Unit 1

Content

1.16 Measuring the Goodness of fit

1.17 Log functional form

- The Log-normal distribution

1.18 Testing normality of the error terms

1.19 Changing the scale of the data

Intoduction: The simple linear regresion model

Firstable we have to answer to the question "what is econometrics?"

Economerics is a field that concerns with the application of mathematical statistics and the tools of statistical inference to the empirical measerument o economic relationships.

Example: Denabd funtion for edible chicken

$$q_i^d = f(p_i)$$

$$ln(q_i^d) = \beta_1 + \beta_2 ln(p_i) + e_i$$

$$q_i^d = exp(\beta_1 + \beta_2 ln(p_i) + e_i)$$

$$\frac{dlnq^d}{dlnp} = \beta_2 \approx \frac{\Delta\%q^d}{\Delta\%p}$$

if cumulative growth rate = 4.25% in 6 years, then annual growth rate is equal to

$$[[(1 + 0.425)^{(1/6)}] - 1] \cdot 100$$

Theory is fundamental of an econometric model

For example:
The wage-income function (Mincer)

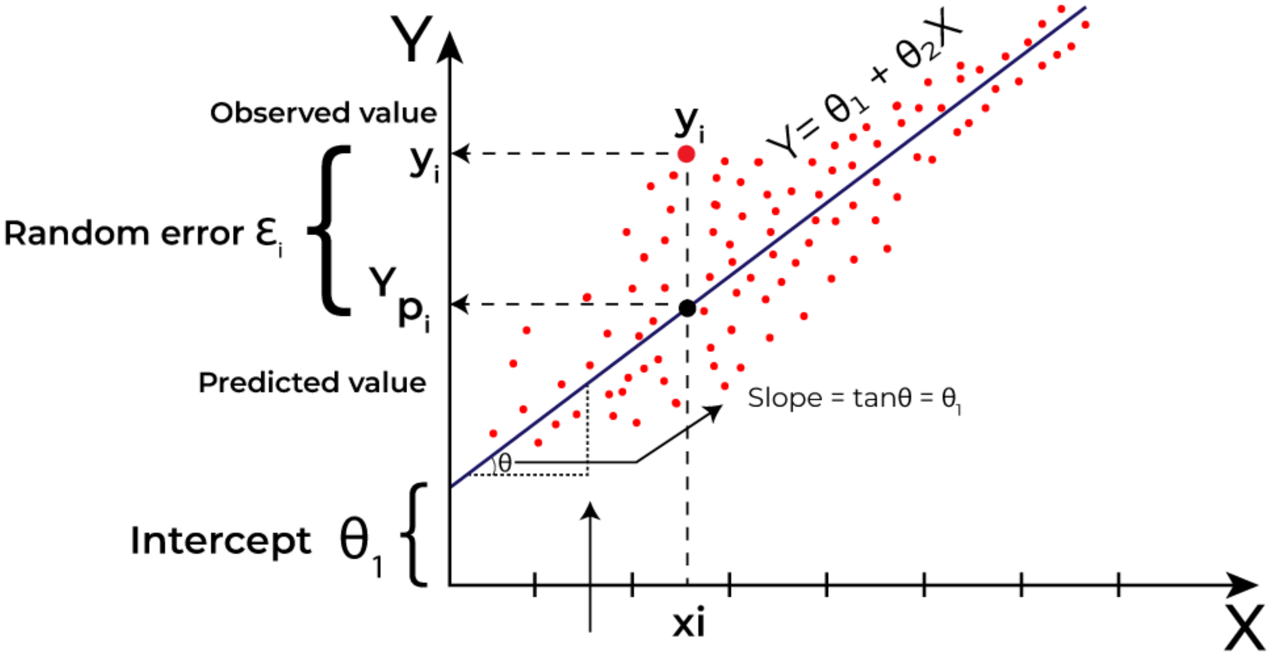
$$W = f(educ, exper)$$

Return of share function (return on financial asset)

$$Rcemex_t = (Rm_t, h_t)$$

What is an econometric model?

it begins with a set of theorical propositions. The theory specifies a set of precise, deterministic relationships among variables



The econometric model to estimate the consumption function is:

$$C = \beta_1 + \beta_2 X + e_i$$

it is form of two components, the systematic one that is by itself an "economic model" and the no systemtic component (random component) which makes the model an economeric model.

Systematic component	$\beta_1 + \beta_2 X$
Random component	e_i

What type of data may we have?

1. **Cross sectional data**: It refers to observations of variables for diferent individuals, entities, firms, countries, etc., but observed at the same moment in time.
2. **Time series data**: It refers to collected data for a specific variable over time
3. **Panel data**: It consists in a combination of cross-sectional data and time series.

An econometric model is defined as follows:

$$\underbrace{c_t}_{\text{Dependent variable}} = \underbrace{\beta_1 + \beta_2 y_t}_{\text{Systematic component}} + \underbrace{e_t}_{\text{Random error}}$$

$$\underbrace{c_t}_{\text{Dependent variable}} = \underbrace{\beta_1 + \beta_2 \underbrace{y_t}_{\text{explanaory variable}}}_{\text{Deterministic part given by theory (systematic)}} + \underbrace{e_t}_{\text{Stochastic part (random)}}$$

Assumptions of Simple Linear Regression Model:

1. **Linear Model**: All data (y_i, x_i) collected from the population satisfy the relathionship

$$y_i = \beta_1 + \beta_2 x_i + e_i \qquad i = 1, 2, 3, \ldots, N$$

2. **Strict Exogeneity**: The conditional expected value of the random error term (e_i) is zero, that is:

```
$$
E(e_i|x_i)=0 \quad \forall  \quad i
$$
That means that we cannot use $x_i$ to predict $e_i$
Taking expected value:$$
```

E(y*i|x_i)=E[\beta_1+\beta_2x_i +e_i|x_i]=\beta_1+\beta_2x_i +\underbrace{E(e_i|x_i)}*\text{0}

$$Then$$

y_i=E(y_i|x_i)+e_i
$$

3. **Conditional Homoskedasticity**: The conditional variance of the error term is constant

$$Var(e_i|x_i) = \sigma^2$$

Because e_i is the random part of y, we may say that the latter may determine the statistical properties of the former. As a consequence, the homoskedasticity assumption can also be expressed in terms of the random variable y.

$$Var(y_i|x_i) = \sigma^2$$

4. **Conditionally Uncorrelated Errors**: The conditional covariance of random errors e_i and e_j is zero; that is:

$$Cov(e_i, e_j|x) = 0 \quad \forall \quad i \neq j$$

```
where $x=x_1,x_2,...,x_n$
```

5. **Explanatory variables must vary**: This means that the explanatory variables must take on at least two values; in other words, an explanatory variable cannot be a constant

$$x_i \neq c$$

6. **Error Normality**: The conditional distribution of the random errors is normal with zero mean and variance σ^2; that is:

$$e_i|x \sim N(0, \sigma^2)$$

```
Because $e_i$ is the random part in the regression funcition,
```

$$y_i = \beta_1 + \beta_2 x_i + e_i$$

Then:

$$y_i | x_i \sim N[\beta_1 + \beta_2 x_i; \sigma^2]$$

If we take continous change in x_i

Supuesto	Consecuencia si no se cumple	Ejemplo sencillo
Linealidad	El modelo está mal especificado, las predicciones son erróneas.	El gasto en comida crece al inicio con el ingreso, pero luego se estanca (curva en vez de recta).
Exogeneidad estricta	Los estimadores son **sesgados y engañosos**.	Familias con más ingreso también tienen más gusto por comer fuera; el efecto se confunde con el ingreso.
Homoscedasticidad	Estimadores correctos en promedio, pero las pruebas estadísticas e intervalos de confianza dejan de ser confiables.	Ingresos bajos: poco margen de gasto. Ingresos altos: gasto muy variable (unos en lujos, otros no).
Errores no correlacionados	Los estimadores pierden eficiencia y los test estadísticos se vuelven dudosos.	En series de tiempo, el gasto de este mes depende del gasto del mes pasado (errores arrastrados).
Variación en x	Imposible estimar la pendiente del modelo.	Todas las familias tienen el mismo ingreso, no se puede ver cómo cambia el gasto.
Normalidad de los errores	Con muestras grandes no pasa nada; con muestras pequeñas los test e intervalos pueden ser incorrectos.	Una familia gasta muchísimo más o menos de lo esperado (errores con colas muy pesadas).

The Least Square Principle

It consists on fitting a line to the data values so that the sum of the squares of the vertical distances from each point to the line is as small as possible.

The distances are squared to prevent large positive distances from being canceled by large negative distances.

The estimators obtaine through this method are known as least squares estimators.

$$\min_{\beta_1,\beta_2} = \sum_{i=1}^{n} e_i^2$$

The estimators obtained through this methodology are called **Ordinary Least Squares Estimators (OLS)**

Note that

$$e_i = y_i - \hat{y_i} \rightarrow e_i = y_i - (b_1 + b_2 x_i) \rightarrow e_i = y_i - b_1 - b_2 x_i$$

Sustituting e_i into the objective function:

$$\min_{\beta_1,\beta_2} \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - \beta_1 - \beta_2 x_i]^2$$

First Condition Orders:

$$\frac{S(\beta_1,\beta_2)}{\beta_1} = 2\sum_{i=1}^{n} [y_i - \beta_1 - \beta_2 x_i](-1) = 0 \quad (1)$$

$$\frac{S(\beta_1,\beta_2)}{\beta_2} = 2\sum_{i=1}^{n} [y_i - \beta_1 - \beta_2 x_i](-x_i) = 0 \quad (2)$$

This values b_1 and b_2 that solve he system of 2 equations (1) and (2) are the Least Squares EStimators.

That is taking (1) we solve for b_1

$$\sum_{i=1}^{n} \quad [y_i - b_1 - b_2 x_i] = 0$$

$$\sum_{i=1}^{n} y_i - b_1 \underbrace{\sum_{i=1}^{n}(1)}_{\text{n}} - b_2 \sum_{i=1}^{n} x_i = 0$$

$$\sum_{i=1}^{n} y_i - n b_1 - b_2 \sum_{i=1}^{n} x_i = 0$$

$$\sum_{i=1}^{n} y_i - b_2 \sum_{i=1}^{n} x_i = n b_1$$

Hence

$$b_i = \underbrace{\frac{\sum_{i=1}^{n} y_i}{n}}_{\bar{y}} - b_2 \underbrace{\frac{\sum_{i=1}^{n} x_i}{n}}_{\bar{x}}$$

Therefore

$$b_1 = \bar{y} - b_2 \bar{x} \Rightarrow \text{LS estimator for } \beta_1$$

From eqtn (a) we solve for b_2

$$\sum_{i=1}^{n} [y_i - b_1 - b_2 x_i] x_i = 0$$

$$\sum_{i=1}^{n} x_i y_i - b_1 \sum x_i + b_2 \sum x_i^2 = 0$$

but we know that

$$\boxed{b_1 = \bar{y} - b_2 \bar{x}}$$

and

$$\frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}$$

$$\Rightarrow n\overline{x} = \sum x_i$$

Hence:

$$\sum x_i y_i - (\bar{y} - b_2 \bar{x}) n\overline{x} - b_2 \sum x_i^2 = 0$$

$$\sum x_i y_i - n\overline{x}\overline{y} + \underbrace{b_2 n\overline{x}^2 - b_2 \sum x_i^2}_{-b_2\left(\sum x_i^2 - n\overline{x}^2\right)} = 0$$

We get

$$\sum x_i y_i - n\overline{x}y - b_2 \left(\sum x_i^2 - n\overline{x}^2\right) = 0$$

$$b_2 = \frac{\sum x_i y_i - n\overline{x}\overline{y}}{\sum x_i^2 - n\overline{x}^2} \Rightarrow \text{LS estimator for } \beta_2$$

There are two more algebraic expressions for (b_2), that allow us to work easily when making proofs.

The first re-expression of the numerator of (b_2) is

$$\sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y} = \sum_{i=1}^{n} \left[(x_i - \overline{x})(y_i - \overline{y})\right]$$

Proof

$$\sum \left[(x_i - \overline{x})(y_i - \overline{y})\right]$$

$$= \sum \left[x_i y_i - \overline{y} x_i - \overline{x} y_i + \overline{x}\overline{y}\right]$$

$$= \sum [x_1 y_1] - \overline{y} \sum x_i - \overline{x} \sum y_i + -n\overline{x}\overline{y_1}$$

$$= \sum [x_i y_i] - n\overline{x}\overline{y} - n\overline{x}\overline{y} + -n\overline{x}\overline{y}$$

Therefore

$$\sum[(x_i-\overline{x})(y_i-\overline{y})]=\sum x_iy_i-n\overline{x}\overline{y}$$

The denominator of b_2 can also be re-expressed as:

$$\sum x_i^2-n\overline{x}^2=\sum_{i=1}^{n}(x_i-\overline{x})^2$$

Proof

$$\begin{aligned}\sum(x_i-\overline{x})^2&=\sum\left[x_i^2-2\overline{x}x_i+\overline{x}^2\right]\\&=\sum x_i^2-2\overline{x}\sum x_i+n\overline{x}^2\\&=\sum x_i^2-2n\overline{x}^2+n\overline{x}^2\\&=\sum x_i^2-n\overline{x}^2\end{aligned}$$

So, the 2nd expression for (b_2) is

$$b_2=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

But there is a 3^rd expression for the numerator of b_2:

$$\begin{aligned}\sum(x_i-\overline{x})(y_i-\overline{y})&=\sum[(x_i-\overline{x})y_i-(x_i-\overline{x})\overline{y}]\\&=\sum(x_i-\overline{x})y_i-\overline{y}\sum(x_i-\overline{x})\end{aligned}$$

But we know that

$$\begin{aligned}\sum(x_i-\overline{x})&=\left(\sum x_i\right)-n\overline{x}\\&=n\overline{x}-n\overline{x}=0\end{aligned}$$

Hence

$$\sum(x_i-\overline{x})(y_i-\overline{y})=\sum(x_i-\overline{x})y_i$$

Therefore, our 3^{rd} expression for b_2 is

$$\boxed{b_2=\frac{\sum(x_i-\overline{x})y_i}{\sum(x_i-\overline{x})^2}}$$

$$e_i=y_i-\underbrace{E(y_i|x_i)}_{\beta_1+\beta_2x_i}\quad\Rightarrow\quad e_i=y_i-\beta_1-\beta_2x_i$$

$$\min_{\beta_1,\beta_2}\sum_{i=1}^{n}(y_i-\beta_1-\beta_2x_i)^2=\sum e_i^2$$

$$\underbrace{estimators}_{\text{unction}}\neq\underbrace{estimates}_{\text{Values}}$$

$$\boxed{b_1 = \bar{y} - b_2\bar{x}}$$

$$\boxed{b_2 = \frac{\sum(x_i - \overline{x})y_i}{\sum(x_i - \overline{x})^2}}$$

We may report our results as

$$\hat{E(y_i|x_i)} = b_1 + b_2x_i \quad \Rightarrow \quad \text{dust regression line}$$

If $x_i = n$, then, find the predicted value of y_i

$$\hat{y_i} = \quad \dots$$

- b_1 is the estimator for β_1
- b_2 is the estimator for $\beta_2 \Rightarrow$ Marginal Effect of x on $E(y|x)$

b_2 is the estimated slope of the regression function, and it represent the change in y given a unit change in x.
And because b_2 is the derivative of y with respect to x, we may interpret b_2 as the marginal effect of x on y.

Using the Estimates to calculate Elasticities

The estimated linear regression model has another charateristics; the elasticities are different at each point on the regression line
The elasticity of y with respect to x is defined as follows:

$$\varepsilon = \frac{\Delta\%y}{\Delta\%x} = \frac{\Delta y/y}{\Delta x/x} = \frac{\Delta y}{\Delta x}\frac{x}{y}$$

En el contexto de LRM:

$$(y_i|x_i) = \beta_1 + \beta_2x_i$$

La elasticidad de $(y|x)$ con respecto a x:

$$\varepsilon = \frac{\Delta(y|x)}{\Delta x_i}\left[\frac{x_i}{(y_i|x_i)}\right]$$

Pero sabemos que:

$$\frac{\Delta(y_i|x_i)}{\Delta x_i} = \beta_2$$

Por lo tanto:

$$\varepsilon_i = \beta_2\left[\frac{x_i}{\beta_1 + \beta_2x_i}\right]$$

Elasticity changes every time we change x_i
Elasticity Estimator will be:

$$\hat{\varepsilon}_i = b_2 \left[\frac{x_i}{b_1 + b_2 x_i} \right]$$

EXAMPLE:

$$\begin{aligned} y_i &= 980.39 + (0.1397)(1500) \\ &= 1189.94 \\ \text{Para } x_i &= 1500 \\ \hat{\varepsilon}_i &= (0.1397)\frac{1500}{1189.94} \\ \hat{\varepsilon} &= 0.17 \end{aligned}$$

Interpretation: Given a 1% change, given an income of 1500, there will be a change in 0.17% in food expenditure.

