Empirical Problem Set # 2
Graduate Industrial Organization, Fall 2005
Glenn Ellison and Stephen Ryan

1 Demand Estimation

The intent of this problem set is to get you familiar with Stata, if you are not already, and to demonstrate some real-world uses of the techniques that we talked about in class. After finishing this problem set, you should be familiar with how to load data sets into Stata, manipulate the variables in basic ways, and perform simple instrumental variables regressions. Stata has a lot of powerful tools, and it is in your own best interest to learn how to exploit them fully if you are going to do empirical work.

First, download the data sets from the class web page. The first data set is quantity, price, cost, and demographic variables on broiler chickens over 40 years in the United States, and is called "broiler.csv". The data is taken from Dennis Epple and Bennett McCallum's paper: "Simultaneous Equation Econometrics: The Missing Example." The data is in comma-delimited format, a common format that is supported by just about every data-processing application out there. It is completely portable across operating systems, and has the additional benefit of being human-readable. Inspecting the data, you will see the following column headers: year, q, y, pchick, pbeef, pcor, pf, cpi, qproda, pop, meatex, and time. The cryptic names are common in empirical work, even appearing in a dataset that is intended to be used for instructional purposes. To decode, "y" is per-capita real disposable income, "pchick" stands for "price of chicken", "pbeef" is "price of beef", "pcor" is "price of corn", "pf" is "price of chicken feed", "cpi" is "consumer price index", "qproda" is "aggregate production of chicken in pounds", "pop" is "population of the US", and "meatex" is "exports of beef, veal, and pork in pounds".

We are interested in using this data to estimate the demand curve for broiler chickens. The first step is to upload the data file to the Stata server that you are going to be using, and then start Stata.[1] Once Stata is up and running, you are ready to load your first data

[1] When I tried this using blackmarket.mit.edu, I was successful in uploading to the /tmp directory. Once you are connected, fire up Stata. You will have to figure out how to run Stata in graphical mode (xstata). This usually requires turning on X-Win and using an SSH terminal window that supports X tunnels. The benefits to having a graphical interface to Stata far outweigh the fixed costs of getting everything working

set. Go to "File → Import → ASCII Data Created By A Spreadsheet". Navigate to the directory you uploaded your data set to, and import it.[2]

Your first task once you have loaded the data is to generate some variable transformations. The model that we are interested is:

$$q = \alpha + \beta' p + \gamma' X + \epsilon, \tag{1}$$

which is the demand curve relating quantity demanded to prices and other covariates. What should enter X? The answer is everything in our data set that we think is going to shift demand for chickens, such as the price of substitutes (pbeef), population (pop), and income (y). In practice, economic theory is helpful in suggesting what kinds of variables might go here. We want to estimate a constant elasticity of demand model, so we are going to take logs of all of our variables. To do this, we generate new variables using the `gen` command: "`gen logq = log(q)`", and so on, for all the variables in this data set. Recall that this allows us to interpret the coefficients as elasticities. For clarity, all variables from this point forward are assumed to be in logs.

Question 1 *Regress price on quantity, leaving out all X's, using OLS (either use the menu system to find linear regression, or use the command* `regress`.[3]*) Interpret and report your results (coefficient and standard errors are sufficient). What exactly are you recovering by running OLS on such an equation? What is the interpretation of the coefficient on price, and what do you make of its sign?*

First-year econometrics, and economic common sense, tells us that there is something fishy with this regression, and it is due to the fact that we have endogenous right-hand side variables. Explicitly, price is jointly determined in a system of equations with quantity, at the point where supply equals demand. We don't have a supply equation here, but the beauty of exclusion restrictions is that we don't even need to estimate such a curve. The exclusion restriction here is something which is going to shift around the quantity supplied, and not shift around the demand curve. By doing so, we can trace out the demand curve, which is held constant, conditioning on observed covariates. What does economic theory

the first time.

[2]Make sure that you select ".csv" as the type of file to import, or it will not automatically show up.

[3]Typing in "`help XYZ`" in Stata brings up the interactive manual pages. This is a valuable resource; learn how to use it.

tells us is going to work on the supply side? The typical supply-side instruments shift costs of production, as these can be reasonably argued to shift supply, and not shift demand. This dataset has two shifters: the price of chicken feed, and the price of corn. Regress one on the other to convince yourself that they are not perfectly correlated.

