

T2. El modelo lineal simple

Ana J. López y Rigoberto Pérez

Dpto Economía Aplicada. Universidad de Oviedo

Curso 2010-2011



Índice

- 1 Planteamiento e hipótesis básicas
- 2 Estimación de los parámetros de regresión
- 3 Propiedades de los estimadores
 - Teorema de Gauss-Markov
 - Estimación con Gretl
- 4 Intervalos de confianza
- 5 Contrastes asociados a un modelo
 - ANOVA
 - Evaluación de la capacidad explicativa
- 6 Predicción
 - Evaluación de predicciones



El modelo lineal simple

Competencias

El modelo lineal simple ya ha sido estudiado en la asignatura "Introducción a la estadística económica", si bien entonces se adoptaba una óptica descriptiva y ahora se completa con el análisis inferencial, incluyendo la construcción de intervalos de confianza y la realización de contrastes asociados a un modelo.

Una vez superado este tema los alumnos serán capaces de:

- Estimar e interpretar los parámetros de un modelo lineal simple.
- Enunciar y resolver el contraste de significación del modelo.
- Utilizar las opciones de estimación de Grel e interpretar correctamente el output.



Especificación de un modelo

Teoría económica	$Y = f(X)$	Supuestos teóricos
Componente aleatoria	u	Conductas humanas Errores de medida Factores no medibles
Modelo econométrico	$Y = f(X) + u$	

Teoría Keynesiana: $C_t = \beta_1 + \beta_2 R_t$

Hipótesis: $\beta_1 > 0$, $0 < \beta_2 < 1$

Componente errática del consumo: u

Modelo econométrico del Consumo: $C_t = \beta_1 + \beta_2 R_t + u_t$

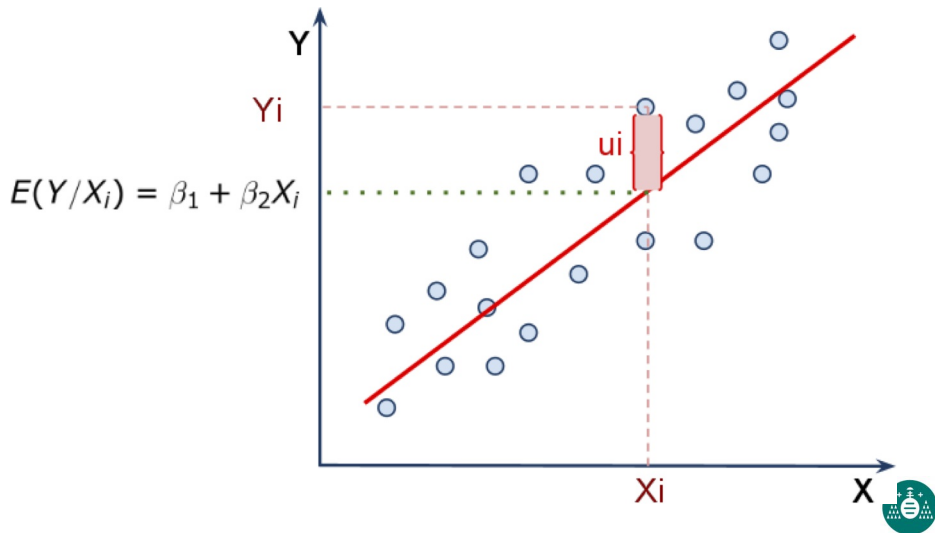
Hipótesis básicas

	Hipótesis sobre u	Supuesto	Hipótesis sobre Y
Esperanza	$E(u_i) = 0$ $\forall i = 1, \dots, n$	Esperanza de la perturbación nula	$E(Y/X_i) = \beta_1 + \beta_2 X_i$ $\forall i = 1, \dots, n$
Varianza	$Var(u_i) = \sigma^2$ $\forall i = 1, \dots, n$	Homocedasticidad	$Var(Y/X_i) = \sigma^2$ $\forall i = 1, \dots, n$
Correlación	$Cov(u_i, u_j) = 0$ $\forall i \neq j = 1, \dots, n$	No autocorrelación	$Cov(Y/X_i, Y/X_j) = 0$ $\forall i \neq j = 1, \dots, n$
Distr.Prob.	$u_i \approx \mathcal{N}(0, \sigma)$	Normalidad	$Y/X_i \approx \mathcal{N}(\beta_1 + \beta_2 X_i, \sigma)$



