, Research and Analysis II

**Assignment #3: Control variables with regression[[1]](#footnote-1)**

**Introduction**

Commentators, particularly conservatives, often worry about non-marital childbearing. In this assignment, you will try to learn whether **non-marital child-bearing has a causal effect on the birth weight of the baby—and if so, how large the effect is**.

You will use a dataset on births in NYC in 2010, described further below. Your assignment should take the form of a brief essay understandable to a policy analyst who understands regression but is focused on real world policy issues.

Although you will need to use multiple regression and interpret statistics: they are not the main focus of the assignment; **the idea of control variables and what you learn with different controls is the focus**. Remember that the goal is to learn if there is a causal effect of non-marital child-bearing on birthweight (and the magnitude of effect), but you want to be completely honest about how well the results at each state do or do not provide such evidence.

**Specific topics to be explored and covered in essay (Note: The essay does not necessarily have to be in this order. It might make more sense to start with an introduction, then theory (path diagrams), then the analysis (naïve effect, and then controls), and conclusion.**

**1. Naïve “effects”**

First examine the apparent effect of non-marital childbearing on birthweight using regression. This will be the naïve analysis: the effects cannot be considered causal at this point.

Put your results in the first column of your regression results table. (See below for details on results table.)

Interpret your results. Be sure to address each of these statistical points:

* What is the adjusted R2 (or just R2) from the regression and what does it tell you?
* What (in prose) does the regression coefficient tell you?
* What are the units of the regression coefficient?
* Is the sign of the coefficient (+ or -) what you expected?
* Is the coefficient statistically significant?
* Is it practically significant?

But remember that the goal is to learn if there is a causal effect of non-marital child-bearing on birthweight. How much have you learned about the causal effect?

**2. Theory and possible omitted variables bias**

Develop some theories of how birth weight and non-marital childbearing could be related. First, consider theories of how non-marital childbearing could causally drive low birth weight: what would plausible mechanisms be?

Second, consider alternative theories that could explain the correlation but are inconsistent with non-marital childbearing causing low birth weight, such as common causes of both birth weight and non-marital childbearing. Create path diagrams for these theories (There should be presented as separate theories. But in the real world, both could be true.) In addition, think of other variables likely correlated with your independent variable and dependent variable. Why might they be related? Remember that you want control variables that are related to both independent and dependent variables **but are *not* intervening variables** (not part of the mechanism)—complex or unknown common causes.

Predict what bias omitting those common causes would create and why (Remember to try to predict the direction of bias, if possible.)

Using the dataset, try to find the common cause variables or proxies for them in the dataset. You will likely not be able to get the common causes (including complex and unknown common causes) that you want, and so think of proxies for the control variables you want. Even with proxies, you may not be able to find the controls you want in the dataset. That is the reality of this kind of research. You may wish to find additional alternative theories, as well as accepting the poor proxies available.

**3. Better causal effect estimates with control variables**

Perform a regression or regressions with the same dependent variable and independent variable of interest but now using the control variables you selected (You may need to create some new dummy variables or transform some variables.)

You should include at least two new specifications (i.e., models, separate multiple regressions), but no more than four new specifications at the absolute most. One of these should be your best specification, with all the relevant controls and no others. The results of each specification should go in separate columns of the results table. This table is similar to Table 13.1 in the book.

Interpret the results of each regression as follows:

* Interpret the new main coefficient of non-marital childbearing.
  + Interpret it in prose, including its units, similarly to how you learned in statistics.
  + Is the sign of the coefficient (+ or -) what you expected?
  + Is the implied association with birth weight statistically significant?
  + Is it practically significant?
* Most important:
  + What does the coefficient say about the causal effect of non-marital childbearing?
  + Try to interpret the coefficient of non-marital child-bearing as a causal effect. On the one hand, don’t forget the goal is a causal effect estimate. On the other hand, don’t over-claim. Using your theories, is there remaining bias in the coefficient of interest as a causal effect estimate?
  + What happened to the coefficient of interest compared to the naïve regression and other specifications with control variables?
    - What does this say about bias in interpreting the naïve regression coefficient as a causal estimate?
* Interpret the coefficients of each control variable, in at least one specification. (prose, units, sign, practical significance, statistical significance, reasonableness)
* What is the adjusted R2 (or just R2) from at least one of the additional regressions and what does it tell you?

Again: In addition to the simple (naïve) regression in part 1, you should have at least two regressions (two specifications), one of which should be your final preferred regression. However, you should have, at the *absolute most*, four additional regressions, for a total of five.

**4. Conclusions**

Look over all the results. Assume that your main interest is the effect of marriage on birthweight. You may have some interest in the effects of other interventions but it’s not the main focus. And you are not interested at all, in this essay, on other outcomes. What are the policy implications of your results? Make sure that your conclusions stem from your results. Try to be clear, organized and specific. Feel free to comment on any lessons you think you have learned from these analyses. Conclusions about what is not known are also useful. One paragraph should be sufficient, but you may somewhat say more if you wish.

