Tópicos de Economía Aplicada

Estimaciones

Regresión

Tabla de contenidos

Regresión

Para qué?

- Recordemos
- Que pasa si...
- Regresión, causalidad, problemas
- Cómo?

Regresión lineal. Para qué?

La regresión lineal de y en X tiene diferentes motivaciones

1. Para **predecir**. Por ejemplo: quiero predecir el nivel de ingresos de los egresados universitarios. Entonces me conviene tener un (el mejor) predictor para proyectar $\hat{y}$. La regresión es la mejor predicción lineal.
2. Para **testear** una teoría. Por ejemplo: CAPM dice que y (retorno observado del activo) depende sólo de x_1 (retorno del porfolio de mercado) sin intercepto, y el CAPM define qué es el porfolio de mercado. Puedo testear esto incluyendo en la regresión más variables y analizando si son significativas (además analizando si la constante es distinta de cero).
3. **Medir** diferencias en variables entre grupos. Por ejemplo, quiero medir las diferencias salariales entre hombres y mujeres.
4. Para **estimar efectos**. Por ejemplo: quiero saber cuál es el efecto (promedio) de un año más de educación. Al intentar identificar un efecto (causal) hace falta ir con más cuidado...

Regresión lineal. Por qué?

Si tenemos una variable y y un vector de variables X una regresión nos aproxima la relación entre estas variables con una recta.
En general, querríamos saber la función de esperanza condicional

$$\mathbb{E}\left[y|X\right] = \mu\left(X\right)$$

Por ejemplo, en una relación univariada, querría saber cómo cambia el salario con los años de educación (univariado)

$$\mathbb{E}\left[y|x_1\right] = \mu\left(x_1\right)$$

Esta función puede tener cualquier forma, en principio. La regresión lineal nos brinda la mejor aproximación lineal a esta función

Regresión lineal, recordemos lo básico

- Dado un modelo

$$y = X'\beta + u$$

donde X y β son vectores de k posiciones, el estimador de Mínimos Cuadrados Ordinarios (MCO) de β será el que minimice la suma de los errores al cuadrado[1]

- Empezamos por $k = 1$, $y = \alpha + \beta x + u$. Definimos la media del cuadrado de los errores como

$$MEC(a, b) = \mathbb{E}\left[(y - a - bx)^2\right]$$

y haciendo el FOC para minimizar $MEC(a, b)$ nos da

$$\alpha = \mathbb{E}[y] - \beta\mathbb{E}[x] \tag{1}$$

$$\beta = \frac{cov(y, x)}{V(x)} \tag{2}$$

[1] Por ahora dejamos de lado la posible endogeneidad asumiendo $x_i \perp u$.

- Con $k > 1$, $y = X'\beta + u$, donde X es un vector $k \times 1$, y β es un vector $k \times 1$, podemos definir

$$MEC\left(\beta\right) = \mathbb{E}\left[\left(y - X'\beta\right)^2\right]$$

cuya derivada con respecto a β brinda

$$\hat{\beta}_{MCO} = \mathbb{E}\left[XX'\right]^{-1}\mathbb{E}\left[Xy\right]$$

El efecto de cada variable - Anatomía de la regresión

- En un modelo MCO multivariada el $\hat{\beta}_k$ asociado a x_k depende del resto de las variables en X, y se puede estimar como

$$\hat{\beta}_k = \frac{cov\left(y, \widetilde{x}_k\right)}{v\left(\widetilde{x}_k\right)}$$

  es decir de una regresión univariada de y en $\widetilde{x}_k$, que a su vez es el residuo de la regresión de x_k en el resto de las variables en X. En otras palabras, el $\hat{\beta}_k$ se estima por lo que aporta x_k a la información provista en X, es decir, a la parte de x_k no correlacionada con el resto de las variables en X.
- Vamos a verificar esta fórmula. Primero supongamos que predecimos x_k a partir de una regresión lineal sobre el resto de las variables en X: $\hat{x}_k = X_{-k}\hat{\pi}$, y encontremos entonces la siguiente relación: $x_k = \hat{x}_k + \widetilde{x}_k$. La variable $\widetilde{x}_k$ entonces no está relacionada con las variables X_{-k}.
- Como segundo paso usamos la definición de y en la fórmula de $\hat{\beta}$

$$\begin{aligned}
\hat{\beta}_k &= \frac{cov\left(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k + ... + u, \widetilde{x}_k\right)}{v\left(\widetilde{x}_k\right)} = && (3)\\
&= \frac{cov\left(\beta_0 + \beta_1 x_1 + ... + u, \widetilde{x}_k\right) + cov\left(\beta_k x_k, \widetilde{x}_k\right)}{v\left(\widetilde{x}_k\right)} && (4)\\
&= \frac{0 + \beta_k cov\left(\hat{x}_k + \widetilde{x}_k, \widetilde{x}_k\right)}{v\left(\widetilde{x}_k\right)} = \beta_k \frac{cov\left(\widetilde{x}_k, \widetilde{x}_k\right)}{v\left(\widetilde{x}_k\right)} = \beta_k && (5)
\end{aligned}$$

Regresión lineal, recordemos lo básico: el estimador MCO

- Veamos ahora el análogo en una muestra de la fórmula MCO.
- Con $y = \mathbf{X}\beta + u$, donde y es un vector $1 \times N$, $\mathbf{X}$ es una matriz de datos $N \times k$, y β es un vector $k \times 1$, podemos definir

$$\hat{\beta}_{MCO} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'y$$

Regresión lineal, recordemos el error estándar

- La estimación de β por medio de $\hat{\beta}_{MCO}$ implica :

$$\hat{\beta}_{MCO} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + u) = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'u$$

  por lo que $\hat{\beta}_{MCO}$ tendrá una distribución centrada en β y su error estándar dependerá de u.