Modelo de regresión

Línea de regresión poblacional



Estimación del modelo

Información estadística

- Muestras temporales
- Muestras de corte transversal
- Muestras de panel

Métodos de estimación

- Método de mínimos cuadrados
- Método de máxima verosimilitud
- Método de los momentos

Análisis de los estimadores



Estimación

Objetivos:

Estimar un modelo lineal $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ que aproxime lo mejor posible los valores observados de Y .

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

Valores estimados

$$Y_i$$

Valores observados

$$\hat{u}_i = Y_i - \hat{Y}_i$$

Errores de estimación o residuos



Estimación mínimo cuadrática

Función a minimizar

$$\sum_{i=1}^n \hat{u}_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

Estimadores mínimo cuadráticos (EMC)

$$\hat{\beta}_2 = \frac{S_{XY}}{S_X^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

Estimadores mínimo cuadráticos

Propiedades descriptivas

$$\sum_{i=1}^n \hat{u}_i = 0$$

$$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}$$

$$\sum_{i=1}^n X_i \hat{u}_i = 0$$

$$\sum_{i=1}^n \hat{Y}_i \hat{u}_i = 0$$



Estimación máximo verosímil

$$u_i \approx \mathcal{N}(0, \sigma) \quad \Rightarrow \quad Y/X_i \approx \mathcal{N}(\beta_1 + \beta_2 X_i, \sigma)$$

$$f(y_i) = f(y_i, \beta_1, \beta_2, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \frac{(y_i - \beta_1 - \beta_2 x_i)^2}{\sigma^2}}$$

$$\begin{aligned} L(y_1, \dots, y_n, \beta_1, \beta_2, \sigma^2) &= \prod_{i=1}^n f(y_i, \beta_1, \beta_2, \sigma^2) \\ &= \prod_{i=1}^n \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \frac{(y_i - \beta_1 - \beta_2 x_i)^2}{\sigma^2}} \right) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n e^{-\frac{1}{2} \sum_{i=1}^n \frac{(y_i - \beta_1 - \beta_2 x_i)^2}{\sigma^2}} \end{aligned}$$



Estimación máximo verosímil

Función a maximizar

$$\ln L(y_1, \dots, y_n, \beta_1, \beta_2, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2} \sum_{i=1}^n \frac{(y_i - \beta_1 - \beta_2 x_i)^2}{\sigma^2}$$

Estimadores máximo verosímiles (EMV)

$$\hat{\beta}_2 = \frac{S_{XY}}{S_X^2} \quad ; \quad \hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{u}_i^2}{n}$$

Características de los estimadores

Estimadores	Esperanzas	Varianzas
$\hat{\beta}_1$	$E(\hat{\beta}_1) = \beta_1$	$Var(\hat{\beta}_1) = \frac{\sigma^2 \sum_{i=1}^n X_i^2}{n \sum_{i=1}^n (X_i - \bar{X})^2}$
$\hat{\beta}_2$	$E(\hat{\beta}_2) = \beta_2$	$Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$

Propiedades de los estimadores:

Insesgados, Consistentes, Óptimos

Teorema de Gauss-Markov

Dentro de la familia de estimadores lineales e insesgados, los EMC son óptimos en el sentido de que presentan mínima varianza



Distribución de los estimadores

Distribución de $\hat{\beta}_1$ y $\hat{\beta}_2$

Bajo la hipótesis de normalidad de las perturbaciones $u \approx \mathcal{N}(0, \sigma)$ se garantiza la normalidad de los estimadores:

$$\hat{\beta}_1 \approx N\left(\beta_1, \sigma_{\hat{\beta}_1}\right) \quad ; \quad \hat{\beta}_2 \approx N\left(\beta_2, \sigma_{\hat{\beta}_2}\right)$$

