ECO310: Empirical Industrial Organization

Tutorial 1: Review of Econometrics

Francis Guiton
Email: francis.guiton@mail.utoronto.ca

December 11, 2021

Outline

- Introduction
- Linear Regression Model
- Data
- Estimation: Ordinary Least Squares
- Hypothesis Testing
- Analysis of Variance
- Example

Introduction

- **Econometrics** uses statistical methods to produce estimates of economic parameters.

- **Parameters** - Quantitative measure of some feature of the population or model

- **Estimates** - Statistical inferences of the unknown parameters of model
  - At the very least want estimators to be consistent and unbiased
  - We are satisfied when they are efficient (low standard errors)

- **Standard Errors** - Measure of the imprecision in our estimates.
  - Our parameter estimates will always contain some error:
    1. Sampling error.
    2. Omitted variables bias.

Experiments and Sample Space

- **Experiment** - Any process of observation that can be conceptually repeated and has an uncertain outcome
  - Toss two coins
  - Measure average height
  - Measure the effect of policy on housing price

- **Sample Space** - The set of all possible outcomes of an experiment
  - Toss two coins: {HH, HT, TH , TT }
  - Height: $(0, \infty)$
  - Policy effect: $(-\infty, \infty)$

Events and Random Variables

- **Event** - A subset of the sample space
  - Toss two coins: {HH, HT, TH } "toss at least one head"
  - Height: "between 150 and 180 cm" , "greater than 190 cm"
  - Policy effect: "positive effect on housing price"

- **Random Variable** - A function that assigns a numerical value to each outcome
  - Toss two coins: $X \in \{0, 1, 2\}$ =number of heads
  - Height: $X \in \{0, \infty\}$
  - Policy effect: $X \in \{-\infty, \infty\}$

Random Variables and their Distribution

- Let Y be a **random variable** (r.v.)
  - That is, the value of Y is subject to variations due to chance
  - As such, there is uncertainty involved in its value.

- The set of possible values of Y, and the probability at which it takes on these values is described by the **distribution** of Y

Random Variables and their Distribution

- The **distribution function** denoted $F(y)$ describes the probability that the r.v. Y takes on a value less than or equal to the number y.

$$F(y) = \Pr\{Y \leq y\}$$

- The **mean** μ of Y is the expected value of the distribution of Y
- The **variance** σ^2 of Y measures the spread in the distribution of Y.

$$\mu = E[Y] \qquad \text{and} \qquad \sigma^2 = E[(Y-\mu)^2]$$

Random Variables and their Distribution

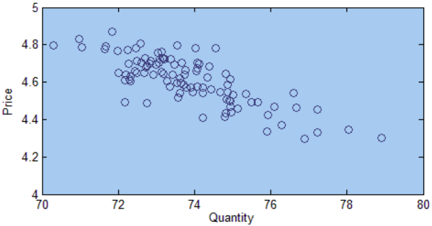
- We often deal with r.v.'s that are generated from an unknown distribution.
- Let $\{y_i : i = 1, ..., N\}$ be a random sample of observations on Y
- Estimators of the population mean and variance are

$$\text{Sample Mean :} \qquad \overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$\text{Sample Variance :} \qquad s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{y})^2$$

Estimating Causal Relationships

- In economics, we are often interested in the causal relationship between an explanatory variable x and an outcome variable y
- A **scatter-plot** is useful way of depicting the relationship between two r.v.'s



Estimating Causal Relationships Cont.

- The **sample covariance** is a useful statistic to describe this relationship

$$cov(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$$

- But covariance/correlation does not imply causation!

Estimating Causal Relationships Cont.

- In other words, we are interested in the **causal relationship** between a set of explanatory variables $x_1, x_2, ..., x_k$ and a **dependent variable** y

- We hypothesize that there is a systematic causal relationship between $x_1, x_2, ..., x_k$ and y through the equation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

