ECO 310: Empirical Industrial Organization
Tutorial 1 - Review of Econometrics

Victor Aguirregabiria

September 8, 2022

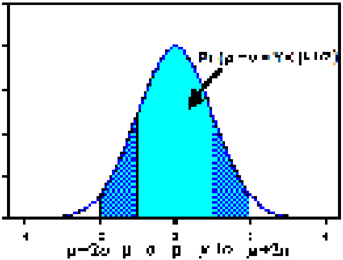
References

- 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

Random Variables and their Distribution [1/3]

- Let Y be a **random variable** (r.v.)
  - That is, the value of Y is subject to variations due to chance
  - Example: Y = Price of gasoline (in gas stations in the Great Toronto Area)

- The set of possible values of Y, and the probability at which it takes on these values is described by the **density function** of $f(Y)$:



Random Variables and their Distribution [2/3]

- The **mean** μ of Y is the expected value of the distribution of Y

$$\mu = \mathbb{E}[Y] = \sum_{y} y \ f(y)$$

- The **variance** σ^2 of Y measures the spread in the distribution of Y.

$$\sigma^2 = \mathbb{E}[(Y - \mu)^2] = \sum_{y} (y - \mu)^2 \ f(y)$$

- We often deal with r.v.'s that are generated from an unknown distribution. We do not know μ and σ^2.

- Then, we use a random sample of Y to obtain estimates of μ and σ^2.

Random Variables and their Distribution [3/3]

- Let $\{y_i : i = 1, ..., N\}$ be a random sample of observations on Y.
- Example: Random sample of $N = 100$ gas stations in the GTA.

- Estimators of the population mean and variance are

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

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

- **Law of Large Numbers (LLN):** $\overline{y}$ is a consistent estimator of μ.
  As $N \rightarrow \infty$, $\Pr(|\overline{y} - \mu| > 0) \rightarrow 0$.

Estimating Causal Relationships [1/2]

- In economics, we are often interested in the causal relationship between an explanatory variable X and an outcome variable Y.

- **Example:** $X =$ Price of gasoline in a gas station; $Y =$ Quantity sold of gasoline in the gas station.

  Suppose that we can change the price from $X = \$1.0/liter$ to $X = \$1.5/liter$. What is the effect on Y?

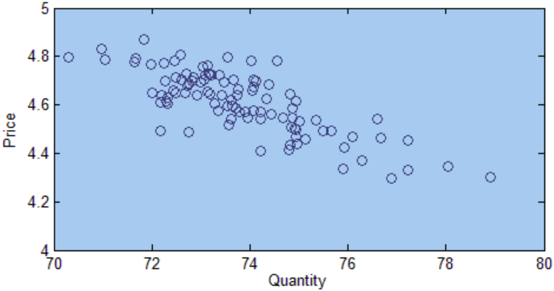
- Let $\{x_i, y_i : i = 1, ..., N\}$ be a random sample of observations on (X, Y).

  Example: A random sample of price and quantity at $N = 100$ gas stations in the GTA.

- We use this sample to estimate the effect of price (X) on quantity (Y).

Estimating Causal Relationships [2/2]

- A **scatter-plot** of the sample $\{x_i, y_i : i = 1, ..., N\}$ is a useful way of depicting the relationship between two r.v.'s



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

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

- **But covariance/correlation does not imply causation !!!**

The Linear Regression Model

- The **Linear Regression Model**

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

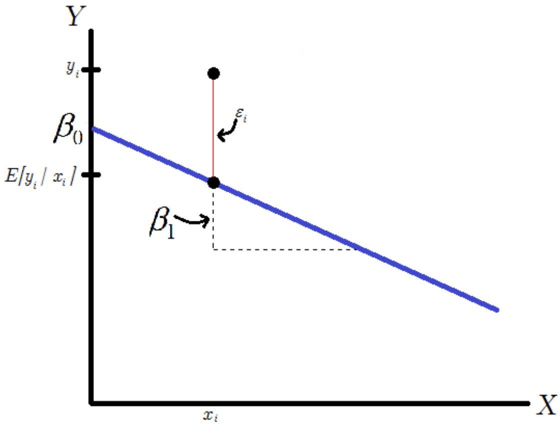
- **Parameter** β_k = causal effect of x_k on y, holding *all* other vars. fixed

- ε captures all other factors that affect y aside from $x_1, x_2, ..., x_k$
- This error term is included because:
  - Some relevant variables are unobservable.
  - Even if observable, might be subject to Measurement Error.

The Simple Linear Regression Model

- The Simple Linear Regression Model

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



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

Functional Forms

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

LRM and Data

- In this **Linear Regression Model**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \qquad \varepsilon \sim (0, \sigma^2)$$

$\beta_0, \beta_1, \beta_2, ..., \beta_k$ and σ^2 are unknown parameters.

