

ECON 623
Fall 2011
Pablo Salinas Macario and John Rust
Final Exam

Do Question 1 and 2 out of 3 of the remaining questions below.

1. [200 points] Probability question. A casino uses what it believes to be fairly balanced dice in gambling. The casino is considering introducing a new gambling game where a pair of dice are rolled 50 times. On each roll of the pair of dice, the customer (“roller”) gets 1 point if the sum of the numbers on the dice is even, or 0 points otherwise. After the 50 draws, if the roller has more 40 points or higher, they get a prize of \$100. You are asked to calculate an entry fee E so that each time a casino customer chooses this game, the casino expects to make a \$10 expected profit on the gamble.

- A **20 points** Can you calculate the entry fee exactly? If so, calculate the exact entry fee that yields the casino a \$10 expected profit each time this gamble is played.
- B **20 points** Suppose Joey Breakaleg, the casino manager, is not so good at probability theory but he vaguely recalls that there was something called the “cento limit serum” that had something to do with a “bar bell curve” from his days back at Wharton, when he was taking classes with Donald Trump. Can you explain to Joey Breakaleg that he was probably thinking of the *central limit theorem* and show how it can be used to approximate the probability that this prize would be paid and hence the entry fee? Compare the CLT approximate probability and implied entry fee to the exact probability and entry calculated in part A above. How much more or less than \$10 in profits per game can the casino expect to get using the approximated rather than true probability of paying out the prize? (Extra credit: plot the exact CDF of the number of points entered in 50 rolls, and on the same plot, show the approximate CDF based on the Central Limit Theorem approximation to this distribution: is the CLT approximation uniformly close to the actual CDF, or are there parts of the distribution where it is a poor approximation?)
- C **20 points** Suppose Joey is actually a really bad student of probability theory and cannot even recall how to apply the central limit theorem here, and he distrusts probability theory (because “only da really bad guys — dem Wall Street scamsters — do dat stuff”) but he does trust computer simulations so he hires you to write a Matlab program that can conduct a monte carlo experiment to estimate the probability of the payoff and hence the entry fee. But because you fear that Joey might break your leg if you get your monte carlo calculation wrong, and you realize there will be some Monte carlo sampling error in your estimate, you set the number of Monte carlo replications used to estimate this probability to a sufficiently large number so that the probability of an error in the estimated probability greater (in absolute value) of more than 0.00001 (i.e. 1 in 100,000) is less than 0.001 (i.e. one tenth of 1%). What is the smallest number of Monte Carlo replications N that you need to achieve this degree of accuracy?
- D **20 points** Joey wants to know how many people you expect to select this gamble and wants you to give him a ball park estimate. Suppose you believe that people coming into the casino are expected utility maximizers and have logarithmic utility functions. Suppose the alternative to this gambling game is to play a “1 armed bandit” at a price of 50 cents per pull. The one armed bandit pays out \$100 in quarters with probability 1 in 100,000 (i.e. probability .00001). Which would a person

with \$100,000 in net worth prefer: the dice game described above, or play the one armed bandit? What price does the dice game have to be adjusted to (up or down) in order for this customer to be indifferent?

E **50 points** You are assigned to test the fairness of dice using a “dice roller machine”. This is an automatic machine that mechanically rolls a single die 6000 times in a row. If it is fair, we expect 1000 of the draws to be 1’s, 1000 to be 2’s etc. Suppose the actual count for a particular test of a single die is 1: 991, 2: 975, 3: 997, 4: 1053, 5: 1005 and 6: 979. Can you devise a test of the hypothesis that this die is fairly balanced? If so, can you reject H_o : *this die is fair* at the 1% level of significance using your test and these data?

F **70 points** The file `dice-draws.txt` contains the actual results from the 6000 draws of the dice roller machine on the die tested above. You suspect that there might be a mechanical problem in the machine so that the rolls produced by this machine are not actually independently distributed draws, i.e. you suspect some sort of *dependence* in the results of successive draws from this machine. Can you think of a way of testing your suspicion using regression or some other means? Given whatever test you think up, test the hypothesis H_o : *draws from the machine are IID draws from the die*. Can you reject this hypothesis using your test at the 1% significance level? If you find dependence, how does this affect your conclusions in part E above?

