STA 431s23 Assignment Five[1]

For the Quiz on Friday Feb. 17th, please bring a printout of your full R input and output for Question 7. The other problems are not to be handed in. They are practice for the Quiz.

---

1. In the following regression model, the explanatory variables X_1 and X_2 are random variables. The true model is

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i,$$

independently for $i = 1, \ldots, n$, where $\epsilon_i \sim N(0, \sigma^2)$.

The mean and covariance matrix of the explanatory variables are given by

$$E\begin{pmatrix} X_{i,1} \\ X_{i,2} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \text{and} \quad Var\begin{pmatrix} X_{i,1} \\ X_{i,2} \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{pmatrix}$$

Unfortunately $X_{i,2}$, which has an impact on Y_i and is correlated with $X_{i,1}$, is not part of the data set. Since $X_{i,2}$ is not observed, it is absorbed by the intercept and error term, as follows.

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i \\ &= (\beta_0 + \beta_2\mu_2) + \beta_1 X_{i,1} + (\beta_2 X_{i,2} - \beta_2\mu_2 + \epsilon_i) \\ &= \beta_0' + \beta_1 X_{i,1} + \epsilon_i'. \end{aligned}$$

The last line represents the 'true" model. The primes just denote a new β_0 and a new ϵ_i. It was necessary to add and subtract $\beta_2\mu_2$ in order to obtain $E(\epsilon_i') = 0$. And of course there could be more than one omitted variable. They would all get swallowed by the intercept and error term, the garbage bins of regression analysis.

   (a) Make a path diagram of the model with $X_{i,1}$ and $X_{i,2}$.

   (b) What is $Cov(X_{i,1}, \epsilon_i')$?

   (c) Make a path diagram of the model with the primes.

   (d) Calculate the variance-covariance matrix of $(X_{i,1}, Y_i)$ under the true model.

   (e) Suppose we want to estimate β_1. The usual least squares estimator is

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (X_{i,1} - \overline{X}_1)(Y_i - \overline{Y})}{\sum_{i=1}^n (X_{i,1} - \overline{X}_1)^2}.$$

   You may just use this formula; you don't have to derive it. Is $\widehat{\beta}_1$ a consistent estimator of β_1 (meaning for all points in the parameter space) if the true model holds? Answer Yes or No and show your work. Remember, X_2 is not available, so you are doing a regression with one explanatory variable. You may use the consistency of the sample variance and covariance without proof.

(f) Are there *any* points in the parameter space for which $\widehat{\beta}_1 \xrightarrow{p} \beta_1$ when the true model holds?

2. A useful way to write a fixed-x regression model is $y_i = \boldsymbol{\beta}^\top \mathbf{x}_i + \epsilon_i$, where $\mathbf{x}_i$ is a $p \times 1$ vector of constants. Of course usually the explanatory variables are best modeled as random variables. So, the model really should be $y_i = \boldsymbol{\beta}^\top \boldsymbol{\mathcal{X}}_i + \epsilon_i$, and the usual model is conditional on $\boldsymbol{\mathcal{X}}_{\boldsymbol{i}} = \mathbf{x}_i$.

   In what way does the usual conditional linear regression model imply that (random) explanatory variables have zero covariance with the error term? For notational convenience, assume $\boldsymbol{\mathcal{X}}_i$ as well as ϵ_i continuous. What is the conditional distribution of ϵ_i given $\boldsymbol{\mathcal{X}}_{\boldsymbol{i}} = \mathbf{x}_i$?

3. In a regression with one explanatory variable, show that $E(\epsilon_i | X_i = x_i) = 0$ for all x_i implies $Cov(X_i, \epsilon_i) = 0$, so that *a standard regression model without the normality assumption still implies zero covariance* (though not necessarily independence) *between the error term and the explanatory variables.* Hint: If you get stuck, the matrix version of this calculation is in the text. We are in Chapter Zero.

4. Independently for $i = 1, \ldots, n$, let $y_i = \beta x_i + \epsilon_i$, where $x_i \sim N(\mu_x, \sigma_x^2)$, and $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. Because of omitted variables, x_i and ϵ_i are not independent. $Cov(x_i, \epsilon_i) = c$.

   (a) The usual fixed-x estimator of β is $\widehat{\beta}_n = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$. Is $\widehat{\beta}_n$ a consistent estimator of β? Answer Yes or No and prove it.

   (b) Another estimator you have seen before is $\tilde{\beta}_n = \frac{\overline{y}_n}{\overline{x}_n}$. Suppose $\mu_x \neq 0$. Do we have $\tilde{\beta}_n \xrightarrow{p} \beta$? Answer Yes or No and show your work.