- The random component of Y is captured by the **error term** ε with

$$E[\varepsilon] = 0 \quad \text{and} \quad V[\varepsilon] = \sigma^2$$

The Linear Regression Model

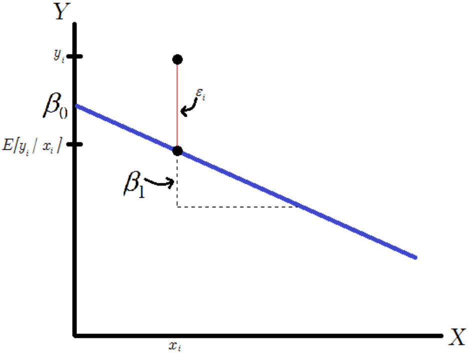
$$\underbrace{y_i}_{\text{Dependent Variable}} = \overbrace{\beta_0}^{\text{Constant Term}} + \overbrace{\beta_1}^{\text{Slope Coefficients}} \underbrace{x_{1i}}_{\text{Independent Variables}} + \overbrace{\beta_2}^{\text{Slope Coefficients}} \underbrace{x_{2i}}_{\text{Independent Variables}} + \cdots + \overbrace{\beta_k}^{\text{Slope Coefficients}} \underbrace{x_{ki}}_{\text{Independent Variables}} + \underbrace{\varepsilon_i}_{\text{Error Term}}$$

The Linear Regression Model Cont.

- **The parameter** β_k measures causal effect of x_k on y, holding *all* other vars. fixed
- ε captures all other factors that affect y aside from $x_1, x_2, \ldots, x_k$
- This error term is included because:
  - Some relevant variables are unobservable.
  - Even if observable, impossible to collect data on everything.
  - Even if collectable, might be subject to Measurement Error.

The Simple Linear Regression Model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$



The Simple Linear Regression Model Cont.

- **Constant** β_0 - "autonomous" level of y.
- **Slope** β_1 - causal effect of a marginal increase in x on y.
- **Error term** - deviation from the systematic model.

The LRM is flexible: it allows for many functional forms – it is only linear in **parameters**, not in **variables**:

Functional Forms

- In a **Linear** specification

$$y = \beta_0 + \beta_1 x + \varepsilon$$

β_1 is the # of units change in y from a 1-unit change in x

- In a **Log-Log** specification

$$\ln y = \beta_0 + \beta_1 \ln x + \varepsilon$$

β_1 is the % change in y from a 1% change in x

- In a **Log-Linear** specification

$$\ln y = \beta_0 + \beta_1 x + \varepsilon$$

$100 * \beta_1$ is the % change in y from a 1-unit change in x

The Data

- The purpose of our econometric analysis is to estimate parameters $\beta_0, \beta_1, \beta_2, ..., \beta_k$ and σ^2
- Towards this end, suppose we have collected a **random sample** of data

$$\{y_i, x_{1i}, x_{2i}, ..., x_{ki} : i = 1, 2, ..., N\}$$

  - Each individual in the population has an equal chance of being chosen at each draw of our sample.
  - This ensures that sample is representative of the underlying population

The Data Cont.

- Data for econometric analysis comes in a variety of types
  - **Cross Section** - observe many individuals for one period

$$Q_i = \beta_0 + \beta_1 P_i + \varepsilon_i \qquad \textit{for } i = \textit{City } 1, \ldots, \textit{City } N$$

  - **Time Series** - observe one individual over successive time periods, e.g.

$$Q_t = \beta_0 + \beta_1 P_t + \varepsilon_t \qquad \textit{for } \; t = \textit{Year } 1, \ldots, \textit{Year } T$$

  - **Panel Data** - observe many individuals over multiple periods, e.g.

$$Q_{it} = \beta_0 + \beta_1 P_{it} + \varepsilon_{it} \qquad \begin{array}{l} \textit{for } \; i = \textit{City } 1, \ldots, \textit{City } N \\ \textit{and } \; t = \textit{Year } 1, \ldots, \textit{Year } T \end{array}$$

The Data Cont.

Cross Section

City	Price	Quantity
Toronto	99.99	1.75 mil
Montreal	103.50	1.65 mil
⋮		
Cranbrook	123	10,000

The Data Cont.

Time Series

Montreal - Year	Price	Quantity
1990	87.50	1.03 mil
1991	87.99	1.02 mil
⋮		
2010	103.50	1.65 mil

The Data Cont.

Panel

City	Year	Price	Quantity
Toronto	1990	87.50	0.9 mil
Toronto	2010	99.99	1.75 mil
Montreal	1990	87.50	1.03 mil
Montreal	2010	103.50	1.65 mil
$\vdots$			
Cranbrook	1990	86.00	1,000
Cranbrook	2010	123	10,000

Assumptions

- The multiple regression model states that, in the population:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i$$

  - The number of parameters is $k + 1$
  - The observation index is i: Cross-sectional data
  - Parameter β_k measures causal effect of x_k on y **holding all other vars fixed**
  - Error term ε is an unobservable capturing all *other* factors that effect y

Assumptions Cont.