2. **[200 points]** Suppose a consumer has an inventory of groceries at home and shopping trips occur when the consumer want to *adjust their desired inventory of groceries upward*. Let the desired *adjustment* to the level of inventories of groceries be given by the regression equation

$$y_{i,t} = X_{i,t}\beta + \varepsilon_{i,t} \quad (1)$$

where $y_{i,t}$ is household i ’s desired adjustment to their stock of groceries in week t and X contains explanatory variables affecting the desired adjustment, including household income, the household’s estimated current stock of groceries, number of people in the household, with breakouts for number of teenagers in the household, dummy variables for whether there are significant sales occurring during the week, and so forth.

A. **50 points** It is natural to assume that if $y_{i,t} < 0$, i.e. the household is “overstocked” with groceries and would actually like to reduce its stock of groceries, that the household might actually throw away of older stale or spoiled food or groceries, but when $y_{i,t} > 0$ then the household would have a motive to go shopping to replenish its stock of groceries. Assume that the survey we have does not record when household throws away groceries, so we do not observe cases where $y_{i,t} < 0$ and further, due to positive transaction and hassle costs of going shopping, household i will not actually go shopping unless $y_{i,t} > K_i$ where $K_i > 0$ is some threshold that must be passed to make it worthwhile for someone in the household to go out shopping to increase the stock of groceries. If the survey then only records the $(y_{i,t}, X_{i,t})$ observations for households that have actually gone out shopping (and thus filled out a diary recording the amount spent, $y_{i,t}$, as well as any other information that is changing at the weekly level, $X_{i,t}$ such as whether there was a significant sale that week, and what the inventories of groceries were at home that week, etc), will a regression limited to just the data on the $(y_{i,t}, X_{i,t})$ observations recorded in the consumer diaries and reported in the survey result in consistent estimates of β and the K_i parameters? Please explain your answer as carefully as possible to receive full credit.

- B. **50 points** If the length of time we observe a particular household i is relatively short, say for 4 weeks, but we have a large number of households i , say N households where $N > 1000$, do you think it will be possible to consistently estimate both β and the N household-specific thresholds K_i , $i = 1, \dots, N$? If you believe it is possible sketch an estimator and an argument for the consistency of your estimator. If you think it is not possible, describe as carefully as you can why you think it is impossible to consistently estimate these parameters by regression or any other means.
- C. **50 points** Suppose we are willing to make additional *distributional assumptions*. In particular if you are willing to assume that $\{\varepsilon_{it}\}$ are *i.i.d.* $N(0, \sigma^2)$ random variables and that the K_i are also $N(\mu, \gamma^2)$ random variables and that K_i and ε_{it} are independently distributed for each i , and that for any household i $\varepsilon_{i,t}$ and $\varepsilon_{i,s}$ are independently distributed for $s \neq t$, and finally, that the random variables $(K_i, \{\varepsilon_{i,t}\}_{t=1}^T)$ are distributed independently of the random variables $(K_j, \{\varepsilon_{j,t}\}_{t=1}^T)$ for any two households $i \neq j$, can you show how you can use these extra prior assumptions to construct a consistent estimator of the parameters $\theta = (\beta, \sigma^2, \mu, \gamma^2)$?
- D. **50 points** Suppose the household diary records $y_{i,t}$ on *every* week that the household is in the survey by asking a member of the household to directly report their *subjective assessment of their desired adjustment of groceries* $y_{i,t}$ in each week t . If you had these data, would ordinary linear regression (OLS) of the grocery inventory adjustment model (1) result in consistent estimates of β and the $\{K_i\}_{i=1}^N$ assuming that we have a relatively large number of households N in our sample but we assume these households only for 4 consecutive weeks? If you think OLS might not be able to consistently estimate all of these $4N + K + 1$ parameters (where K is the number of regression parameters β and the extra 1 is for the unknown $\sigma^2 = \text{var}(\varepsilon_{i,t})$), can you think of some other estimator that would be able to at least estimate the parameters (β, σ^2) , but assuming that you are NOT willing to impose a normality assumption on the $\{K_i\}$ or the $\{\varepsilon_{i,t}\}$? If you think it is possible, do you need to impose any assumptions on the independence of the error terms $\{\varepsilon_{i,t}\}$ across households i or across time, t ?