Si $x_i = \overline{x} = 3,641.63$

Nota que si $x_i = \overline{x}$

$$\hat{y_i} = b_1 + b_2\overline{x}$$

pero

$$b_1 = \overline{y} - b_2\overline{x}$$

Sustituyendo b_1 en y_i:

$$\begin{aligned} \hat{y_i} &= (\overline{y} - b_2\overline{x}) + b_2\overline{x} \\ \Rightarrow \hat{y_i} &= \overline{y} \end{aligned}$$

Hence, the estimated elasticity at $x_i = \hat{x}$ will be:

$$\boxed{\hat{\varepsilon}_{\overline{x}} = b_x \left[\frac{\overline{x}}{\overline{y}} \right] \Rightarrow \text{Elasticity at the average}}$$

$$\hat{\varepsilon}_{\overline{x}} = (0.1397) \left[\frac{3,641.67}{1,489.31} \right] = 0.3395$$

And **Average Elasticity** is estimator will be:

$$\boxed{\hat{\overline{\varepsilon}} = \frac{\sum_{i=1}^{n} \left[b_2 \left[\frac{x_i}{b_1 + b_2 x_i} \right] \right]}{n} = \frac{\sum \hat{\varepsilon}_i}{n}}$$

The Log-Log Model: constant elasticity

We call log-log model a model which the logarhm apear on both sides of the equation:

$$y_i = e^{[\beta_1 + \beta_2 ln(x_i) + e_i]}$$

$$ln(y_i) = \beta_1 + \beta_2 ln(x_i) + e_i$$

This model has the charactheristics that β_2 is the elasticity of y with respect to x, because

$$\beta_2 = \frac{dln(y_i)}{dln(x_i)} \approx \frac{\Delta\%y}{\Delta\%x} = \varepsilon \Rightarrow \text{Constant Elasticity Model}$$

What is the slope of the regression function for the log-log model
Taking the derivative of y with respect to x

$$\frac{dy_i}{dx_i} = e^{[\beta_1+\beta_2 ln(x_i)+e_i]}[\beta_2\Big(\frac{1}{x_i}\Big)] \Rightarrow y_i\beta_2\Big(\frac{1}{x_i}\Big) = \beta_2\Big(\frac{y_i}{x_i}\Big)$$

Said it in other words:

$$\frac{dE(y_i|x_i)}{dx_i} = \beta_2\frac{E(y_i|x_i)}{x_i}$$

Using the estimators

$$\frac{dE(\hat{y_i}|x_i)}{dx_i} = b_2 \cdot \frac{\hat{y_i}}{x_i} \Rightarrow \text{Slope, Marginal effect of x on y}$$

$$\text{where } \hat{y_i} \approx \exp(b_1 + b_2 \ln(x_i))$$

Expresing the results as the fitted regression line we have:

$$\ln(y_i) = 3.279987 + 0.4876635 \ln(x_i)$$

Interpretation:

$b_2 = 0.488$ A one percent (1%) increase in the weekly family income, will increase the weekly food expenditure by 0.488 percent (0.488%).

We must clarify that what we have estimated is the log of y, that is, we've got $\ln(y_i)$.

In order to obtain the estimator of y it is necessary to apply the exponential function to both sides of the equation:

$$E(y|x) \approx \exp(\ln(y_i)) = \exp(b_1 + b_2 \ln(x_i))$$

Propiedades de los Estimadores de Mínimos Cuadrados

1) Los estimadores MCO son lineales

Los estimadores MCO son funciones lineales de la variable aleatoria y. Esto significa que b_1 y b_2 pueden expresarse como promedios ponderados de los valores y_i.

Demostración:

$$b_2 = \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2}$$

Sea $w_i = \frac{x_i-\bar{x}}{\sum(x_i-\bar{x})^2}$, entonces b_2 puede expresarse como:

$$b_2 = \sum_{i=1}^{N} w_i y_i$$

Dado que w_i es solo una función de las x_i (que, hasta ahora, se consideran no aleatorias), podemos decir que b_2 es un promedio ponderado de las y_i; y un promedio ponderado es una combinación lineal. Por lo tanto, b_2 es un estimador lineal.

2) Los estimadores MCO son insesgados

Un estimador es insesgado si su valor esperado es igual al parámetro verdadero que queremos estimar.

Para los estimadores MCO, esto significa que:

$$E(b_1|x) = \beta_1$$

$$E(b_2|x) = \beta_2$$

Demostración:

Para mostrar que los estimadores MCO son insesgados, necesitamos definir sus valores esperados correspondientes.

Para b_2, comenzamos con su definición:

$$b_2 = \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2}$$

$$b_2 = \frac{\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}\underbrace{(\beta_1 + \beta_2 x_i + e_i)}_{y_i} = \beta_1 \underbrace{\frac{\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}}_{0} + \beta_2 \underbrace{\frac{\sum(x_i - \bar{x})x_i}{\sum(x_i - \bar{x})^2}}_{1} + \frac{\sum(x_i - \bar{x})e_i}{\sum(x_i - \bar{x})^2}$$

Por lo tanto:

$$b_2 = \beta_2 + \frac{\sum(x_i - \bar{x})e_i}{\sum(x_i - \bar{x})^2}$$

Y recordando que $w_i = \frac{x_i - \bar{x}}{\sum(x_i - \bar{x})^2}$ podemos definir:

$$b_2 = \beta_2 + \sum w_i e_i$$

Entonces, el valor esperado condicional de b_2 es:

$$E(b_2|x) = \beta_2 + E(\sum w_i e_i|x) = \beta_2 + \sum w_i E(e_i|x)$$

Como hemos mencionado, w_i depende solo de x_i, por lo tanto, es un término no aleatorio. En consecuencia, w_i saldrá del operador de expectativa matemática.

Considerando el supuesto de exogeneidad estricta (supuesto #2 del MCRL) sabemos que $E(e_i|x) = 0$; así, si este supuesto se cumple:

$$E(b_2|x) = \beta_2 \quad \text{por lo que } b_2 \text{ es un estimador insesgado}$$

Following the same reasoning, we may perform the same proof for b_1):

$$b_1 = \bar{y} - b_2\bar{x}$$

$$b_1 = \frac{1}{N}\sum y_i - b_2\bar{x}$$

$$b_1 = \frac{1}{N}\sum y_i - b_2\bar{x} = \frac{1}{N}\sum(\beta_1 + \beta_2 x_i + e_i) - b_2\bar{x} = \beta_1 + \beta_2\frac{\sum x_i}{N} + \frac{\sum e_i}{N} - b_2\bar{x}$$

$$b_1 = \beta_1 + \beta_2\bar{x} + \frac{1}{N}\sum e_i - b_2\bar{x}$$

$$E(b_1|x) = \beta_1 + \beta_2\bar{x} + \frac{1}{N}\sum E(e_i|x) - \bar{x}E(b_2|x)$$

Once more, considering the strict exogeneity assumption we know that $E(e_i|x) = 0$; thus, if this assumption holds:

$$E(b_1|x) = \beta_1$$

b_1 is an unbiased estimator.

\end{document}

$$E(b_2|x) = E[\beta_2|x] + E\left[\frac{\sum((\underbrace{x_i}_{\text{deterministic}} - \bar{x})\underbrace{e_i}_{\text{random}})}{\sum(x_i - \bar{x})^2}\right]$$

$$E(b_2|x) = \beta_2 + \left[\frac{\sum\big((x_i - \bar{x})\underbrace{E(e_i|X)}_{0}\big)}{\sum(x_i - \bar{x})^2}\right] \Rightarrow \beta_2 \Rightarrow b2 \text{ is an unbiased estimator}$$

This is held by the strict exogenety assumption.

3) Los estimadores MCO son eficientes

Un estimador es **EFICIENTE** si su varianza es la más pequeña en relación con otros estimadores comparables.

Para los estimadores MCO, eficiencia significa que:

$$\text{Var}(b_1|x)$$

$$\text{Var}(b_2|x)$$

Son las más pequeñas en comparación con otros estimadores lineales e insesgados para β_1 y β_2. Esto es precisamente lo que muestra el **Teorema de Gauss-Markov**.

Demostración:
Para mostrar que los estimadores MCO son eficientes, debemos comenzar encontrando una expresión para sus varianzas.

The Variance of b_2

Recall that

$$b_2 = \beta_2 + \sum w_i e_i$$

Hence:

$$\text{Var}(b_2|x) = \text{Var}\{[\beta_2 + \sum w_i e_i]|x\} = \text{Var}\{(\sum w_i e_i)|x\}$$

And the variance of a linear combination of random variables multiplied by constants (considering that w_i is non-random) is:

$$\text{Var}\left\{\left(\sum w_i e_i\right)|x\right\} = \sum w_i^2 \text{Var}(e_i|x) + \sum\sum_{i \neq j} w_i w_j \text{Cov}(e_i e_j|x)$$

The Variance of b_2

Now, taking into account assumptions 3 and 4 of the SLRM (homoskedasticity and uncorrelated errors):

$$Var(e_i|x_i) = \sigma^2 \quad \forall i \quad \text{and} \quad Cov(e_i, e_j|x) = 0 \quad \forall i \neq j$$
$$Var(b_2) = \sum w_i^2 Var(e_i|x) + \sum\sum_{i \neq j} w_i w_j Cov(e_i e_j|x)$$

Hence:

$$Var(b_2|x) = \sigma^2 \sum w_i^2 = \sigma^2 \sum \left[\frac{x_i - \bar{x}}{\sum(x_i - \bar{x})^2}\right]^2 = \sigma^2 \frac{\sum(x_i - \bar{x})^2}{[\sum(x_i - \bar{x})^2]^2}$$

Finally:

$$Var(b_2|x) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

Following the same reasoning we may develop the (\text{Var}(b_1|x)).

$$Var(b_1|x) = Var(\bar{y} - b_2\bar{x})$$

But we know that the variance of the sample mean is $Var(\bar{y}) = \frac{\sigma^2}{N}$ and the variance of a random variable X multiplied by a constant a is $Var(aX) = a^2 Var(X)$. Consequently:

$$Var(b_1|x) = Var(\bar{y}) + (-\bar{x})^2 Var(b_2|x) = \frac{\sigma^2}{N} + \bar{x}^2 \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

Or

$$Var(b_1|x) = \sigma^2 \left[\frac{1}{N} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right]$$

This expression for $Var(b_1|x)$ will be used later in the context of the forecast error (so...keep it handy)

But because

$$\frac{1}{N}+\frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}=\frac{\sum(x_i-\bar{x})^2+N\bar{x}^2}{N\sum(x_i-\bar{x})^2}=\frac{\sum x_i^2}{N\sum(x_i-\bar{x})^2}$$

Hence, another convenient way of expressing the conditional variance of b_1 is:

$$Var(b_1|x)=\frac{\sigma^2\sum x_i^2}{N\sum(x_i-\bar{x})^2}$$

Similarly we may develop the estimators' covariance equation (given that b_1 and b_2 are correlated random errors).

The process to get the expression for Cov $(b_1, b_2|x)$ may be reviewed in Hill, et al (2018, cap.2) or Greene (2003).

Summarizing, the variances and covariance of the OLS estimators are expressed as:

$$\mathrm{Var}(b_1|x)=\sigma^2\frac{\sum x_i^2}{N\sum(x_i-\bar{x})^2}$$

$$\mathrm{Var}(b_2|x)=\frac{\sigma^2}{\sum(x_i-\bar{x})^2}$$

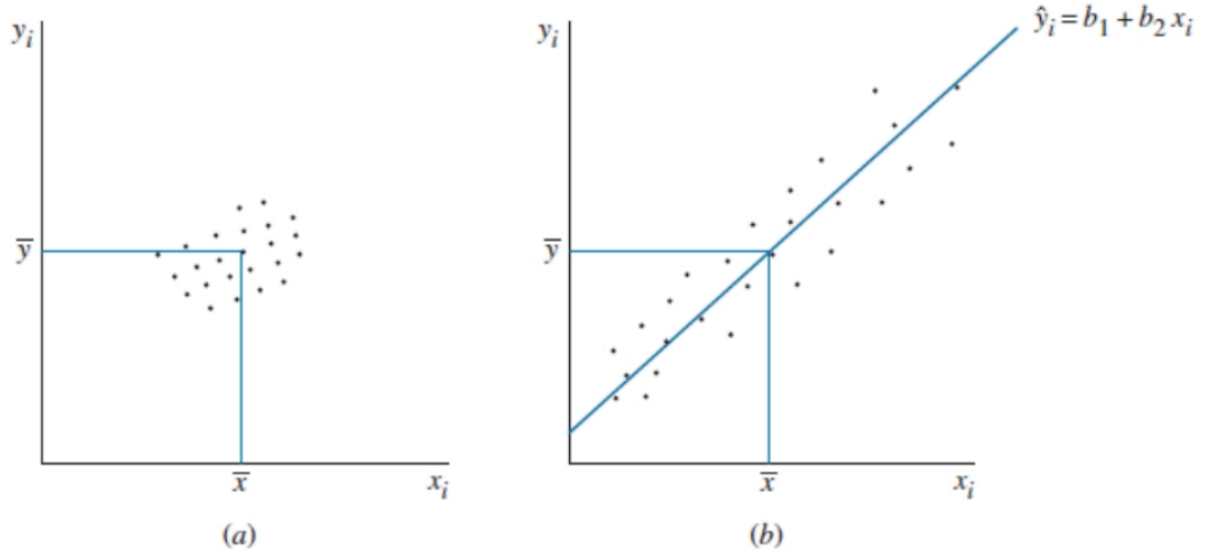
$$\mathrm{Cov}(b_1,b_2|x)=\sigma^2\left[\frac{-\bar{x}}{\sum(x_i-\bar{x})^2}\right]$$

The greater the variance σ^2, the information about β_1 and β_2 will be less precise and the larger the variances and covariances of the estimators.

The larger the sum of squares

$$\sum(x_i-\bar{x})^2$$

(variation of the explanatory variable around its mean), the smaller the conditional variances of the least squares estimators.



The Gauss-Markov Theorem

Under assumptions 1 to 5 of the linear regression model, the OLS estimators b_1 and b_2 have the smallest variance of all linear and unbiased estimators of β_1 and β_2. They are the Best Linear Unbiased Estimators (BLUE) of β_1 and β_2.

b_1 and b_2 are the best estimators because they have the smallest variance given the previously specified algebraic expressions, and this is the reason why they are efficient

The efficiency of the OLS estimators does not depend on the normality assumption of the error term.

Proof in class

4) The OLS estimators are Consistent

Consistency is an asymptotic property; that is, consistency is a property of estimators in the context of large samples.

This property refers to the fact that, as the sample size gets larger, the probability that the LS estimators b_1 and b_2 deviate from their true parameter values β_1 and β_2, approaches to zero.

