ASSIGNMENT 1

ETHAN LIGON

You should complete all exercises. You are strongly encouraged to work as a team, and to turn in a single assignment for grading. The principal deliverable you turn in should be a pdf.

1. Admin

Some steps to help us share code via `git`:

(1) Your team should "fork" the repository `https://github.com/ligonteaching/ARE212_Materials`, and each team member should have "write" access to the repo.
(2) When submitting code for assignments, please provide links to that code in the pdf.

2. Exercises

(1) From ARE210, recall (Section 9 in Mahajan's "Handout 1") the rule for computing the distribution of certain transformations of random variables (The "inverse Jacobian rule").

Let (x, y) be independently distributed continuous random variables possessing densities f_x and f_y. Let $z = x + y$. Use the rule to obtain an expression for the distribution of z.

(2) We've discussed ways to program a convolution of random variables in a Jupiter notebook [ipynb] [datahub]. As in the notebook, consider a discrete random variable s and a continuous random variable x. Prove that the convolution of s and x (or, informally, $x + s$) has a continuous distribution, as suggested by the figure at the end of the notebook, **or** establish that the figure is wrong or misleading.

(3) Let $\boldsymbol{A}$ be an $m \times n$ matrix. A matrix $\boldsymbol{A}^-$ is a *generalized inverse* of $\boldsymbol{A}$ if $\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}$. Such a generalized inverse can be shown to always exist. If $\boldsymbol{A}$ is a matrix of zeros, what can we say about $\boldsymbol{A}^-$?

(4) Econometricians spend a great deal of time writing down linear regressions relating an object "Why" to an object "Ex", but

---

Date: Due April 3, 2023.

sometimes use quite distinct notations to express this regression. Following our discussion in class, suggest a notation for each of the three following cases:
   (a) "Why" is a scalar random variable, while "Ex" is a vector random variable;
   (b) "Why" is a single *realization* of a scalar random variable, while "Ex" is similarly a single *realization*;
   (c) "Why" is a *vector* of N realizations, while "Ex" is similarly a *matrix* of realizations.

(5) Let $\boldsymbol{A}$ be an $m \times n$ matrix.

Moore-Penrose Inverse

A matrix $\boldsymbol{A}^+$ is a "Moore-Penrose" generalized inverse if:

- $\boldsymbol{A}\boldsymbol{A}^+\boldsymbol{A} = \boldsymbol{A}$;
- $\boldsymbol{A}^+\boldsymbol{A}\boldsymbol{A}^+ = \boldsymbol{A}^+$;
- $\boldsymbol{A}^+\boldsymbol{A}$ is symmetric; and
- $\boldsymbol{A}\boldsymbol{A}^+$ is symmetric.

Full rank factorization

Let $\boldsymbol{A}$ be an $n \times m$ matrix of rank r. If $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{R}$, where $\boldsymbol{L}$ is a $n \times r$ full column rank matrix, and $\boldsymbol{R}$ is a $r \times m$ full row rank matrix, then $\boldsymbol{L}\boldsymbol{R}$ is a *full rank factorization* of $\boldsymbol{A}$.

Fact

Provided only that $r > 0$, the Moore-Penrose inverse $\boldsymbol{A}^+ = \boldsymbol{R}^\top(\boldsymbol{L}^\top\boldsymbol{A}\boldsymbol{R}^\top)^{-1}\boldsymbol{L}^\top$ exists and is unique.

(1) If $\boldsymbol{A}$ is a matrix of zeros, what is $\boldsymbol{A}^+$?
(2) Show that if $\boldsymbol{X}$ has full column rank, then $\boldsymbol{X}^+ = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ (this is sometimes called the "left inverse), and $\boldsymbol{X}^+\boldsymbol{X} = \boldsymbol{I}$.
(3) Use the result of (2) to solve for $\boldsymbol{b}$ in the (matrix) form of the regression $\boldsymbol{y} = \boldsymbol{X}b + \boldsymbol{u}$ if $\boldsymbol{X}^\top\boldsymbol{u} = 0$.

3. CONVOLUTIONS

We've discussed ways to program a convolution of random variables in a Jupyter notebook [ipynb] [datahub].