1 **Linearity**: each predictor variable x is linearly related to y.
- Means no non-linearities in parameters - cannot have $y_i = \beta_0 + x_i^{\beta_1} + \varepsilon$.
- However, the x and y variables can be non-linear transformations - can have $\ln y_i = \beta_0 + \beta_1 \ln x_i + \varepsilon_i$ or $y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon$

2 **Zero Mean**: Error terms have a mean of zero. $E[\varepsilon_i] = 0$
- Can be made without loss of generality if constant β_0 has been included

3 **Exogeneity**: Each x_k is unrelated with the error term. $cov(x_k, \varepsilon_i) = 0$.
- Means no "lurking variables". – i.e. any omitted variable do not have confounding effects on both x's and y.
- Crucial is random sampling, so variation in x's is independent of variation in ε

4 **Independence**: Error terms are independently distributed.
$cov(\varepsilon_i, \varepsilon_j) = 0$

5 **Homoscedasticity**: Error terms have a constant variance.
$var(\varepsilon_i) = \sigma_\varepsilon^2$

6 **Normality**: Error terms are normally distributed. $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$

Estimation

- With a sample of data $\{y_i, x_{1i}, ...x_{ki} : i = 1, ..., N\}$ we can estimate the unknown population parameters $\beta_0, \beta_1, ...\beta_k$

- Let $b_0, b_1, ...b_k$ be the estimated parameters from our sample of data.
- Based on these estimates, the **fitted value** or **predicted value** of y_i given $x_{1i}, x_{2i}, ...x_{ki}$ is

$$\widehat{y}_i = b_0 + b_1 x_{1i} + b_1 x_{2i} + ... + b_k x_k$$

- The difference between observed value of y_i and predicted value $\widehat{y}_i$ is the **residual**

$$e_i = y_i - \widehat{y}_i$$

  and *can be thought* of as a measure of how close our prediction is to the true value

Estimation - Some Ideas

- We want to choose our estimates such that the error is small

1. Choose parameters to minimize the sum of residuals $\sum_{i=1}^{n}(y_i - \hat{y}_i)$
   - Doesn't account for errors of opposite sign
   - Any line that passes through the point $(\bar{x}, \bar{y})$ will have this sum equal to 0 (non unique solution)
2. Choose parameters to minimize $\sum_{i=1}^{n} |(y_i - \hat{y}_i)|$
   - "Least absolute value regression" - this is seldom used
3. Choose parameters to minimize $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
   - This type of estimator is called a **Least Squares Estimator**
   - One of the most common estimators in econometrics
   - Easy to compute and provides a unique solution
   - Best Linear Unbiased Estimator (BLUE)

Estimation - Ordinary Least Squares

- The most common estimator in econometrics is **Ordinary Least Squares**
  - We do not observe the error term ε_i.
  - But given estimates of the β parameters, we can construct an estimate of it.
  - The **residuals**

$$e_i = y_i - \hat{y}_i = y_i - b_1 x_{1i} - b_2 x_{2i} - ... - b_k x_{ki}$$

  - The **OLS Estimator** is the value of the $b's$ which minimizes the sum of squared residuals

$$b = \arg\min \sum_{i-1}^{N} e_i^2$$

Estimation - Ordinary Least Squares Cont.

- For the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

the OLS estimator for the slope parameter has a simple expression

$$b_1 = \frac{\sum_{i-1}^{N}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i-1}^{N}(x_i - \overline{x})^2} = \frac{cov(x, y)}{var(x)}$$

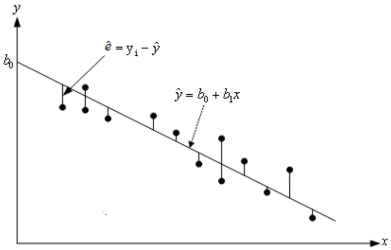
$$b_0 = \bar{y} - b_1 \bar{x}$$

- And our estimator for the error variance σ^2 is given by

$$s^2 = \frac{1}{N-2} \sum_{i-1}^{N} e_i^2$$

Interpretation

- How do we interpret the estimated parameter?



