Herramientas Econometricas
Modelo de Proyección (Regresión) Lineal

Pasquini, Ricardo

UCA

October 16, 2024

Modelo Lineal (Proyección Lineal)

- Para presentar el modelo de regresión (múltiple) es útil la notación:

$$\mathbf{x} = \left\{ \begin{array}{c} x_1 \\ x_2 \\ \cdots \\ x_{k-1} \\ 1 \end{array} \right\} \boldsymbol{\beta} = \left\{ \begin{array}{c} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_{k-1} \\ \beta_k \end{array} \right\}$$

entonces

$$\mathbf{x}'\boldsymbol{\beta} = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{k-1} x_{k-1} + \beta_k x_k$$

Modelo CEF lineal o Regresion Lineal

- El modelo de Regresion Lineal queda definido por:

$$y = \boldsymbol{x'}\boldsymbol{\beta} + \varepsilon$$

$$E[\varepsilon|\boldsymbol{x}] = 0$$

- Usar este modelo en vez del CEF implica de hecho optar por una proyección lineal (para cualquier x).
- Nuestro modelo pierde flexibilidad, pero también tiene ventajas: nos permite realizar predicciones donde no tenemos datos de x.

Modelo CEF lineal o Regresion Lineal

- Un supuesto adicional define el Modelo de Regresion Homoscedastico:
$$E[\varepsilon^2|\boldsymbol{x}] = \sigma^2$$
(La varianza es constante independiente de **x**)
- Notar que este supuesto se utiliza para simplificar el análisis de los modelos, pero no es algo que podemos suponer en general. Al contrario, en general esperamos que exista variabilidad condicional a x.

Encontrando el Mejor Predictor Lineal

En general, también existe un mejor modelo lineal, en el sentido que existe un modelo lineal que minimiza el error cuadrático medio.

Proof.

$$argmin_{b \epsilon \Re^k} S(\beta) = E[(y - \mathbf{x}'\beta)^2]$$

$$= E[y^2] - 2\beta' E[\mathbf{x}y] + \beta' E[\mathbf{x}\mathbf{x}']\beta$$

$$0 = \frac{\partial S(\beta)}{\partial \beta} = -2E[\mathbf{x}y] + 2E[\mathbf{x}\mathbf{x}']\beta$$

□

$$\Rightarrow \beta = (E[\mathbf{x}\mathbf{x}'])^{-1} E[\mathbf{x}y]$$

Completamos el Modelo de Proyeccion Lineal

Remark

El modelo lineal de menor error será:

$$y = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon$$

$$E[\mathbf{x}'\varepsilon] = 0$$

$$\beta = (E[\mathbf{x}\mathbf{x}'])^{-1}E[\mathbf{x}y]$$

Variables Categoricas y Dummys

Si los regresores en **x** toman un set finito de valores, entonces el CEF se puede escribir como un modelo lineal. En otras palabras, una regresión lineal con un número finito de valores de x (i.e. categorías) coincide con la estimación del CEF. Supongamos

$$E[y|\text{sexo}) = \begin{cases} \mu_0 & \text{si sexo=hombre} \\ \mu_1 & \text{si sexo=mujer} \end{cases}$$

Variables Categoricas y Dummys

Definimos

$$x_1 = \begin{Bmatrix} 1 & \text{si sexo=hombre} \\ 0 & \text{si sexo=mujer} \end{Bmatrix}$$

$$E[y|x] = \beta_1 x_1 + \beta_2$$

Notar que $\beta_1 = \mu_0 - \mu_1$ y $\beta_2 = \mu_1$

Supongamos

$$E[y|\text{sexo}) = \begin{Bmatrix} \mu_{00} & \text{si sexo=hombre soltero} \\ \mu_{01} & \text{si sexo=hombre casado} \\ \mu_{10} & \text{si sexo=mujer casada} \\ \mu_{11} & \text{si sexo=mujer soltera} \end{Bmatrix}$$

Definimos

$$x_1 = \begin{Bmatrix} 1 & \text{si casado/a} \\ 0 & \text{si soltero/a} \end{Bmatrix}, x_2 = \begin{Bmatrix} 1 & \text{si hombre} \\ 0 & \text{si mujer} \end{Bmatrix}$$

$$E[y|x_1] = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4$$

Notar que $\beta_4 = \mu_{11}$, $\beta_2 = \mu_{00} - \mu_{11}$, $\beta_3 = \mu_{01} - \mu_{00} - \mu_{10} - \mu_{11}$, $\beta_1 = \mu_{10} - \mu_{11}$,

CEF y Linealidad

- El CEF es lineal, siempre y cuando las variables explicativas tomen un número finito de categorías.
- Cuando tengo una variable con I categorías puedo reducirlo a un CEF líneal, siempre y cuando traduzca las categorías en $I - 1$ variables dummies.
- Algunas variables que toman un número no muy grande de valores pueden categorizarse. En ese caso una proyección lineal y el CEF serían equivalente.