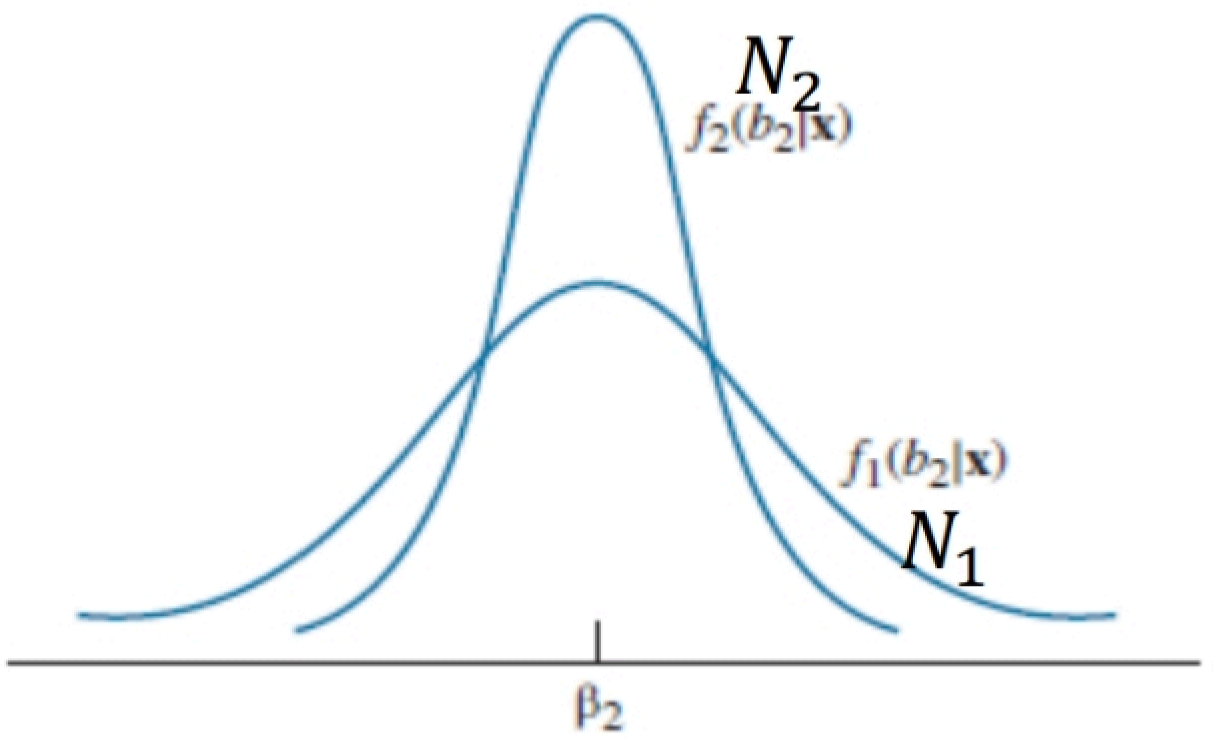
More specifically, the LS estimators b_1 and b_2 are said to be consistent if they converge in probability to β_1 and β_2. Algebraically, for any $\varepsilon > 0$:

$$\lim_{N\rightarrow} \Pr[|b_1 - \beta_1| > \varepsilon] = 0$$

$$\lim_{N\rightarrow} \Pr[|b_2 - \beta_2| > \varepsilon] = 0$$

$$\operatorname{plim}(b_1) = \beta_1$$

$$\operatorname{plim}(b_2) = \beta_2$$



This also means that, as the sample size gets larger, the probability distribution of the estimators becomes concentrated at points close to the true parameter values, implying that:

$$\lim_{N\to} Var(b_1|x) = 0$$

$$\lim_{N\to} Var(b_2|x) = 0$$

Probability Distributions of the Ordinary Least Squares Estimators

The sampling properties of the OLS estimators do not depend on the normality assumption 6 (Hill, et al, 2018). That is, we do not require the error terms to be normally distributed for the OLS estimators to be linear, unbiased, efficient and consistent.

If we include the assumption that the random errors e_i are normally distributed:

$$e_i \mid x \sim N(0, \sigma^2)$$

Then, the conditional probability distributions of the LS estimators are also normal (*ibid.*):

$$b_1 \mid x \sim N\left(\beta_1, \sigma^2 \frac{\sum x_i^2}{N\sum(x_i - \bar{x})^2}\right)$$

$$b_2 \mid x \sim N\left(\beta_2, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)$$

REMEMBER THAT THE FIRST ONE IS THE MEAN AND THE SECOND ONE IS THE VARIANCE

Knowing that (b_1) and (b_2) are jointly distributed variables correlated with each other, hence the conditional probability distribution of the LS estimators is a bivariate normal distribution and can be expressed as:

$$\begin{bmatrix} b_1 \mid x \\ b_2 \mid x \end{bmatrix} \sim N \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \begin{matrix} \sigma^2 \frac{\sum x_i^2}{N\sum(x_i-\bar{x})^2} & \sigma^2 \left[\frac{-\bar{x}}{\sum(x_i-\bar{x})^2}\right] \\ \sigma^2 \left[\frac{-\bar{x}}{\sum(x_i-\bar{x})^2}\right] & \frac{\sigma^2}{\sum(x_i-\bar{x})^2} \end{matrix}$$

Where the matrix (2×2 in this case)

$$\begin{matrix} \sigma^2 \frac{\sum x_i^2}{N\sum(x_i-\bar{x})^2} & \sigma^2 \left[\frac{-\bar{x}}{\sum(x_i-\bar{x})^2}\right] \\ \sigma^2 \left[\frac{-\bar{x}}{\sum(x_i-\bar{x})^2}\right] & \frac{\sigma^2}{\sum(x_i-\bar{x})^2} \end{matrix}$$

Is known as the variance-covariance matrix of the OLS estimators; this is a symmetric matrix whose main diagonal contains the variance of the coefficients and off main diagonal elements are the covariances.

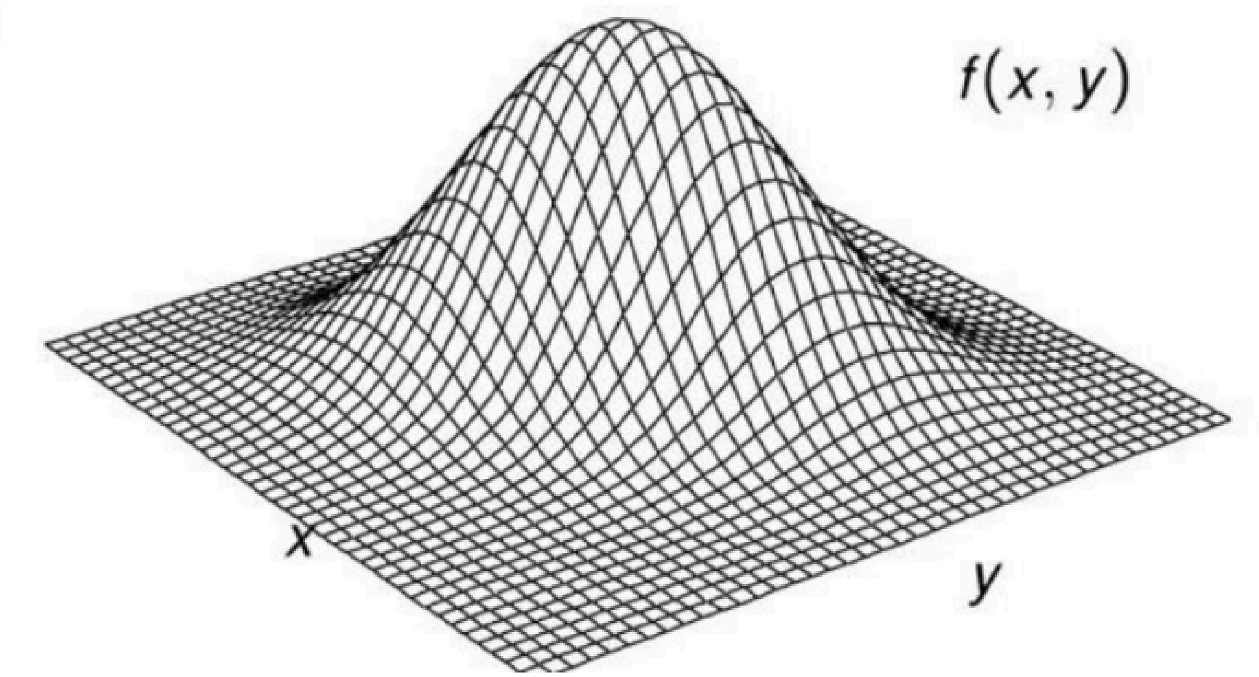
NOTE: Regarding the bivariate normal probability distribution

Definition 5.4
Two random variables X and Y are said to have a **bivariate normal distribution** with parameters $\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2$, and , if their joint PDF is given by

$$f_{XY}(x,y) = \frac{1}{2\sigma_X\sigma_Y\sqrt{1-{}^2}} \cdot \exp\left\{-\frac{1}{2(1-{}^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - 2\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right\}$$

(5.24)

where $\mu_X, \mu_Y \ \ , \sigma_X, \sigma_Y > 0$ and $\ (-1,1)$ are all constants.



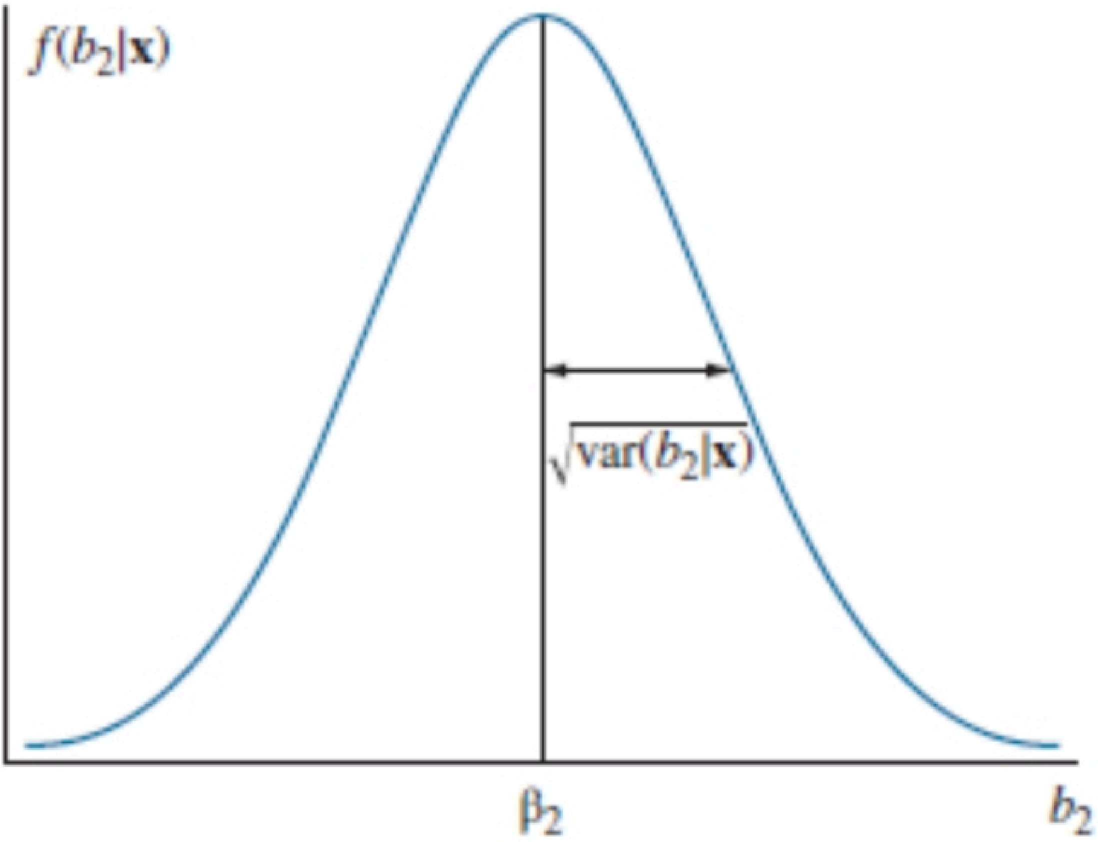
But, what happens if the random errors are not normally distributed? In such a case, what would be the probability distribution of the LS estimators?

The Central Limit Theorem allows us to say that:

If the sample size N is sufficiently large, then the least squares estimators have a distribution that approximates the normal distribution.

Knowing the probability distribution of the LS estimators is necessary for the following:

- Interval estimation
- Hypothesis testing



The variance of the error term

Previously, we have defined the algebraic expressions of the OLS estimators b_1 and b_2 and their corresponding variances and covariances. However, we have not defined how to estimate the variance of the error term σ^2, which appears in the LS variances and covariance expression.

In this section, we define an estimator for σ^2 which will be used to define the estimators of $Var(b_1|x)$, $Var(b_2|x)$ and $Cov(b_1, b_2|x)$.

We start with the general definition of the conditional variance of a random variable: its second moment minus its first moment squared; that is:

$$Var(e_i|x) = E(e_i^2|x) - [E(e_i|x)]^2$$

Now, if the second assumption of the linear regression model holds $E(e_i|x) = 0$; therefore; If strict exogeneity holds then:

$$Var(e_i|x) = E(e_i^2|x) = \sigma^2$$

Estimating the variance of the error term

By definition, the expected value is an average –weighted by probabilities–, then, the σ^2 estimator might be an average of the squared residuals (obtained after LS estimation):

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{N-2}$$

OLS estimator of the error term variance.

The degrees of freedom for this expression are $N-2$, since we are using estimated residuals which have two estimated parameters (we do not know the true error terms); that is, to calculate $\hat{e}_i$ we use two estimators (b_1) and (b_2), then, we lose two degrees of freedom:

$$\hat{e}_i = y_i - \hat{y}_i = y_i - b_1 - b_2 x_i$$

Estimating the variances and covariance of b_1 and b_2

Once the estimator of σ^2 has been defined, we may define the estimators of the variances and covariance of b_1 and b_2:

$$\boxed{\widehat{Var(b_1|x)} = \hat{\sigma}^2 \frac{\sum x_i^2}{N\sum(x_i-\bar{x})^2}}$$

$$\boxed{\widehat{Var(b_2|x)} = \frac{\hat{\sigma}^2}{\sum(x_i-\bar{x})^2}}$$

$$\boxed{\widehat{Cov(b_1,b_2|x)} = \hat{\sigma}^2 \left[\frac{-\bar{x}}{\sum(x_i-\bar{x})^2}\right]}$$

The square roots of the estimated variances are the **standard errors of the OLS coefficients**:

$$se(b_1) = \sqrt{\widehat{Var(b_1|x)}}$$

$$se(b_2) = \sqrt{\widehat{Var(b_2|x)}}$$

Estimating the variances and covariance of the LS estimators in the food-expenditure example

For our food-expenditure example, we may estimate the variance of the error term first, then used to estimate the variances of b_1 and b_2:

The estimated residuals are simply calculated by the difference between the observed and predicted food-expenditure:

$$\hat{e}_i = y_i - \hat{y}_i$$

The sum of the square residuals is:

$$\sum \hat{e}_i^2 = 14,280,944$$

And considering that we have N=30 sample observations, the estimated error term variance is:

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{N-2} = \frac{14,280,944}{30-2} = 510,033.715$$

The estimated variances of b_1 and b_2 are:

$$\widehat{Var(b_1|x)} = \hat{\sigma}^2 \frac{\sum x_i^2}{N\sum(x_i-\bar{x})^2} = (510033.715) \times \frac{541050450}{30 \times (143198366.7)} = 64235.828$$

$$se(b_1) = \sqrt{64235.828} = 253.44788$$

$$\widehat{Var(b_2|x)} = \frac{\hat{\sigma}^2}{\sum(x_i-\bar{x})^2} = \frac{510033.715}{143198366.7} = 0.00356173$$

$$se(b_2) = \sqrt{0.00356173} = 0.05968023$$

And the estimated covariance is:

$$\widehat{Cov(b_1,b_2|x)} = \hat{\sigma}^2 \left[\frac{-\bar{x}}{\sum(x_i-\bar{x})^2}\right] = (510033.715) \times \left[\frac{-3641.666667}{143198366.7}\right] = -12.970628$$

Note: Recall the symmetry of the covariance: $\mathrm{Cov}(b_1, b_2|x) = \mathrm{Cov}(b_2, b_1|x)$

Prueba de hipótesis

- 1.11 Interval estimation
- 1.12 Hypothesis testing
- 1.13 Confidence interval for a linear combination of parameters
- 1.14 Hypothesis testing for a linear combination of parameters

1.11 Interval estimation

Interval estimation for the SLRM parameters

In previous sections we have seen how to estimate the parameters of the SLRM:

$$y_i = \beta_1 + \beta_2 x_i + e_i$$

Where the estimators for β_1 and β_2 are:

$$b_1 = \bar{y} - b_2\bar{x}$$

$$b_2 = \frac{\sum(x_i-\bar{x})y_i}{\sum(x_i-\bar{x})^2}$$

These are the OLS point estimators

Now, including our knowledge of the conditional probability distribution of the estimators, we may define the confidence interval estimators for β_1 and β_2.

Hence, recalling that

$$b_1|x \sim N\left(\beta_1, \sigma^2 \frac{\sum x_i^2}{N\sum(x_i - \bar{x})^2}\right)$$

$$b_2|x \sim N\left(\beta_2, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)$$

We may work on probabilistic expressions to define Interval estimators for β_1 and β_2.

