Exam #2

Economics 435: Quantitative Methods

Fall 2011

Please answer the question I ask - no more and no less - and remember that the correct answer is often short and simple.

1 Short answers

a) Prove that if $E(u|x) = 0$ then $cov(x,u) = 0$.

b) Suppose $\bar{x}$ is the sample average from a random sample with mean μ and finite variance σ^2. Prove that $e^{\bar{x}} \to^p e^{\mu}$.

c) Suppose $\sqrt{n}(\hat{\theta} - \theta) \to^D N(0, \sigma^2)$. Then we can use Slutsky's theorem to prove a commonly-used result known as the *delta method*: for any function $g(.)$ that is differentiable at θ we have:

$$\sqrt{n}(g(\hat{\theta}) - g(\theta)) \to^D N(0, [g'(\theta)]^2\sigma^2)$$

where $g'(\theta)$ is just the derivative of $g(\theta)$. Take this result as given.

Suppose $\bar{x}$ is the sample average from a random sample with mean μ and finite variance σ^2. Find the asymptotic distribution of $\sqrt{n}(e^{\bar{x}} - e^{\mu})$.

d) Why is the Central Limit Theorem a useful result? What do we use it for?

2 The 2011 Census

Canada conducts a Census every 5 years. The Census aims to be a complete enumeration of every person living in Canada. Most households receive the "short form," a quick survey that asks just a few basic questions. In most recent Censuses, 20% of households would be randomly selected to receive a "long form" that includes much more detailed questions. Participation in the Census (both the short form and the long form) is mandatory, and failure or refusal to return one's Census form is punishable by fines and (at least in theory) jail. The participation rate for the long form Census is usually about 95%.

The Government of Canada dropped the mandatory long-form Census in 2011, and replaced it with a voluntary survey (known as the National Household Survey) that asks many of the same questions. The participation rate for the NHS has been estimated at 50%. The resulting decline in sample size due to nonparticipation has been partly remedied by sending the NHS to 1/3 of households. This question will be about the statistical consequences of this policy change.

We start by constructing a model. The population of interest[1] has size N. Each individual i in the population is characterized by the random variables of interest y_i and x_i. Individual i is surveyed ($s_i = 1$) or not ($s_i = 0$), and chooses whether to respond ($r_i = 1$) or not ($r_i = 0$). Let $S = E(s_i)$ and $R = E(r_i)$ be the sampling rate and response rate respectively. Throughout this question we will assume that sampling is random:

$$\Pr(s_i = 1|x_i, y_i) = \Pr(s_i = 1) = S$$

but we do not necessarily assume that response is random. We will treat S and R as known quantities rather than quantities that need to be estimated.

To introduce some notation, let $E(x_i) = \mu_x$, let $var(x_i) = \sigma_x^2$, let $E(y_i) = \mu_y$, let $var(y_i) = \sigma_y^2$, let $\bar{x}$ and $\bar{y}$ be the usual sample averages, and let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the coefficients from an OLS regression of y_i on x_i.

a) *For parts (a) and (b) of this question only*, assume response is random:

$$\Pr(r_i = 1|x_i, y_i) = \Pr(r_i = 1) = R \tag{1}$$

Find $var(\bar{y})$ in terms of (N, S, R, σ_y^2).

b) By how much (in percentage terms) does the move from the mandatory LF ($S = 0.2, R = 0.95$) to the voluntary NHS ($S = 0.333, R = 0.5$) raise the standard deviation of $\bar{y}$?

c) In practice, the sample size is not a crucial issue: most analyses using Census data have very low standard errors due to the very large sample size (20% of 30 million is 6 million observations). So for the remainder of this question, we will drop the assumption of random response (equation 1) and will focus on identification issues.

Suppose we are interested in estimating $E(y_i)$. Now, we know that:

$$E(y_i) = E(y_i|r_i = 1)\Pr(r_i = 1) + E(y_i|r_i = 0)\Pr(r_i = 0)$$

and that:

$$\bar{y} \rightarrow^P E(y_i|r_i = 1)$$

Without further assumptions, is $E(y_i)$ identified?

d) Now suppose that we can place upper and lower bounds on y_i (and thus on $E(y_i|r_i = 0)$):

$$a \leq y_i \leq b$$

Then we can identify upper and lower bounds on $E(y_i)$. State those upper and lower bounds in terms of $(a, b, R, E(y_i|r_i = 1))$

e) Using that result, state a consistent estimator of the upper and lower bounds in terms of $(a, b, R, \bar{y})$

[1] We might be interested in Canada as a whole, in which case N is about 30 million, or we might be interested in some smaller subpopulation such as first-generation immigrants living in Halifax.

f) Suppose that you find that 10% of respondents are poor. What bounds can you place on the percentage in the population who are poor, assuming that this data came from the long-form Census ($S = 0.2, R = 0.95$)?

g) What bounds can you place on the percentage in the population who are poor, assuming that this data came from the NHS ($S = 0.333, R = 0.5$)?