Proyeccion Lineal Vs CEF. Ejemplos

1. Modelo con Interaccion vs. sin interaccion. Caso Sexo y condicion de inmigrante

El CEF es equivalente a la especificación con interacciones:

```
. regress logingreso mujer inmigrante mujerinmigrante
```

Source	SS	df	MS
Model	369.089695	3	123.029898
Residual	7129.97081	10,107	.705448779
Total	7499.06051	10,110	.741746835

Number of obs	=	10,111
F(3, 10107)	=	174.40
Prob > F	=	0.0000
R-squared	=	0.0492
Adj R-squared	=	0.0489
Root MSE	=	.83991

logingreso	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mujer	-.295399	.0194459	-15.19	0.000	-.3335168	-.2572813
inmigrante	-.2435053	.0280165	-8.69	0.000	-.2984232	-.1885875
mujerinmigrante	-.0276682	.0381717	-0.72	0.469	-.1024924	.047156
_cons	9.50513	.0140199	677.97	0.000	9.477648	9.532612

En el caso de la mujer inmigrante (versus otras mujeres) deberia sumarse -0.02 a los -0.24 del efecto inmigrante. Un total de -0.26.

Proyeccion Lineal Vs CEF. Ejemplos

1. Modelo con Interaccion vs. sin interaccion

Al estimar sin interacción obtenemos:

```
. regress logingreso mujer inmigrante
```

Source	SS	df	MS
Model	368.719064	2	184.359532
Residual	7130.34144	10,108	.705415655
Total	7499.06051	10,110	.741746835

Number of obs	=	10,111
F(2, 10108)	=	261.35
Prob > F	=	0.0000
R-squared	=	0.0492
Adj R-squared	=	0.0490
Root MSE	=	.83989

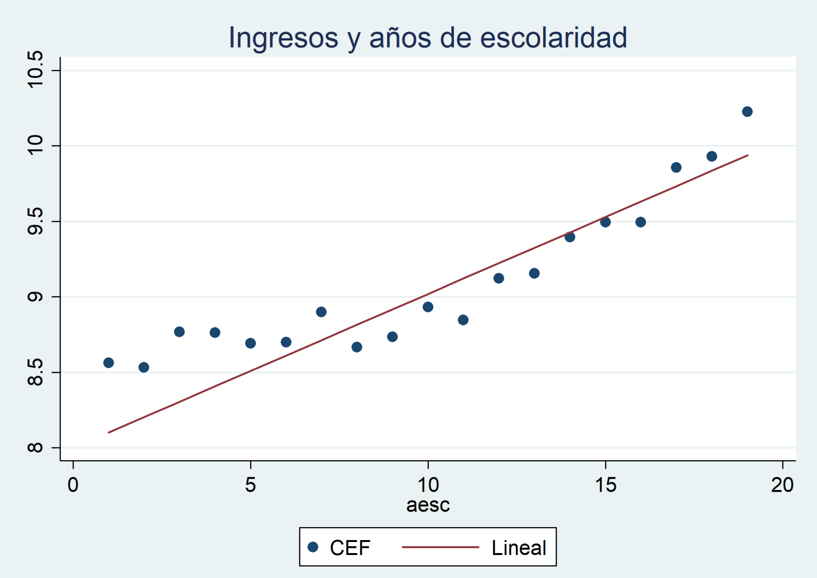
logingreso	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mujer	-.3025795	.016733	-18.08	0.000	-.3353795	-.2697795
inmigrante	-.25841	.0190282	-13.58	0.000	-.295709	-.221111
_cons	9.508862	.0130398	729.22	0.000	9.483302	9.534423

La intuición es que la proyeccion lineal promedia los casos que no surgen en la interaccion. En este caso el efecto inmigrantes (sin diferenciar sexo es -0.25, aun cuando ya incorporamos sexo como explicativa)