**Data to be used**

The STATA dataset NYCbirths01\_revised.dtav contains data on singleton (no twins, triplets, etc.) live first births in 2010 in NYC to women who are NYC residents. The data set is a 3% random sample. The source of these data is the National Natality Files compiled by the National Center for Health Statistics (from birth certificates); the extract and some recoding was performed by Professor Korenman for PAF 9172.

Because some mother’s education has some missing values, a new variable **meducrec** has been created in which missing mother’s education values have been replaced with 12 years of schooling. This is *not* good practice, but it does not affect these results and we can’t afford to a smaller sample size. *These data are approved for use in homework exercises only.*

More information is in the “mini codebook” (NYCbirths01.doc). This file and the data file (nycbirths01\_revised.dta) are in the assignment folder under assignment #3 on blackboard.

**Product to produce**

**Your assignment should take the form of a brief essay understandable to a policy analyst who understands regression but is focused on real world policy issues.**

In order to assess the magnitude of differences in birth weight, you must have some point of comparison. The federally-funded WIC (Women, Infants & Children) Food Program has an effect, on average, of raising birth weight by about 100 grams.

***Use birth weight in grams for this assignment****. You may convert to pounds once or twice in addition, if desired, but otherwise, stick to grams.*

Your essay should **include a *single* table that shows the results of *all* of your regressions**. It will look like the table below in which each column corresponds to a different model (specification). If a particular variable is not used in a model, the box corresponding to the row for that variable remains blank. See Table 13.1 on page 408, which is a good illustration, except that it only has stars for p-values and no SE or CI in parentheses.

You may add more than one control per model, just remember that Model 5, should be your completed, full model, and include all controls needed. You can have up to as many controls as you think necessary, again, making sure they are **not intervening variables.**

**If you control for a dummy variable, for example race, you must leave one category out, this would be your “reference category.”** The reference category is usually either the largest or the group that would make sense to have others compared to for other reasons, for example: if you have the following dummy variables for race: white, black, Hispanic, “other,” it makes sense to have white as the reference category. You would include in your regression black, Hispanic and “other,” and leave white out. Now, the results you get for black is in reference to white, the results you get to Hispanic is also in reference to white, and the results you get for “other” is also in reference to whites.   
So, let’s pretend the coefficient for black is -1020, you would interpret this as: “Babies born to black women weight on average 1020 grams *less* than babies born to **white women (reference category)**.”

Keep in mind that you must include all dummies for a variable EXCEPT for the reference category.

Table 1: (add descriptive title here, including what the dependent variable that you are describing)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Mother’s characteristics | Coefficient Estimates  (standard errors, confidence intervals &/or p-values in parentheses) | | | | |
| Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
| Married |  |  |  |  |  |
| Control variable 1 |  |  |  |  |  |
| Control variable 2 |  |  |  |  |  |
| Control variable 3 |  |  |  |  |  |
| … |  |  |  |  |  |
| … |  |  |  |  |  |
| … |  |  |  |  |  |
| Constant |  |  |  |  |  |
| R-sq.  (or Adj R-sq) |  |  |  |  |  |
| Sample Size |  |  |  |  |  |

Notes:

1) You must create your own versions of tables from the STATA output. Do not turn in raw STATA output as part of your assignment. (If you want me to look at your STATA output because you are confused or because you are interested in learning more, then include it as an appendix. But it is just for me to help you understand; it is not part of the assignment.)

2) This will take you quite a bit of time. Make sure you start well in advance.

3) You should *not* have all three of: standard errors, confidence intervals or p-values. At most, have SEs or CIs, in addition to p-values. A better approach might be to include SEs (or CIs) only, with stars to indicate the level of statistical significance, if any. The note above in the table, below coefficient estimates, should be edited to indicate which inferential statistics you have chosen to include.

4) **Remember that this is an *essay* for someone interested in policy.** Don’t just list statistical definitions. The point is to make the statistical analysis relevant and useful. Remember your audience. You do not need to interpret the p-value, for example. Your reader knows how to interpret p-values.

