

A.1 You wish to quantify the effect of cannabis consumption on student performance. You carry out a survey asking a random sample of your fellow students about their average mark after two years of studies and number of times they have consumed cannabis in the last 30 days. Let  $AM_i$  and  $SM_i$  be student  $i$ 's self-reported average mark and number of times used,  $i = 1, \dots, n$ , where  $n$  is the number of students in the sample.

- (a) Suppose that  $AM$  is observed with measurement error while  $SM$  is observed without. That is,  $AM_i = AM_i^* + v_i$ , where  $AM_i^*$  is the actual average mark and  $v_i$  is the measurement error. The measurement error is assumed to be fully independent of  $(SM_i, u_i)$  with  $E[v_i] = 0$ ,  $i = 1, \dots, n$ . Suppose that the actual average mark satisfies

$$AM_i^* = \beta_0 + \beta_1 SM_i + u_i, \quad (1)$$

and that SLR.1-SLR.5 are satisfied in the above model. Derive the (conditional on  $SM_1, \dots, SM_n$ ) mean and variance of the OLS estimator of  $\beta_1$  obtained by regressing  $AM$  on  $SM$ .

- (b) You use the following estimator of the variance of the OLS estimator  $\hat{\beta}_1$  as described in (a),

$$\widehat{\text{Var}}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{n\hat{\sigma}_{SM}^2}, \quad \hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2, \quad \hat{\sigma}_{SM}^2 = \frac{1}{n} \sum_{i=1}^n (SM_i - \overline{SM})^2,$$

where  $\hat{u}_i = AM_i - \hat{\beta}_0 - \hat{\beta}_1 SM_i$ ,  $i = 1, \dots, n$ . Is this a consistent estimator of the variance of  $\hat{\beta}_1$ ? Explain.

- (c) Consider the reverse situation: You observe the actual mark average  $AM^*$  but now instead of  $SM$  you observe  $\widetilde{SM}_i = SM_i + v_i$  where  $v_i$  still satisfies the assumptions stated in (a),  $i = 1, \dots, n$ . Derive the probability limit of the OLS estimator of  $\beta_1$  obtained by regressing  $AM^*$  on  $\widetilde{SM}_i$ .
- (d) You obtain a consistent estimator  $\hat{\sigma}_v^2$  of  $\sigma_v^2 = \text{Var}(v)$ . Use  $\hat{\sigma}_v^2$  to develop a consistent estimator of  $\beta_1$ .
- (e) Still considering the scenario in (c), discuss how realistic the following two assumptions are,  $E[v_i] = 0$  and  $v_i$  fully independent of  $(SM_i, u_i)$ , when the measurement error is due to incorrect reporting of cannabis consumption.
- (f) Suppose that you observe  $SM$  and  $AM$  without measurement error. However, some of the students that you asked to participate in the survey refused. Is this a concern regarding the validity of SLR.1-SLR.5?

A.2 You are interested in estimating the effect of per-student spending on math performance. For that purpose, you use a data set on 408 schools in the UK. For each school, the data set contains *math*, the percentage of students receiving a passing mark in a standardized math test, together with *spend*, per-student spending, and *enroll*, number of students enrolled.

- (a) You obtain the following regression results,

$$\widehat{math} = -69.24 + 11.13 \log(spend) + 0.22 \log(enroll), \quad R^2 = .0297.$$

(26.72)    (3.30)                      (.615)

If *spend* increases by 10% what is the (approximate) estimated percentage change in *math*?

- (b) Test the hypothesis that *math* does not change with *spend* against the alternative that it does increase with *spend*. Perform the test at a 5% and 1% level. Conclude.
- (c) You conjecture that family background has an effect on student performance and would like to include *poverty*, the percentage of students in a given school that live in poverty, in your regression. However, this variable is not in the data set and you instead decide to include *meal*, the percentage of students eligible for free school meals, as an additional regressor. Is this a sensible strategy? Explain.
- (d) Including *meal* you obtain the following results,

$$\widehat{math} = -23.14 + 7.75 \log(spend) - 1.26 \log(enroll) - .324meal, \quad R^2 = .1893.$$

(24.99)    (3.04)                      (.580)                      (.036)

Explain why the effect of spending on *math* is lower in this new regression compared to the one in (a).

- (e) Interpret the coefficients on  $\log(enroll)$  and *meal*.
- (f) What do you make of the increase in  $R^2$  from the regression in (a) to the regression in (d)?

B.1 Schumpeterian growth theory implies that the threat of technologically advanced entry spurs innovation incentives in sectors close to the technology frontier, where successful innovation allows incumbents to survive the threat, but discourages innovation in laggard sectors, where the threat reduces incumbents' expected rents from innovating. In “The Effects of Entry on Incumbent Innovation and Productivity,” (*The Review of Economics and Statistics*, Vol.91, No.1, 2009), Philippe Aghion, Richard Blundell, Rachel Griffith, Peter Howitt and Susanne Prantl study the effects of firm entry on labour productivity — more specifically, the real output per employee in the firm — and innovation — more specifically, the count of patents issued to the firm — taking into account how far the industry of interest is from the technological frontier. The authors use data from the United Kingdom and measure distance to the technological frontier by comparing the labour productivity in the industry in the United Kingdom to labour productivity in the same industry in the United States.