Proyeccion Lineal Vs CEF. Ejemplos

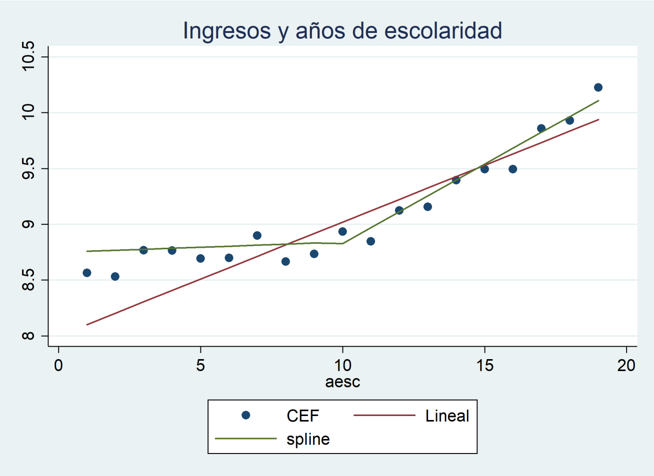
2. Lineal cuando el ajuste no es bueno

A veces la relacion lineal es buena solo en un segmento



Proyeccion Lineal Vs CEF. Ejemplos

2. Lineal cuando el ajuste no es bueno

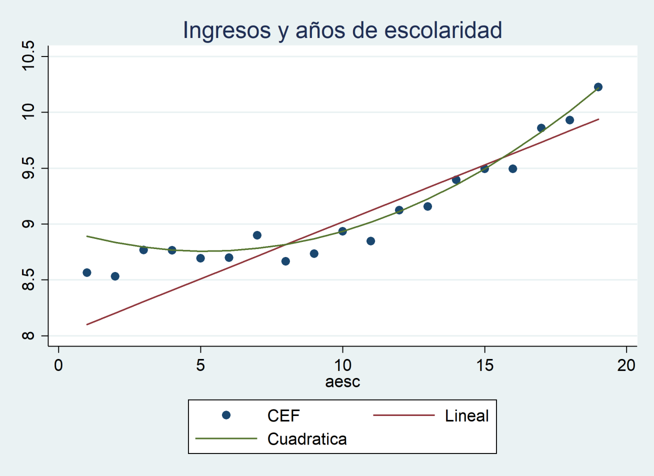


$$P(log(ingreso)|esc) = \beta_1 + \beta_2 esc + \beta_3(esc - 9) * 1(esc >= 9)$$

Proyeccion Lineal Vs CEF. Ejemplos

3. Proyeccion Cuadratica

$$P(log(ingreso)|experiencia) = \beta_1 + \beta_2 experiencia$$
$$P(log(ingreso)|experiencia) = \beta_1 + \beta_2 experiencia + \beta_3 experiencia^2$$



Bondad de Ajuste - R^2

- El coeficiente R^2 se define como la proporcion de la varianza de la variable dependiente Y que es explicable por la variable X.
- Paso 1: Calcular los valores predichos de Y utilizando la ecuación de regresión:

$$\hat{Y}_i = \mathbf{x}'\hat{\beta}$$

  Donde $\hat{Y}_i$ es el valor predicho de Y para la i-ésima observación, y $\hat{\beta}_0$ y $\hat{\beta}_1$ son los coeficientes estimados obtenidos del análisis de regresión.
- Paso 2: Calcular la suma total de cuadrados (TSS), que mide la variabilidad total en la variable dependiente:

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

  Donde $\bar{Y}$ es la media de los valores observados de Y.

Bondad de Ajuste - R^2

- Paso 3: Calcular la suma de cuadrados residual (RSS), que mide la variabilidad que no es explicada por el modelo de regresión:

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

- Paso 4: Calcular la suma de cuadrados explicada (ESS), que mide la variabilidad explicada por el modelo de regresión:

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

- Paso 5: Calcular R^2 utilizando la fórmula:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

  Esta fórmula representa la proporción de la variabilidad total en Y que es explicada por el modelo de regresión.

R^2 Ajustado

- $R^2 = 1 - \frac{RSS/n}{TSS/n}$ puede pensarse como un estimador de

$$\rho = 1 - \frac{\sigma_u^2}{\sigma_y^2}$$

  donde σ_u^2 y σ_y^2 son las varianzas poblacionales
- RSS y TSS son estimadores sesgados, y es necesario hacer una transformación a la medida en relación al número de variables explicativas utilizadas.

$$R_a^2 = 1 - \frac{RSS/(n-k-1)}{TSS/(n-1)}$$

- R_a^2 incorpora una penalidad proporcional al número de variables explicativas.

Error Cuadrático Medio (MSE)

- El MSE es una medida de la calidad de un modelo de regresión.
- Indica el promedio de los cuadrados de los errores, es decir, la diferencia entre los valores predichos por el modelo y los valores reales de la variable dependiente.

