

An Introduction to Linear Regression: Lecture II

Charles B. Moss ¹

¹University of Florida

January 10, 2012

- 1 Readings
- 2 Basic Linear Regression Model
- 3 Assumptions of the Linear Model
 - Linear Formulations in Economics
 - Formal Demand Systems
 - Household Production Model
 - Full Rank
 - Exogeneity
 - Spherical Errors - Homoscedasticity
 - Normality
- 4 Mathematical Content Area Test
- 5 Readings for Lecture III

Readings

- * Greene, W.H. 2012. *Econometric Analysis* Seventh Edition. Prentice Hall (Chapter 2).
- * Popper, K. 2010. *The Logic of Scientific Discovery* Routledge Classics (Chapter 2: On the Problem of a Theory of Scientific Method, 27-34).

Basic Linear Regression Model

- The most basic formulation of the linear model is

$$\begin{aligned} y &= f(x_1, x_2, \dots, x_k) + \epsilon \\ &= \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon \end{aligned} \tag{1}$$

Its simplicity conceals several logical steps.

- Weintraub (2002) presents the development of economics as a mathematical science.
- This occurred in phases.
- The rigor of the nineteenth century would have demanded that observable measures for empirical models (i.e., derivatives of a utility function would require some measurement of utility).
- Under this approach, theory is developed from axioms which yield a synthesis or proof of an economic theory.

Assumptions of the Linear Model

- A1 Linearity
- A2 Full Rank
- A3 Exogeneity of the independent variables
- A4 Homoscedasticity and nonautocorrelation
- A5 Data generation
- A6 Normal distribution

Linear Formulations in Economics

- While economic theory yields valid restrictions (i.e., homogeneity and symmetry of consumer and derived demand functions), it seldom yields exact formulations for these economic relationships.
- Take the standard consumer demand model

$$\left. \begin{array}{l} \max_{x_1, x_2} U(x_1, x_2) \\ \text{s.t. } p_1 x_1 + p_2 x_2 \leq Y \end{array} \right\} \Rightarrow \left\{ \begin{array}{l} x_1^*(p_1, p_2, Y) \\ x_2^*(p_1, p_2, Y) \end{array} \right. \quad (2)$$

where x_1 and x_2 are consumption goods, p_1 and p_2 are prices, and Y is the level of consumer income.

Linear Formulations in Economics - Continued

- If we restrict our analysis to a specific functional form (such as the Cobb-Douglas function)

$$\left. \begin{aligned} L &= x_1^\alpha x_2^{1-\alpha} - \lambda(Y - p_1 x_1 - p_2 x_2) \\ \frac{\partial L}{\partial x_1} &= \alpha \frac{U}{x_1} - \lambda p_1 = 0 \\ \frac{\partial L}{\partial x_2} &= (1 - \alpha) \frac{U}{x_2} - \lambda p_2 = 0 \\ \frac{\partial L}{\partial \lambda} &= Y - p_1 x_1 - p_2 x_2 = 0 \end{aligned} \right\} \Rightarrow \quad (3)$$

• Continued

$$\left\{ \begin{array}{l} \frac{\alpha}{1-\alpha} \frac{x_2}{x_1} = \frac{p_1}{p_2} \\ x_2 = \frac{1-\alpha}{\alpha} \frac{p_1}{p_2} x_1 \\ Y - p_1 x_1 - p_2 \left[\frac{1-\alpha}{\alpha} \frac{p_1}{p_2} x_1 \right] = 0 \\ Y - p_1 \left[\frac{\alpha + (1-\alpha)}{\alpha} \right] x_1 = 0 \\ Y - p_1 \frac{1}{\alpha} x_1 = 0 \Rightarrow x_1(p_1, p_2, Y) = \frac{\alpha Y}{p_1} \\ x_2(p_1, p_2, Y) = \left[\frac{1-\alpha}{\alpha} \frac{p_1}{p_2} \right] \left[\frac{\alpha Y}{p_1} \right] = \frac{(1-\alpha) Y}{p_2} \end{array} \right. \quad (4)$$

The assumption of a specific functional form is specious

Formal Demand Systems

- Demand systems such as the Rotterdam demand system are derived around unspecified utility functions.
- Alternatively, demand systems such as the Almost Ideal Demand System (AIDS) are formulated around general specifications of the expenditure functions.
- However, the demand functions as specified in Equation 4 can be formulated in several ways.

