgroups. Again, however, merely not having looked at such variation doesn't mean that so looking ought to teach us a lot about life, the universe, and everything (let alone about some important bit of scientific understanding).

I do *not* mean that you have to talk about Rousseau and Hobbes (or Douglas Adams) necessarily, but that you have to talk about the theoretical and practical/policy reasons why anyone should be reading your paper. If the real reason you wrote the paper is "Because (for some reason) I want a Ph.D. and this paper was required for a class." then you should consider writing a different paper. There is very little room in graduate school (let alone life) for you to write papers that you don't care about and that you can't convince other people to care about, too.

Answer "So what?" constantly Even if you have a page or two at the beginning of your paper explaining how it contributes important knowledge about important topics (raises important questions, generates important new concepts, etc..), readers should not have to ask *why* they are reading a giving sentence, paragraph, or phrase. That is, the need for motivation is not over once you've written a couple of paragraphs. Every piece of the paper should be directly and clearly intelligible as contributing to the "So what?" question answer.

Even if I am not a specialist in a particular area, I should still be able to say why this paper matters, why this section contributes to the paper (in a way that is linked to it being important), why this paragraph is here, why this sentence, and why this word. That is, the motivation for the paper in general ought to permeate the entire paper. Each sentence should impel the narrative forward. People should *want* to read the next sentence. They should be grateful for the next section since they are so desperate to find out the answer to the important mystery you posed at the beginning. And they should be convinced that this next section is precisely the perfect section to be reading at exactly that moment in the paper.

Think about clarity rather than sophistication Again, I refer you to Becker (1986). Let us say that you write something like, "X can explain Y." What does this mean? Are you referring to a particular kind of statistical model? If so, are you talking descriptively? Predictively? Causally? What about X "explains" what about Y? What does it mean for some thing to "explain" something else? Is X like a professor trying to give a lecture about something complicated called Y? This is *jargon*. And worse, it is *unclear*.

What you probably mean is that X and Y are things called variables — by which mean columns of numbers — and that when X has higher values, Y also tends to have higher values. Another way to say this is that they co-vary. Or, using scientific language in a clear way, you can say that, the R^2 from a linear regression of Y on X is substantial (XX%). Perhaps you mean that if we could change the values of X for some unit of observation then we would almost surely see, as a result, a change in the values of Y for that unit. Notice that *scientific language* is not a problem. But when does scientific language becomes jargon? It is fine to use technical language and theoretical language, but it is not fine to use such language either incorrectly or obscurely.

Hannah Pitkin has said that she writes for college freshmen. This means that she aims to talk to smart and motivated people who do not know a lot of particular technical language or particular literature. I like this rule of thumb. If I can say "A bright and interested but inexperienced person can understand this sentence", then I know that my writing has a fighting chance at clarity.

Interpret, Interpret, Interpret and Explain, Explain, Explain A regression table or a plot doesn't talk about itself. That is your job. And your job is to *interpret* them — that is, *make them meaningful* to the reader and the field at large. And, you might need to *explain* them a bit as you interpret.

This exhortation also holds for claims and arguments that you state. You should be citing appropriate literature, explicating and explaining and defending and justifying the things that you say. (Notice that if I write "means what?" or "So?" next to some text then it is both unmotivated as well as unexplained, undefended, unjustified, and thus uninterpreted.)

Do not focus on "statistical significance". Focus on the Real Story (1) When it comes to OLS, the assumptions required to justify hypothesis testing (and thus the production of p-values and confidence intervals) are much more stringent than those required to justify the estimation of coefficients. Thus, focusing on statistical significance rests your argument on the weakest parts of your statistical model. (2) In addition, statistical significance is not the real story. The real story is that if people switch from saying X=1 to X=2 then, on average, they also tend to say Y=2 rather than Y=1 — and this difference is the difference between elections won and lost, between people fed and starving, between views of the world, etc... You should never *never* present a regression table without talking about the substantive implications of the predictions and coefficients of the model. Never. NEVER. I would much much prefer that you not mention the results of hypothesis tests rather than you not mention the substantively important predictions, effects, differences implied by your model. (3) Talking about statistical significance allows you to avoid talking about your research design at precisely the moment when your design ought to be helping you make your case. Talking about

the predicted changes and effects and differences that come out of your model as they relate to your research design allows you to use multiple strong arguments together — rather than just one weak argument. (4) In general, focusing on statistical significance in current Political Science is a sign of lack of substantive knowledge and lack of theoretical engagement let alone lack of methodological depth.[5]

Excessive passive voice allows sloppiness and cowardice Again, read Becker (1986) on this. By using the passive voice you allow someone other than you to act: you abdicate responsibility for your words without attribution. Of course, sometimes passive voice is fine. The problem occurs with excessive or ill-placed passivity (For example, when the crucial sentences justifying your decisions suddenly become passive, then the reader knows that you don't feel very confident about your assertions).

Logical Paper Structure should be Evocatively Denoted This means that section headings should describe and suggest. "Literature Review" does not impart forward motion to your story. "A common misunderstanding about war and democracy, and how it can be clarified" is much more exciting and foreshadows your argument.

Tables and figures should stand alone A lazy reader should be able to look at the table or figure, and from the notes, caption, and label be able to make sense of it. You shouldn't tend to force readers to comb through the text to understand a figure or table.

All specifications of all models should be written out in equation format There is just too much possibility for slippage between a table and what is actually produced at the end of some computational algorithm. And talking at length in a paragraph can also be misinterpreted. If you are using a linear model (be it logistic, Poisson, or OLS), a simple sketch of what goes into your linear specification ($\mathbf{X}\boldsymbol{\beta}$) really clarifies exactly what you are estimating — especially since not all coefficients from all models must be presented in all tables.