Estimador de la varianza

La varianza σ^2 es desconocida y por tanto también lo serán: $\sigma_{\hat{\beta}_1}^2$ y $\sigma_{\hat{\beta}_2}^2$

$$S^2 = \frac{\sum_{i=1}^n \hat{u}_i^2}{n-2}$$

$$E(S^2) = \sigma^2$$

Estimación de las varianzas

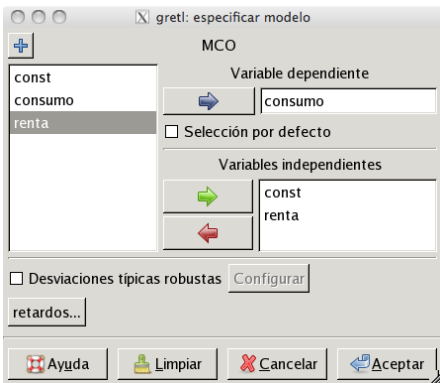
$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2 \sum_{i=1}^n X_i^2}{n \sum_{i=1}^n (X_i - \bar{X})^2}$$

$$S_{\hat{\beta}_1}^2 = \frac{S^2 \sum_{i=1}^n X_i^2}{n \sum_{i=1}^n (X_i - \bar{X})^2}$$

$$S_{\hat{\beta}_2}^2 = \frac{S^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$



Estimación con Gretl



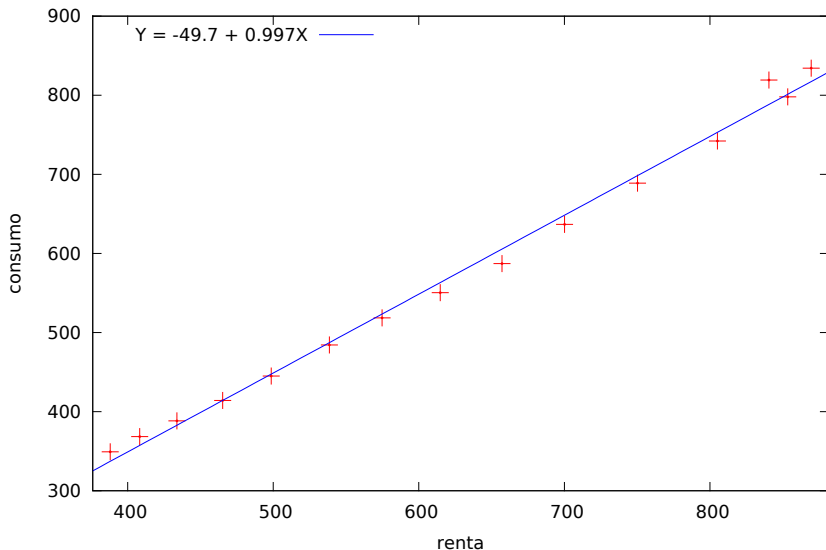
Modelo 1: MCO, usando las
observaciones 1995--2009 ($T = 15$)
Variable dependiente: consumo

	Coefficiente	Desv. típica
const	-49,7299	13,9325
renta	0,997079	0,0215101
	$=\hat{\beta}_2$	$=s_{\hat{\beta}_2}$



Estimación con Gretl

consumo con respecto a renta (con ajuste mínimo-cuadrático)



Análisis inferencial

Para los parámetros de regresión

$$\hat{\beta} \approx \mathcal{N}(\beta, \sigma_{\hat{\beta}}) \Rightarrow \frac{\hat{\beta} - \beta}{\sigma_{\hat{\beta}}} \approx \mathcal{N}(0, 1)$$

$$\Rightarrow \boxed{\frac{\hat{\beta} - \beta}{S_{\hat{\beta}}} \approx t_{n-2}}$$

Para la varianza poblacional

$$\boxed{d_{S^2} = \frac{(n-2)S^2}{\sigma^2} \approx \chi_{n-1}^2}$$

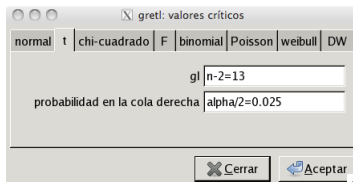
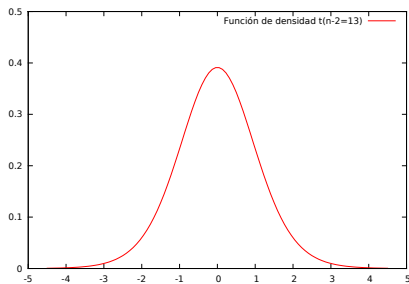