- (a) To study the relationship between entry, distance to the frontier and patent counts, the authors use a Poisson model. Suppose you decide to estimate a similar (i.e., Poisson model) where the expected number of patents is given by:

$$\mathbb{E}(P_j|D_j, E_j^F) = \exp(\beta_0 + \beta_1 E_j^F + \beta_2 D_j + \beta_3 D_j \times E_j^F),$$

where  $P_j$  is the count of patents for firm  $j$  in a given year,  $E_j^F$  measures the entry rate of foreign firms in firm  $j$ 's industry in the previous year and  $D_j$  measures the distance from the technological frontier. Both  $D_j$  and  $E_j^F$  are continuous. Write down the expression for the (log-)likelihood used to compute the Maximum Likelihood Estimator. In their estimates (which uses a somewhat more sophisticated version of the model above), the authors estimate  $\beta_2$  to be between 0.582 and 0.852 (depending on the specification used). Does this imply that the partial effect at the average (PEA) for distance to the technological frontier is positive? Please elaborate on your answer.

*Hint: If  $Y$  follows a Poisson distribution with parameter  $\lambda > 0$ , its probability mass function is*

$$\mathbb{P}(Y = k) = \frac{\lambda^k \exp(-\lambda)}{k!}$$

for  $k = 0, 1, 2, \dots$

- (b) The authors note that “entry can be endogenous to innovation and productivity growth” and consider a set of instrumental variables related to policy reforms related to entry:

“reforms at the European level and reforms at the U.K. level that changed the entry costs and effected entry differentially across industries and time.” The European reforms were undertaken as part of the Single Market Programme and deemed to reduce medium or high entry barriers. The U.K. reforms include, for instance, privatization cases which resulted in opening up markets to firm entry. Consider then the following simple linear regression model for labour productivity growth,  $\Delta LP_j$ , as it relates to entry,  $E_j^F$ :

$$\Delta LP_j = \alpha_0 + \alpha_1 E_j^F + U_j, \quad (2)$$

where  $U_j$  is an unobserved error. Suppose you have at your disposal one instrumental variable  $Z_j$  that consolidates information about the implementation of the reforms alluded to above. Describe how you would implement the TSLS estimator in this context. How would you argue for the validity of this instrument?

- (c) How can you use the estimates from (2) above to test whether  $E_j^F$  is endogenous?
- (d) Let  $\hat{E}_j^F = \hat{\pi}_0 + \hat{\pi}_1 Z_j$ , where  $\hat{\pi}_0$  and  $\hat{\pi}_1$  are OLS estimates from a regression of  $E_j^F$  on a constant and  $Z_j$ . If one uses  $\hat{E}_j^F$  as an instrumental variable instead of  $Z_j$  how would the estimates compare with those obtained in the previous item? Elaborate.  
*Hint: Since  $\hat{\pi}_0$  and  $\hat{\pi}_1$  are obtained by OLS,  $E_j^F = \hat{E}_j^F + V_j = \hat{\pi}_0 + \hat{\pi}_1 Z_j + V_j$  and  $\sum_{j=1}^n (\hat{E}_j^F - \overline{\hat{E}_j^F}) V_j = 0$ . Furthermore,  $\overline{E_j^F} = \overline{\hat{E}_j^F}$ .*

- (e) Imagine you have time series data for a *single* firm and estimate the following time-series regression by OLS:

$$\Delta LP_t = \alpha_0 + \alpha_1 E_t^F + \alpha_2 \Delta LP_{t-1} + U_t,$$

Would the estimator be unbiased? Under what conditions would it be consistent? Elaborate on your answers.

B.2 To study alcohol consumption in the UK, James Collis, Andrew Grayson and Surjinder Johal (“Econometric Analysis of Alcohol Consumption in the UK”, *HRMC Working Paper 10*, December 2010) use data from the Expenditure and Food Survey (2001-2006) to estimate the following model:

$$\begin{aligned} Y_j^* &= \mathbf{X}_j^\top \beta + \epsilon_j \\ Y_j &= \max\{Y_j^*, 0\} \end{aligned}$$

where  $Y_j$  is the proportion of total expenditure on a particular category of alcohol by household  $j$  and the explanatory variables  $\mathbf{X}_j$  include (log) prices for *all* alcohol categories, (log) income and other controls. The alcohol categories analyzed were beer, wine, spirits, cider and ready-to-drink (RTDs, also known as ‘alcopops’). Each category was also subdivided into on-trade (pubs and restaurants) and off-trade (supermarkets and off-licences).

- (a) Assume that  $\epsilon \sim \mathcal{N}(0, \sigma^2)$ . Provide the (log-)likelihood function for the above model.