Use full and informative labels in all figures, tables and in text Acronyms like EducCat3 or CONFDUR obfuscate. Those kinds of labels work fine for memos between collaborators but not for papers for general reading and publication. The general rule is, as I see it, to not force your readers to do extra work. Be grateful that anyone is reading your paper at all. Make life easy and exciting and pleasant for them — or else they will stop reading, or if forced to continue reading (because they are reviewing your article for publication, tenure, or funding) they will skim, or worse, get annoyed.

Assumptions of the statistical model should be directly engaged I mean this for every paper — not just papers for this class. Great writers and scholars can do this in one paragraph and a couple of footnotes. Every statistical procedure requires a set of logical, mathematical, and algorithmic commitments in addition to commitments about how the data at hand arrived into your hands (the research design): you should be up-front about such commitments and what they mean for the substantive interpretation of your results and for the status of your contribution to our shared understanding about the world.

Formatting (1) Papers should have abstracts. (2) Fonts should not be less than 11 pt font, with no less than 1 inch margins (I like 1.5 inch margins since I write in the margins rather than in between lines). (3) Tables within papers should have as few vertical lines as possible and never use double rules (some might say, never use vertical lines) of Chicago (2003) [§13.51–§13.53]. (4) If at all possible, ensure that you don't have annoying justification or footnoting problems in your manuscripts (i.e. single lines with excessive space between words because the justification algorithm has failed; footnotes ill-placed; etc. . .)

Consider your Ink to Information Ratio See Tufte (1983) for more on this. Big areas of your figures ought not consist of color or other ink that does not carry information. For example, the default Stata graphics with a big colorful border has a very low ink to information ratio: What does that border mean? If it is not meaningful, why is it there? Why present it to me? Default Excel graphics with the body of the graph in gray are even worse on this metric since the ink is all over the plotting area itself. And you already know about 3D bar charts: Ugh!

Figures and Tables and Notes should be in the text itself I know that some journals like tables and figures at the end of the paper. Some even like endnotes. I hate this. It is terrible for the reader to have bounce back and forth through your document.. When you give stuff to me to read, please put the tables and figures in the text itself and avoid endnotes in favor of footnotes.

Decide on number of digits based on substance If your substantive problem allows for distinguishing between .001 and .002, then, by all means use 3 digits, if not, then use the number of significant digits that is substantively meaningful. I have never seen a political science problem where one could detect substantively important differences at 5 (let alone 8) decimal places. But, if you are really interpreting your modeling results and your summary statistics keeping the

[5] *Note:* For PS531 we are engaging directly with questions about statistical significance and the meaningfulness of such statements.

theory and substance of your problem in the forefront of your reader's attention, then, even 8 decimal places will make sense if they are in fact sensible. (If you are cursory and hasty and technical in your interpretations, of course, you will just annoy and distract your reader with your misplaced precision.)

Understand Dashs and Hyphens A hyphen (-) combines two words into one word or breaks words at end of lines: short- or long-term; pick-me-up. The em dash (—) and the en dash (–) are forms of punctuation. Use the en dash to indicate ranges like 1990–2000 or relationships like Student–teacher. And use the em dash to indicate a parenthetical thought — as suggested in of Chicago (2003)[§ 6.80–§ 6.96].

> The hyphen, the en dash, and the em dash are the most commonly used [of the hyphens and dashes] and must be typeset correctly; an en dash appearing where a hyphen is called for bespeaks editorial or typographic confusion. of Chicago (2003) [§ 6.80].

"Impact" does not produce much impact This from *New Oxford American Dictionary* on my computer:

> USAGE The phrasal verb **impact on**, as in: *when produce is lost, it always impacts on the bottom line*, has been in the language since the 1960s. Many people disapprove of it despite its relative frequency, saying that **make an impact on** or other equivalent wordings should be used instead. This may be partly because, in general, new formations of verbs from nouns (as in the case of impact) are regarded as somehow inferior. As a verb, impact remains rather vague and rarely carries the noun's original sense of forceful collision. Careful writers are advised to use more exact verbs that will leave their readers in no doubt about the intended meaning. In addition, since the use of impact is associated with business and commercial writing, it has a peripheral status of 'jargon,' which makes it doubly disliked.

References

Achen, C. H. (1982). *Interpreting and Using Regression*. Sage, Newbury Park, CA.

Achen, C. H. (1986). *The Statistical Analysis of Quasi-Experiments*. University of California Press, Berkeley, CA.

Achen, C. H. (2002). Toward A New Political Methodology: Microfoundations and Art. *Annual Review of Political Science*, 5:423–450.

Achen, C. H. (2004). Let's put garbage—can regressions and garbage—can probits where they belong. Prepared for presentation at the annual meeting of the Peace Science Society, Rice University, Houston, Texas, November 12-14, 2004.

Adcock, R. and Collier, D. (2001). Measurement validity: A shared standard for qualitative and measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, 95(3):529–546.

Becker, H. S. (1986). *Writing for Social Scientists : How to Start and Finish Your Thesis, Book, or Article (Chicago Guides to Writing, Editin*. University Of Chicago Press.

Berk, R. (2004). *Regression Analysis: A Constructive Critique*. Sage.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

of Chicago, T. U. (2003). *The Chicago Manual of Style*. The University of Chicago press, 15th edition.

Tufte, E. (1983). *The Visual Display of Quantative Information*. Graphics Press, Cheshire, CT.