Econ 6200: Econometrics II
Prelim, April 11^{th}, 2024

This exam consists of ten questions, not of equal length or difficulty, grouped into three exercises. The questions are only partly cumulative. Each question is worth 10 points. Remember to always explain your answer.

Good luck!

1. Hint: This question is really about the algebra of linear regression. It can be solved without a calculator. Remember that R^2 and estimated regression coefficients can be written in terms of sample variances and covariances.

Background/Motivation (not really part of question): In rank-rank regression, covariate and outcome are expressed in terms of their rank in their respective marginal distribution. Rank-rank regression is popularly applied in analysis of intergenerational mobility, e.g. by regressing income percentiles of children (observed at adult age) on those of their parents.

Question: I recently collected observations on $n = 999$ children and their mothers. Each child was assigned an outcome $Y_i \in \{1, \ldots, 999\}$ based on household income, where $Y_i = 1$ corresponded to the poorest child and so on. Similarly, each child's mother was assigned a rank position $X_i \in \{1, \ldots, 999\}$ corresponding to the rank position of her household income. There was no overlap in parents and no tie in household income, so that both X_i and Y_i took each value from 1 to 999 exactly once.

I wanted to regress Y on X. However, I first accidentally regressed X on Y, then poured coffee over my laptop, and then the dog ate my data. The only thing I remember is that the estimated slope was exactly 3/5.

1.1 What value did the estimated intercept take?

1.2 What value did R^2 take?

1.3 What values did intercept, slope, and R^2 take in the *intended* (i.e., reverse) regression?

2 This question is about scalar random variables (Y, X, Z). Researchers observe i.i.d. realizations $(Y_i, X_i, Z_i)_{i=1}^n$ and are interested in the causal effect of X on Y. Assume homoskedasticity throughout.

2.1 Researcher 1 uses the following two-step procedure:

1. Regress X on Z.
2. Regress Y on the fitted values $\hat{X}$ of the first regression.

Name the estimator and state standard assumptions under which this estimator will be appropriate in some asymptotic sense.

2.2 Researcher 2 uses the following two-step procedure:

1. Regress X on Z.
2. Regress Y on the residuals $\hat{\eta} := X - \hat{X}$ of the first regression.

Name the estimator and state standard assumptions under which this estimator will be appropriate in some asymptotic sense.

2.3 Suppose all assumptions that you stated hold. What estimator (of the above or other) would you propose?

3 Consider the two-wave panel

$$\begin{aligned} Y_{i1} &= X_i'\beta + W_{i1}'\gamma + U_i + \epsilon_{i1} \quad (1) \\ Y_{i2} &= X_i'\beta + W_{i2}'\gamma + U_i + \epsilon_{i2}. \quad (2) \end{aligned}$$

Assume i.i.d. observations $(Y_{i1}, Y_{i2}, X_i, W_{i1}, W_{i2})_{i=1}^n$ are available; other random variables are unobserved. Suppose also that $(\epsilon_{i1}, \epsilon_{i2})$ is uncorrelated with any other variable.

3.1 Suppose furthermore that (W_{i1}, W_{i2}) and U_i are uncorrelated. Can γ be estimated by pooled OLS?

3.2 Can γ be estimated (under otherwise standard conditions) by OLS in a regression of $Y_{i2} - Y_{i1}$ on $W_{i2} - W_{i1}$?

3.3 Define the OLS estimator for the demeaned version of (1)-(2) and show that it is numerically the same estimator as the one from the previous question.

3.4 Does the observation from 3.3 generalize to panels with more than 2 waves? If yes, please provide an argument. If not, then which, if any, estimator is preferred under which circumstances?

Brief Answers

1.1 Recalling that, in OLS, the fitted line must pass through the sample averages, we have

$$\overline{X} = \hat{\beta}_0 + \hat{\beta}_1 \overline{Y} \implies \hat{\beta}_0 = \overline{X} - \hat{\beta}_1 \overline{Y} = 500 - 3/5 \cdot 500 = 200.$$

Unfortunately, this gets much messier if you don't spot, e.g. from symmetry considerations, that the sample averages are 500.

1.2 Here and later, it is important to spot that X and Y by construction have the same sample variance. It is not recommended to evaluate that variance!

$$R^2 = \frac{(\hat{\text{cov}}(Y,X))^2}{\hat{\text{var}}(Y)\hat{\text{var}}(X)} = \frac{(\hat{\text{cov}}(Y,X))^2}{(\hat{\text{var}}(Y))^2} = \hat{\beta}_1^2 = \frac{9}{25}.$$

1.3 For the same reasonm (i.e., sample variances are the same), all numbers are the exact same in the reverse regression.

2.1 That's the IV estimator, expressed as just-identified TSLS estimator.

2.2 That's the OLS coefficient on X in the regression $Y = \beta_0 + \beta_1 X + \beta_2 X + \varepsilon$, where the two-stage algorithmic description is Frisch-Waugh-Lovell.

2.3 If assumptions for both hold, we can of course use both. But what should we do?

Both of the above estimators are consistent, and their asymptotic variances correspond to the variatuion in X used. That is, if X and Z are very highly correlated, most variation in X is "explained by" Z and the IV estimator will have low variance, whereas the OLS estimator will have high variance. Indeed, in the limit as $X = Z$, the IV estimator becomes the simple OLS estimator with according variance, whereas the multivariate OLS estimator is not defined. As X and Z approach independence, the opposite comparison holds.

However, is either the best we can do? No; under the intersection of all of these assumptions, we could also do simple OLS of Y on X and that would have smaller variance than either.

And yet that's still not the best because (as, to my shame, a tester of this exam had to point out to me) orthogonality of Z and ε is an overidentifying assumption that can be used to estimate this by what's basically TSLS.

3.1 No! X may be correlated with U!

3.2 Yes, this is classic fixed effects.

(If you answered "yes" on the first question, this part should have been a red flag.)

3.3 There are at leats two ways to see that. First, and maybe easiest,

$$Y_{2i} - \overline{Y}_i = Y_{2i} - \tfrac{1}{2}(Y_{1i} + Y_{2i}) = \tfrac{1}{2}(Y_{2i} - Y_{1i}),$$

so the demaned variables equal the first-differenced ones up to scale (and up to sign-flipped replication as we demean Y_{1i}, but that will not affect the estimator and we'd drop one wave after demeaning anyway).

But you can also compare the demeaning and first-differencing matrices. The former equals

$$Q = I_2 - \mathbf{1}\underbrace{(\mathbf{1}'\mathbf{1})^{-1}}_{=1/2}\mathbf{1}' = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix},$$

whereas the first differencing matrix is

$$D = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

Again, these are the same up to scale.

3.4 No, that "demeaning equals first differencing" is strictly a curiosity for $T = 2$; note that they are treated as variants in the lecture.