5. The following is the general instrumental variables regression model for observed variables. Independently for $i = 1, \ldots, n$,

$$\mathbf{y}_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{x}_i + \boldsymbol{\epsilon}_i, \text{ where}$$

   - $\mathbf{y}_i$ is an $q \times 1$ random vector of observable response variables, so the regression is multivariate; there are q response variables.
   - $\mathbf{x}_i$ is a $p \times 1$ observable random vector; there are p explanatory variables.
   - $E(\mathbf{x}_i) = \boldsymbol{\mu}_x$ and $cov(\mathbf{x}_i) = \boldsymbol{\Phi}_x$.
   - $\boldsymbol{\beta}_0$ is a $q \times 1$ vector of unknown constants.
   - $\boldsymbol{\beta}_1$ is a $q \times p$ matrix of unknown constants. These are the regression coefficients, with one row for each response variable and one column for each explanatory variable.
   - $\boldsymbol{\epsilon}_i$ is a $q \times 1$ unobservable random vector with expected value zero and unknown variance-covariance matrix $cov(\boldsymbol{\epsilon}_i) = \boldsymbol{\Psi}$.
   - $cov(\mathbf{x}_i, \boldsymbol{\epsilon}_i) = \boldsymbol{\mathcal{C}}$, a $p \times q$ matrix of covariances that arise from omitted variables.
   - There are at least p instrumental variables. Put the best p in the random vector $\mathbf{z}_i$.
   - $E(\mathbf{z}_i) = \boldsymbol{\mu}_z$ and $cov(\mathbf{z}_i) = \boldsymbol{\Phi}_z$.
   - $cov(\mathbf{x}_i, \mathbf{z}_i) = \boldsymbol{\mathcal{K}}$, $p \times p$ matrix of covariances. Assume $\boldsymbol{\mathcal{K}}$ has an inverse.

(a) Calculate the expected value and variance-covariances matrix of $\left(\begin{array}{c} \mathbf{x}_i \\ \hline \mathbf{y}_i \\ \hline \mathbf{z}_i \end{array}\right)$ in terms of the model parameters. Your answers are partitioned matrices.

(b) Writing $\boldsymbol{\Sigma} = \left(\begin{array}{c|c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{13} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} & \boldsymbol{\Sigma}_{23} \\ \hline \boldsymbol{\Sigma}_{31} & \boldsymbol{\Sigma}_{32} & \boldsymbol{\Sigma}_{33} \end{array}\right)$,

i. Solve for the crucial parameter matrix $\boldsymbol{\beta}_1$ in terms of $\boldsymbol{\Sigma}_{ij}$ matrices.

ii. Give a method of moments estimator for $\boldsymbol{\beta}_1$.

iii. Indicate how it is possible to solve for the other parameter matrices that appear in $\boldsymbol{\Sigma}$. You don't have to give the complete solutions in terms of $\boldsymbol{\Sigma}_{ij}$ matrices. For example, you have $\boldsymbol{\Phi}_x = \boldsymbol{\Sigma}_{11}$, and you also have a solution for $\boldsymbol{\beta}_1$. So, you can just write $\boldsymbol{\Ç} = \boldsymbol{\Sigma}_{12} - \boldsymbol{\Phi}_x \boldsymbol{\beta}_1^\top$.

6. Instrumental variables are a powerful solution to the problem of omitted variables, but they are not easy to find. One suggestion is that cigarette taxes could be an instrumental variable for testing the connection between smoking and lung cancer. This relationship that is hardly open to question, but still, it is contaminated by many omitted variables – except in experimental studies where animals are exposed to cigarette smoke in a controlled way. So consider a study in which the n cases are U.S. states (provinces), and the variables are

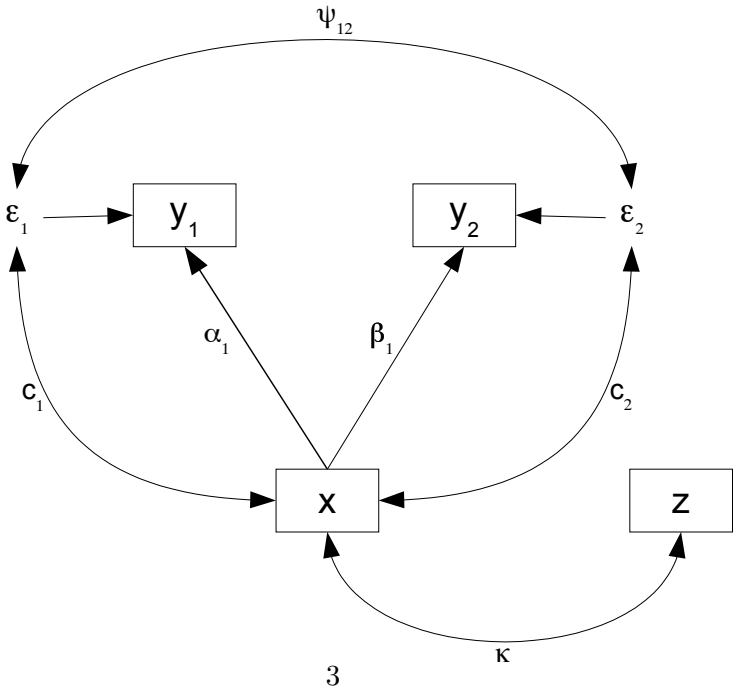
z: State tax on a pack of cigarettes (there's a federal tax too, but it's the same in all states).