Stata has an enormous amount of built-in features that are useful to economists. We are going to make use of the `ivreg` command.[4] IV regression can be found under "Statistics → Linear regression and related → Multiple equation models → Instrumental variables & two-stage least squares."

Question 2 *Estimate the IV model, without any covariates, using pf and pcor as instruments. Report your results, as before. How do the regression estimates change?*

So...that wasn't very insightful, was it? What is the problem? Well, for one, you would rightly reject the model out of hand as being misspecified, as we haven't incorporated any of our covariates yet. They may improve the fit of the model.

Question 3 *Re-estimate the IV model, this time using beef prices as an additional covariate. Report the changes in the price coefficient, and the estimate of the price of beef's effect on chicken demand. What was happening in the previous regression which we are now accounting for?*

Our model was misspecified, as we were missing some information about the effect of beef prices on the slope of the demand curve. We still aren't there, yet, as the coefficient on price is still pointing the wrong direction. Note, however, that the 95% confidence interval now contains the right sign on price, although it also contains zero.

Question 4 *Add population and income to the model. Report the coefficients and standard deviations, along with your comments about what happened with the inclusion of more variables. What has happened to the fit of the model overall? What about the precision of the individual estimates? Do coefficients have the right signs? What are the units for your coefficients?*

[4]People doing applied work in their dissertations are highly encouraged to use Stata to download and install `ivreg2`, a much-improved version that incorporates many cutting-edge econometric extensions to the basic IV regression program in Stata.

We are closer to seeing coefficients of the right sign, and of reasonable magnitudes. We have one more variable that we can include here, CPI.

Question 5 *Repeat the regression, including all previous variables plus CPI. The value for the elasticity of demand that Epple and McCallum report is -0.40, with a standard deviation of 0.086. How do your results compare? What is the intuition for including logged CPI as a variable in the demand for chicken? Looking at the adjusted R^2 across the two equations, are you better off including that variable or keeping it out? This data set was chosen as having good characteristics for textbook supply and demand; what is your opinion of that claim? After seeing how your results change (play around with adding and dropping various covariates, or using the non-logged versions of the variable to estimate a linear demand curve), what do you think is the one dimension of this data set which is most lacking?*

If you can get it working, note that `ivreg2` has a number of diagnostic statistics built into the regression output. If you plan on doing empirical work, make sure that you familiarize yourself with using this tool, and understand how the first-stage F statistic and J-statistic work. Quantitative backup that your instrument are correlated with the endogenous regressor (F) and orthogonal to the error term (J-statistic) can defuse arguments in your job talk, and are an order of magnitude more convincing than verbal hand-waving. Discussing instruments is a topic that economists are expert at, and expect to account for your own instruments and functional form decisions. The skills that you can develop and hone in your demand analysis will have beneficial spillovers in other areas of your work.

2 The Berry Logit

The second set of exercises concerns the estimation of the so-called "Berry Logit". The data set "berryLogit.csv" contains observations on the market shares of three products and the outside good in 100 markets. The underlying utility model is:

$$u_{ijm} = X'_{jm}\beta + \xi_j + \epsilon_{ijm}, \tag{2}$$

where ξ_j is an unobserved, product-specific covariate, and ϵ_{ijm} is an idiosyncratic error shock which is distributed iid across markets, individuals, and products. The X vary across products and markets, while the product-specific characteristics are fixed across markets.

The object is to recover β and ξ_j. We assume that the model is generated with extreme value errors, which in turn generate multinomial logit choice probabilities.

There are several special properties using observations on market shares. The most important, which we will leverage, is that when the number of individuals grows large enough, market shares are measured without error. Under the assumption that consumers have identical preferences, these market shares directly reflect the choice probabilities: $s_{jm} = I_m Pr(jm)$, where I_m is the size of the market. As the market shares are measured without error, it is possible to obtain exact fits of the observed data.

Question 6 *Using the data in "berryLogit.csv", recover the underlying parameter vector β, and estimate of the product-specific unobservable ξ_j for each product. Assume that the deterministic portion of utility for the outside good is equal to 0.*

Hint: You can use a spreadsheet to rearrange the data into a form which makes this estimation very simple.

Question 7 *Suppose that the error term was not distributed extreme value, but was some other known distribution. How would you procede, in principle, with the recovery of the unknown parameters? Suppose the error term is unknown; how do you interpret the parameter estimates recovered using a Berry logit? What can you say about the confidence interval of the unknown parameters when you fit the data perfectly, and yet the model is misspecified?*