New York University
Applied Statistics and Econometrics I
Fall 2015

Problem Set 3

Cristian Fuenzalida
cristian.fuenzalida@nyu.edu

(I) This is application 1 in Greene (2012). *You need to provide codes and explanations for this problem.* Data on U.S. gasoline consumption for the years 1953 to 2004 are given in *TableF2.2.txt* (see attached file). Note that the consumption data appear as total expenditure. To obtain the per capita quantity variable, divide GASEXP by GASP times Pop. The other variables do not need transformation.

1. Compute in R the multiple regression of per capita consumption of gasoline on per capita income, the price of gasoline, all the other prices and a time trend. Report all results. Do the signs of the estimates agree with your expectations?
2. Test the hypothesis that at least in regard to demand for gasoline, consumers do not differentiate between changes in the prices of new and used cars.
3. Estimate the own price elasticity of demand, the income elasticity, and the cross price elasticity with respect to changes in the price of public transportation. Do the computations at the 2004 point in the data.
4. Re-estimate the regression in logarithms so that the coefficients are direct estimates of the elasticities. (Do not use the log of the time trend.) How do your estimates compare with the results in the previous question? Which specification do you prefer?
5. Compute the simple correlations of the price variables.Would you conclude that multicollinearity is a "problem" for the regression in part 1 or part 4?
6. Notice that the price index for gasoline is normalized to 100 in 2000, whereas the other price indices are anchored at 1983 (roughly). If you were to re-normalize the indices so that they were all 100.00 in 2004, then how would the results of the regression in part 1 change? How would the results of the regression in part 4 change?
7. This exercise is based on the model that you estimated in part 4. We are interested in investigating the change in the gasoline market that occurred in 1973. First, compute the average values of log of per capita gasoline consumption in the years 1953-1973 and 1974-2004 and report the values and the difference. If we divide the sample into these two groups of observations, then we can decompose the change in the expected value of the log of consumption into a change attributable to change in the regressors and a change attributable to a change in the model coefficients, as shown in Section 4.5.3. Using the Oaxaca-Blinder approach described there, compute the decomposition by partitioning the sample and computing separate regressions. Using your results, compute a confidence interval for the part of the

change that can be attributed to structural change in the market, that is, change in the regression coefficients.

(II) This is application 2 in Greene (2012). The generalized Cobb-Douglas cost function has the form:

$$\ln(C) = \alpha + \beta \ln(Q) + \gamma \left(0.5 \left(\ln(Q)\right)^2\right) + \delta_k \ln(P_k) + \delta_l \ln\left(P_l\right) + \delta_f \ln\left(P_f\right) + \epsilon$$

P_k , P_l, and P_f indicate unit prices of capital, labor, and fuel, respectively, Q is output and C is total cost. To conform to the underlying theory of production, it is necessary to impose the restriction that the cost function be homogeneous of degree one in the three prices. This is done with the restriction $\delta_k + \delta_l + \delta_f = 1$.

Inserting this result in the cost function and rearranging produces the estimating equation,

$$\ln(C/P_f) = \alpha + \beta \ln(Q) + \gamma \left(0.5 \left(\ln(Q)\right)^2\right) + \delta_k \ln(P_k/P_f) + \delta_l \ln(P_l/P_f) + \epsilon$$

The purpose of the generalization was to produce a U-shaped average total cost curve. We are interested in the efficient scale, which is the output at which the cost curve reaches its minimum.

1. Data on 158 firms extracted from Christensen and Greene?s study are given in *TableF4.4.txt* Using all 158 observations, write an R code to compute the estimates of the parameters in the cost function and the estimate of the asymptotic covariance matrix.
2. Note that the cost function does not provide a direct estimate of δ_f . Compute this estimate from your regression results, and estimate the asymptotic standard error.
3. Compute an estimate of Q^* using your regression results and then form a confidence interval for the estimated efficient scale.
4. Examine the raw data and determine where in the sample the efficient scale lies. That is, determine how many firms in the sample have reached this scale, and whether, in your opinion, this scale is large in relation to the sizes of firms in the sample. Christensen and Greene approached this question by computing the proportion of total output in the sample that was produced by firms that had not yet reached efficient scale.

(III) You have been entrusted with the mission to determine the proportion of NYU who have never been in the Bobst library (θ). Due to the reluctance of some students to give a straight answer, the designed procedure to interview them is the following: first, they are told to flip a biased coin secretly. Then if the result is heads (which happens with probability p), they answer the question “Have you ever been in the library?”. If not, they answer the question “Were you born in the NY State?”. Denote by Z_i the response given by the student, where $Z_i = 1$ if the answer is “Yes”. You do not observe which question the student answered.

1. The NYU Registrar Office tells you that a fraction q of the students were born in the New York State. Calculate $\mathbb{P}\{Z_i = 1\}$ as a function of θ.
2. Propose a consistent estimator $\hat{\theta}$ of θ.
3. Use the Central Limit Theorem to construct a confidence region for θ valid in large samples with a 99% confidence level.

(IV) Let $X_1, \ldots, X_N$ an i.i.d. sample from a $N(\mu, 1)$ distribution. A statistician tells us we can estimate $\theta := \mu^2$ with:

$$Z_N = \bar{X}^2$$

1. Prove that $\sqrt{N}(Z - \theta) \xrightarrow{d} N(0, 4\theta)$ if $\mu \neq 0$.
2. Let $\mu = 0$. What is the limit behaviour of $N(Z - \theta)$?

(V) Does a change in the unit of measurement for the dependent variable affect R^2? In order to make the comparison, analyse the change of R^2 when going from dependent variable y to the transformed λy. How about when there is a change in the measurement unit of the regressors? Compare now the R^2 when the regressors are a matrix X versus λX. Here, λ is a scalar number that transforms the data from certain units of measurement to another.

(VI) Consider a linear model with dependent variable y and $\hat{y} = X(X'X)^{-1}X'y$. Prove that R^2 is the square of the correlation between y and $\hat{y}$.

(VII) Prove the **Gauss-Markov Theorem**: the OLS is the best linear unbiased estimator of the coefficients. That is, its variance is the lowest among the class of linear estimators of the coefficients. Work in the general case of an arbitrary number of regressors. Given that the variance is a matrix in this case, prove that the difference matrix is positive semi-definite when comparing with another variance matrix.

(VIII) We want to estimate the model:

$$Y = X\beta + \epsilon$$

In this model, we have $\epsilon_t \sim \sigma(e_t - 1)$, where e_t is a standard exponential variable with mean 1. Assume that X are independent of ϵ. Suppose that (x_t, ϵ_t) are i.i.d. across t.

1. Do the Gauss-Markov assumptions hold for this model? Check one by one.
2. Consider the least squares estimator $\hat{\beta}$. Compute $\mathbb{E}(\hat{\beta})$ and $Var(\hat{\beta})$. Is $\hat{\beta}$ normally distributed in finite samples, conditional on X?
3. Is $\hat{\beta}$ the best linear unbiased estimator for this model?
4. Suppose we want to estimate the following effect:

$$\mathbb{E}(y_t|x_t = x_0) - \mathbb{E}(y_t|x_t = x_1) = (x_0 - x_1)\hat{\beta}$$

   Give an economic example where this effect could be interesting. Be careful in defining the whole model and the specific effect to be measured. Is $(x_0 - x_1)\hat{\beta}$ the best linear unbiased estimator for this effect?
5. What is the maximum likelihood estimator of this model?