Intervalos de confianza para los parámetros de regresión

IC para β con un nivel de confianza $1 - \alpha$

$$P\left(\left|d_{\hat{\beta}}\right| \leq k_{\alpha}\right) = 1 - \alpha \Rightarrow P\left(\left|\frac{\hat{\beta} - \beta}{S_{\hat{\beta}}}\right| \leq k_{\alpha}\right) = 1 - \alpha$$

$$P\left(\hat{\beta} - k_{\alpha}S_{\hat{\beta}} \leq \beta \leq \hat{\beta} + k_{\alpha}S_{\hat{\beta}}\right) = 1 - \alpha \Rightarrow \left[\hat{\beta} - kS_{\hat{\beta}}, \hat{\beta} + kS_{\hat{\beta}}\right]$$



probabilidad a dos colas = 0.05

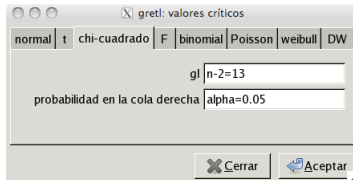
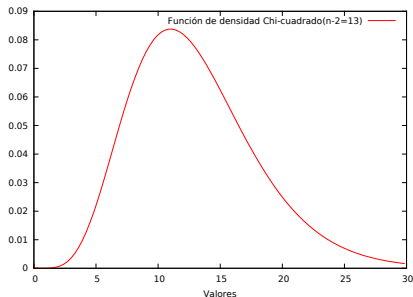
Valor crítico $k_{\alpha} = 2.16037$ 

Intervalos de confianza para la varianza poblacional

IC para σ^2 con un nivel de confianza $1 - \alpha$

$$P\left(\frac{(n-2)S^2}{\sigma^2} < k_1\right) = P\left(\frac{(n-2)S^2}{\sigma^2} > k_2\right) = \frac{\alpha}{2}$$

$$\Rightarrow \left[\frac{(n-2)S^2}{k_2}, \frac{(n-2)S^2}{k_1} \right]$$



probabilidad en la cola derecha = 0.05

Valor crítico $k_{\alpha} = 22.362$



Contrastes de significación: $Y_i = \beta_1 + \beta_2 X_i + u_i$

Contraste básico: ¿explica X los cambios de Y?

- Valor muestral:

$$\begin{array}{l} H_0 : \beta_2 = 0 \\ H_1 : \beta_2 \neq 0 \end{array}$$

$$d_{\hat{\beta}_2}^* = \frac{\hat{\beta}_2}{S_{\hat{\beta}_2}}$$

Contraste individual t de Student

$$\frac{\hat{\beta}_2 - \beta_2}{S_{\hat{\beta}_2}} \approx t_{n-2}$$

- Nivel crítico:

$$p = P\left(|t_{n-2}| > |d_{\hat{\beta}_2}^*|\right)$$

- **Conclusión:** Para p bajo se rechaza la hipótesis (por tanto se concluye que X tiene sentido para explicar Y)

Contraste de significación con Gretl

Modelo 1: MCO, usando las observaciones 1995--2009 ($T = 15$)

Variable dependiente: consumo

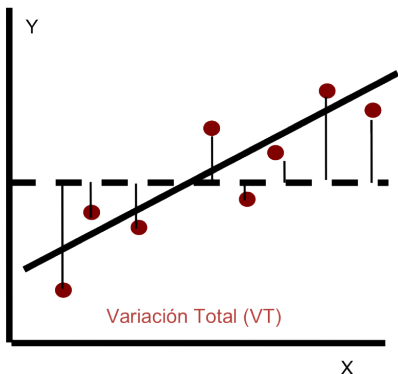
	Coefficiente	Desv. Típica	Estadístico t	Valor p	
const	-49.7299	13.9325	-3.5694	0.0034	***
renta	0.997079	0.0215101	46,3539	0,0000	***
$\frac{\hat{\beta}_2 - 0}{S_{\hat{\beta}_2}} = \frac{0,997079}{0,0215101} \quad p=P(t_{13} > 46,3539)$					