*Hint: The cumulative distribution function for  $\epsilon$  is  $F_\epsilon(e) = \Phi(e/\sigma)$  and its probability density function  $f_\epsilon(e) = \phi(e/\sigma)/\sigma$  where  $\Phi(\cdot)$  and  $\phi(\cdot)$  are, respectively, the cumulative distribution function and the probability density function for the standard normal distribution.*

- (b) To assess the adequacy of the Tobit, the authors compare estimates of  $\beta/\sigma$  (where  $\sigma$  is the standard deviation of  $\epsilon_j$ ) to estimates of the coefficients from a Probit where the dependent variable is whether expenditure on alcohol (for particular categories) is zero or positive. Part of the table is reproduced below (for the purposes of the exam, it is irrelevant whether the table cells or numbers are shaded or not):

| regressors                 | (beer on)            |        | (wine on)            |        | (cider on)           |        | (spirit on)          |        | (RTD on)             |        |
|----------------------------|----------------------|--------|----------------------|--------|----------------------|--------|----------------------|--------|----------------------|--------|
|                            | tobit $\beta/\sigma$ | probit | tobit $\beta/\sigma$ | probit | tobit $\beta/\sigma$ | probit | tobit $\beta/\sigma$ | probit | tobit $\beta/\sigma$ | probit |
| $\ln P(\text{beer on})$    | -0.500               | -0.890 | 0.265                | 0.336  | -0.159               | -0.170 | 0.016                | 0.016  | 0.090                | 0.073  |
| $\ln P(\text{wine on})$    | -0.034               | -0.051 | -0.224               | -0.754 | -0.029               | -0.027 | 0.048                | 0.066  | 0.000                | 0.006  |
| $\ln P(\text{cider on})$   | 0.102                | 0.144  | 0.041                | 0.126  | -0.319               | -0.452 | 0.111                | 0.145  | 0.119                | 0.128  |
| $\ln P(\text{spirits on})$ | -0.085               | -0.186 | -0.020               | -0.023 | -0.116               | -0.135 | -0.540               | -0.742 | -0.269               | -0.317 |
| $\ln P(\text{RTDs on})$    | -0.034               | 0.024  | 0.122                | 0.129  | -0.087               | -0.101 | -0.032               | 0.019  | -0.358               | -0.604 |
| $\ln P(\text{beer off})$   | -0.119               | -0.119 | 0.102                | 0.180  | 0.029                | 0.043  | 0.032                | 0.034  | -0.060               | -0.062 |
| $\ln P(\text{wine off})$   | -0.017               | -0.002 | 0.061                | 0.066  | -0.043               | -0.066 | -0.048               | -0.048 | -0.179               | -0.201 |
| $\ln P(\text{cider off})$  | -0.025               | 0.006  | 0.143                | 0.199  | -0.130               | -0.126 | 0.016                | 0.028  | 0.030                | 0.061  |
| $\ln P(\text{spirit off})$ | 0.034                | 0.088  | 0.082                | 0.115  | -0.043               | -0.032 | -0.063               | -0.036 | -0.045               | -0.036 |
| $\ln P(\text{RTD off})$    | 0.008                | -0.024 | 0.041                | 0.025  | -0.014               | -0.016 | 0.079                | 0.063  | -0.015               | -0.056 |
| $\ln \text{income}$        | 0.203                | 0.466  | 0.449                | 0.611  | 0.246                | 0.310  | 0.254                | 0.364  | 0.269                | 0.342  |

Explain why this comparison might be useful.

- (c) The researchers are ultimately interested in the elasticities with respect to prices (own- and cross-) and to income. Since those variables are entered as logarithms, you decide to estimate those as:

$$\epsilon = \partial E(Y|\mathbf{X} = \mathbf{x}) / \partial x_k / y$$

where  $x_k$  is the relevant variable for the elasticity of interest (i.e., log of own price, log of substitute category or log of income). Explain how you would estimate the elasticity for a particular household. Suggest a measure of elasticity for the general population and explain how you would estimate it.

- (d) Suppose that instead of individual data, you have access to data on the market shares for off-trade beer (i.e., beer bought in supermarkets and off-licences) and prices for each of the alcohol categories in several local markets in the United Kingdom. Consider then the following model for the market share for off-trade beer:

$$\log S_m = \beta_0 + \beta_1 \log E_m + \beta_2 \log P_m + \epsilon_m \quad (3)$$

where  $S_m$  is the market share for off-trade beer in market  $m$ ,  $E_m$  is the expenditure on alcohol in market  $m$  and  $P_m$  is the price for off-trade beer in market  $m$ . (Assume that off-trade beer prices are uniform within a market.) Since the market share for off-trade beer depends not only on the variables above, but also on other variables not included in the model (e.g., prices for other alcohol categories), you decide to use a variable  $Z_m$  encoding the distribution costs of supermarkets or off-licences for beer (e.g., average distance to beer producers) as an instrumental variable for  $\log P_m$ . (Assume that  $\log E_m$  is uncorrelated with  $\epsilon_m$ .) Describe how you would implement the TSLS estimator in this context. How would you argue for the validity of this instrument?

- (e) Consider now equation (3) for a single market but across many periods  $t$  and suppose there are no endogeneity issues:

$$\log S_t = \beta_0 + \beta_1 \log E_t + \beta_2 \log P_t + \epsilon_t$$

Explain how you would test whether there is serial correlation in  $\epsilon_t$ . Would serial correlation imply that OLS is inconsistent?