- The purpose of our econometric analysis is to estimate these parameters

- Towards this end, suppose we have collected a random sample of data

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

Data

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

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

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

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

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

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

The Data Cont.

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

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

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
⋮			
Cranbrook	1990	86.00	1,000
Cranbrook	2010	123	10,000

Assumptions

- We want to *estimate* the causal effect of k explanatory variables $x_1, x_2, ..., x_k$ on the dependent variable y.
- 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. Notationally, we use
    - i for cross-sectional data
    - t for time series data
    - it for panel 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 (key assumption)**: 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. **No perfect collinearity between explanatory variables**

Estimation

- $\beta_0, \beta_1, ...\beta_k$ are unknown population parameters.
- But, if we have a sample of data $\{y_i, x_{1i}, ...x_{ki} : i = 1, ..., N\}$ can estimate them
- 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
- Choose parameters to minimize the sum of residuals $\sum_{i=1}^{n}(y_i - \hat{y}_i)$
  - Doesnt 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)
- Choose parameters to minimize $\sum_{i=1}^{n} |(y_i - \hat{y}_i)|$
  - "Least absolute value regression" - this is seldom used
- 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

- Our goal is to **estimate** the unknown parameters of our model.
- 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.

- Our goal is to **estimate** the unknown parameters of our model.
- The most common estimator in econometrics is **Ordinary Least Squares**
  - 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

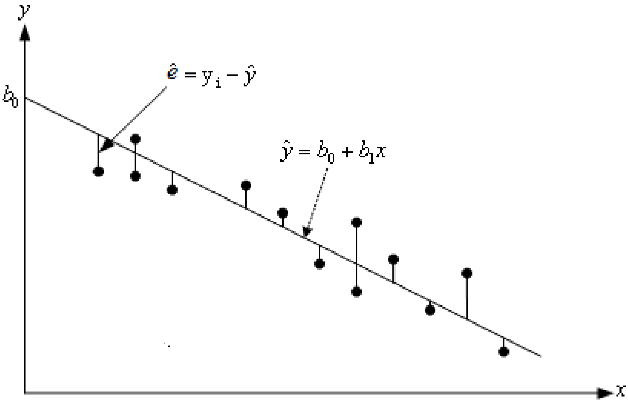
$$b_1 = \frac{\sum_{i-1}^{N}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i-1}^{N}(x_i - \overline{x})^2} \qquad b_0 = \bar{y} - b_1 \bar{x}$$

  - 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 that a line "best fits" the data in the least squares sense.
  - This results in a slope parameter of

$$b_1 = \frac{cov(x, y)}{var(x)}$$

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.
- 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**
  - 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 **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.

- **Hypothesis Testing Cont.:**
- Where 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.

- **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 of the T-distribution as determined by the computed value of the t-stat.
  - It measures the strength of the evidence against the Null Hypothesis.
  - Thus, we can equivalently reject the Null in favour of the Alternative if the P-Value of the test is less than our level of significance

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

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 variable 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 that 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

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

  - By construction

$$SST = SSR + SSE$$

Goodness of Fit

- How much of y is explained by x_1, x_2, ... ,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 =      5
-------------+------------------------------           F(  2,    47) =    9.3
       Model |  24.0503982     2  12.0251991           Prob > F      =  0.000
    Residual |  60.2333272    47  1.28156015           R-squared     =  0.285
-------------+------------------------------           Adj R-squared =  0.254
       Total |  84.2837254    49  1.72007603           Root MSE      =  1.132

------------------------------------------------------------------------------
         lnQ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnP |  -.9464336   .5006762    -1.89   0.065    -1.953664     .06079
         lnY |   1.806263   .4239203     4.26   0.000     .9534459    2.65908
       _cons |   10.70829   .6591015    16.25   0.000     9.382345    12.0342
------------------------------------------------------------------------------
```

An Example - Results

```
. reg lnQ lnP lnY

      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(  2,    47) =     9.3
       Model |  24.0503982     2  12.0251991           Prob > F      =  0.000
    Residual |  60.2333272    47  1.28156015           R-squared     =  0.285
-------------+------------------------------           Adj R-squared =  0.254
       Total |  84.2837254    49  1.72007603           Root MSE      =  1.132

------------------------------------------------------------------------------
         lnQ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval
-------------+----------------------------------------------------------------
         lnP |  -.9464336   .5006762    -1.89   0.065    -1.953664     .06079
         lnY |   1.806263   .4239203     4.26   0.000     .9534459    2.65908
       _cons |   10.70829   .6591015    16.25   0.000     9.382345    12.0342
------------------------------------------------------------------------------
```

b

se(b)

An Example - Results