Conclusión

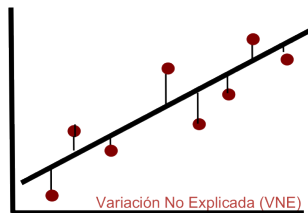
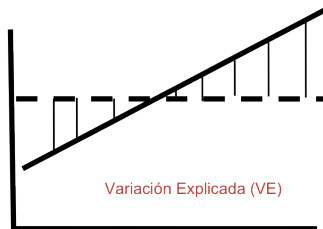
Se rechaza la nulidad del coeficiente de la renta y por tanto ésta es una variable relevante para explicar el consumo



Análisis de la varianza $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$



$$VT = \sum_{i=1}^n (Y_i - \bar{Y})^2$$



$$VE = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 ; VNE = \sum_{i=1}^n \hat{u}_i^2$$



Análisis de la varianza (ANOVA)

$$(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

ANOVA en Gretl

Análisis de Varianza:

	Suma de cuadrados	gl	Media de cuadrados
Regresión (VE)	402180	1	402180
Residuo (VNE)	2433.27	13	187.175
Total (VT)	404614	14	28901

$$R^2 = 402180 / 404614 = 0.993986$$

$$F(1, 13) = 402180 / 187.175 = 2148.69 \text{ [Valor p } 7.98\text{e-16]}$$



Análisis de varianza (ANOVA)

	Variabilidad	g.l.	Ratios
VT	$\sum_{i=1}^n (Y_i - \bar{Y})^2$	n-1	$\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}$
VE	$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_2^2 \sum_{i=1}^n (X_i - \bar{X})^2$	1	$\hat{\beta}_2^2 \sum_{i=1}^n (X_i - \bar{X})^2$
VNE	$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n \hat{u}_i^2$	n-2	$S^2 = \frac{\sum_{i=1}^n \hat{u}_i^2}{n-2}$

$$\frac{\frac{VE}{1}}{\frac{VNE}{n-2}} = \frac{\hat{\beta}_2^2 \sum_{i=1}^n (X_i - \bar{X})^2}{S^2} \approx F_{n-2}^1$$

$$R^2 = 1 - \frac{VNE}{VT} = 1 - \frac{\sum_{i=1}^n \hat{u}_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$



Contraste F

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

Contraste

$$H_0 : \beta_2 = 0$$

$$H_1 : \beta_2 \neq 0$$

$$\frac{\hat{\beta}_2^2 \sum_{i=1}^n (X_i - \bar{X})^2}{S^2} \approx F_{n-2}^1$$

Si el modelo propuesto es adecuado la variación explicada será muy superior a la no explicada, con lo que el ratio F adoptará un valor elevado y su nivel crítico será reducido.

En el modelo lineal simple este contraste es equivalente al de la t de Student ya que se cumple: $F_{n-2}^1 = (t_{n-2})^2$



Medidas de bondad de un modelo

Coeficiente de determinación

Proporción de la variación de Y que viene explicada por X

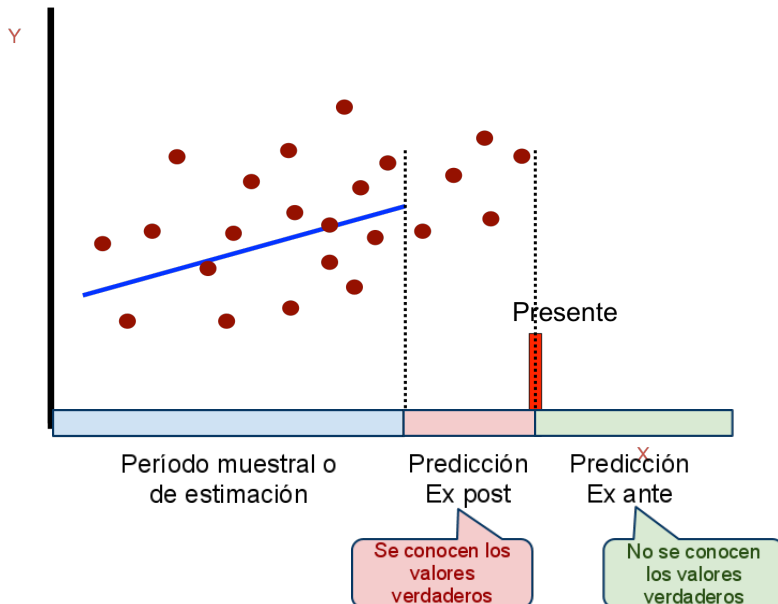
$$R^2 = 1 - \frac{\sum_{i=1}^n \hat{u}_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