- The principle behind OLS is to estimate the model parameters by drawing a line that "best fits" the data in the least squares sense.

Interpretation Cont.

- How do we interpret the estimated parameter?
- The estimated value b_k measures the *typical* (i.e. average) change in y associated with a one unit change in x_k, holding the other included x variables fixed.
  - You can think of b_k as the "partial correlation" between x_k and y – i.e. the correlation between x_k and y *after* controlling for the other *included* x's
  - NB: partial-correlation is not the same thing as correlation. E.g., it is possible to observe positive correlation between x_k and y, and then get a negative estimate b_k.

Interpretation Cont.

- However, (Partial) Correlation does not imply Causation
  - Because of the possibility of latent or ommitted variables (violation of Exogeneity) – b_k is not necessarily an estimate of the causal effect of x_k on y.
  - That is, due to the possibility of **Endogeneity**, **we cannot** say that b_k measures the change in y associated with a one unit change in x_k, holding **all** variables fixed.

Hypothesis Testing

- Under Assumption 1-6, b_k is an estimate of the (partial) effect of x_k on y based on our *sample of data*.
- We can use it to do **inference** about the value of β_k, the (partial) effect of x_k on y in the *population*.

Hypothesis Testing Cont.

- **Hypothesis Testing**
  - Suppose we wanted to answer the question "Is the (partial) effect of x_k on y in the *population* equal to (the number) β?"
  - We maintain the **Null Hypothesis** that β_k is indeed equal to β in the population

$$H_0: \ \beta_k = \beta$$

    and we ask the data to show us otherwise – i.e. our **Alternative Hypothesis**

$$H_1: \ \beta_k \neq \beta$$

  - The **test-statistic** for this test is the **t-statistic**

$$t = \frac{b_k - \beta}{s_{b_k}}$$

Hypothesis Testing Cont.

- s_{b_k} is the standard error of our estimator b_k. In a simple linear regression $y_i = \beta_0 + \beta_1 x_i + \epsilon$ this is given by

$$s_{b_1} = \frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Under the null hypothesis, H_0, our test statistic follows a T **Distribution** with $n - k - 1$ deg. of freedom

Hypothesis Testing Cont.

- At significance level α, let $t_{\alpha/2}$ be the **critical value** from the T-distribtion that leaves probability mass $\alpha/2$ in the tails.
- We reject H_0 in favour of H_1 if t-statistic is greater than $t_{a/2}$ in absolute value:

$$\text{Reject if} \quad t > t_{\alpha/2} \quad \text{or} \quad t < -t_{\alpha/2}$$

- The **P-value** of the test is the prob. in the tails, as determined by the computed value of the t-stat, under the T-distribution.
- It measures the strength of the evidence against the Null Hypothesis.
- Thus, equivalently:

$$\text{Reject if} \quad P\text{-}value < \alpha$$

Test of Statistical Significance

- A particularly important question is whether x_k indeed has an effect of y. We call this a **Test of Statistical Significance** or just a "Significance Test"
- Our Null Hypothesis and Alternative Hypothesis are

$$H_0 : \ \beta_k = 0 \qquad \text{vs} \qquad H_1 : \ \beta_k \neq 0$$

- The test-statistic for this test is a special case of our usual t-statistic

$$t = \frac{b_k}{s_{b_k}}$$

and under the Null-Hypothesis, $t \sim T(n - k - 1)$.

- Rule of thumb: we can reject H_0 if t is greater than 2 in absolute value.

Analysis of Variance

- The linear regression model is designed to explain the variation of y:

$$s_y^2 = \frac{\sum_i (y_i - \overline{y})^2}{n-1}$$

- **Analysis of Variance** (ANOVA): How the total variability of y is related to the variation in the x's versus the variation in ε
  - Define the Total Sum of Squares as

$$SST = \sum_i (y_i - \overline{y})^2$$

  - The Sum of Squares of the Regression (SSR) is the part of the variation in y that is explained by our regression model
  - The Sum of Squares of the Errors (SSE) is that part left unexplained

$$SSR = \sum_i (\widehat{y}_i - \overline{y})^2 \qquad SSE = \sum_i (y_i - \widehat{y}_i)^2$$

$$SST = SSR + SSE$$

Goodness of Fit

- How much of y is explained by $x_1, x_2, \dots, x_k$?
- The **R-Squared** of the regression is that fraction of the total variation in y that has been explained by the variation in the x's

$$R^2 = \frac{SSR}{SST} \qquad \text{or equivalently} \qquad R^2 = 1 - \frac{SSE}{SST}$$

  - R^2 is a number between 0 and 1.
  - The higher is R^2 the greater is the percent of the variation of y explained by our model.