3. [200 points] Suppose that you have just landed a job at a top economic consulting firm and that you are having a disagreement with your boss about an econometric model. You think that the data are generated by

$$y = X\beta_o + u \tag{2}$$

where $\beta_o \in R^k$, X is $n \times k$, $y \in R^n$ and $\{u_i\}$ are *i.i.d.* random variables with mean 0 and variance σ_o^2 . On the other hand, your boss says that years of experience point her to the model

$$y = X\beta_o + Z\rho_o + u \tag{3}$$

where Z is also $n \times k$ since, after all, “it can’t hurt to add more variables to the model”. You are not sure about that so you set out to investigate her claim.

- A. **[65 Points]** Suppose that your model is the correct one but that you estimate the parameters by applying OLS to the competing model. Derive an expression for $\hat{\beta}$ and show that this estimator is unbiased, stating clearly any assumptions that you make. Then derive an expression for the variance of this estimator. Apply OLS to the correct model and call the estimator $\tilde{\beta}$. Repeat the previous steps.

- B. **[30 Points]** How can we compare our two estimators? Invoke a well-known theorem to make your argument and then show explicitly the difference in efficiency between $\hat{\beta}$ and $\tilde{\beta}$. Which estimator is more efficient? When will there be no loss of efficiency? Start with the one-parameter case and then extend your argument to the k parameter case.
- C. **[85 Points]** Now, on the contrary, suppose that it is your boss that has the correct model. Repeat the calculations you performed above. Suggest some criteria by which to evaluate the performance of your estimator and derive an expression to compare to the variance of the correctly specified model. Is one estimator unambiguously better than the other? Discuss.
- D. **[20 Points]** What do you conclude about your boss' claim? Do your conclusions depend in any way on the size of the sample to be analyzed?

4. **[200 points]** Let y be $n \times 1$, X be $n \times (k + 1)$ and suppose that $E(y|X) = X\beta$. Consider the linear programming problems

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \left\{ \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^k x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^k \beta_j^2 \right\} \quad (4)$$

and

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^k x_{ij}\beta_j)^2, \quad (5)$$

subject to

$$\sum_{j=1}^k \beta_j^2 \leq t \quad (6)$$

- A. **[20 Points]** Show that there is a one-to-one correspondence between the parameters λ and t above. Offer an interpretation of these parameters and compare the objective functions above to the one for OLS.
- B. **[15 Points]** Provide a convincing argument for why the intercept is not penalized in the problems above.
- C. **[60 Points]** Now suppose the data has been centered so that the data matrix X has k columns. Rewrite the first problem above in matrix form and show that the solution $\hat{\beta}$ is a linear function of y . Be careful to provide conditions that allow $\hat{\beta}$ to be well defined.
- D. **[10 Points]** Is the problem well defined if $X'X$ is *not* of full rank?
- E. **[60 Points]** Find $E\{\hat{\beta}|X\}$. Use the conditions you outlined in part C above to show that $\hat{\beta}$ is biased for β unless $\beta = 0$.
- F. **[15 Points]** Find the probability limit of $\hat{\beta}$ as $n \rightarrow \infty$. Is $\hat{\beta}$ consistent for β ? What happens to $\hat{\beta}$ as $\lambda \rightarrow 0$?
- G. **[20 Points]** Now let $\lambda = an$ where $a > 0$ is fixed and $n \rightarrow \infty$. Find the probability limit of $\hat{\beta}$ as $n \rightarrow \infty$.