Homework 8

36-467, Fall 2018

Due at 6 pm on Wednesday, 31 October 2018

AGENDA: Confronting your fear of the unknown and mysterious, by taming statistical inference for dependent data.

1. The file demoruns contains 30 different simulations of a stationary AR(1) model, with intercept 0, one run per column. The same parameters were used in all runs, which differ only in their random choices of innovations.

   (a) (5) Using the first 10 observations from each run, estimate the slope for each run. (You should get 30 slopes.) Report the mean and standard deviation of these slopes, and plot their over-all distribution. (I recommend using either `ar.ols` or `lm` to estimate the slope. Since the intercept is 0, you will find it helpful to enforce that in your estimation.)

   (b) (3) Repeat Problem 1a, using the first 100 observations from each run.

   (c) (1) Repeat Problem 1a, using the first 1,000 observations from each run.

   (d) (1) Problem 1a, using all 10,000 observations for each run.

   (e) (5) Make a plot showing the slope estimates against sample size. Use a logarithmic scale for the horizontal, sample-size axis. Connect points coming from the same simulation run by lines. (There should be 30 lines; don't label them.) Are the estimates converging? How can you tell?

   (f) (4) Make a plot showing the variance of the slope estimates against sample size. Use a logarithmic scale for both axes. Explain whether this is compatible with the idea that $\mathrm{Var}\left[\hat{b}_n\right] \propto 1/n$.

   (g) (3) What's your best guess at the slope used in the simulations?

2. *Estimation by minimizing* Suppose X is generated by a stationary AR(1) model with expectation 0,

$$X(t+1) = \beta X(t) + \epsilon(t+1)$$

with $\mathrm{Var}\left[\epsilon(t)\right] = \tau^2$, and $\epsilon(t)$ being statistically independent of $\epsilon(s)$ for all $s \neq t$, and statistically independent of $X(s)$ for all $s < t$. Since X is stationary, we know $-1 < \beta < 1$.

We can estimate the slope b by minimizing the mean-squared error,

$$M_n(b) = \frac{1}{n-1}\sum_{t=1}^{n-1}(X(t+1) - bX(t))^2$$

(a) (2) Show that $X(t+1) - bX(t) = (\beta - b)X(t) + \epsilon(t+1)$.

(b) (2) Show that $\mathbb{E}\left[X(t+1) - bX(t)\right] = 0$, for all b.

(c) (4) Show that $m(b) \equiv \mathbb{E}\left[(X(t+1) - bX(t))^2\right] = \tau^2\left(1 + \frac{(\beta-b)^2}{1-\beta^2}\right)$. *Hints:* Use $\mathbb{E}\left[Z^2\right] = (\mathbb{E}\left[Z\right])^2 + \mathrm{Var}\left[Z\right]$, and a result on $\mathrm{Var}\left[X(t)\right]$ for stationary AR(1) models.

(d) (2) Show that $\frac{dm}{db} = 0$ if and only if $b = \beta$.

(e) (2) Show that $\frac{d^2m}{db^2} = \frac{2\tau^2}{1-\beta^2}$.

(f) (3) Show that

$$\frac{dM_n}{db}(b) = \frac{-2}{n-1}\sum_{t=1}^{n-1}X(t)(X(t+1) - bX(t))$$

(g) (3) Show that if $b = \beta$, then

$$\frac{dM_n}{db}(\beta) = \frac{-2}{n-1}\sum_{t=1}^{n-1}X(t)\epsilon(t+1)$$

(h) (4) Show that $\mathrm{Var}\left[X(t)\epsilon(t+1)\right] = \frac{\tau^4}{1-\beta^2}$. *Hint:* Use independence.

(i) (4) Show that $\mathrm{Cov}\left[X(t)\epsilon(t+1), X(t+h)\epsilon(t+h+1)\right] = 0$ for all $h > 0$. *Hint:* Use the law of iterated expectations.

(j) (3) Show (using Lecture 15) that $\frac{dM_n}{db}(\beta) \to 0$.

(k) (3) Show (using Lecture 14 and the previous problem) that $\hat{b}_n \to \beta$.

(l) (4) Show that $\mathrm{Var}\left[\frac{dM_n}{db}(\beta)\right] = \frac{4}{(n-1)}\frac{\tau^4}{1-\beta^2}$.

(m) (4) In Lecture 14, we showed that, for estimators of this kind, for large n,

$$\mathrm{Var}\left[\hat{b}_n\right] \approx \left(\frac{d^2m}{db^2}\right)^{-2}\mathrm{Var}\left[\frac{dM_n}{db}(\beta)\right]$$

Use this, and the previous parts of this problem, to find an expression for $\mathrm{Var}\left[\hat{b}_n\right]$. Simplify until your expression involves only n, τ^2, β, and numerical constants (like 1 or -3). *Hint:* If you do it right, one of the parameters will disappear. (Don't look up which one.)

(n) (4) Apply your formula for Var $\left[\hat{b}_n\right]$ to the best estimate of the slope you got in problem 1. Do the standard errors it predicts at $n = 10, 100, 1000, 10000$ match the standard deviations you found across simulation runs? Explain whether or not they *should* match.

(o) (2) $\frac{d^2 m}{db^2}$ measures how sharply curved $m(b)$ is around its minimum. Why does the minimum become more sharply curved when there is *more* noise from the innovations?

3. *Somewhat more practical sandwich variances* The file remoruns contains 30 simulation runs of a model which is *not* AR(1), but is stationary with expectation 0.

   (a) (3) Fit an AR(1) model, without intercept, to each of the runs. (You should get 30 distinct slopes.) Report the mean and standard deviation of the estimated slopes, and plot their distribution.

   (b) (4) Using the notation from Problem 2, show that

   $$\frac{d^2 M_n}{db^2}(b) = \frac{2}{n-1}\sum_{t=1}^{n-1} X^2(t)$$

   Why does b not appear in this formula?

   (c) (3) Calculate $\frac{d^2 M_n}{db^2}$ for the first simulation run. What is it?

   (d) (4) Explain why the result of Problem 2f still holds, even though the data are not from an AR(1).

   (e) (4) For the first simulation run, estimate Var $\left[\frac{dM_n}{db}(\beta)\right]$ by $(n-1)^{-2}\sum_{t=1}^{n-1}(-2X(t)(X(t+1) - \hat{b}_n X(t)))^2$. Explain why this is a reasonable (though perhaps not ideal) estimate.

   (f) (4) Using your answers to Problems 3c, 3e, and Lecture 14, estimate Var $\left[\hat{b}_n\right]$. Does this match the variance you found across simulation runs? Explain whether or not they *should* match.

Rubric (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Questions which ask for a plot or table are answered with both the figure itself and the command (or commands) use to make the plot. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is properly integrated with a tool like R Markdown or knitr, and both the knitted file and the source file are submitted. The code is indented, commented, and uses meaningful names. All code is relevant, without dangling or useless commands. All parts of all problems are answered with coherent sentences, and raw computer code or output are only shown when explicitly asked for.