An Example

- Is the demand for gasoline inelastic?
- Suppose we collected a sample of 50 towns in Ontario during 2013
  - Q_i - the quantity of gasoline sold in that town last year
  - P_i - the (average) price of gasoline in that town
  - Y_i - median household income in that town
- Economic theory gives us a valid regression model of the Demand for Gasoline

$$\ln Q_i = \beta_0 + \beta_1 \ln P_i + \beta_2 \ln Y_i + \varepsilon_i$$

- In STATA, the syntax for regression is: **regress y x1 x2 ...xk**

An Example - Results

```
. reg lnQ lnP lnY

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  2,    47) =    9.38
       Model |  24.0503982     2  12.0251991           Prob > F      =  0.0004
    Residual |  60.2333272    47  1.28156015           R-squared     =  0.2854
-------------+------------------------------           Adj R-squared =  0.2549
       Total |  84.2837254    49  1.72007603           Root MSE      =  1.1321

------------------------------------------------------------------------------
         lnQ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnP |  -.9464336   .5006762    -1.89   0.065    -1.953664     .060797
         lnY |   1.806263   .4239203     4.26   0.000     .9534459    2.659081
       _cons |   10.70829   .6591015    16.25   0.000     9.382345    12.03423
------------------------------------------------------------------------------
```

An Example - Results

```
. reg lnQ lnP lnY

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  2,    47) =    9.38
       Model |  24.0503982     2  12.0251991           Prob > F      =  0.0004
    Residual |  60.2333272    47  1.28156015           R-squared     =  0.2854
-------------+------------------------------           Adj R-squared =  0.2549
       Total |  84.2837254    49  1.72007603           Root MSE      =  1.1321

------------------------------------------------------------------------------
         lnQ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnP |  -.9464336   .5006762    -1.89   0.065    -1.953664     .060797
         lnY |   1.806263   .4239203     4.26   0.000     .9534459    2.659081
       _cons |   10.70829   .6591015    16.25   0.000     9.382345    12.03423
------------------------------------------------------------------------------
```

b

se(b)

An Example - Results

```
. reg lnQ lnP lnY

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  2,    47) =    9.38
       Model |  24.0503982     2  12.0251991           Prob > F      =  0.0004
    Residual |  60.2333272    47  1.28156015           R-squared     =  0.2854
-------------+------------------------------           Adj R-squared =  0.2549
       Total |  84.2837254    49  1.72007603           Root MSE      =  1.1321

------------------------------------------------------------------------------
         lnQ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnP |  -.9464336   .5006762    -1.89   0.065    -1.953664     .060797
         lnY |   1.806263   .4239203     4.26   0.000     .9534459    2.659081
       _cons |   10.70829   .6591015    16.25   0.000     9.382345    12.03423
------------------------------------------------------------------------------
```

t-Statistic & P-Value for
H0: beta = 0 vs H1: beta ≠ 0

An Example - Results

```
. reg lnQ lnP lnY

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  2,    47) =    9.38
       Model |  24.0503982     2  12.0251991           Prob > F      =  0.0004
    Residual |  60.2333272    47  1.28156015           R-squared     =  0.2854
-------------+------------------------------           Adj R-squared =  0.2549
       Total |  84.2837254    49  1.72007603           Root MSE      =  1.1321

------------------------------------------------------------------------------
         lnQ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnP |  -.9464336   .5006762    -1.89   0.065    -1.953664     .060797
         lnY |   1.806263   .4239203     4.26   0.000     .9534459    2.659081
       _cons |   10.70829   .6591015    16.25   0.000     9.382345    12.03423
------------------------------------------------------------------------------
```

```
95% CI for beta
  LB = b - t.025*se(b)
  UB = b + t.025*se(b)
```