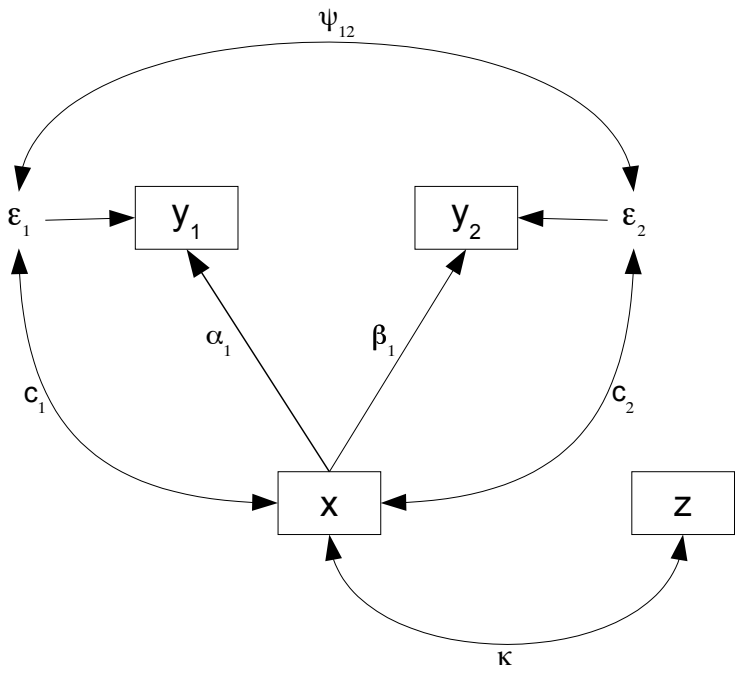
x: Smoking rate in percent.

y_1: Age-adjusted rate of new Lung and Bronchus cancers (per 100k population).

y_2: Age-adjusted rate of new Brain and other nervous system cancers (per 100k population).

The "age-adjusted" business is some kind of regression correction. I believe we are getting residuals plus a constant. Here is a picture of an instrumental variables model for these data.





This model does not stand up to close examination. It has lots of flaws, and listing them (with discussion) would be enlightening. However, for now let's just pretend to believe it, and proceed with the homework problem.

(a) Write down the model equations and the other details of the model, following the notation indicated in the path diagram. Use $Var(x) = \phi_x$ and $Var(z) = \phi_z$ The regression equations have intercepts.

(b) Referring to the general model in Question 5, give the following matrices in terms of the model you have just written down: $\mathbf{y}_i$, $\boldsymbol{\epsilon}_i$, $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$, $\boldsymbol{\Psi}$, $\boldsymbol{\zeta}$, $\boldsymbol{\kappa}$. Half marks off if you give the transpose.

(c) Calculate $cov \begin{pmatrix} x_i \\ y_{i,1} \\ y_{i,2} \\ z_i \end{pmatrix}$.

(d) Give method of moments estimates of α_1 and β_1. Compare them to your answer to 5(b)i.

(e) There are unique solutions for the other parameters as well. How do you know this without doing the calculations?

(f) For the model you described in Question 6a, what is the parameter vector $\boldsymbol{\theta}$? It should consist only of the unique parameters. How many parameters are there?

(g) How many moments (unique expected values, variances and covariances) are there?

(h) How do you know that for this model, the method of moments estimators are also the maximum likelihood estimators?

7. The data set described in Problem 6 is available at

   https://www.utstat.toronto.edu/brunner/openSEM/data/CancerTax2.data.txt

   (a) Fit your model from Problem 6 (meaning estimate the parameters) with `lavaan`. My standard error for $\widehat{\psi}_{12}$ was 1.048.

   (b) Using the `var` function with `na.rm=TRUE`, calculate your method of moments estimate of α_1 from Problem 6d.

      i. Does it agree with the MLE?

      ii. Are you surprised?

      iii. Why does it not matter that `var` uses $n-1$ in the denominator, while the maximum likelihood estimates use n?

   (c) The output of `summary` includes a test of $H_0 : \alpha_1 = 0$.

      i. Give the value of the test statistic. It is a number from your printout.

      ii. Give the p-value. It is a number from your printout.

      iii. In terms of the influence of smoking on cancer (which is the point of all this), what do you conclude from this test? If a conclusion is justified, draw a *directional* conclusion.

   (d) The output of `summary` includes a test of $H_0 : \beta_1 = 0$.

      i. Give the value of the test statistic. It is a number from your printout.

      ii. Give the p-value. It is a number from your printout.

      iii. In terms of the influence of smoking on cancer, what do you conclude from this test? If a conclusion is justified, draw a *directional* conclusion.

Please bring a printout of your full R input and output for Question 7 to the quiz.