Acotación: $0 \leq R^2 \leq 1$

Error estándar de la regresión

$$S = \sqrt{\frac{\sum_{i=1}^n \hat{u}_i^2}{n - 2}}$$

Predicción ex-post y ex-ante



Predicción

Predicciones condicionadas

Los modelos econométricos estimados permiten obtener **predicciones condicionadas** a determinados valores de la variable explicativa.

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0$$

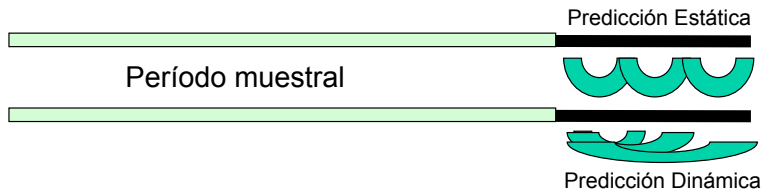
Horizonte de predicción

En modelos temporales, considerando horizontes de predicción 1, 2, 3...T las predicciones se obtendrán sustituyendo en el modelo estimado los correspondientes valores de la variable X en esos periodos.



Predicciones estáticas y dinámicas

Generalmente realizaremos predicciones estáticas, condicionadas a los valores registrados de X y con horizonte de predicción 1. Cuando intervienen como explicativas variables endógenas retardadas es posible realizar predicciones dinámicas, que a medida que aumenta el horizonte de predicción irán condicionadas a las predicciones anteriores.



Elaboración de predicciones

Predicción de Y para un valor X_0

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0$$

Error de predicción

$$e_{\hat{Y}_0} = Y_0 - \hat{Y}_0 = \underbrace{Y_0 - E(Y/X_0)}_{\text{Error poblacional}} + \underbrace{E(Y/X_0) - \hat{Y}_0}_{\text{Error muestral}}$$

Varianza del error de predicción

$$\text{Var}(e_{\hat{Y}_0}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)$$

Elaboración de predicciones

Intervalo de confianza al nivel $1 - \alpha$ para la predicción de Y cuando $X = X_0$

$$\left[\hat{Y}_0 - kS \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}, \hat{Y}_0 + kS \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}} \right]$$

siendo k el valor tal que $P(|t_{n-2}| > k) = 1 - \alpha$



Predicción con Gretl

Gretl: En la salida del modelo, Análisis → Predicciones ...

gretl: predicción

Dominio de predicción:

Inicio	Final
2008	2012

☐ predicción automática (dinámica fuera de la muestra)

☐ predicción dinámica

☒ predicción estática

☐ corriendo las predicciones k pasos hacia adelante: k = 1

Número de observaciones a representar anteriores a la predicción 11

☒ Mostrar los valores ajustados para el rango anterior a la predicción

Representar el intervalo de confianza usando **barras de error**

Predicción con Gretl

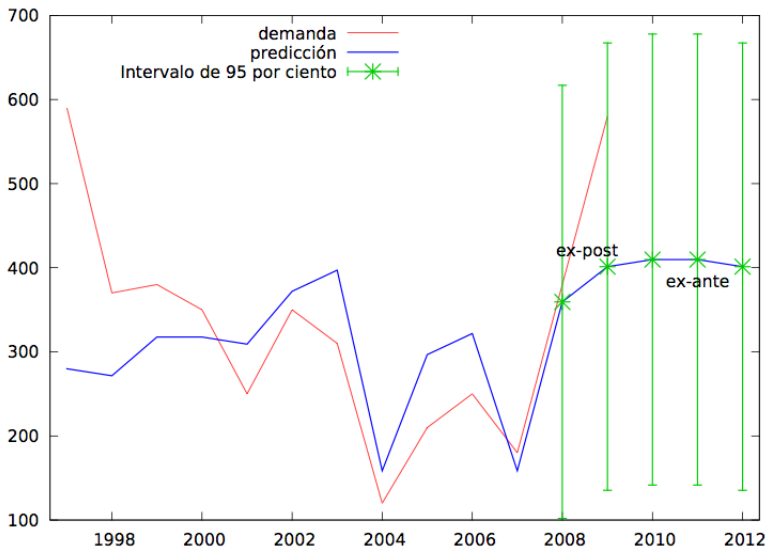
Para intervalos de confianza 95%, $t(22, .0.025) = 2.074$