Taking the distribution of the random variable b_2, the standardized expression is:

$$Z = \frac{b_2 - \beta_2}{\sqrt{\frac{\sigma^2}{\sum(x_i - \bar{x})^2}}} \sim N(0, 1)$$

However, since we do not know the value of σ^2, we must use its estimator $\hat{\sigma}^2$, and this changes the distribution of the standardized variable to student-t distribution with $(N-2)$ degrees of freedom. Hence:

$$t = \frac{b_2 - \beta_2}{\sqrt{\frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2}}} \sim t_{(N-2)}$$

Or also:

$$t = \frac{b_2 - \beta_2}{se(b_2)} \sim t_{(N-2)}$$

The probabilistic expression we start off to define the Interval estimator for β_2 is:

$$Pr\left[-t_{\alpha/2} \leq \frac{b_2 - \beta_2}{se(b_2)} \leq t_{\alpha/2}\right] = 1 - \alpha$$

Reorganizing the expression we have the $(1-\alpha)$% confidence Interval for β_2:

$$Pr\left[b_2 - t_{\alpha/2} se(b_2) \leq \beta_2 \leq b_2 + t_{\alpha/2} se(b_2)\right] = 1 - \alpha$$

Lower Confidence Limit (LCL)

Upper Confidence Limit (UCL)

NOTE: It is very important to recognize that, in this expression, the variables are the lower and upper limits. When we change the sample, the limits will also change because b_2 and $se(b_2)$ change. This fact has important implications for the correct interpretation of the confidence Interval.

The confidence Interval can be expressed in the usual form:

$$(b_2 - t_{\alpha/2}se(b_2); b_2 + t_{\alpha/2}se(b_2)) \Rightarrow$$

or also:

$$(b_2 \pm t_{\alpha/2}se(b_2))$$

Following the same reasoning, we may define the corresponding interval estimator (confidence interval) for β_1:

$$Pr[b_1 - t_{\alpha/2}se(b_1) \leq \beta_1 \leq b_1 + t_{\alpha/2}se(b_1)] = 1 - \alpha$$

Alternatively:

$$(b_1 \pm t_{\alpha/2}se(b_1))$$

For our food-expenditure example and considering the obtained point estimates and their corresponding standard errors, we can calculate the interval estimates for β_1 and β_2 using a 95% confidence level:

Recalling that

$$b_1 = 980.39 \quad se(b_1) = 253.45$$

$$b_2 = 0.1397492 \quad se(b_2) = 0.0596802$$

Now, for $\alpha = .05$ and $N - 2 = 28$ degrees of freedom the critical value of the t statistic is:

$$t_{\alpha/2} = t_{0.05/2} = t_{0.025} = 2.0484071$$

Then:

Confidence Interval for β_1:

$$\begin{array}{l} (b_1 \pm t_{0.025}se(b_1)) \\ (980.39 - 2.0484071 \times 253.45;\ 980.39 + 2.0484071 \times 253.45) \\ (461.22;\ 1499.56) \end{array}$$

Interpretation: With a 95% confidence level, the limits (461.22; 1499.56) contain the intercept of the food-expenditure function.

Confidence interval for β_2:

$$(b_2 \pm t_{0.025}se(b_2))$$

$$(0.1397492 - 2.0484071 \times 0.0596802; 0.1397492 + 2.0484071 \times 0.0596802)$$

$$(0.01749985; 0.26199855)$$

Interpretation: With a 95% confidence level, the limits (0.0175 ; 0.262) contain the slope of the food-expenditure function.

Another interpretation: With 95% probability, the marginal effect of income on food-expenditure is contained within the limits 0.0175 and 0.262

Hypothesis Testing: another tool of inferential statistics

From inferential statistics courses we have learned that hypothesis testing is a **tool used to validate theories or beliefs** regarding the relationship between variables and the influence that the explanatory variable might have on the dependent variable.

So, for decision making in Economics, particularly for financial and business decisions and public policy, it is common to question specific values of certain parameters or a combination of them.

For example, in our food-expenditure model, we may ask if β_2 is larger than 0.15, implying that, for each 100 pesos increase in weekly family income, food expenditure would increase by more than 15 pesos.

In the case of a production function, we may ask, for example, if this function shows constant returns to scale.

In order to answer these questions, **we use the hypothesis testing procedure**, which takes information from the sample (empirical evidence) to contrast it with the theory or belief and obtain a conclusion: decision making.

Hypothesis Testing: Contrasting beliefs with evidence

Theory or Belief

VS

Empirical Evidence (sample data)

1.12 Hypothesis testing

The elements of hypothesis testing

Recall that in all hypothesis testing procedure we find 5 elements:

1) The Null Hypothesis

For example

$$\text{Ho: } \beta_2 \geq c$$

2) The Alternative Hypothesis

For example

$$\text{H1: } \beta_2 < c$$

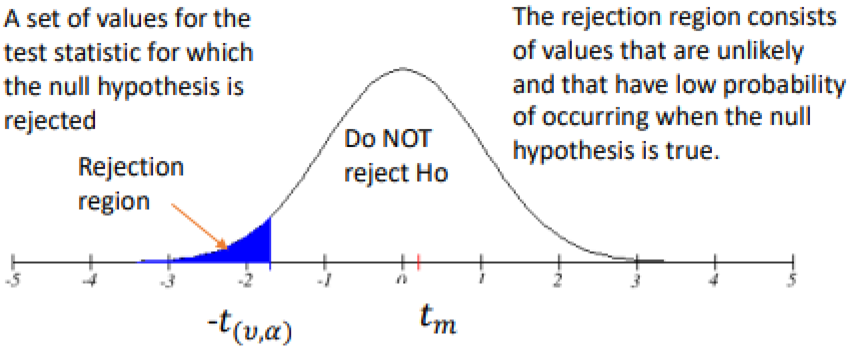
3) The test statistic and its distribution

For example, taking Ho as true, then $\beta_2 = c$

And:

$$t = \frac{b_2 - c}{se(b_2)} \sim t(N-2)$$

4) Rejection region of the Null Hypothesis



A set of values for the test statistic for which the null hypothesis is rejected

The rejection region consists of values that are unlikely and that have low probability of occurring when the null hypothesis is true.

5) Conclusion

For example:
There is not enough evidence to reject Ho.

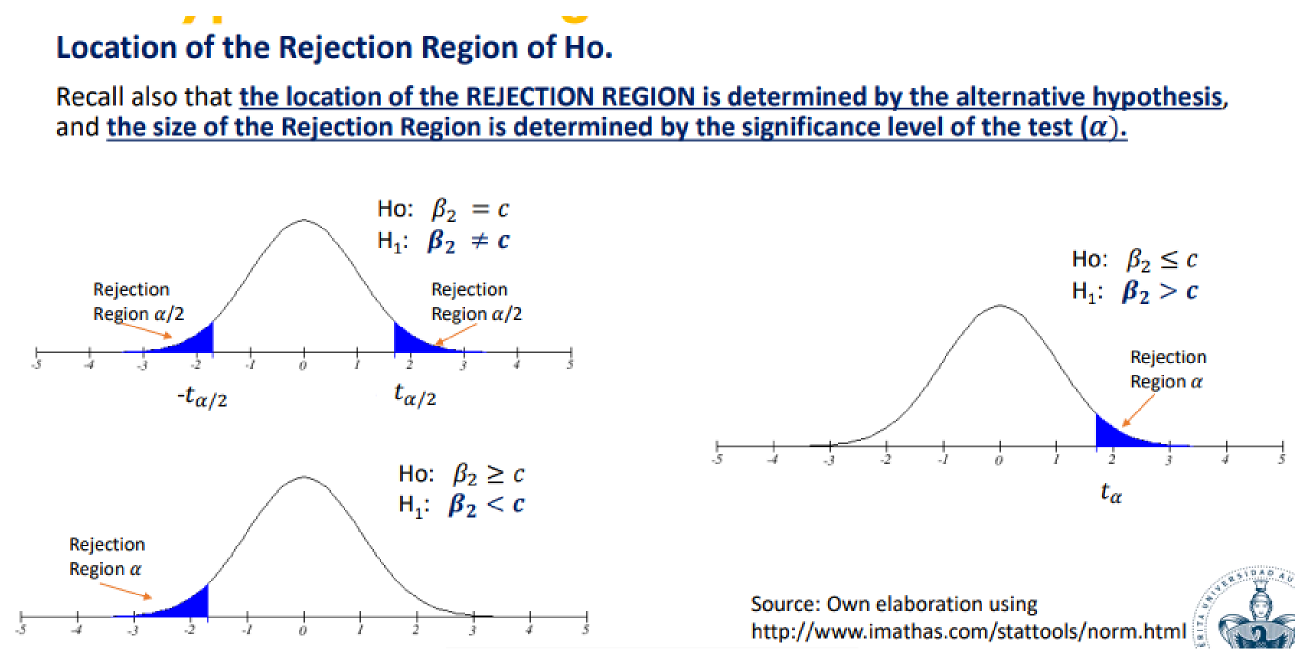
Hypothesis Testing

Recall that, in general, the alternative hypothesis (H1) is the hypothesis of interest (the one we want to test). We say "in general" because the alternative hypothesis MUST be composite. That is, the alternative hypothesis must be like:

$$\begin{aligned} &\mathrm{H}_1: \quad \beta_2 \neq c \quad \overset{\text{Two tailed test}}{\longrightarrow} \\ &\text{or:} \\ &\mathrm{H}_1: \quad \beta_2 \leq c \quad \overset{\text{Lower-tailed test}}{\longrightarrow} \\ &\text{or:} \\ &\mathrm{H}_1: \quad \beta_2 > c \quad \overset{\text{Upper-tailed test}}{\longrightarrow} \end{aligned}$$

If the hypothesis of interest is NOT composite (that is, it is simple: $\beta_2 = c$), then it will become the Null hypothesis.

On the other hand, we must remember that the test statistic and its distribution must be known under the Null hypothesis assumption (assuming the null is true).



Location of the Rejection Region of Ho.

Recall also that the location of the REJECTION REGION is determined by the alternative hypothesis, and the size of the Rejection Region is determined by the significance level of the test (α).

Ho: $\beta_2 = c$
H$_1$: $\beta_2 \neq c$

Ho: $\beta_2 \leq c$
H$_1$: $\beta_2 > c$

Ho: $\beta_2 \geq c$
H$_1$: $\beta_2 < c$

Source: Own elaboration using http://www.imathas.com/stattools/norm.html

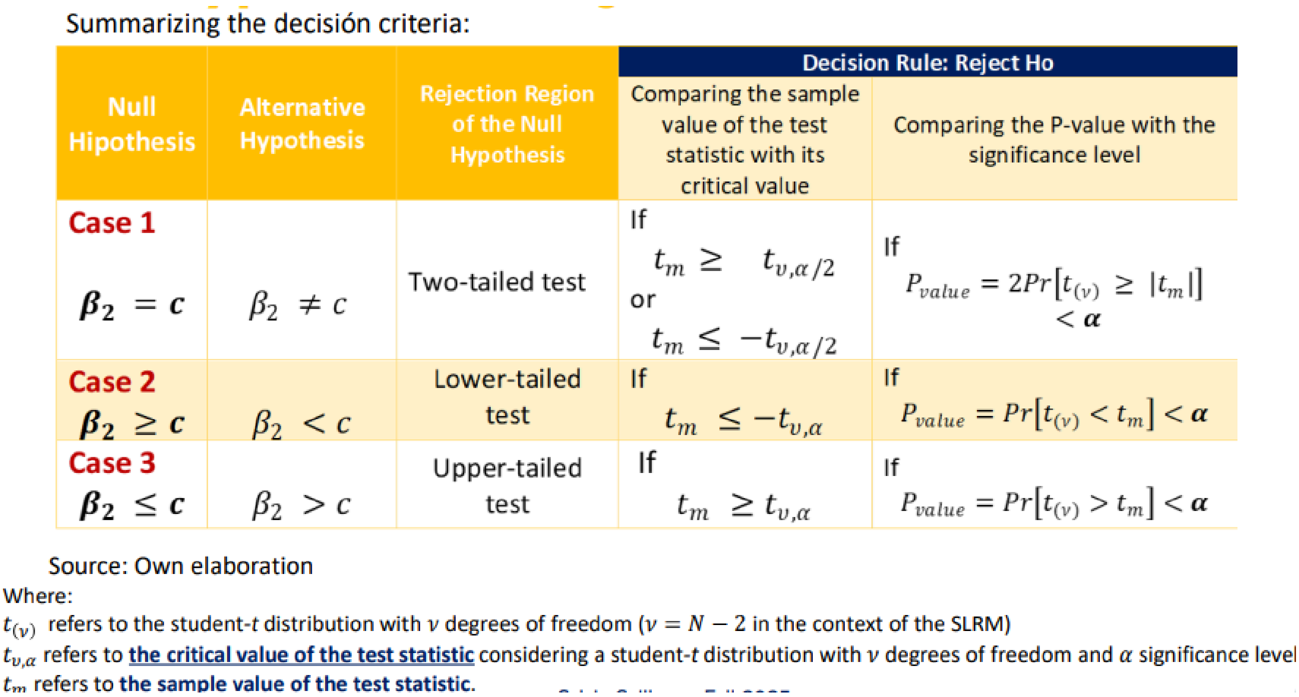
The test conclusion

As a final step of the hypothesis testing procedure, it is necessary to state a conclusion, which must be specified, taking into account the context of the analyzed problem.

The question to be answered is if we reject the null hypothesis or not. This question can be answered based on the following criteria:

1. Comparison between the sample value and critical value of the test statistic.
2. Comparison between the P-value and significance level of the test.

NOTE: It is important to highlight that if the evidence favors the null hypothesis, this does not mean the null is accepted, but we say there is not enough evidence to reject the null hypothesis.



Summarizing the decisión criteria:

Null Hipothesis	Alternative Hypothesis	Rejection Region of the Null Hypothesis	Decision Rule: Reject Ho	
			Comparing the sample value of the test statistic with its critical value	Comparing the P-value with the significance level
Case 1 $\beta_2 = c$	$\beta_2 \neq c$	Two-tailed test	If $t_m \geq t_{v,\alpha/2}$ or $t_m \leq -t_{v,\alpha/2}$	If $P_{value} = 2Pr[t_{(v)} \geq \lvert t_m \rvert] < \alpha$
Case 2 $\beta_2 \geq c$	$\beta_2 < c$	Lower-tailed test	If $t_m \leq -t_{v,\alpha}$	If $P_{value} = Pr[t_{(v)} < t_m] < \alpha$
Case 3 $\beta_2 \leq c$	$\beta_2 > c$	Upper-tailed test	If $t_m \geq t_{v,\alpha}$	If $P_{value} = Pr[t_{(v)} > t_m] < \alpha$

Source: Own elaboration

Where:
$t_{(v)}$ refers to the student-t distribution with v degrees of freedom ($v = N - 2$ in the context of the SLRM)
$t_{v,\alpha}$ refers to the critical value of the test statistic considering a student-t distribution with v degrees of freedom and α significance level
t_m refers to the sample value of the test statistic.

Example: the food-expenditure function

Once we have estimated the model, we usually test if the coefficients are statistically different from zero (hypothesis of interest). The objective is to show evidence that the explanatory variable influences the dependent variable. In the context of our food-expenditure example, we may set up the following hypotheses:

$$H_0: \quad \beta_2 = 0 \Rightarrow \text{Income does not influence food expenditure}$$
$$H_1: \quad \beta_2 \neq 0 \Rightarrow \text{Income does influence food-expenditure}$$

Under H_0 true, the test statistic and its distribution is:

$$t = \frac{b_2 - 0}{se(b_2)} \sim t_{(N-2)}$$

Hypothesis Testing Example

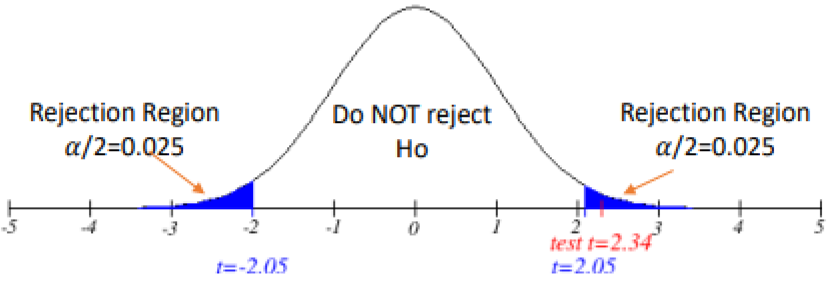
The sample value of the test statistic is:

$$t_m = \frac{.1397492 - 0}{.0596802} = 2.34$$

Taking a 5% confidence level ($\alpha = 0.05$) and $N - 2 = 28$ degrees of freedom, and considering a two-tailed test, the critical value of the test statistic is:

$$t_{v,\alpha/2} = t_{28,0.05/2} = t_{28,0.025} = 2.05$$

Graphically, it is easy to locate the rejection region of the null hypothesis (two tails):



Hypothesis Testing Conclusion

Following the decision criteria previously specified we have that:

$$t_m = 2.34 > t_{28,0.025} = 2.05$$

Or following the P-value criterion:

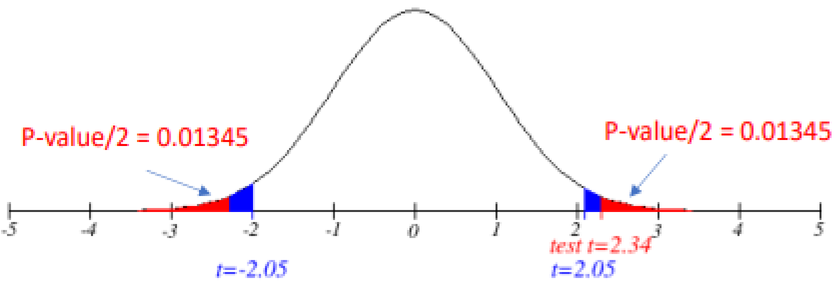
$$P_{value} = 2Pr[t_{(28)} \geq |2.34|] = 0.0269 < 0.05$$

CONCLUSION:

There is evidence to reject H_0;

That is, we reject $\beta_2 = 0$

This means that the income DOES influence food-expenditure



Another example using the food-expenditure model

Assume we want to test $\beta_2 > 0.15$; that is, we want to test that an increase in income by one peso, will increase the food expenditure by more than 0.15 pesos (15 centavos).