h) Now we will consider the implications for OLS regressions. To keep the model tractable, assume that both x_i and y_i are binary. The parameter we would like to estimate is:

$$\theta = E(y_i|x_i = 1) - E(y_i|x_i = 0)$$

We can show that:

$$\hat{\beta}_1 \rightarrow^p E(y_i|x_i = 1, r_i = 1) - E(y_i|x_i = 1, r_i = 1) = \beta_1$$

Find a lower bound and upper bound on θ in terms of (a, b, R, β_1)

i) Find a consistent estimator for the lower bound and upper bound on θ in terms of $(a, b, R, \hat{\beta}_1)$

j) Suppose that x_i was an indicator variable for being in a female-headed household, and y_i was an indicator variable for living in poverty. Suppose we find no relationship in the data between living in a female-headed household and poverty ($\hat{\beta}_1 = 0$). What bounds can we place on the relationship in the population (θ), assuming that the data came from the LF Census ($S = 0.2, R = 0.95$)?

k) What bounds can we place on the relationship in the population (θ), assuming that the data came from the NHS ($S = 0.333, R = 0.5$)?

3 Cluster data with fixed effects

The Government of British Columbia recently instituted[2] universal full-day kindergarten (FDK) in all B.C. public schools. That is, the kindergarten school day runs from 9:00 AM to 3:00 PM. Previously, schools had half-day kindergarten (HDK), i.e., some students attended from 9:00 to noon, and others attended from noon to 3:00. The introduction of FDK was staggered. In 2009, all schools had HDK. About half of schools began to operate FDK in 2010, and the rest began in 2011.

Suppose our data set consists of a random sample of B.C. kindergarten students each year from 2009 to 2011. Students are indexed by $i = 1, 2, \ldots, n$, the schools are indexed by $s = 1, 2, \ldots, S$, and time is indexed by $t = 2009, 2010, 2011$.

For each student we measure an outcome y_i, the school he or she attended in kindergarten $s(i)$ and the year he or she attended kindergarten $t(i)$. For each school s and year t we observe:

$$FDK_{st} = \begin{cases} 1 & \text{if school } s \text{ had FDK in year } t \\ 0 & \text{if school } s \text{ had HDK in year } t \end{cases}$$

[2] Just to be sure I'm not giving out false information, please be aware that I've simplified the actual policy for the purpose of this question. In particular, a few schools already had FDK in 2009, and the data set I've described doesn't quite exist.

The model we wish to estimate is:

$$y_i = a_{s(i)} + \beta_1 FDK_{s(i)t(i)} + \beta_2 x_i + u_i \tag{2}$$

where a_s is a fixed effect for school s, β_1 is the parameter of interest (the effect of FDK on the outcome) x_i is some vector of individual-specific control variables for person i, and u_i is an unobserved individual-specific factor that is assumed to obey a form of strict exogeneity:

$$E\left(u_i | a_{s(i)}, \left\{FDK_{s(i)t}\right\}_{t=2009}^{2011}, \left\{x_j\right\}_{j:s(j)=s(i)}\right) = 0 \tag{3}$$

Let n_{st} be the number of observations from school s in year t, let n_s be the total number of observations from school s, and let:

$$\begin{aligned}
\bar{y}_{st} &= \frac{1}{n_{st}} \sum_{i:s(i)=s,t(i)=t} y_i & \bar{y}_s &= \frac{1}{n_s} \sum_{i:s(i)=s} y_i \\
\bar{x}_{st} &= \frac{1}{n_{st}} \sum_{i:s(i)=s,t(i)=t} x_i & \bar{x}_s &= \frac{1}{n_s} \sum_{i:s(i)=s} x_i \\
\bar{u}_{st} &= \frac{1}{n_{st}} \sum_{i:s(i)=s,t(i)=t} u_i & \bar{u}_s &= \frac{1}{n_s} \sum_{i:s(i)=s} u_i \\
& & \overline{FDK}_s &= \frac{1}{n_s} \sum_{i:s(i)=s} FDK_{s(i),t(i)}
\end{aligned}$$

These various averages are potentially useful in working with this model. Finally, assume that x_i has all of the necessary variation across schools and time.

a) Show that:

$$(y_i - \bar{y}_{s(i)}) = \beta_1(FDK_{s(i)t(i)} - \overline{FDK}_{s(i)}) + \beta_2(x_i - \bar{x}_{s(i)}) + (u_i - \bar{u}_{s(i)}) \tag{4}$$

and that:

$$E\left(u_i - \bar{u}_{s(i)} | (FDK_{s(i)t(i)} - \overline{FDK}_{s(i)}), (x_i - \bar{x}_{s(i)})\right) = 0 \tag{5}$$

b) Show that:

$$\bar{y}_{st} - \bar{y}_s = \beta_1(FDK_{st} - \overline{FDK}_s) + \beta_2(\bar{x}_{st} - \bar{x}_s) + (\bar{u}_{st} - \bar{u}_s) \tag{6}$$

and that:

$$E\left(\bar{u}_{st} - \bar{u}_s | (FDK_{st} - \overline{FDK}_s), (\bar{x}_{st} - \bar{x}_s)\right) = 0 \tag{7}$$

c) Describe a method for consistently estimating β_1 using individual-level data (all data with subscript i, s, or st).

d) Describe a method for consistently estimating β_1 using school-level data (all data with subscript s or st).