Formal Demand Systems - Continued

- First, we could take the natural logarithm of the demand equations

$$\left. \begin{aligned} x_1(p_1, p_2, Y) &= \frac{\alpha Y}{p_1} \\ x_2(p_1, p_2, Y) &= \frac{(1 - \alpha) Y}{p_2} \end{aligned} \right\} \Rightarrow$$

$$\left\{ \begin{aligned} \ln(x_1) &= \beta_{01} + \beta_{11} \ln(p_1) + \beta_{21} \ln(p_2) + \beta_{31} \ln(Y) \\ \ln(x_2) &= \beta_{02} + \beta_{21} \ln(p_1) + \beta_{22} \ln(p_2) + \beta_{23} \ln(Y) \end{aligned} \right. \quad (5)$$

Formal Demand Systems - Continued

- Alternatively, we could formulate the demand system from the relationships in Equation 4 using a linear approximation

$$\begin{aligned}
 x_1^*(p_1, p_2, Y) &\approx x_1(p_1^0, p_2^0, Y^0) + \left. \frac{\partial x_1(p_1, p_2, Y)}{\partial p_1} \right|_{p_1 \rightarrow p_1^0} (p_1 - p_1^0) + \\
 &\quad \left. \frac{\partial x_1(p_1, p_2, Y)}{\partial p_2} \right|_{p_2 \rightarrow p_2^0} + \left. \frac{\partial x_1(p_1, p_2, Y)}{\partial Y} \right|_{Y \rightarrow Y^0} \\
 x_2^*(p_1, p_2, Y) &\approx x_2(p_1^0, p_2^0, Y^0) + \left. \frac{\partial x_2(p_1, p_2, Y)}{\partial p_1} \right|_{p_1 \rightarrow p_1^0} (p_1 - p_1^0) + \\
 &\quad \left. \frac{\partial x_2(p_1, p_2, Y)}{\partial p_2} \right|_{p_2 \rightarrow p_2^0} + \left. \frac{\partial x_2(p_1, p_2, Y)}{\partial Y} \right|_{Y \rightarrow Y^0}
 \end{aligned} \tag{6}$$

Household Production Model

- Using a slightly more elaborate formulation, consider the household production model where the consumer purchases a variety of different inputs (in this case six food groups $[x_i \ i = 1, \dots, 6]$) in order to produce a vector of consumption goods (in this case two food outputs $[y_j \ j = 1, 2]$).
- This formulation can be written as

$$\begin{aligned} \max_{y_1, y_2, x_1, x_2, x_3, x_4, x_5, x_6} \quad & U(y_1, y_2) \\ \text{s.t.} \quad & F(y_1, y_2, x_1, x_2, x_3, x_4, x_5, x_6) = 0 \end{aligned} \tag{7}$$

$$p_1 x_1 + p_2 x_2 + p_3 x_3 + p_4 x_4 + p_5 x_5 + p_6 x_6 \leq Y$$

I am interested in this formulation from several perspectives.

Household Production Model - Continued

- First, one of the arguments to the utility and the production function could be household labor.
- This decision has consequences such as prepared meals (wheat versus bread), or food away from home.
- An additional issue involves the cost obtaining healthy diets (specifically given that the health indices may be designed by federal government).
- Specifically, what is the impact of the decisions from Equation 7 on a health index (H_1) defined as

$$H_1(x_1, x_2, x_3, x_4, x_5, x_6) = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4 + \alpha_5 x_5 + \alpha_6 x_6. \quad (8)$$

Basically, how does the linear relationship in Equation 8 relate to the production function and utility function in Equation 7.

Household Production Model - Continued

- Specifically, suppose that the United States Department of Agriculture develops a dietary guideline for healthy eating defined as

$$H_1(x_1, x_2, x_3, x_4, x_5, x_6) = -0.03x_1 + 0.73x_2 + 0.23x_3 - 0.38x_4 - 0.47x_5 \quad (9)$$

For example, we could assume that x_3 are red meats and x_5 are calorie laden foods such as candy while x_2 are fruits and vegetables, and x_3 are complex carbohydrates.

- Hence, the question becomes: What is the relationship between the consumer's choice and the health index.
- Would it make any sense to regress H_1 on income, food prices, or ethnic group?

Household Production Model - Continued

- Looking forward slightly, we can take the dominant eigenvalue from the covariance matrix for the optimum choice of outputs and inputs to yield

$$u_8 = \begin{bmatrix} 0.22 \\ 0.06 \\ -0.03 \\ 0.73 \\ 0.23 \\ -0.38 \\ -0.47 \\ -0.01 \end{bmatrix} \Leftrightarrow \begin{bmatrix} y_1 \\ y_2 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} \quad (10)$$

or the USDA's health index is a significant factor spanning the consumer's choice space.

Full Rank

- The basic idea of the full rank of the matrix of independent variables $X \in M_{n \times K}$ such that $n \gg k$ involves the independence of regression variables (i.e., that each of the variables contains a kernel of information not contained in the other variables).
- As a starting point, let's assume $X \in M_{6 \times 3}$.
- Starting with the first vector or independent variable (noted $X_{\cdot 1}$)

$$X_{\cdot 1} = \begin{bmatrix} 1 \\ 9 \\ 7 \\ 3 \\ 2 \\ 4 \end{bmatrix} \quad (11)$$

- Next, assume that the second independent variable is roughly one half of the first variable

$$X_{.2} = \frac{1}{2}X_{.1} + \sigma_2\epsilon_2 = \frac{1}{2} \begin{bmatrix} 1 \\ 9 \\ 7 \\ 3 \\ 2 \\ 4 \end{bmatrix} + \sigma_2 \begin{bmatrix} \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{24} \\ \epsilon_{26} \\ \epsilon_{25} \end{bmatrix} \quad (12)$$

where ϵ_2 is a random vector where $\epsilon_{2i} \sim N(0, 1)$. As $\sigma_2 \rightarrow 0$ $X_{.2}$ becomes a linear function of $X_{.1}$.