The hypotheses are:

$$H_0: \quad \beta_2 \leq 0.15 \Rightarrow \text{The marginal effect of income on food-expenditure is T greater than } 0.15$$
$$H_1: \quad \beta_2 > 0.15 \Rightarrow \text{The marginal effect of income on food-expenditure IS greater than } 0.15$$

Under H_0, the test statistic is:

$$t = \frac{b_2 - 0.15}{se(b_2)} \sim t_{(N-2)}$$

Hypothesis Testing Calculation

The sample value of the test statistic is:

$$t_m = \frac{0.1397492 - 0.15}{0.0596802} = -0.17176216$$

Taking a 5% confidence level ($\alpha = 0.05$) and $N - 2 = 28$ degrees of freedom, and considering that this is an upper-tailed (right-tail) test, the critical value of the test statistic is:

$$t_{v,\alpha} = t_{28,0.05} = t_{28,0.05} = 1.70$$

Graphically, it is easy to locate the rejection region of the null hypothesis, which is in the right tail of the distribution:

Hypothesis Testing Conclusion

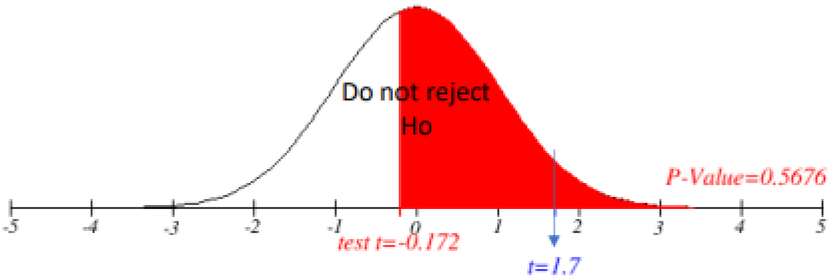
Following the decision criteria:

$$t_m = -0.172 < t_{28,0.05} = 1.7$$

**There is NOT enough evidence to reject H_0

Or, following the P-value criteria:

$$P_{value} = Pr[t_{(28)} \geq -0.172] = 0.5676 > 0.05$$



Evidently the P-value is greater than $\alpha = 0.05$, and greater than $\alpha = 0.1$. We conclude that there is NOT enough evidence to reject H_0

1.13 Confidence Interval for a linear combination of parameters

Interval estimation of a linear combination of parameters

In previous sections, we have mentioned that one of the objectives of the regression analysis is to estimate the conditional expected value of the dependent variable y_i; that is, we want to estimate:

$$E(y_i|x_i) = \beta_1 + \beta_2 x_i \Rightarrow \text{This is a linear combination of parameters}$$

We also know that **the point estimate of $E(y_i|x_i)$ is**:

$$\widehat{E(y_i|x)} = b_1 + b_2 x \Rightarrow \text{This is a linear combination of estimators}$$

Hence, for a specific value of the explanatory variable, $x_i = x_0$, where x_0 is a constant, the point estimate of the conditional expected value of y, is:

$$E\widehat{(y_i|x_0)} = b_1 + b_2 x_0$$

Variance of the Point Estimate

What is the variance of this point estimate, which represents a linear combination of two estimators b_1 and b_2? To answer this question, we must remember that b_1 and b_2 are two jointly distributed random variables.

Recall that (as shown in the statistics course), if X, Y are two correlated random variables, and c_1, c_2 are two constants, then:

$$Var(c_1 X + c_2 Y) = c_1^2 Var(X) + c_2^2 Var(Y) + 2c_1c_2 Cov(X, Y)$$

Hence, the variance of

$$E\widehat{(y_i|x_0)} = b_1 + b_2 x_0$$

Representing a linear combination of two correlated random variables ($Cov(b_1, b_2|x) \neq 0$), where $c_1 = 1$, and $c_2 = x_0$, is the following:

$$Var(\widehat{E(y_i|x_0)}|x) = Var(c_1 b_1 + c_2 b_2|x)$$

$$= c_1^2 Var(b_1|x) + c_2^2 Var(b_2|x) + 2c_1 c_2 Cov(b_1, b_2|x)$$

Estimator of the Variance

And the estimator for this variance is found by substituting the corresponding variances and covariances estimators of b_1 and b_2. That is:

$$\widehat{Var}\left(\widehat{E(y_i|x_0)}|x\right) = c_1^2 \widehat{Var(b_1}|x) + c_2^2 \widehat{Var(b_2}|x) + 2c_1 c_2 \widehat{Cov(b_1, b_2}|x)$$

Substituting the values of the constants $c_1 = 1$, and $c_2 = x_0$ we have:

$$\widehat{Var}\left(\widehat{E(y_i|x_0)}|x\right) = \widehat{Var(b_1}|x) + x_0^2 \widehat{Var(b_2}|x) + 2x_0 \widehat{Cov(b_1, b_2}|x)$$

And the standard error of $\widehat{E(y_i|x_0)}$ is:

$$se(\widehat{E(y_i|x_0)}) = \sqrt{\widehat{Var}(\widehat{E(y_i|x_0)}|x)} = \sqrt{\widehat{Var}(c_1 b_1 + c_2 b_2|x)}$$

Distribution and Interval Estimator

With the definition of the standard error of $\widehat{E(y_i|x_0)}$ and assuming the normally distributed errors, the distribution of the estimator for the conditional mean of y is:

$$\widehat{E(y|x)} \sim N[\beta_1 + \beta_2 x, \quad Var(\widehat{E(y|x_0)}|x)]$$

How can we obtain an interval estimator for $E(y|x) = \beta_1 + \beta_2 x$?

➔ Starting from the sampling distribution of the standardized estimator $\widehat{E(y|x_0)}$ we have:

$$t = \frac{\widehat{E(y|x)} - E(y|x_0)}{se(\widehat{E(y|x)})} \sim t_{(N-2)}$$

Note that we are using the estimated variance of $\widehat{E(y|x_0)}$:

Confidence Interval for Conditional Mean

Hence, the probabilistic expression needed to derive the $(1-\alpha)\%$ confidence interval for $E(y|x)$ is the following:

$$Pr\, -t_{\alpha/2} \leq \frac{\widehat{E(y|x)} - E(y|x)}{se(\widehat{E(y|x)})} \leq t_{\alpha/2} = 1 - \alpha$$

Reordering:

$$Pr[\widehat{E(y|x)} - t_{\alpha/2}\ se(\widehat{E(y|x)}) \leq E(y|x) \leq \widehat{E(y|x)} + t_{\alpha/2}\ se(\widehat{E(y|x)})] = 1 - \alpha$$

Confidence Interval for Conditional Mean

Another form of expressing the confidence Interval estimator for $E(y|x)$ is:

$$(\widehat{E(y|x_0)} - t_{\alpha/2}\ se(\widehat{E(y|x_0)});\ \widehat{E(y|x_0)} + t_{\alpha/2}\ se(\widehat{E(y|x_0)}))$$

Alternatively

$$\widehat{E(y|x_0)} \pm t_{\alpha/2}\ se(\widehat{E(y|x_0)})$$

DO NOT FORGET THAT $\widehat{E(y|x_0)} = b_1 + b_2 x_0$

Example:
For our food-expenditure model calculate the following:
a) The point estimate of food expenditure for a family with a weekly income of 2,500 pesos:

$$\widehat{E(y|x_0)} = 980.39 + 0.1397492 \times 2500 = 1329.763$$

b) The 95% confidence interval for a family with 2,500 pesos weekly income:

Standard Error Calculation

To answer these requests, we need to calculate the standard error of $\widehat{E(y|x_0)}$, given $x_0 = 2500$ and taking the point estimates b_1 and b_2

In this case $c_1 = 1$, and $c_2 = 2500$

$$\widehat{Var}\left(\widehat{E(y_i|x_0)}|x\right) = \widehat{Var}(b_1|x) + x_0^2 \widehat{Var}(b_2|x) + 2x_0 \widehat{Cov}(b_1, b_2|x)$$

Recall that the estimated variances and covariance for b_1 and b_2 are:

Covariance matrix of coefficients of regression model

e(V)	x	_cons
x	0.00356173	
_cons	-12.970628	64235.828

Variance and Standard Error Calculation

Entonces:

$$\widehat{Var}\left(\widehat{E(y_i|x_0)}|x\right) = 64235.828 + (2500^2)(0.00356173) + (2)(2500)(-12.970628) = 21643.501$$

$$se(\widehat{E(y_i|x_0)}) = \sqrt{\widehat{Var}\left(\widehat{E(y_i|x_0)}|x\right)} = \sqrt{21643.501} = 147.1173$$

Confidence Interval Calculation

Now, for $1-\alpha = 0.95$, $\alpha = 0.05$; and considering $N-2 = 28$ degrees of freedom, the critical value of the statistic is

$$t_{28,0.025} = 2.05$$

Therefore, the 95% confidence interval for $E(y|x)$ is:

$$\left(\widehat{E(y|x_0)} - t_{\alpha/2} \cdot se(\widehat{E(y|x_0)}); \widehat{E(y|x_0)} + t_{\alpha/2} \cdot se(\widehat{E(y|x_0)})\right)$$

$$(1329.763 - 2.05 \times (147.1173); 1329.763 + 2.05 \times (147.1173))$$

$$(1028.1725; 1631.3535)$$

Interpretation: With 95% probability, the limits 1028.1725 and 1631.3535 contain the expected food expenditure for a family with 2,500 pesos weekly income.

1.14 Hypothesis testing for a linear combination of parameters

Hypothesis testing for a linear combination of parameters

The hypotheses testing done so far, refer to testing a single parameter of the regression function. We now extend the procedure to the case of a linear combination of parameters. In particular, in this section we analyze hypothesis testing for the conditional expectation function.

Let $c_1\beta_1 + c_2\beta_2$ be a linear combination of β_1 and β_2 where c_1 and c_2 are constants; the objective is to test

the **hypothesis of interest**

$$c_1\beta_1 + c_2\beta_2 = c_0$$

Then, the corresponding null and alternative hypothesis may be reexpressed in a more convenient form:

$$\begin{aligned} H_0: &\quad c_1\beta_1 + c_2\beta_2 - c_0 = 0 \\ H_1: &\quad c_1\beta_1 + c_2\beta_2 - c_0 \neq 0 \end{aligned}$$

Two-tailed test

Hypothesis Test Calculation

Under H_0 taken as true, the test statistic is:

$$t = \frac{b_1 + 2500b_2 - 1000}{se(b_1 + 2500b_2)} \sim t_{N-2}$$

To calculate the sample value of the test-statistic we must calculate $se(b_1 + 2500b_2)$ first. Since we have previously calculated this standard error, we know that:

$$\widehat{Var}(b_1 + 2500b_2|x) = 21643.501$$

$$se(b_1 + 2500b_2) = \sqrt{21643.501} = 147.1173$$

Hence, the sample value of the test statistic is:

$$t_m = \frac{980.39 + (2500) \times 0.1397492 - 1000}{147.1173} = 2.2414971$$

Hypothesis Test Conclusion

For $\alpha = 0.05$; and considering $N - 2 = 28$ degrees of freedom:

$$t_{28,0.05} = 1.7$$

We may observe that

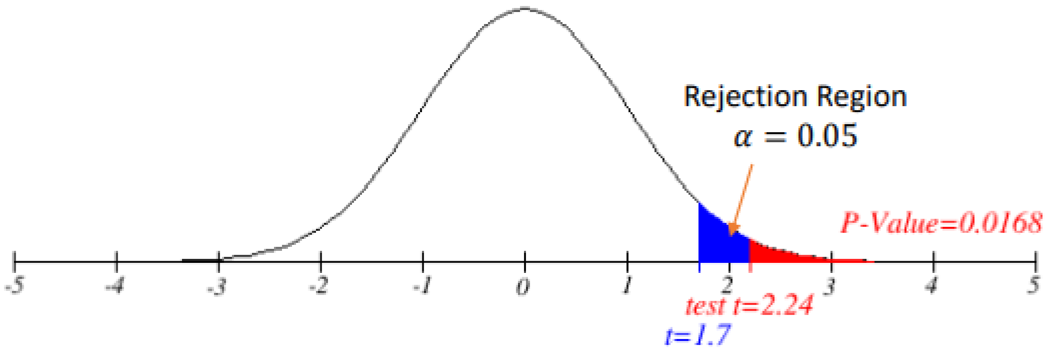
$$t_m = 2.24 > t_{28,0.05} = 1.7$$

hence, we reject H_0 in favor of the alternative hypothesis: we expect the food-expenditure for a family with 2500 pesos weekly income, to be greater than 1000 pesos.

The P-value for the test is

$$P_{value} = Pr[t_{(28)} > 2.24] = .0168 < \alpha = 0.05$$

We may conclude that, we reject H_0 at 5% significance level.



1.15 Prediction Interval

Prediction VS Estimation

In previous sections, we have talked about how the least squares estimates of the linear regression model provide a way to estimate the value of y given the value of x.

What is the difference between estimation and prediction?

Following Hill, et al (2018), chapter 4, prediction is related to estimation. Let us see:

If we want to predict the value of y for a specific value of x, say x_0, using the regression model, we must consider that y has a deterministic part $(\beta_1 + \beta_2 x_0)$ and a stochastic part.

$$y_0 = E(y|x_0) + e_0 = \beta_1 + \beta_2 x_0 + e_0$$

Hence we may say that prediction is related to estimation; we estimate the systematic part of y_0 using

$$\boxed{\widehat{E(y|x_0)} = b_1 + b_2 x_0}$$

and we add an estimate of the random component e_0 which is

$$E(e_0|x_0)$$

which has implications for the variance of the forecast error.

Recalling that

$$E(e_0) = 0$$

and

$$Var(e_0) = \sigma^2$$

the estimate for the least squares predictor of y_0 is specified as:

$$\hat{y}_0 = \widehat{E(y|x_0)} + 0 = b_1 + b_2 x_0$$

How Good is the predictor?

In order to answer this question, we must calculate the forecast error, which is the difference between the predictor and its estimator; let us call the forecast error f, following Hill, et al (2018):

$$f = y_0 - \hat{y}_0$$

Hence, what are the properties of the forecast error?

Expected Value of the forecast error in the SLRM:

$$E(f|x) = E[(y_0 - \hat{y}_0)|x] = E[(\beta_1 + \beta_2 x_0 + e_0) - (b_1 + b_2 x_0)|x]$$

Reordering terms:

$$E(f|x) = \beta_1 + \beta_2 x_0 + E(e_0|x) - E[(b_1 + b_2 x_0)|x] = \beta_1 + \beta_2 x_0 + 0 - E(b_1|x) - x_0 E(b_2|x)$$

But, **we know that b_1 and b_2 are unbiased estimators** of β_1 and β_2, hence:

$$E(b_1|x) = \beta_1 \text{ and } E(b_2|x) = \beta_2$$

Properties of the Forecast Error

Therefore:

$$E(f|x) = \beta_1 + \beta_2 x_0 - \beta_1 - x_0\beta_2 = 0$$

That is, on average, the forecast error is zero, meaning that $\hat{y}_0$ is an unbiased predictor of the conditional expected value of y

But also, in addition for $\hat{y}_0$ to be an unbiased predictor, we want this predictor to have the smallest variance as possible.