```
. reg lnQ lnP lnY

      Source |       SS       df       MS              Number of obs =      5
-------------+------------------------------           F(  2,    47) =    9.3
       Model |  24.0503982     2  12.0251991           Prob > F      =  0.000
    Residual |  60.2333272    47  1.28156015           R-squared     =  0.285
-------------+------------------------------           Adj R-squared =  0.254
       Total |  84.2837254    49  1.72007603           Root MSE      =  1.132

------------------------------------------------------------------------------
         lnQ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnP |  -.9464336   .5006762    -1.89   0.065    -1.953664     .06079
         lnY |   1.806263   .4239203     4.26   0.000     .9534459    2.65908
       _cons |   10.70829   .6591015    16.25   0.000     9.382345    12.0342
------------------------------------------------------------------------------
```

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 =       5
-------------+------------------------------           F(  2,    47) =     9.3
       Model |  24.0503982     2  12.0251991           Prob > F      =  0.000
    Residual |  60.2333272    47  1.28156015           R-squared     =  0.285
-------------+------------------------------           Adj R-squared =  0.254
       Total |  84.2837254    49  1.72007603           Root MSE      =  1.132

------------------------------------------------------------------------------
         lnQ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval
-------------+----------------------------------------------------------------
         lnP |  -.9464336   .5006762    -1.89   0.065    -1.953664     .06079
         lnY |   1.806263   .4239203     4.26   0.000     .9534459    2.65908
       _cons |   10.70829   .6591015    16.25   0.000     9.382345    12.0342
------------------------------------------------------------------------------

    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     .06079
         lnY |   1.806263   .4239203     4.26   0.000     .9534459    2.659081
       _cons |   10.70829   .6591015    16.25   0.000     9.382345    12.0342
------------------------------------------------------------------------------
```

k

n−k−1

n

SSR SSE SST

```
. reg lnQ lnP lnY

      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(  2,    47) =     9.3
       Model |  24.0503982     2  12.0251991           Prob > F      =  0.000
    Residual |  60.2333272    47  1.28156015           R-squared     =  0.285
-------------+------------------------------           Adj R-squared =  0.254
       Total |  84.2837254    49  1.72007603           Root MSE      =  1.132

------------------------------------------------------------------------------
         lnQ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval
-------------+----------------------------------------------------------------
         lnP |  -.9464336   .5006762    -1.89   0.065    -1.953664     .06079
         lnY |   1.806263   .4239203     4.26   0.000     .9534459    2.65908
       _cons |   10.70829   .6591015    16.25   0.000     9.382345    12.0342
------------------------------------------------------------------------------
```

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

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

An Example - Results

```
. reg lnQ lnP lnY

      Source |       SS       df       MS              Number of obs =      5
-------------+------------------------------           F(  2,    47) =    9.3
       Model |  24.0503982     2  12.0251991           Prob > F      =  0.000
    Residual |  60.2333272    47  1.28156015           R-squared     =  0.285
-------------+------------------------------           Adj R-squared =  0.254
       Total |  84.2837254    49  1.72007603           Root MSE      =  1.132

------------------------------------------------------------------------------
         lnQ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnP |  -.9464336   .5006762    -1.89   0.065    -1.953664     .06079
         lnY |   1.806263   .4239203     4.26   0.000     .9534459    2.65908
       _cons |   10.70829   .6591015    16.25   0.000     9.382345    12.0342
------------------------------------------------------------------------------
```

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

R-Sq and Adj R-Sq

An Example - Results

```
. reg lnQ lnP lnY
```

Source	SS	df	MS
Model	24.0503982	2	12.0251991
Residual	60.2333272	47	1.28156015
Total	84.2837254	49	1.72007603

Number of obs	=	5
F(2, 47)	=	9.3
Prob > F	=	0.000
R-squared	=	0.285
Adj R-squared	=	0.254
Root MSE	=	1.132

lnQ	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnP	-.9464336	.5006762	-1.89	0.065	-1.953664	.06079
lnY	1.806263	.4239203	4.26	0.000	.9534459	2.65908
_cons	10.70829	.6591015	16.25	0.000	9.382345	12.0342

- 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 controling 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

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

- Consider the case of the demand function for a good.
- Suppose we wanted to estimate the relationship between quantities demanded Q, price P, 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

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

    - 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"
- Had we not included the time trend, and had our included regressor variables P_t and Y_t been "trending" themselves, we could have **spuriously** attributed that change in Q_t generated by these secular effects mistakenly to P_t and Y_t.

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

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

  - 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