Error Cuadrático Medio (MSE)

Fórmula

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- Donde Y_i es el valor real de la variable dependiente para la i-ésima observación.
- $\hat{Y}_i$ es el valor predicho por el modelo de regresión para la i-ésima observación.
- n es el número total de observaciones.

Error Cuadrático Medio (MSE)

- **Paso 1:** Calcular los errores para cada observación:

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

- **Paso 2:** Elevar al cuadrado cada error:

$$\hat{\varepsilon}_i^2 = (Y_i - \hat{Y}_i)^2$$

- **Paso 3:** Calcular el promedio de los errores al cuadrado:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

Ausencia de Sesgo del Estimador OLS

- Definición de estimador sin sesgo
- El estimador OLS no tiene sesgo:

$$E(\hat{\beta}_j) = \beta_j, j = 0, 1, \ldots k$$

- Comprensión intuitiva de la ausencia de sesgo: En promedio, el estimador OLS es igual al verdadero parámetro de población

Supuestos de OLS para garantizar que no hay sesgo

- Supuestos necesarios para la ausencia de sesgo del estimador OLS:
  1. Linealidad en los parámetros (el modelo poblacional es $y = \beta_0 + \beta_1 X + \varepsilon$)
  2. Muestreo aleatorio (los valores (x_i, y_i) son variables aleatorias del modelo poblacional)
  3. La esperanza condicional del error es cero ($E(\varepsilon|X) = 0$)

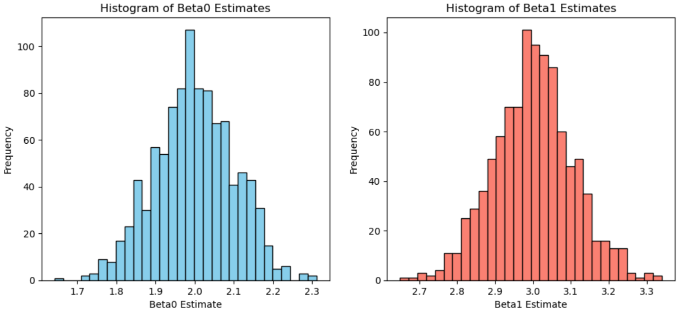
Ausencia de Sesgo del Estimador OLS

- Usando los supuestos, se puede demotrar el resultado.
- Es util hacerlo en dos pasos.
- Primero mostrar que $\hat{\beta}_1 = \beta_1 + \frac{\sum(x_i - \bar{x})\varepsilon_i}{SST_x}$
- Luego $E[\hat{\beta}_1|x] = \beta_1 + \frac{1}{SST_x}\sum(x_i - \bar{x})E[\varepsilon_i|x] = \beta_1$

Idea de la Simulación

- **Objetivo**: Mostrar que los coeficientes del modelo OLS son insesgados.
- **Pasos de la simulación**:
  - **Definición de parámetros**: Establecemos los valores verdaderos de los coeficientes asumidos verdaderos (β_0 y β_1), tamaño de muestra y número de simulaciones.
  - **Iteración**: Realizar múltiples simulaciones.
    - Generamos pares (x,y) provenientes de la poblacion (cumplen $y = \beta_0 + \beta_1 X + \epsilon$)
    - Ajustamos el modelo de regresión lineal y obtenemos los estimadores de los coeficientes.
    - Almacenar los valores estimados de los coeficientes.
  - **Visualización**: Construimos histogramas de los valores estimados de los coeficientes.
  - **Estadísticas descriptivas**: Calculamos media y desviación estándar de los valores estimados de los coeficientes.

Ausencia de Sesgo del Estimador OLS



Varianza del Estimador OLS

- Denotamos la varianza del estimador OLS para la variable x_j como $Var(\hat{\beta}_j)$
- Asumiendo homocedasticidad ($\sigma^2 = Var(\varepsilon)$), se puede derivar que la varianza del estimador OLS esta dada por

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2(1 - R_j^2)}$$

- Interpretación: mide la dispersión o variabilidad de la distribución muestral del estimador OLS.

Regresión Múltiple - Motivación

- Nos interesa agregar al modelo una cantidad arbitraria de variables explicativas. Por eso vamos a estudiar los modelos del tipo:

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

  donde $x_1, x_2, ..., x_k$ representan k variables explicativas.