An Example - Results

```
. reg lnQ lnP lnY

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  2,    47) =    9.38
       Model |  24.0503982     2  12.0251991           Prob > F      =  0.0004
    Residual |  60.2333272    47  1.28156015           R-squared     =  0.2854
-------------+------------------------------           Adj R-squared =  0.2549
       Total |  84.2837254    49  1.72007603           Root MSE      =  1.1321

------------------------------------------------------------------------------
         lnQ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnP |  -.9464336   .5006762    -1.89   0.065    -1.953664     .060797
         lnY |   1.806263   .4239203     4.26   0.000     .9534459    2.659081
       _cons |   10.70829   .6591015    16.25   0.000     9.382345    12.03423
------------------------------------------------------------------------------
```

k

n

n−k−1

SSR SSE SST

An Example - Results

```
. reg lnQ lnP lnY

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  2,    47) =    9.38
       Model |  24.0503982     2  12.0251991           Prob > F      =  0.0004
    Residual |  60.2333272    47  1.28156015           R-squared     =  0.2854
-------------+------------------------------           Adj R-squared =  0.2549
       Total |  84.2837254    49  1.72007603           Root MSE      =  1.1321

------------------------------------------------------------------------------
         lnQ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnP |  -.9464336   .5006762    -1.89   0.065    -1.953664     .060797
         lnY |   1.806263   .4239203     4.26   0.000     .9534459    2.659081
       _cons |   10.70829   .6591015    16.25   0.000     9.382345    12.03423
------------------------------------------------------------------------------
```

$$S_e^2 = \frac{SSE}{n-k-1} \qquad S_e = \sqrt{\frac{SSE}{n-k-1}}$$

An Example - Results

```
. reg lnQ lnP lnY

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  2,    47) =    9.38
       Model |  24.0503982     2  12.0251991           Prob > F      =  0.0004
    Residual |  60.2333272    47  1.28156015           R-squared     =  0.2854
-------------+------------------------------           Adj R-squared =  0.2549
       Total |  84.2837254    49  1.72007603           Root MSE      =  1.1321

------------------------------------------------------------------------------
         lnQ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnP |  -.9464336   .5006762    -1.89   0.065    -1.953664     .060797
         lnY |   1.806263   .4239203     4.26   0.000     .9534459    2.659081
       _cons |   10.70829   .6591015    16.25   0.000     9.382345    12.03423
------------------------------------------------------------------------------
```

F-Stat and P-Value for
H0: $beta_1$ = $beta_2$ = 0 vs
H1: At least on ≠ 0

R-Sq and Adj R-Sq

An Example - Results

- In a "typical" (i.e. average) market, a 1% increse in Price is associated with a 0.95% decrease in quantity demanded, after controlling for Income.
- The P-value for a significance test is 0.065. Thus, at α=10%. we reject null hypothesis that, even after controlling for income, price has no effect on demand.
- The R-Square for this model is 0.2854.
- This means 28.54% of the variation in (log) Quantities Demanded in our sample is explained by the variation in (log) Prices and Income.

Functional Forms

- As we have just seen, the multiple regression model is much more flexible than it appears – It can be used to estimate non linear relationships between y and the $x's$
- The linearity assumption only means that the parameters enter linearly
- Some common functional forms involve
  - Logarithms
  - Quadratics
  - Interaction Terms
  - Dummy Variables
  - Time Series Models: Trends
  - Panel Data Model: Fixed Effects

Functional Forms - Logarithms

- Suppose we wanted to estimate the relationship between quantities demanded Q for a good, price P of the good, and income Y.
  - In the **Log-Log Model**

    $$\ln Q_i = \beta_0 + \beta_1 \ln P_i + \beta_2 \ln Y_i + \varepsilon_i$$

    β_1 is interpreted as $\%\triangle$ in Q from a 1% $\triangle$ in P, conditional on (log) Y; That is, b_1 is an estimate of the Price-Elasticity of Demand
  - In the **Log-Linear Model**

    $$\ln Q_i = \beta_0 + \beta_1 P_i + \beta_2 Y_i + \varepsilon_i$$

    $\beta_1 * 100$ is interpreted as $\%\triangle$ in Q from a 1 unit $\triangle$ in P, conditional. on Y.