Obs.	consumo	predicción	Desv. Típica	Intervalo de confianza 95%		
2005	210.00	296.65				
2006	250.00	321.77				
2007	180.00	158.49				
2008	380.00	359.44	124.182	101.91	-	616.98
2009	580.00	401.31	128.245	135.35	-	667.27
2010		409.68	129.293	141.55	-	677.82
2011		409.68	129.293	141.55	-	677.82
2012		401.31	128.245	135.35	-	667.27

2008-2009 predicción ex-post, 2010-2012 predicción ex-ante



Predicción con Gretl



Evaluación de predicciones

Error medio (EM)	$\frac{1}{T} \sum_{t=1}^T (Y_t - \hat{Y}_t)$
Error cuadrático medio (ECM)	$\frac{1}{T} \sum_{t=1}^T (Y_t - \hat{Y}_t)^2$
Raiz del error cuadrático medio (RECM)	$\sqrt{\frac{1}{T} \sum_{t=1}^T (Y_t - \hat{Y}_t)^2}$
Error absoluto medio (EAM)	$\frac{1}{T} \sum_{t=1}^T Y_t - \hat{Y}_t $
Porcentaje de error medio	$\sum_{t=1}^T \frac{(Y_t - \hat{Y}_t)}{TY_t} 100$
Porcentaje de error absoluto medio	$\sum_{t=1}^T \frac{ Y_t - \hat{Y}_t }{TY_t} 100$



Índice de Theil

U de Theil	$U = \sqrt{\frac{\frac{1}{T} \sum_{t=1}^{T-1} \left(\frac{Y_{t+1} - \hat{Y}_{t+1}}{Y_t} \right)^2}{\frac{1}{T} \sum_{t=1}^{T-1} \left(\frac{Y_{t+1} - Y_t}{Y_t} \right)^2}}$
------------	--

El índice de Theil puede ser interpretado como el ratio entre las raíces del error cuadrático medio asociadas al modelo propuesto y a un modelo "naive" o ingenuo que asignase como predicción el valor actual ($\hat{Y}_{t+1} = Y_t$).

Predicciones ingenuas	$\hat{Y}_{t+1} = Y_t$	U=1
Predicciones perfectas	$\hat{Y}_{t+1} = Y_{t+1}$	U=0



Índice de Theil

Además Theil propone una descomposición de los errores cuadráticos de predicción en tres términos, denominados respectivamente de sesgo, de regresión y de perturbación:

Proporción de sesgo	$\frac{(\bar{\hat{Y}} - \bar{Y})^2}{ECM}$
Proporción de regresión	$\frac{(S_{\hat{Y}} - r_{Y\hat{Y}} S_Y)^2}{ECM}$
Proporción de error	$\frac{(1 - r_{Y\hat{Y}}^2) S_Y^2}{ECM}$

Es deseable que las proporciones de sesgo y de regresión sean lo más pequeñas posibles.



Evaluación de predicciones con Gretl

Gretl: En la salida del modelo, Análisis → Predicciones ...

Estadísticos de evaluación de la predicción

Error medio	99.623
Error cuadrático medio	16177
Raíz del Error cuadrático medio	127.19
Error absoluto medio	99.623
Porcentaje de error medio	18.109
Porcentaje de error absoluto medio	18.109
U de Theil	0.89346
Proporción de sesgo, UM	0.61353
Proporción de regresión, UR	0.38647
Proporción de perturbación, UD	2.7453e-16