The variance of the forecast error in the SLRM

$$Var(f|x) = Var[(y_0 - \hat{y}_0)|x] = Var[(\beta_1 + \beta_2 x_0 + e_0) - (b_1 + b_2 x_0)|x]$$

$$Var(f|x) = Var(e_0|x) + (-1)^2 Var(b_1|x) + (-x_0)^2 Var(b_2|x) + 2(-1)(-x_0)Cov(b_1, b_2|x)$$

Variance of Forecast Error

Given the proofs shown in section 1.7, we have that:

$$Var(b_1|x) = \sigma^2 \left[\frac{1}{N} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right]$$

$$Var(b_2|x) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

$$Cov(b_1, b_2|x) = \sigma^2 \left[\frac{-\bar{x}}{\sum(x_i - \bar{x})^2}\right]$$

Substituting this expression into the $Var(f|x)$ and considering $Var(e_0|x) = \sigma^2$, we have:

$$Var(f|x) = \sigma^2 + \sigma^2 \left[\frac{1}{N} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right] + x_0^2 \frac{\sigma^2}{\sum(x_i - \bar{x})^2} + 2x_0\sigma^2 \left[\frac{-\bar{x}}{\sum(x_i - \bar{x})^2}\right]$$

Variance of Forecast Error (Simplified)

Factorizing we have:

$$Var(f|x) = \sigma^2 \left\{1 + \frac{1}{N} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2} + \frac{x_0^2}{\sum(x_i - \bar{x})^2} - \frac{2x_0\bar{x}}{\sum(x_i - \bar{x})^2}\right\}$$

$$Var(f|x) = \sigma^2 \left\{1 + \frac{1}{N} + \frac{\bar{x}^2 + x_0^2 - 2x_0\bar{x}}{\sum(x_i - \bar{x})^2}\right\}$$

$$\boxed{Var(f|x) = \sigma^2 \left\{1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right\} \Rightarrow}$$

This expression allows us to see that the more distant is x_0 from the sample mean of x (center of the sample data), the larger the forecast error variance will become.

Finally, given that σ^2 is not known and must be estimated, the estimated variance of the forecast error is:

$$\widehat{Var(f|x)} = \hat{\sigma}^2 \left\{ 1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right\}$$

Predictions that are closer to the sample mean are better.

Prediction Interval

And the standard error of the forecast is:

$$se(f|x) = \sqrt{\widehat{Var(f|x)}}$$

Prediction interval

Previously, in section 1.13, we defined the confidence interval for the conditional expected value of y. Now, we may define the prediction interval for y. First, we start with the expression for the standardized predictor:

$$t = \frac{f}{\sqrt{\widehat{Var(f|x)}}} = \frac{\hat{y}_0 - y_0}{\sqrt{\widehat{Var(f|x)}}} = \frac{\hat{y}_0 - y_0}{se(f|x)} \sim t_{(N-2)}$$

Hence, the initial probabilistic expression for the prediction interval of y_0 is the following:

$$Pr\left[-t_{\alpha/2} \leq \frac{\hat{y}_0 - y_0}{se(f|x)} \leq t_{\alpha/2} \right] = 1 - \alpha$$

Prediction Interval

And the standard error of the forecast is:

$$se(f|x) = \sqrt{\widehat{Var(f|x)}}$$

Prediction interval

Previously, in section 1.13, we defined the confidence interval for the conditional expected value of y. Now, we may define the prediction interval for y. First, we start with the expression for the standardized predictor:

$$t = \frac{f}{\sqrt{\widehat{Var(f|x)}}} = \frac{\hat{y}_0 - y_0}{\sqrt{\widehat{Var(f|x)}}} = \frac{\hat{y}_0 - y_0}{se(f|x)} \sim t_{(N-2)}$$

Hence, the initial probabilistic expression for the prediction interval of y_0 is the following:

$$Pr\left[-t_{\alpha/2} \leq \frac{\hat{y}_0 - y_0}{se(f|x)} \leq t_{\alpha/2}\right] = 1-\alpha$$

Prediction Interval (Continued)

Reordering:

$$Pr[\hat{y}_0 - t_{\alpha/2}\, se(f|x) \leq y_0 \leq \hat{y}_0 + t_{\alpha/2}\, se(f|x)] = 1-\alpha$$

Another way of expressing the interval:

$$\boxed{(\hat{y}_0 - t_{\alpha/2}\, se(f|x);\ \hat{y}_0 + t_{\alpha/2}\, se(f|x)) \Rightarrow \text{Prediction interval}}$$

Clearly, we can conclude that the more distant the prediction is from the sample mean of the data $(\bar{x})$, the larger will be the standard error of the forecast and the wider will be the prediction interval (the less reliable the prediction will be).

Example: The food-expenditure function

Consider the summary statistics of the weekly family income variable x:

Variable	Obs	Mean	Std. Dev.	Min	Max
x	30	3641.667	2222.133	750	10000

Now, assume you want a prediction of food expenditure for a family with weekly income $x_0 = 3641.667$ pesos (sample mean of x)

The point prediction is:

$$\hat{y}_0 = b_1 + b_2 x_0 = 980.39 + (0.1397492) \times (3641.667) = 1489.31$$

Prediction Interval Calculation

To obtain the **prediction interval** we require to calculate the estimated variance of the forecast and corresponding standard error.

The estimated variance of the forecast error at $x_0 = 3641.667$ is:

$$\widehat{Var(f|x)} = \hat{\sigma}^2 \left\{1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right\}$$

$$\widehat{Var(f|x)} = 510033.715 \times \left\{1 + \frac{1}{30} + \frac{(3641.667 - 3641.667)^2}{143198367}\right\} = 527034.84$$

And the standard error of the forecast is:

$$se(f|x) = \sqrt{527034.84} = 725.9717$$

The 95% confidence interval ($t_{(28,0.025)} = 2.05$) is the following:

Prediction Interval Calculation

$$(\hat{y}_0 - t_{\alpha/2}\ se(f|x);\ \hat{y}_0 + t_{\alpha/2}\ se(f|x))$$

$$(1489.31 - (2.05 \times 725.9717);\ 1489.31 + (2.05 \times 725.9717))$$

$$(1.068015;\ 2977.552)$$

This is a very wide interval, but is due to the magnitude of $\hat{\sigma}^2 = 510033.715$ (even though we are predicting at $x_0 = \bar{x}$)

Let us see now what happens if we want to predict the food expenditure for a family with a weekly income of 8,000 pesos. This income level is far higher than the sample mean of income (3641.667).

Prediction at Extreme Value

The point prediction at $x_0 = 8000$ pesos is:

$$\hat{y}_0 = b_1 + b_2 x_0 = 980.39 + (0.1397492) \times (8000) = 2098.384$$

The estimated variance of the forecast is:

$$\widehat{Var(f|}x) = 510033.715 \times \left\{1 + \frac{1}{30} + \frac{(8000 - 3641.667)^2}{143198367}\right\} = 594690.21$$

And the corresponding standard error is:

$$se(f|x) = \sqrt{594690.21} = 771.1616$$

Therefore, the prediction interval is:

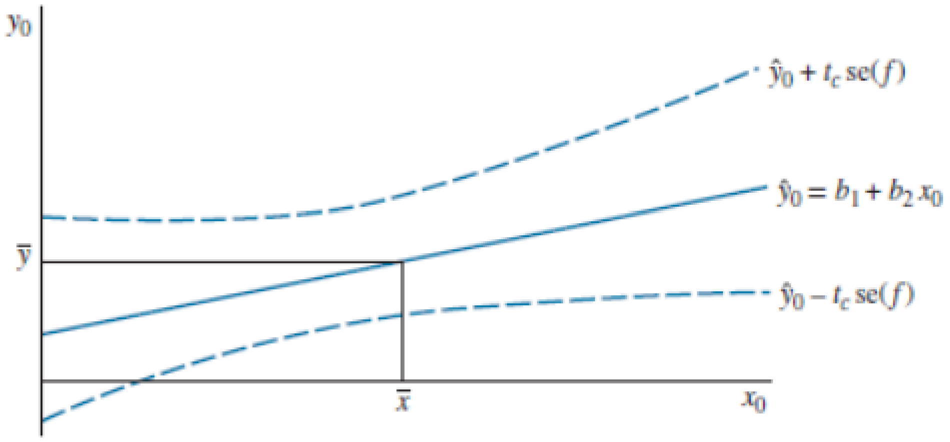
$$(2098.384 - (2.05 \times 771.1616);\ 2098.384 + (2.05 \times 771.1616))$$

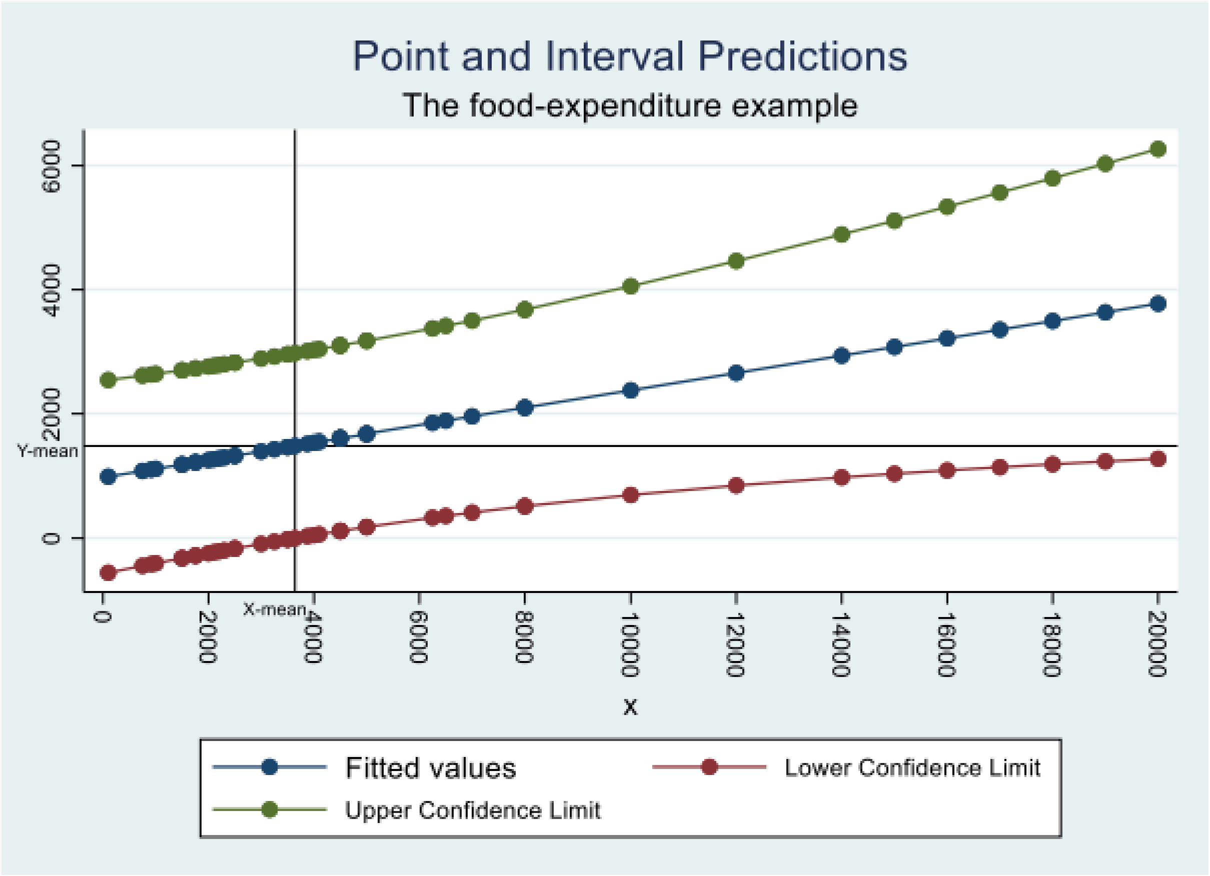
$$(517.50272; 3679.2653)$$

We may observe that the interval is wider compared to the one we calculated at $\bar{x}$.

Effect of Predicting Far from Sample Mean

The effect of predicting far from the sample mean may be observed in the following graph:


y_0
$\bar{y}$
$\hat{y}_0 + t_c \text{se}(f)$
$\hat{y}_0 = b_1 + b_2 x_0$
$\hat{y}_0 - t_c \text{se}(f)$
$\bar{x}$
x_0



Using the sample data for our food expenditure example, we may observe what happens to the prediction interval as we move far from the sample mean of x.

Measuring the goodness of fit

Another interest derived from our econometric analysis is the one related to using the explanatory variable x_i to explain the variation of the dependent variable as much as possible.

How can we explain how much of the variation on y is explained by variation on x?

In answering this question, we obtain the goodness of fit of the model; to do so, it is necessary to separate y into its two components:

$$y_i = \underbrace{E(y_i|x)}_{\text{Systematic Component (Explained)}} + \underbrace{e_i}_{\text{Unsystematic component (unexplained)}}$$

Goodness of fit

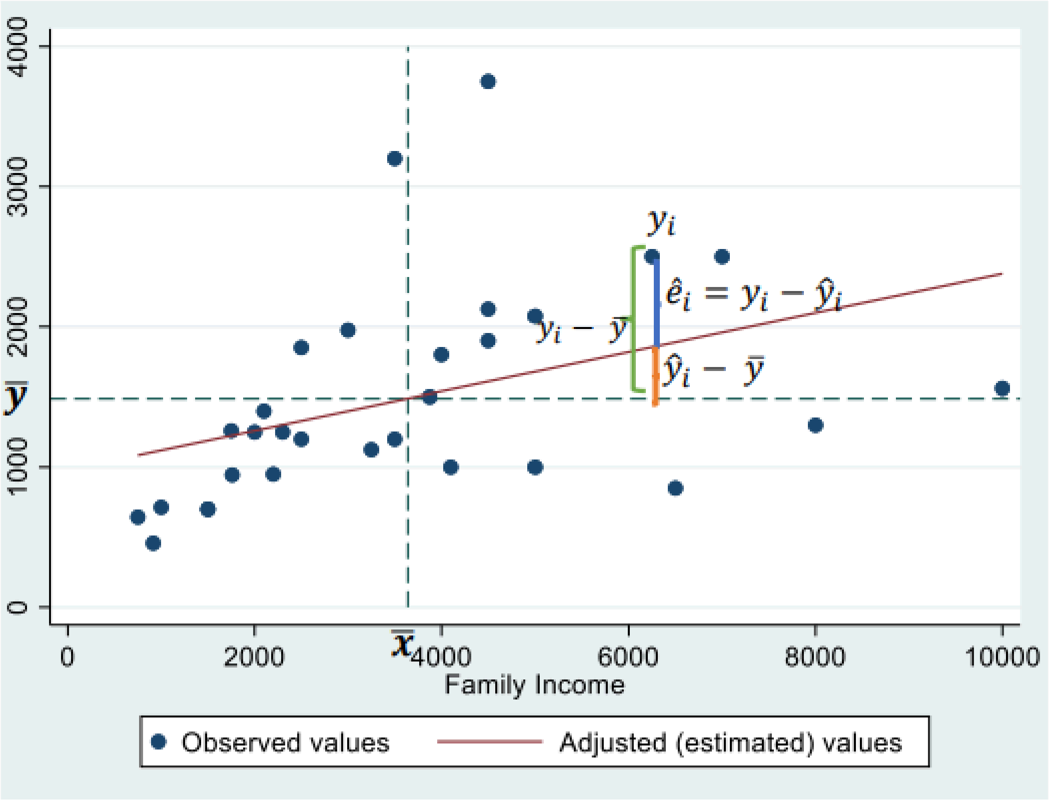
Because none of those components are observed or known, we use their corresponding estimators, which added, must be equal to the observed value of y; that is:

$$y_i = \hat{y}_i + \hat{e}_i$$

Where

$$\hat{y}_i = b_1 + b_2 x_i \quad \text{and also} \quad \hat{e}_i = y_i - \hat{y}_i$$

These components can be identified graphically such that we may observe what the part is explained by the model (estimated regression line) and what part is not.