Functional Forms - Quadratic

- One might assume that people are more price-elastic at higher prices
- In this case, the price elasticity of demand is dependent on price

- A model of demand with a **Quadratic term** in price

$$\ln Q_i = \beta_0 + \beta_1 \ln P_i + \beta_2 \ln Y_i + \beta_3 \ln {P_i}^2 + \varepsilon_i$$

- The price-elasticity of demand is

$$\frac{\partial \ln Q}{\partial \ln P} = \beta_2 + 2\beta_3 \ln P_i$$

and thus price-elasticty changes as the price level changes

Functional Forms - Interaction Terms

- One might assume that markets with higher income are less price-elastic than those with lower income
- In this case, the price elasticity is dependent on the level of income

- A model of demand with an **Interaction term** between price and income

$$\ln Q_i = \beta_0 + \beta_1 \ln P_i + \beta_2 \ln Y_i + \beta_3 \ln P_i * \ln Y_i + \varepsilon_i$$

- The price-elasticity of demand is

$$\frac{\partial \ln Q}{\partial \ln P} = \beta_1 + \beta_3 \ln Y_i$$

and thus price-elasticity changes as income changes

Functional Forms - Dummy Variables

- Suppose we believed demand in cities is higher than demand in towns.
- Define the **Dummy Variable** CITY by

$$CITY_i = \begin{cases} 1 & \text{if market-}i\text{ is a city} \\ 0 & \text{otherwise} \end{cases}$$

- A model of demand with a City-**Dummy**

$$\ln Q_i = \beta_0 + \delta_0 CITY_i + \beta_1 \ln P_i + \beta_2 \ln Y_i + \varepsilon_i$$

  - The regression for towns vs cities

$$\ln Q_i = \beta_0 + \beta_1 \ln P_i + \beta_2 \ln Y_i + \varepsilon_i \quad \text{vs} \quad \ln Q_i = (\beta_0 + \delta_0) + \beta_1 \ln P_i + \beta_2$$

    - β_0 is intercept for towns (**omitted category**).
    - $\beta_0 + \delta_0$ is intercept for cities

Functional Forms - Time Trends

- The use of data with a time component (both Time-Series and Panel Data) allow us to control for unobserved **trending variables** or **secular effects**

- Consider the demand model with time series data

$$\ln Q_t = \beta_0 + \beta_1 \ln P_t + \beta_2 \ln Y_t + \beta_3 t + \varepsilon_t$$

- Recall that the data for this model come from a single market that is observed over successive periods.
- The **time-trend** t, which is nothing more then the obervation number, is included to control unobserved factors that are growing at a constant rate – i.e. trending – over time.
- Such factors – such as population change – are sometimes referred to as "secular effects"

Functional Forms - Fixed Effects

- The use of panel data allows us to control for '**unobserved heterogeneity** when this heterogeneity is time-invariant
- Consider the demand model with panel data

$$\ln Q_{it} = \beta_0 + \beta_1 \ln P_{it} + \beta_2 \ln Y_{it} + u_i + \varepsilon_{it}$$

  where u_i is an unobserved component that affects market i and is constant over time. We call u_i the **Fixed Effect** of market i. Since u_i is unobserved, it cannot be directly controlled.

- However, since we observe each market i at multiple points in time, we can include a series of dummy variables – one for each market – to indirectly serve as controls for these Fixed Effects
- Define the market-j dummy by:

$$D_{it}^{j} = \begin{cases} 1 & \text{if observation } i, t \text{ is from market-}j \\ 0 & \text{otherwise} \end{cases}$$

Functional Forms - Panel Data Model: Fixed Effects

- The Fixed Effects model

$$\ln Q_{it} = \beta_0 + \beta_1 \ln P_{it} + \beta_2 \ln Y_{it} + u_1 D^1_{it} + u_2 D^2_{it} + \ldots + u_M D^M_{it} + \varepsilon_{it}$$

  - That is, the Fixed Effects model allows each market to have its own intercept
  - Formally, the effects from the unobserved heterogeneity are treated as the coefficients of the market-specific dummy variable.
  - Intuitition: each market serves as a control for itself
    - Since the u_i varies over markets but not over time the identity of market i is sufficient to control for u_i
    - Thus, unobserved heterogeneity will be absorbed by the market dummies
  - Had we not accounted for these fixed effects, we could have attributed the change in Q_t generated by this unobserved heterogeneity mistakenly to P_t and Y_t, leading to endogeneity bias

References

- Wooldridge (2008). *Introductory Econometrics: A Modern Approach, 4th Edition*. South-Western College Publishers.
    - Chapter 2
    - Chapter 3, Sections 3.1-3.4
    - Chapter 4
    - Chapter 6, Sections 6.1-6.2
    - Chapter 7, Sections 7.1-7.4