- Para la interpretación tomemos un ejemplo. Anteriormente estudiamos el modelo del ingreso (ing) de las personas en base a los años de educación ($educ$). Ahora incorporamos adicionalmente la experiencia (exp):

$$ing = \beta_0 + \beta_1 educ + \beta_2 exp + \epsilon$$

  - Interpretamos β_1: este coeficiente mide el efecto de un año de educación, una vez que *ya tenemos en cuenta* el efecto de los años de experiencia.

Comparación de Modelos

Modelo con Experiencia

$$ing = \beta_0 + \beta_1 educ + \beta_2 exp + \epsilon \tag{1}$$

- Interpretamos β_1: este coeficiente mide el efecto de un año de educación, una vez que *ya tenemos en cuenta* el efecto de los años de experiencia. También podemos interpretar como "manteniendo la experiencia constante".

Modelo sin Experiencia

$$ing = \beta_0 + \beta_1 educ + \epsilon \tag{2}$$

- En este caso, β_1 capturaba el efecto de un año extra de educación. Pero, ¿qué pasaría si la gente que se educa más siempre tiene más experiencia?
- Ya que un año extra de educación también implica mayor experiencia, entonces lo que capturaría β_1 sería el efecto del año extra de educación pero conteniendo el efecto de la experiencia

Regresión Múltiple (cont.)

Estimador y Propiedades Estadísticas

- Las estimaciones del modelo de los parámetros (i.e., $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$) se obtienen mediante Mínimos Cuadrados Ordinarios (OLS).
- ¿Qué sabemos sobre las propiedades estadísticas de estos estimadores?
  1. Sabemos que estos estimadores son *insesgados*.
  2. Sabemos que la varianza toma la siguiente forma:

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

Componentes de la Varianza

Fórmula de la Varianza de $\hat{\beta}_j$

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

donde:

- σ^2 es la varianza del error (n.b., que es desconocida pero estimable como ya vimos anteriormente).
- SST_j es la suma de los cuadrados totales de la j-ésima variable explicativa ($SST_j = \sum_{i=1}^{n}(x_{i,j} - \bar{x}_j)^2$). La suma de los cuadrados totales es una medida de varianza, en este caso de la variable j-ésima.
- R_j^2 es una medida del tipo R^2 de bondad de ajuste que ya vimos. En este caso, el subíndice j indica que esta medida es el R^2 de otra regresión (no la que estamos analizando). R_j^2 se refiere a la bondad de ajuste del modelo que busca explicar x_j en función del resto de las variables explicativas del modelo.

Regresión Múltiple (cont.)

Inclusión de variables irrelevantes

- Incluir una variable irrelevante no implicará un sesgo en la estimación.
- La inclusión de una variable irrelevante implica una pérdida de eficiencia en la estimación.

Inclusión de Variables Irrelevantes

No hay sesgo en la estimación

- Supongamos que el modelo poblacional (el generador de los datos) es:
$$y = \beta_0 + \beta_1 x_1 + \epsilon$$
- Pero al estimar el modelo, agregamos una variable x_2 como explicativa:
$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \epsilon$$
- En realidad, lo que se hizo no implica que se asumió un modelo incorrecto, ya que el modelo poblacional es equivalente al modelo estimado pero cuando el coeficiente de efecto de x_2 es 0:
$$y = \beta_0 + \beta_1 x_1 + 0 x_2 + \epsilon$$
- No hay inconsistencia entre el modelo poblacional y el estimado, y al estimar el modelo extendido, deberíamos esperar que el valor de $\hat{\beta}_2$ sea cercano a 0.

Regresión Múltiple (cont.)

Omision de variables relevantes

- La omisión de una variable relevante implica un sesgo en los coeficientes estimados.
- Se puede demostrar que la estimación estará sesgada.

Omisión de Variables Relevantes

Sesgo en los Coeficientes Estimados

- Supongamos que el modelo poblacional (el generador de los datos) es:
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$
- Pero al estimar el modelo, no contamos con x_2 como explicativa:
$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \epsilon$$
- Se puede demostrar que la estimación estará sesgada.
  - Es relativamente fácil demostrar que:
  $$E[\tilde{\beta}_1] = \beta_1 + \beta_2 \delta_1$$
  donde δ_1 es el coeficiente de la pendiente de una regresión de x_2 sobre x_1.

Omisión de Variables Relevantes

Sesgo en los Coeficientes Estimados

- El sesgo en la estimación de β_1 está determinado por la magnitud de $\beta_2 \delta_1$.
- La magnitud del sesgo se anularía si:
    - x_2 no tiene efecto en y (i.e., en la población $\beta_2 = 0$), o bien,
    - x_2 no tiene relación con x_1. (i.e., $\delta_1 = 0$.)