$y_i - \bar{y} =$ is the variation of y around its mean

$\hat{y}_i - \bar{y} =$ part of variation of y around its mean that is **explained by the model**

$\hat{e}_i = y_i - \hat{y}_i =$ part of variation of y around its mean that is **NOT** explained by the model

Decomposition of Total Variation

That is, the variation of y around its mean can be separated as follows:

$$y_i - \bar{y} = \hat{y}_i - \bar{y} + \hat{e}_i$$

If we sum all the squared variations of all observed points y_i with respect to the mean $\bar{y}$, then we have:

$$\sum_{i=1}^{N}(y_i-\bar{y})^2=\sum_{i=1}^{N}(\hat{y}_i-\bar{y}+\hat{e}_i)^2$$

$$\sum_{i=1}^{N}(y_i-\bar{y})^2=\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{N}\hat{e}_i^2+2\qquad\underbrace{\sum_{i=1}^{N}(\hat{y}_i-\bar{y})\hat{e}_i}_{\text{This term is zero iff the regression model includes intercept}}$$

1.16 Goodness of fit

The analysis of variance (ANOVA) can be summarized as:

$$\underbrace{\sum_{i=1}^{N}(y_i-\bar{y})^2}_{\text{Total Sum of Squares (SST)}}\quad=\quad\underbrace{\left(\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2\right)}_{\text{Sum of Squares explained by the Regression (SSR)}}\quad+\quad\underbrace{\sum_{i=1}^{N}\hat{e}_i^2}_{\text{Sum of Squared Errors (SSE)}}$$

We have that:

SST = is the **Total Sum of Squares** and measures the total variation in y about the sample mean.

SSR = is the **Sum of Squares Explained by the Regression** and measures the part of total variation in y about the sample mean that is explained by the regression.

SSE = is the **Sum of Squared Errors** and measures the part of total variation in y about the sample mean that is **NOT** explained by the regression.

Coefficient of Determination

Finally, in order to know what proportion of the total variation in y about its sample mean is explained by the model R^2, we divide both sides of the previous expression by the Total Sum of Squares (SST).

$$\underbrace{\frac{\sum_{i=1}^{N}(y_i-\bar{y})^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2}}_{1}=\underbrace{\frac{\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2}}_{R^2}+\frac{\sum_{i=1}^{N}\hat{e}_i^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2}\Rightarrow\text{We solve for }R^2\text{ from this expression}$$

Therefore, the goodness of fit measure R^2 is expressed as:

$$R^2=1-\frac{\sum_{i=1}^{N}\hat{e}_i^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2}$$

Where $0\leq R^2\leq 1$ if the model has intercept.

The closer is the R^2 to 1, the closer are the sample values y_i to the fitted line; the better the goodness of fit of the model.

R^2 Is the proportion of total variation in y around its mean, explained by x within the regression model.

Example: Goodness of Fit

Taking our food-expenditure model, we may use the estimation results to calculate the R^2:

$$R^2 = 1 - \frac{14280944}{17077585.2} = 0.16376093$$

Interpretation: 16.37% of total variation in food-expenditure about its mean is explained by income variation.

In this case the R^2 is low; although we should mention that, for cross-section data, high R^2's are not usually observed.

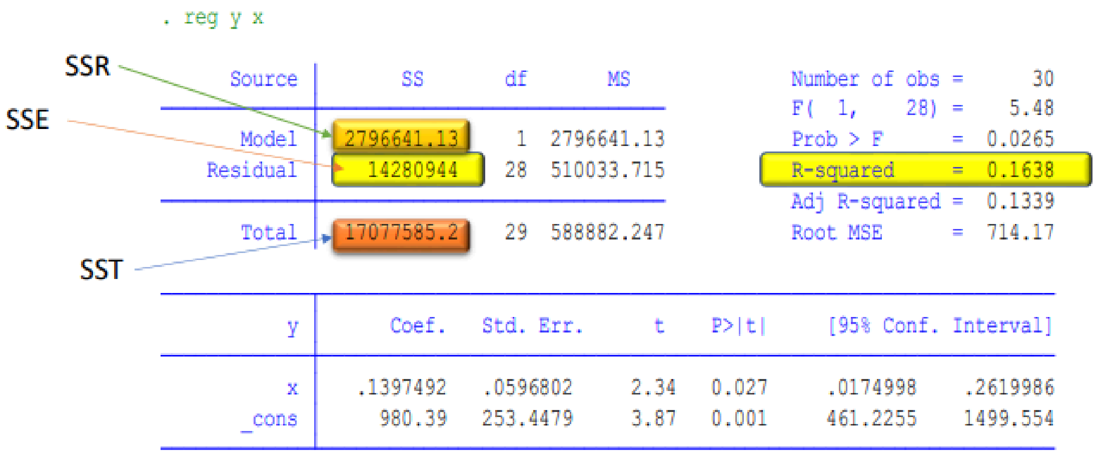
Another way of calculating the goodness of fit: The generalized R^2

The R^2 may also be calculated as the squared correlation between the observed (y) and the fitted values $(\hat{y})$ of the dependent variable:

$$R^2 = (Corr(y, \hat{y}))^2 \Rightarrow \text{generalized } R^2$$

Using Stata

The STATA output provides the ANOVA table and the R². It is important to identify the information in the output; for our food expenditure example we have the following output:



```
. reg y x

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  1,    28) =    5.48
       Model |  2796641.13     1  2796641.13           Prob > F      =  0.0265
    Residual |    14280944    28  510033.715           R-squared     =  0.1638
-------------+------------------------------           Adj R-squared =  0.1339
       Total |  17077585.2    29  588882.247           Root MSE      =  714.17

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .1397492   .0596802     2.34   0.027     .0174998    .2619986
       _cons |     980.39   253.4479     3.87   0.001     461.2255    1499.554
------------------------------------------------------------------------------
```

We are focus on getting a sum of squared errors lower with our estimation.

1.17 Log-linear functional form

The Log-Linear model

We call log-linear model, the model whose dependent variable is expressed in logarithms:

$$\ln(y_i) = \beta_1 + \beta_2 x_i + e_i$$

$$E[\ln(y_i)|x] = \beta_1 + \beta_2 x_i$$

In this model β_2 is the semi-elasticity of y with respect to x, because:

$$\beta_2 = \frac{dE[\ln(y_i)|x]}{dx_i} \approx \frac{\Delta\%E(y|x)}{\Delta x}$$

In this model the slope changes at each point, but the SEMI-ELASTICITY is constant.

We observe that, under this functional form, β_2 is the percentage change in y given a one-unit increase in the explanatory variable x. The semi-elasticity is CONSTANT.

Exponential Form of Log-Linear Model

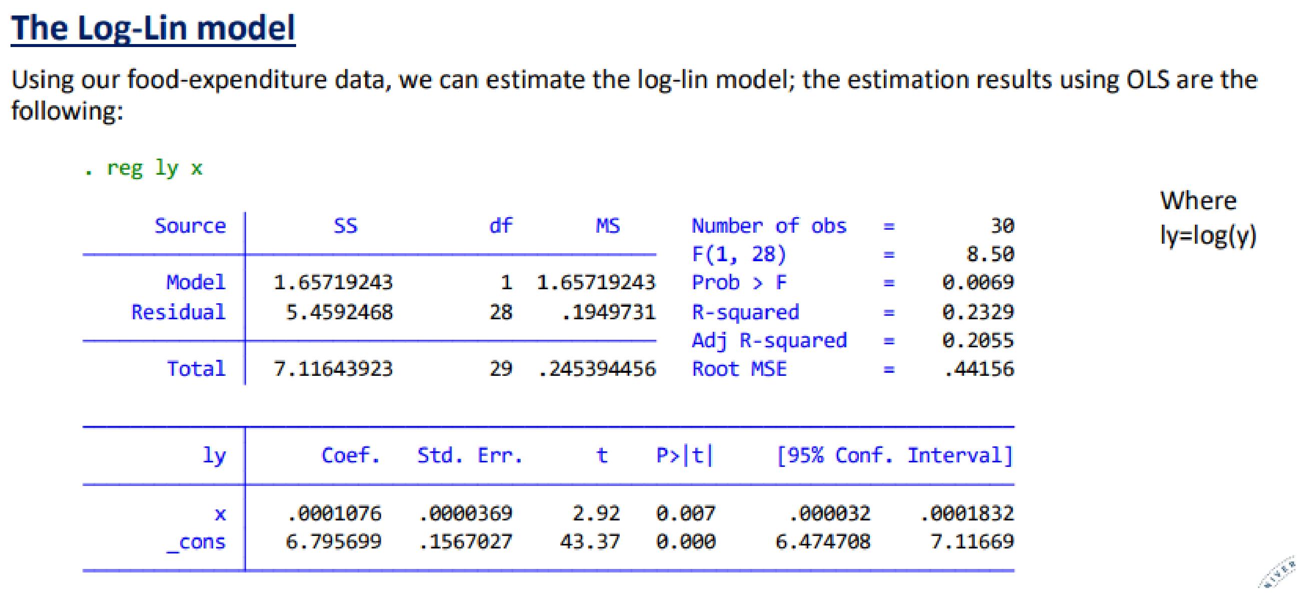
Applying exponential function to both sides of the equation we have:

$$\ln(y_i) = \beta_1 + \beta_2 x_i + e_i$$

$$\exp[\ln(y_i)] = \exp[\beta_1 + \beta_2 x_i + e_i]$$

$$\boxed{y_i = \exp[\beta_1 + \beta_2 x_i + e_i]}$$

The log-linear model is an exponential function



The Log-Lin model

Using our food-expenditure data, we can estimate the log-lin model; the estimation results using OLS are the following:

```
. reg ly x

      Source |       SS           df       MS      Number of obs   =        30
-------------+----------------------------------   F(1, 28)        =      8.50
       Model |  1.65719243         1  1.65719243   Prob > F        =    0.0069
    Residual |   5.4592468        28    .1949731   R-squared       =    0.2329
-------------+----------------------------------   Adj R-squared   =    0.2055
       Total |  7.11643923        29  .245394456   Root MSE        =    .44156

------------------------------------------------------------------------------
          ly |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .0001076   .0000369     2.92   0.007      .000032    .0001832
       _cons |   6.795699   .1567027    43.37   0.000     6.474708     7.11669
------------------------------------------------------------------------------
```

Where
ly=log(y)

The Log-Lin Model

Expressing the estimation results we have:

$$\ln(y_i) = 6.7957 + 0.00011 x_i$$

Interpretation:

$b_2 = 0.00011$

A one-peso increase in weekly family income, increases the food-expenditure by 0.011% (note that it is necessary to multiply the coefficient by 100 in order to express it as a percentage).

Alternatively, a 100-pesos increase in the weekly family income, increases the food expenditure by 1.1%

Now, we must realize that we estimated the log of y, that is, we've got $\ln(y_i)$.

Hence, in order to obtain the estimator of the conditional mean of y is necessary to apply the exponential function to both sides of the expression:

$$\widehat{E(y|x)} \approx \exp(\widehat{\ln(y_i)}) = \exp(b_1 + b_2 x_i)$$

The Log-Linear Model

Why do we say that $\exp(\widehat{\ln(y_i)}) = \exp(b_1 + b_2 x_i)$ is an approximation to $\widehat{E(y|x_i)}$?

In order to answer this question, we MUST know what the distribution of $\ln(y_i)$ is:

What is the $E(\ln(y_i)|x_i)$?

What is the $Var(\ln(y_i)|x_i)$?

The Log-Normal distribution

Let the random variable W:

$$W \sim N(\mu, \sigma^2)$$

Now, if Y is a function of W such that

$$Y = \exp(W) \quad \text{or} \quad Y = e^W$$

Hence

$$W = \ln(Y) \sim N(\mu, \sigma^2)$$

We say that Y has a log-normal distribution.

Mean and Variance of Log-Normal Distribution

What is the mean and variance of a log-normal variable Y?

Recall that the expected value of a non-linear function of a random variable is NOT the function of the expected value of the variable. That is:

$$E(Y) = E(e^W) \neq e^{E(W)}$$

It can be shown that:

$$E(Y) = e^{\left(\mu + \frac{\sigma^2}{2}\right)} = \exp(\mu + \sigma^2/2)$$

And the variance of Y is given by:

$$Var(Y) = e^{(2\mu+\sigma^2)(e^{\sigma^2}-1)} = [\exp(2\mu+\sigma^2)][\exp(\sigma^2)-1]$$

Application to Log-Linear Model

These results are related to our log-lin and log-log models.

Hence, given a log-linear model:

$$\ln(y_i) = \beta_1 + \beta_2 x_i + e_i$$

This implies that:

$$y_i = \exp(\beta_1 + \beta_2 x_i + e_i)$$

Or that:

$$y_i = [\exp(\beta_1 + \beta_2 x_i)] \times [\exp(e_i)]$$

The Log-Normal Distribution

And if we assume that the error terms are normally distributed:

$$e_i|x \sim N(0, \sigma^2)$$

Hence:

$$E(y_i|x) = E\{\exp(\beta_1 + \beta_2 x_i + e_i)|x\}$$

$$= E\{\{\exp(\beta_1 + \beta_2 x_i)|x\} \times [\exp(e_i|x)]\}$$

$$= [\exp(\beta_1 + \beta_2 x_i)] \times E[\exp(e_i|x)]$$

But we have said that, if a random variable has a log-normal distribution, its expected value is $\mu + \sigma^2/2$. Hence, the expected value of $\exp(e_i|x)$ is:

$$E[\exp(e_i|x)] = \exp(E(e_i|x) + \sigma^2/2) = \exp(\sigma^2/2)$$

Corrected Predictor

Then

$$E(y_i|x) = [\exp(\beta_1 + \beta_2 x_i)] \times \exp[\sigma^2/2] = \exp(\beta_1 + \beta_2 x_i + \sigma^2/2)$$

Consequently, when we have a log-linear model, and we want to predict y_i, we must use the corrected predictor of $\hat{y}$; that is:

$$\hat{y}_i(\text{corrected}) = \widehat{E(y_i|x_i)} = \exp(b_1 + b_2 x_i + \hat{\sigma}^2/2)$$

Alternatively:

$$\hat{y}_i(\text{corrected}) = \hat{y}_i \times \exp[\hat{\sigma}^2/2]$$

Where:
b_1, b_2 and $\hat{\sigma}^2$ are obtained using OLS on the log-linear model

Note on Sample Size

NOTE:

- $\hat{y}_i$ (corrected) is a better predictor of the dependent variable when we have large samples
- $\hat{y}_i$ is a better predictor when we have small samples (N<30)

Why?

Because the estimated variance $\hat{\sigma}^2$ adds "noise" to the estimation when we use $\hat{y}_i$ (corrected), leading it to have increased variability relative to $\hat{y}_i$; this can outweigh (eliminate) the benefit of the correction when calculating the $\widehat{E(y_i|x_i)}$ in small samples (Hill et al, 2018: p. 175).

Model Comparison

The Lin-Lin VS Log-Lin functional forms

Can we compare the reported R^2's of a lin-lin VS a log-lin model directly?

The answer is NO because:

The R^2 of the Lin-Lin model is the proportion of total variation in y (about its mean) explained by variation in x.

The R^2 of the Log-linear model is the proportion of total variation in the log of y (about the mean of $\ln(y)$) explained by variation in x.

Then, how can we compare the goodness of fit of two models?

Generalized R² for Log-Lin Model

How can we compare the goodness of fit of two models?