- Mathematically, if we assume that the dependent variable Y is a linear function of X_1 and X_2

$$Y = \alpha_0 + \alpha_1 X_{.1} + \alpha_2 X_{.2} = \alpha_0 + \left(\alpha_1 + \frac{1}{2} \alpha_2 \right) X_{.1} + \alpha_2 \sigma_2 \epsilon_2. \quad (13)$$

The ability to estimate both α_1 and α_2 depends on the $\sigma_2 \neq 0$.

Exogeneity

- The classical problem with exogeneity is the case of market equilibrium. Suppose that researcher is interested is estimating a supply function for sugar in the United States to determine the effect of lifting the Tariff Rate Quota.
- Following the standard approach, the quantity supplied can be hypothesized as a function of the price of sugar and the price of inputs used to produce sugar

$$q_s = \alpha_0 + \alpha_1 p_s + \alpha_2 w_1 + \alpha_3 w_3 + \epsilon \quad (14)$$

where q_s is the quantity of sugar supplied, p_s is the price of sugar, and w_1 and w_2 are the prices of two inputs.

Exogeneity - Continued

- Typically these regressions are doomed if the region is large enough to effect the market. Specifically, the relationship between quantity supplied and the price is affected in part by the demand for sugar (i.e., the market equilibrium)

$$q_s = \beta_0 + \beta_1 p_s + \beta_2 p_1 + \beta_3 p_2 + \nu \quad (15)$$

where p_1 and p_2 are the prices of complementary or substitute goods.

Exogeneity - Continued

- Solving for the system of equations

$$\alpha_0 + \alpha_1 p_s + \alpha_2 w_1 + \alpha_3 w_3 + \epsilon = \beta_0 + \beta_1 p_s + \beta_2 p_1 + \beta_3 p_2 + \nu$$
$$p_s = \frac{1}{\alpha_1 - \beta_1} [-\alpha_0 - \alpha_2 w_1 - \alpha_3 w_3 + \beta_0 + \beta_2 p_1 + \beta_3 p_2 - \epsilon + \nu] \quad (16)$$

focusing on the last two terms of Equation 16 it is clear that $-1/(\alpha_1 - \beta_1)\epsilon$ term in p_s will be correlated with ϵ in Equation 14.

- This correlation undermines straightforward linear regression analysis.

Spherical Errors - Homoscedasticity

- Given the linear model, full rank of the matrix of independent variables, and exogeneity of independent variables, a linear regression exists and the estimates are generally unbiased.
- The next step is typically to prove that ordinary least squares (OLS) estimators are best linear unbiased (BLU or BLUE - the best linear unbiased estimator).
- This result will be established using the Gauss-Markov theorem which adds the assumption that the errors are homoscedastic (or spherical [related to the circle])

$$V(\epsilon) \Rightarrow E[\epsilon\epsilon'] = \sigma^2 I \quad (17)$$

where I is the identity matrix.

Normality

- Finally, the assumption that the residuals are normally distributed contribute the usefulness of small sample properties and the application of t -tests and F -tests.

- Matrix Operations
 - Matrix Addition
 - Matrix Multiplication
 - Matrix Determinant 4×4
 - Row Reduction/Computing the Rank of a Matrix
 - Matrix Inverse
- Calculus

Readings for Lecture III

- * Anderson, T.W. 1984. *An Introduction to Multivariate Statistical Analysis* Second Edition. John Wiley & Sons. (Section 2.5 pp. 35-43).
- * Dhrymes, P.J. 2000. *Mathematics for Econometrics* Third Edition. Springer-Verlag. (Chapter 2 [Section 2.7]).
- Frisch, R. and F.V. Waugh. 1933. Partial Time Regressions as Compared with Individual Trends. *Econometrica* 1(4), 387-401.
- * Greene, W.H. 2012. *Econometric Analysis* Seventh Edition. Prentice Hall. (Appendix A: Matrix Algebra pp. 973-1014).
- * Moss, Charles B. 1997. Returns, Interest Rates, and Inflation: How They Explain Changes in Farmland Values. *American Journal of Agricultural Economics* 79(4), 1311-1318. (<http://www.jstor.org/stable/1244287>).

- * Popper, K. 2010. *The Logic of Scientific Discovery* Routledge Classics [Chapter 3: Theories, 37-56].
- Theil, H. 1971. *Principles of Econometrics* John Wiley & Sons. (Chapter 1 [Sections 1.0 - 1.5]).
- Theil, H. 1983. Chapter 1: Linear Algebra and Matrix Methods in Econometrics. In *Handbook of Econometrics: Volume 1* (eds.) Zvi Griliches and Michael D. Intriligator. North-Holland, 5-65. [Sections 1 - 3]
- Theil, H. 1987. How Many Bits of Information Does an Independent Variable Yield in a Multiple Regression. *Statistics and Probability Letters* 6(2), 107-108.