1. As in the notebook, consider a discrete random variable s and a continuous random variable x. Prove that the convolution of s and x (or, informally, $x + s$) has a continuous distribution, as suggested by the figure at the end of the notebook, **or** establish that the figure is wrong or misleading.
2. The notebook develops a simple class `ConvolvedContinuousAndDiscrete` to allow for the creation and manipulations of (you guessed it) convolutions of a continuous rv with a discrete rv. Can you develop a similar class for convolutions of independent discrete random variables?
3. Same as (2), but convolutions of independent continuous random variables?

4. GENERAL WEIGHTED LINEAR REGRESSIONS

List the main regression estimators you encountered in the first half of the class. For each estimator, establish whether it belongs to the class of general weighted linear regressions

$$T'Y = T'X\beta + T'u. \tag{1}$$

For the estimator you've listed, if it is a general weighted linear regression then what is the form of T? Is T random? If the estimator is *not* in this class, show why.

5. SIMULTANEOUS EQUATIONS

When we defined the general weighted regression, we didn't assume anything about the **dimension** of the different objects except that they were 'conformable.'

So: consider

$$y = X\beta + u, \qquad \text{with } \mathrm{E}T'u = 0, \tag{2}$$

and where $y = [y_1, y_2, \dots, y_k]$, so that if you had a sample of N observations realizations of y would be an $N \times k$ matrix.

1. What does our assumption of conformability then imply about the dimensions of X, β, T, and u?
2. Could you use the estimator we developed in weighted_regression.ipynb to estimate this system of simultaneous equations?
3. Extend the code in weighted_regression.ipynb to actually estimate β in the case with $k = 3$.
4. What additional assumptions are necessary to estimate the distribution of the estimator of β?

6. SUR

Picking up from the discussion of simultaneous equations above, where y is $N \times k$, and

$$y = X\beta + u. \tag{3}$$

If X is $N \times \ell$ and $\mathrm{cov}(u|X) = \Omega$ then this is a generalization of the assumption of homoskedasticity to a multivariate setting; the resulting structure is called a system of "Seemingly Unrelated Regressions" (SUR).

1. If Ω isn't diagonal then there's a sense in which the different equations in the system are dependent, since observing a realization of, say, y_1 may change our prediction of y_2. (This is why the system is called "seemingly" unrelated.) Describe this dependence formally.
2. Adapt the code in weighted_regression.ipynb so that the data-generating process for u can accommodate a general covariance matrix such as Ω, and let $X = T$. Estimate β.
3. How are the estimates obtained from this SUR system different from what one would obtain if one estimated equation by equation using OLS?

7. Food expenditures in India

The NSS surveys in India pioneered (in considerable part due to Mahalanobis) a wide variety of methodological innovations in sampling, questionnaire design, and have been among the most ambitious regularly collected data on household behavior and characteristics until recently.

The most recently publicly released data on household expenditures was the "68th round", collected in 2011–12. (More recent data has been collected, but suppressed for political reasons.) Data on household-level total food (and a few other non-durable) expenditures from the 68th round is available here, in the file `total_expenditures.parquet`. (You can use the `pandas.read_parquet` method to read these files–you may need to install some additional dependencies such as `pyarrow`.).

1. Use these data to produce a figure describing the distribution of non-durable expenditures across households, measured in INR, using a Gaussian kernel and some bandwidth h. What are the strengths and weaknesses of your figure in terms of what it conveys about the underlying distribution? Can weaknesses be addressed by choosing a different bandwidth or kernel? (Nothing formal required here—I encourage you to simply play around.)

(2) Once you've arrived at some favorite kernel & bandwidth (say

$$\hat{f}^h(x)$$

) describing the density of expenditures, can you use the "inverse Jacobian" rule to describe instead the density of *log* expenditures? Write code to produce this figure.

(3) Instead of the route you've taken in (2), choose some kernel & bandwidth to estimate the density of log expenditures directly. How do the approaches of (2) and (3) compare?

8. "Plug-in" Kernel Bias Estimator

In our discussion of bias of the kernel density estimator in lecture we constructed an "Oracle" estimator, which can be implemented when we know the true density

$$f$$

that we're trying to estimate.

Of course, the Oracle estimator is only feasible when we don't need it. What about the idea of using the same expression for bias as in the Oracle case, but replacing f with our estimate $\hat{f}$? Would this tell us anything useful? If so, under what conditions? What pitfalls might one encounter?