Answer:

The generalized R^2 must be calculated for the Log-Lin model

1. Once we have estimated the Log-Lin model, predict $\ln(y_i)$; that is, compute $\widehat{\ln(y_i)}$
2. Use the exponential function on $\widehat{\ln(y_i)}$ to obtain $\hat{y}_i$ (including the correction factor if the sample size is large)
3. Calculate $\widehat{Corr(y, \hat{y})}$
4. Calculate $R_g^2 = (\widehat{Corr(y, \hat{y})})^2$ and compare this with the R^2 of the Lin-Lin model

Graphically we have

$\widehat{E(y|x)} = \hat{y}$

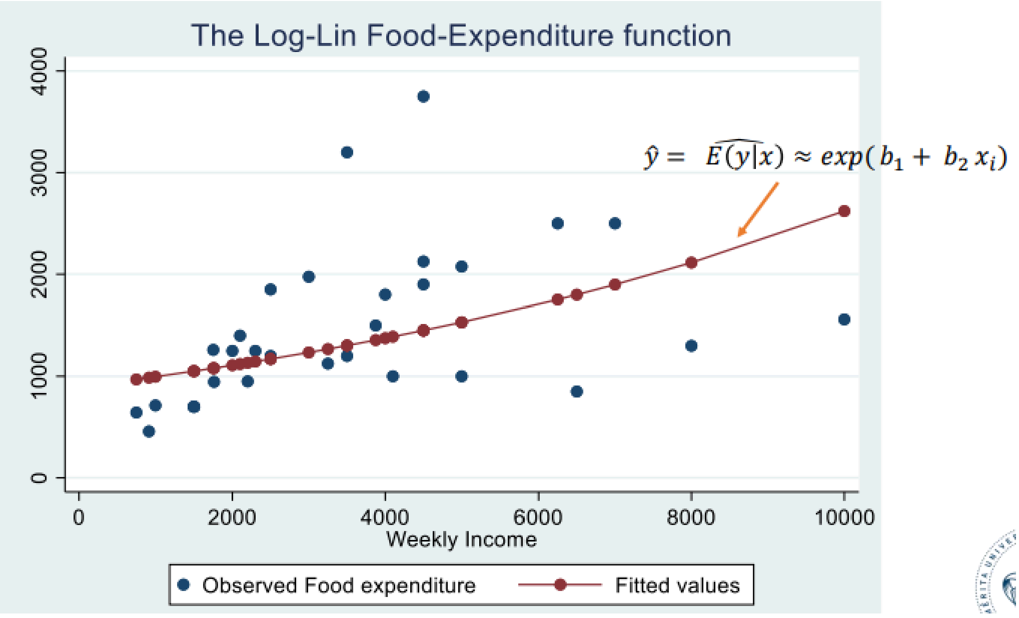
Which is not a line but a curve (adjusted)

The generalized R²

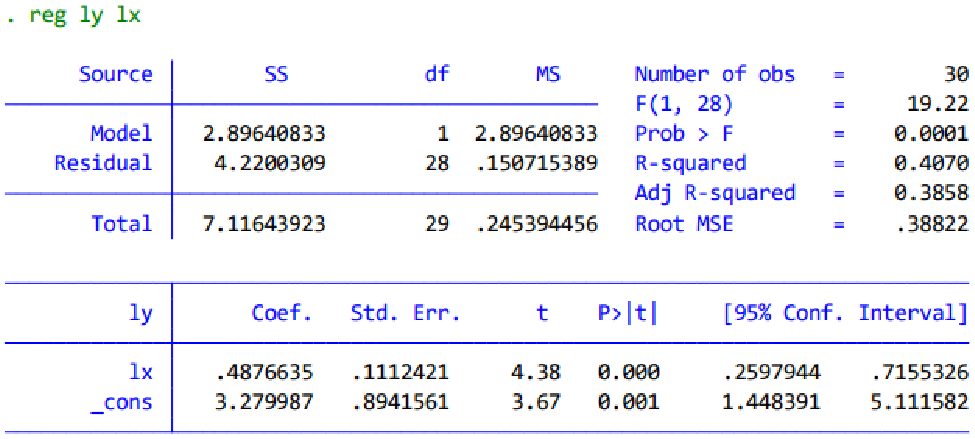
$R_g^2 = (0.3337)^2$

$= 0.11$

Which is smaller compared to the R² of the Lin-Lin model



For our food-expenditure model and using OLS, we get the following results:



```
. reg ly lx

      Source |       SS           df       MS      Number of obs   =        30
-------------+----------------------------------   F(1, 28)        =     19.22
       Model |  2.89640833         1  2.89640833   Prob > F        =    0.0001
    Residual |   4.2200309        28  .150715389   R-squared       =    0.4070
-------------+----------------------------------   Adj R-squared   =    0.3858
       Total |  7.11643923        29  .245394456   Root MSE        =    .38822

------------------------------------------------------------------------------
          ly |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          lx |   .4876635   .1112421     4.38   0.000     .2597944    .7155326
       _cons |   3.279987   .8941561     3.67   0.001     1.448391    5.111582
------------------------------------------------------------------------------
```

Where

ly=log(y)

lx=log(x)

Source: Own elaboration using Stata 17

We may compare the goodness of fit of the **LOG-LOG VS LOG-LIN models**

Graphically, we have the fitted food-expenditure

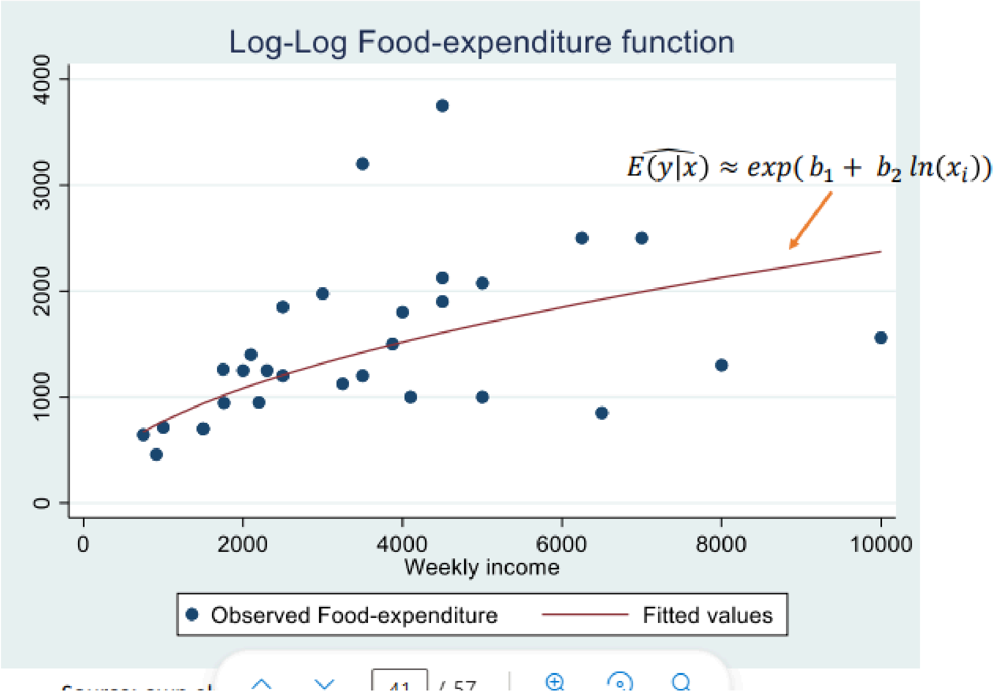
$\widehat{E(y|x)} = \hat{y}$

The generalized goodness of fit measure is:

$R_g^2 = (0.4769)^2$

$= .22743361$

Which **IS bigger** compared to the Lin-Lin model R²



1.18 Testing normality of the errors

The Jarque-Bera Test

- □ One of the assumptions of the SLRM is that the regression errors (and hence the dependent variable) are normally distributed; that is:

$$e_i \mid x \sim N(0, \sigma^2)$$

- □ The hypothesis testing procedure and interval estimation rely on the normality assumption of the regression error.
- □ If the errors are not normally distributed, the confidence intervals and the test statistics for hypothesis testing would be wrong.
- □ For this reason, it is essential to test the normality assumption of the regression errors.
- □ The histogram of the estimated residuals can give us an idea about the sample distribution of the residuals; however, it is necessary to perform a test based on a test statistic.
- □ One of the most commonly used normality tests is the Jarque-Bera test.

The Jarque-Bera test

This test is based on two parameters that characterize the probability distribution function of a random variable:

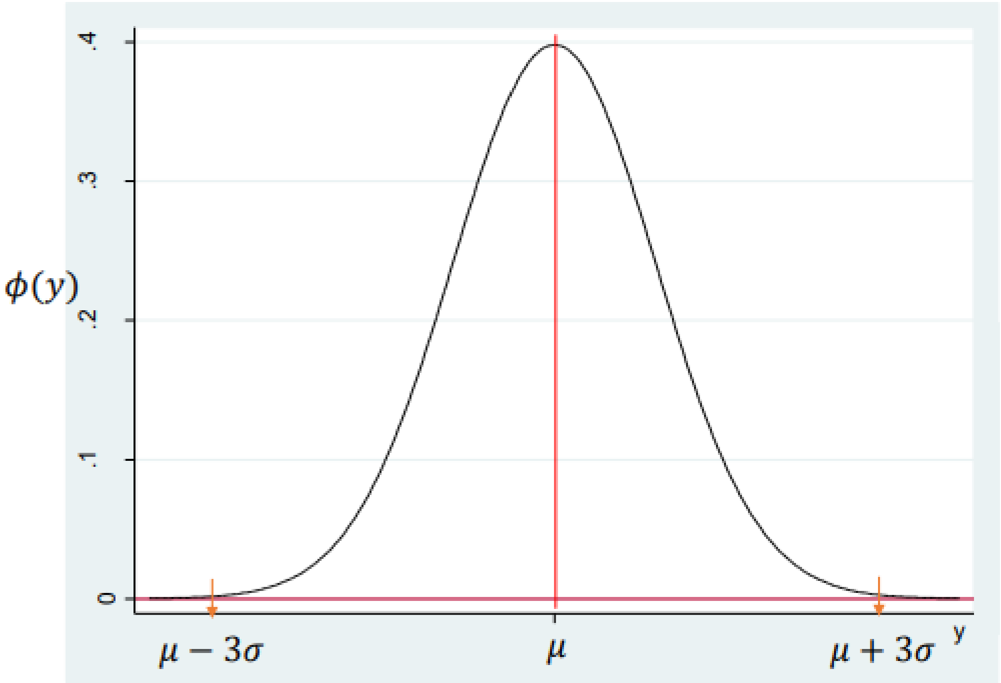
- **Skewness (S)**
- **Kurtosis (K)**

Recall that skewness refers to the asymmetry of the density function of a random variable. The normal distribution is symmetric; hence, its skewness must be zero.

Recall also that kurtosis refers to the concentration of the random variable values about its mean and hence refers to the area under the tails of the distribution.

The normal distribution has kurtosis = 3, meaning that all area under the curve is practically contained between ± 3 standard deviations from the mean.

The Standard Normal Density Function



Source: Own elaboration using Stata 17

The Jarque-Bera Test

The Jarque-Bera test is based on testing that the two conditions of the normal distribution hold for the distribution of any random variable we are testing (the regression errors in this case).

The null and alternative hypotheses are:

$$H_0: \quad S = 0 \quad \text{and} \quad = 3 \quad (= 0) \Rightarrow \text{The distribution is ormal}$$

$$H_1: \quad S \neq 0 \quad \text{and/or} \quad \neq 3 \quad (> 0) \Rightarrow \text{The distribution is T ormal}$$

Test Statistic

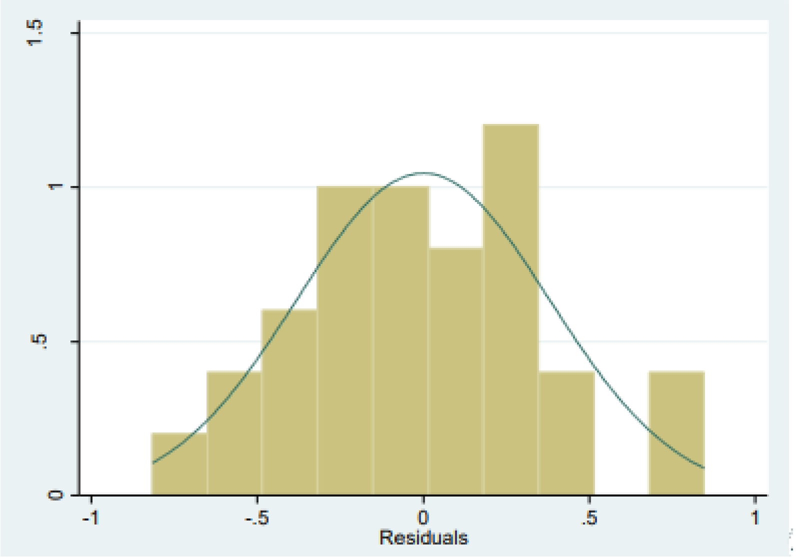
Under H_0 being true, the test statistic is:

$$B = \frac{N}{6}\left(S^2 + \frac{(- 3)^2}{4}\right) \sim {}^{2}_{(2)}$$

Visual Inspection

The log-log model:
We have the histogram of the estimated residuals.

Visually, we may conclude that the distribution is NOT normal; however, we must test if statistically, the distribution is normal.



Jarque-Bera Test Calculation

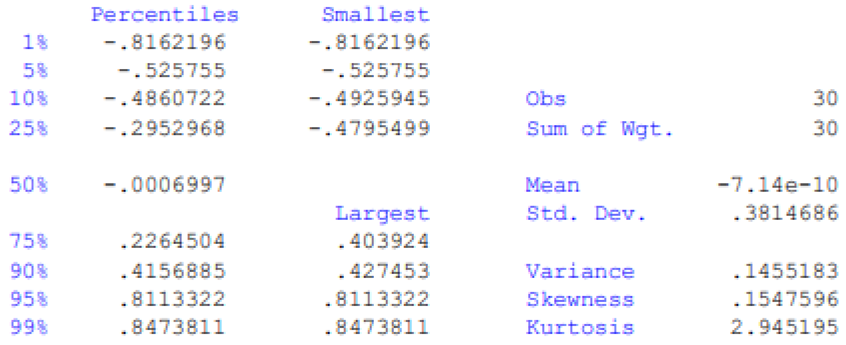
To do so, we may calculate the **JB statistic**; the estimated skewness and kurtosis are reported by STATA using the detailed summary function. For the log-log model we have:

$$B = \frac{30}{6}\left(1.547596^2 + \frac{(2.945195-3)^2}{4}\right)$$

$$= 0.12350715$$

The critical value of the test statistic for a 5% significance level is $^2_{(2)} = 5.9914645$

Given that the sample value of the test statistic is smaller than the corresponding critical value, we may say that there is no evidence to reject the null hypothesis → we cannot reject that the residuals are normally distributed.



```
      Percentiles      Smallest
 1%    -.8162196      -.8162196
 5%     -.525755       -.525755
10%    -.4860722      -.4925945       Obs                  30
25%    -.2952968      -.4795499       Sum of Wgt.          30

50%    -.0006997                      Mean          -7.14e-10
                        Largest       Std. Dev.      .3814686
75%     .2264504        .403924
90%     .4156885        .427453       Variance       .1455183
95%     .8113322       .8113322       Skewness       .1547596
99%     .8473811       .8473811       Kurtosis       2.945195
```

NOTE: Stata calculates the JB statistic if you have the corresponding **ado** file.

Using the JB function

jb ehat

```
Jarque-Bera normality test: .1235 Chi(2) .9401
Jarque-Bera test for Ho: normality:
```

1.19 Changing the scale of the variables

What are the effects of scaling the variables of the regression model?

Changing the scale of the variables = changing the units in which they are measured

1) Changing the scale of the explanatory variable X

If we have a model $y_i = \beta_1 + \beta_2 x_i + e_i$ and we change the scale of x such that:

$$x_i^* = \frac{x_i}{c}$$

Then, in order to keep the equality of the left and right-hand sides, the coefficient of x must be multiplied by c; that is:

$$y_i = \beta_1 + c\beta_2 \frac{x_i}{c} + e_i$$

Hence:

$$y_i = \beta_1 + \beta_2^* x_i^* + e_i$$

Note that $\beta_2^* = c\beta_2$, implying that when scaling the variable x_i, β_2^* is c times larger than β_2.

Therefore, the effects of changing the scale of x_i are:

- b_2 is re-scaled (multiplied) by the factor c
- $se(b_2)$ is re-scaled also (multiplied by c), such that the t-ratio of b_2 does not change

$$t = \frac{b_2}{se(b_2)}$$

The R^2 does NOT change

2) Changing the scale of the dependent variable Y

If we change the scale of y_i, dividing by c for example, we must divide the right-hand side by the same factor c in order for the equation to remain valid:

$$\frac{y_i}{c} = \frac{\beta_1}{c} + \frac{\beta_2}{c} x_i + \frac{e_i}{c}$$

$$y_i^* = \beta_1^* + \beta_2^* x_i + e_i^*$$

$$\beta_2^* = \frac{y_i^*}{x_i}$$

The effects of scaling the dependent variable y_i are:

- All the coefficients are scaled (divided) by the same factor c
- The residuals are also scaled (divided) by the same factor c
- The standard errors of the estimated coefficient, $se(b_1), se(b_2)$ are also scaled (divided by c), such that the t-ratios of b_1 and b_2 do not change:

$$t_{(b_1)} = \frac{b_1}{se(b_1)}; \quad t_{(b_2)} = \frac{b_2}{se(b_2)}$$

- The R^2 DOES NOT CHANGE

3) Changing the scale of both dependent and explanatory variables by the same factor

In this case we have:

$$\frac{y_i}{c} = \frac{\beta_1}{c} + \beta_2 \frac{x_i}{c} + \frac{e_i}{c}$$

$$y_i^* = \beta_1^* + \beta_2 x_i^* + e_i^*$$

The effects of scaling both variables by the same factor are:

- b_2 DOES NOT change scale
- The intercept b_1 changes scale; that is, is divided by c
- The residuals are re-scaled (divided) by the same factor c
- The t-ratios of b_1 and b_2 do NOT change ($t_{(b_1)} = \frac{b_1}{se(b_1)}$; $t_{(b_2)} = \frac{b_2}{se(b_2)}$)
- R^2 DOES NOT CHANGE