OLS Checklist: For estimating cross-sectional models

Here's a list of things to check when you are estimating cross -sectional OLS econometric models.[1]

… What did I miss?

1. Before you do anything: Look at your data! (scatterplots, summary stats, etc)
2. Check **Number of Obs**… the most overlooked stat on the page… Did you drop obs?
3. Goodness of Fit
   a. Check **R-squared**… How much variation in the dependent variable did you explain? … and don't get too excited if it's so high … or so low.
   b. Check **adjusted R-squared**… But as always don't get too excited if it's so high … or so low. This statistic is useful in picking and choosing between models.
   c. Check **MSE** and **Root MSE**… similar to R-squared, and basically the same as adjusted R-squared, Mean Squared Error (MSE) tells you how well the model predicts the values of the dependent variable (RMSE is sort of, but not exactly, an average prediction error)
   d. Remember that goodness of fit metrics tell you how well you have *explained* the variation in the dependent variable. They don't necessarily tell you how close your estimated coefficients are to the true parameter values. Of course, unless we are doing simulations, we typically never know how close we actually are to the true parameters… which is why we focus on goodness of fit metrics, which are observable.
4. Check the **F statistic** and **Prob > F**.[2] If the F probability isn't tiny you have a really bad model! … because you can't reject the Null Hypothesis that your explanatory variables collectively have no explanatory power.
5. Ignore the constant term line… unless you have some special reason to not do so.
6. Do your estimated coefficients have the *right* signs? Remember that these coefficients are estimating *partial* effects (holding everything else in the model constant). If Yes, then Joy!... and if No, then you've got work to do. (But see comment about *Favorite Coefficient* models below.)
7. Are your estimated coefficients statistically significant?
   a. At what significance level? Don't get too excited if the p-value is .0999… or too depressed if it's .1001.
   b. Remember that statistical significance just says you can reject the Null Hypothesis that the true parameter value is zero.

[1] As opposed to time series models.
[2] The list starts at 12 since the SLR list stopped at 11.

c. If some coefficients/variables are not statistically significant then you might rerun the model without those explanatory variables and see what you get.

d. **But**: Sometimes it makes sense to leave statistically insignificant variables in the model. For example: We generally believe that income affects demand. And so if you are estimating a demand function, you might leave income in the model even if the estimated coefficient is not statistically significant.

8. Never forget: Irrespective of whether or not you can reject that Null Hypothesis, your best (***BLUE***) estimate of the slope parameter value is the estimated coefficient.
9. If you want more statistical significance, get more data!
10. There are typically two main culprits driving statistically significant coefficients with wrong signs: multicollinearity and omitted variable bias.[3]
11. **Multicollinearity**: Recall that the estimated coefficients are picking up partial effects, holding everything else constant. If you have highly collinear data, then it may be difficult to estimate these partial effects (since the RHS variables are always moving together). In other words, you may run into trouble if you're trying to estimate an effect that just doesn't exist in the data.

    a. To see if multicollinearity is the issue, run Stata's VIF command to see how much collinearity you have. And then try dropping some of the highly collinear explanatory variables to see if results become more sensible.

    b. A different fix is to just grab more data if you can.

12. If you are working with a ***Favorite Coefficient*** model, then so long as the collinearity and omitted variable bias aren't impacting your favorite RHS variable, don't worry about it. In these circumstances, there's nothing wrong with a ***Kitchen Sink*** model, which is doing all it can to kill the favorited coefficient. And if your critics complain about wrong signs tell them that's just multicollinearity or omitted variable bias at work, and it's not impacting your favorite coefficient… and who cares about any coefficient other than your favorite coefficient anyway?
13. To deal with issues related to functional forms or omitted variables…. well, deal with them! Try out lots of different models and compare results. A standard variation is to run models with both levels and logs of levels, to allow for additive as well as multiplicative effects.
14. Worry about economic/impact significance as well as statistical significance.[4] Remember that if you want to directly compare R^2's across models, it helps to have the same dependent variable, and constant terms in the models. Otherwise, you'll need to make adjustments.

    a. And be careful: If you are dropping obs when estimating different models the R-squareds may not be directly comparable, even if you have the same dependent variable in the different models.

[3] Sometimes this may be a functional form issue as well.

[4] You can use the margins command in Stata to generate the elasticities associated with your SRF.

15. The elasticities (at the means) in an MLR model are somewhat like influence shares: Since $\bar{y} = \hat{\beta}_0 + \sum \hat{\beta}_j \bar{x}_j$, $\frac{1}{\bar{y}} \hat{\beta}_0 + \sum \hat{\beta}_j \frac{\bar{x}_j}{\bar{y}} = 1$. And so $\sum e_j = 1 - \frac{1}{\bar{y}} \hat{\beta}_0$, where the e_j's are the elasticities.[5]
16. Look at (perhaps slices of) the SRF (scatterplot) … I say slices because it's hard to do scatterplots of multi-dimensional SRFs. Remember, the ***eyeball test*** may the best way to determine economic significance![6]
17. If you have a bunch of explanatory variables and are trying to figure out where to start, you might look at the simple correlations between each of the explanatory variables and the dependent variable. Since $\hat{\rho}^2 = R^2$ for SLR models, the explanatory variable most highly correlated with the dependent variable will also give you the SLR model with the highest R^2. That's not where you should stop… but it will get you started.
18. Pairwise correlation tells you something about multicollinearity, but certainly not everything. If you are trying to assess the degree of multicollinearity amongst of set of potential explanatory variables, just run any regression with all of those variables on the RHS, and run Stata's VIF to get the collinearity stats.
19. ***Econometrics is story telling***. Don't just run one MLR model. Run lots of variations and buildups to get a feel for what drives results. Use Stata's **eststo** and **esttab** commands to simplify the comparison of the models. And use F tests to help you decide what linear restrictions to impose or drop.
20. In the end you'll have a favorite model… but no one should be impressed if those are the only results you present. ***Tell the whole story!***

[5] Of course, these aren't actually shares because they can be negative, and will only sum to 1 if the estimated constant coefficient is zero.

[6] *Heat maps* will enable you to have two dimensions for the domain of the SRF.