COLLECTIVE PROBLEM SET 1 — 2025

1. Demand Estimation

The data set *Data_Broiler.mat* contains aggregate data on quantity, price, cost, and demographic variables related to edible meat of young chickens (”broilers”) in the United States from 1962 to 1999.

There are 12 variables in the dataset:

(1) Year
(2) Quantity of broiler chicken
(3) Income
(4) Price of broiler chicken
(5) Price of beef
(6) Price of corn
(7) Price of chicken feed
(8) Consumer Price Index
(9) Aggregate production of chicken in pounds
(10) US population
(11) Exports of beef, veal, and pork in pounds
(12) Time trend

We are interested in estimating the demand for broiler chicken meat using a linear model:

$$q = \alpha + \beta p + \boldsymbol{x}'\boldsymbol{\gamma} + \epsilon$$

(1) Regress the quantity of broiler chicken on the price of broiler chicken (and a constant), excluding all other variables. Interpret and report your results (coefficients, standard errors, and t-statistics). What exactly are you estimating by running OLS on this model? What is the interpretation of the price coefficient, and what do you make of its sign?

(2) Re-estimate the OLS model, now including all available demand covariates. That is, include $\boldsymbol{x}$, which contains: income, price of beef, consumer price index, aggregate production of chicken in pounds, US population, exports of beef/veal/pork in pounds, and time trend. Report the estimated coefficients and standard errors, and comment on what happened after including these additional variables. What happened to the fit of the model overall? What about the precision of the individual estimates? Do the coefficients have the expected signs?

(3) To address the potential endogeneity of price, use as instrumental variables the price of corn, the price of chicken feed, and their interaction. Explain the rationale behind this choice. Estimate the model using two-stage least squares (2SLS) with these three instruments. Comment on your results.

(4) Plot the price elasticity of demand for broiler chicken over the sample years.

VERY IMPORTANT, HOW TO READ COEFFICIENTS FROM LOGIT AND POISSON:
THERE ARE TWO WAYS:

- EXP(BETAS) -1:

FOR LOGIT IT IS THE ODDS RATIO, AND TELLS YOU HOW MUCH THE ODDS OF HAVING P(Y=1|X)/P(Y = 0|X) CHANGES IN PERCENTAGES
FOR POISSON THE INTERPRETATION IS THE SAME

- AVERAGE MARGINAL EFFECTS (AME):

FOR LOGIT IT TELLS YOU HOW MUCH THE EXPECTATION OF PROB(Y=1|X) CHANGE IN POINT PERCENT != PERCENTAGE (IF AME= 0.2 IT MEANS THAT P(Y=1) IS INCREASED BY O.2, LIKE FROM O.17 TO 0.37, NOT MULTIPLIED BY 1.2) AFTER A ONE UNIT INCREASE OF THE REGRESSOR
FOR POISSON IT TELLS YOU HOW MANY UNITS THE EXPECTATION OF Y CHANGES AFTER A ONE UNIT INCREASE IN THE REGRESSOR.

2. Binary Regression

The dataset `Data_Logit.xlsx` contains data from an experiment that studies violations of First Order Stochastic Dominance (FOSD). Subjects were asked to choose between different lotteries across repeated questions.

Recall that given two random variables X and Y with cumulative distribution functions $F_X(x)$ and $F_Y(x)$, we say that X first-order stochastically dominates Y if and only if:

$$\int_{-\infty}^{x} [F_Y(t) - F_X(t)]\, dt \geq 0 \quad \text{for all } x$$

In the context of lotteries, a lottery A first-order stochastically dominates B if A yields higher outcomes with greater probability.

The variables in the dataset are:

- Dummy equal to 1 if the subject violated FOSD, 0 otherwise.
- Cognitive ability score.
- Time used in the cognitive ability test.
- Understanding of the experiment (1 = high, 5 = low).
- Understanding of the experiment (duplicate measure).
- Gender (1 = female, 0 = male).
- Education (1 = less than high school, 2 = high school, 3 = bachelor, 4 = master, 5 = doctorate).
- Response time in the experiment.
- Dummy for ”expected value/risk neutrality” behavioral rule.
- Dummy for ”high risk aversion” behavioral rule.
- Dummy for randomization behavior.

(1) Import the dataset `Data_Logit.xlsx`.
(2) Estimate a logit regression of the FOSD violation dummy on all other variables and a constant. Maximize the log-likelihood using both `fminunc` and `fmincon` in MATLAB. (This is primarily for practice with both optimizers; `fminunc` is recommended.)
(3) Re-estimate the parameters by solving the first-order conditions using `fsolve`.
(4) Re-estimate the parameters and standard errors using MATLAB's `glmfit` function with a logit link.
(5) Estimate a probit model using `glmfit` with a probit link.
(6) Compare numerically the estimated logit and probit coefficients. Do you notice any patterns?
(7) Estimate the average marginal effect of having high risk aversion on the probability of violating FOSD in both the logit and probit models. What do you observe?
(8) Comment on your results.

3. Negative Binomial Regression

The dataset *PoissonDATA.mat* contains the same data used in Lecture 5 for Poisson regression.

Poisson regression imposes the *equidispersion* condition: $\mathbb{E}[y_i] = \text{Var}(y_i)$. However, count data often exhibit *overdispersion*: $\text{Var}(y_i) > \mathbb{E}[y_i]$. The Negative Binomial Regression Model (NBRM) generalizes the Poisson model to allow for overdispersion.

In the NBRM, the conditional mean is $\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$ and:

$$y_i \sim \text{NegBin}(\mu_i, \alpha) \quad \text{with} \quad \text{Var}(y_i \mid \mathbf{x}_i) = \mu_i + \alpha\mu_i^2$$

The conditional density is:

$$f(y_i \mid \mathbf{x}_i) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right)^{\alpha^{-1}} \left(\frac{\mu_i}{\alpha^{-1} + \mu_i}\right)^{y_i}$$

The log-likelihood is:

$$\mathcal{L}(\boldsymbol{\beta}, \alpha) = \sum_{i=1}^{N} \left[\ln\Gamma(y_i + \alpha^{-1}) - \ln\Gamma(y_i + 1) - \ln\Gamma(\alpha^{-1}) + \alpha^{-1}\ln\left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right) + y_i \ln\left(\frac{\mu_i}{\alpha^{-1} + \mu_i}\right)\right]$$

(1) Load the dataset *PoissonDATA.mat*.
(2) Write a MATLAB function for the log-likelihood as a function of parameters and data.
(3) Estimate the parameters of the NBRM using this data.
(4) Report and briefly interpret the results (coefficients, standard errors, t-statistics).
(5) Compute the sample average marginal effect of age on the expected number of doctor visits.
(6) Test the overdispersion hypothesis (i.e., test whether $\alpha = 0$).
(7) Plot the empirical distribution of counts in the sample, along with the distribution of predicted counts from the Poisson and NBRM models.