Please include the rubric with your assignment…

Rubric for Control Variables Assignment:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Component** | **Exceeds Expectation (20)** | **Meets Expectation (18)** | **Falls below Expectation (16)** | **Expectation not met (14)** |
| Execution of regression  (20 points) | Dependent, independent and control variables entered into software correctly  Regression performed correctly  Justification and choice of controls make sense  Regression results correctly extracted into tables | Dependent, independent and control variables entered into software correctly with minor errors possible  Regression performed correctly  Choice of controls make sense  Regression results mostly correctly extracted into tables | Dependent, independent and control variables not entered into software correctly in a significant way  Regression not performed correctly  Choice of controls does not make sense  Regression results not correctly extracted into tables | Dependent, independent and control variables not entered into software correctly at all  Regression not performed correctly  Choice of controls does not make sense and is not justified  Regression results not at all correctly extracted into tables |
|  | **10 points** | **8 points** | **6 points** | **4 points** |
| Interpreting main coefficient of interest in simple regression **(with no controls, model 1)** (10 points) | Main coefficient of interest correctly, clearly and fully interpreted  Causal issues fully and clearly addressed  Practical and statistical significance correctly and clearly discussed | Main coefficient of interest interpreted generally correctly  Causal issues addressed with some problems  Practical and/or statistical significance discussed with some flaws | Main coefficient of interest interpreted with significant flaws  Causal issues not addressed at all or mostly incorrect  Practical and/or statistical significance mostly incorrect or missing | Main coefficient of interest not interpreted at all or interpreted completely incorrectly  Causal issues completely wrong or absent  Practical and/or statistical significance not discussed at all |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Component** | **Exceeds Expectation (20)** | **Meets Expectation (18)** | **Falls below Expectation (16)** | **Expectation not met (14)** |
| Theory and omitted variable bias  (20 points) | Sensible theories, fully explained, with valid path diagram for both causal effect of marriage and alternative theories  Good theories used for selection of controls, noting whether they are proxies  Clear and sensible prediction of direction of bias from omitting common causes | Mostly reasonable theories whose explanation and path diagrams have some flaws,  for both causal effect of marriage and/or alternative theories  Reasonable theories used for selection of controls  Some prediction of direction of bias from omitting common causes, with some explanation | Poor or no theories or poor or no explanations and path diagrams  for both causal effect of marriage and alternative theories  Poor or no theories used for selection of controls  No prediction of direction of bias from omitting common causes | No theory for causal effect of marriage or alternative theories  No theory used to support selection of controls  No prediction of direction of bias from omitting common causes |
|  | **10 points** | **8 points** | **6 points** | **4 points** |
| Interpreting main coefficient of interest in regressions with controls **(Models 2 to 5)** (10 points) | Main coefficient of interest clearly and fully interpreted  Causal issues fully and clearly addressed  Practical and statistical significance correctly and clearly discussed | Main coefficient of interest interpreted generally correctly  Causal issues addressed with some problems  Practical and/or statistical significance discussed with some flaws | Main coefficient of interest interpreted with significant flaws  Causal issues not addressed at all or mostly incorrect  Practical and/or statistical significance mostly incorrect or missing | Main coefficient of interest not interpreted or interpreted completely incorrectly  Causal issues completely wrong or completely absent  Practical and/or statistical significance not discussed at all |
| **Component** | **Exceeds Expectation (10)** | **Meets Expectation (8)** | **Falls below Expectation (6)** | **Expectation not met (4)** |
| Comparing main coefficient of interest with and without controls **(model 1 vs model 5)**  (10 points) | Clear and correct contrast of main coefficient of interest with and without controls (all models)  Contrasts fully interpreted in light of theory  Focus on causal effects maintained but with correct caveats and nuance | Reasonable contrast of main coefficient of interest with and without controls (all models)  Some ties made to theory  Causal conclusions over-stated somewhat or causation issue neglected | No contrast of main coefficient of interest with and without controls (all models) or completely wrong interpretation  No ties made to theory  Causal conclusions vastly overstated or causation issue completely ignored | No discussion of comparison of models with and without controls  No discussion of theory  No discussion of causal issues |
| Interpreting coefficients of control variables and other regression statistics (10 points) | Coefficients of control variables are interpreted clearly and correctly, including notable issues of practical and statistical significance  R-squared of naïve regression and some control regressions interpreted correctly | Coefficients of control variables are interpreted mostly correctly  Practical and statistical significance discussed with some flaws  R-squared of naïve regression interpreted correctly | Coefficients of control variables not interpreted correctly  Practical and statistical significance not discussed or discussed incorrectly  Naïve R-squared not interpreted or incorrectly interpreted | Coefficients of control variables not interpreted  Practical and statistical significance not discussed  No R-squared interpretation |
| **Component** | **Exceeds Expectation (10)** | **Meets Expectation (8)** | **Falls below Expectation (6)** | **Expectation not met (4)** |
| Overall presentation (10 points) | Assignment takes the forms of an essay for stated reader  Tables are clear, correct, easy to read and contain needed notes  Conclusions are clear and relevant to policy | Assignment mostly takes the forms of an essay for stated reader  Tables are mostly clear, correct and easy to read  Some conclusions are drawn but not related to policy | Assignment does not take the form of an essay for stated reader  Tables are fairly incorrect or confusing  Conclusions are not drawn | Assignment takes a totally inappropriate form  Tables are incomplete and very confusing  No conclusions section |
| Writing quality (10 points) | Writing is very clear  Arguments are cogent and persuasive  Organization is quite sensible and clear  Language is correct  Concise | Writing is fairly clear  Arguments are fairly cogent and persuasive  Organization is mostly sensible and clear  Language is mostly correct  Some unnecessary repetition | Writing is somewhat unclear  Arguments are not cogent and persuasive  Somewhat poor organization  Language has a more than a few mistakes  A fair amount of unnecessary repetition | Writing is very unclear  Arguments are not at all cogent or persuasive  No organization  Many language mistakes  Much repetition |

1. Thanks to Professor Sanders Korenman for making this data extract. Thanks to Professors Deborah Balk and Sanders Korenman for developing this assignment. All alterations and the final form are my responsibility alone. [↑](#footnote-ref-1)