- Con $k > 1$, $y = \mathbf{X}\beta + u$, los errores estándar estimados para los coeficientes MCO se calculan a partir de

$$\hat{SE}(\hat{\beta}_j) = \sqrt{s^2 (\mathbf{X}'\mathbf{X})^{-1}_{jj}} \text{ con } s^2 = \frac{\sum_i \hat{u}_i^2}{N - K}$$

  donde se describe el error estándar del coeficiente de la variable j, donde $(\mathbf{X}'\mathbf{X})_{jj}$ es la posición jj de la matriz, y donde se asume homocedasticidad. Es decir que el error estándar de un coeficiente, $SE(\hat{\beta}_k)$, depende de todas las variables en $\mathbf{X}$.

Error estándar de una variable

- Con $k = 1$, $y = \alpha + \beta x + u$, el error estándar es

$$SE\left(\hat{\beta}\right) = \frac{\sigma_u}{\sqrt{N}\sigma_x}$$

donde σ_x es el desvío de la variable x.

Error estándar de una variable - Anatomía de la regresión

- También, se puede volver a recurrir a la idea de pensar que lo relevante para la variable x_k en una regresión multivariada es lo que aporta, la parte no correlacionada con el resto de las variables en $\mathbf{X}$, es decir $\widetilde{x}_k$, y usar

$$SE\left(\hat{\beta}_k\right) = \frac{\sigma_u}{\sqrt{N}\sigma_{\widetilde{x}_k}}$$

Regresión lineal, recordemos el error estándar

- Es importante notar que los errores estándar de un coeficiente MCO
  - se reducen con la muestra (mayor cantidad de observaciones reducirá el error estándar, siempre que la nueva muestra no altere σ_u). (Si la muestra mayor es a expensas de introducir enteramente otra vinculación entre x e y, el error no necesariamente se reduce.)
  - serán más altos cuando σ_u sea mayor (reducir el residuo en la estimación ayuda a estimaciones de β más precisas). (Aunque tenga una estimación insesgada de β_k con sólo x_k, introducir otras variables relevantes explicativas de y puede ayudar a reducir σ_u y el SE.)
  - serán más bajos cuanto mayor sea la variabilidad de x (o $\widetilde{x}_k$). (Supongamos un efecto lineal de una transferencia incondicionada al hogar sobre la oferta laboral. Si quiero medir este efecto con asignaciones aleatorias, con el objetivo de tener precisión en el estimador, conviene que la transferencia vaya de 0$ a 1000$ más que de 0$ a 100$.)

Regresión lineal, recordemos el R^2

- A partir de una regresión,

$$y = \hat{y} + \hat{u}$$

es decir, que la variable y tiene una parte explicada y una parte no explicada

- Consideremos la suma de cuadrados totales, y suma de cuadrados de los residuos, proporcional a la varianza:

$$SC_{TOT} = \sum_i (y_i - \bar{y})^2 \quad (6)$$

$$SC_{RES} = \sum_i (y_i - \hat{y}_i)^2 \quad (7)$$

- Luego, el coeficiente de determinación es

$$R^2 = 1 - \frac{SC_{RES}}{SC_{TOT}} = 1 - \frac{V(\hat{u}_i)}{V(y_i)}$$

donde $V(x)$ es la varianza de x.

- El R^2 se interpreta como la proporción de la variabilidad de y que es explicada por la regresión.

Regresión lineal, recordemos qué pasa si...

1. Heterocedasticidad
2. Multicolinealidad
3. Variables transformadas
4. Interacciones de variables
5. Muestra estratificada

Regresión lineal, recordemos qué pasa si...

- Hay heterocedasticidad? Homocedasticidad implica $\sigma_{ui} = \sigma_u$. Si esto no es así, entonces estamos ante heterocedasticidad. Implica que tenemos que calcular los errores estándar de otra manera, lo que implica que la matriz de varianzas y covarianzas

$$\hat{Var}\left(\hat{\beta}\right) = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{i=1}^{N}\hat{u}_i x_i' x_i\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

Esto modifica los errores estándar (la raíz cuadrada de la diagonal de la matriz) haciéndolos robustos a heterocedasticidad pero no modifica en nada los coeficientes. Es conveniente considerar que puede haber heterocedasticidad. En la práctica es agregar una opción al comando de regresión `.regress y x_1, robust`

- Hay multicolinealidad? Si hay multicolinealidad perfecta en el universo esto es un error conceptual. Si hay multicolinealidad perfecta en la muestra es evidente que hace falta eliminar una variable. Si hay multicolinealidad (alta correlación) entre 2 variables que son controles (cuyo coeficiente no interesa) no es problema grave. Si son coeficientes de interés es un problema (más grave cuanto mayor sea la correlación entre las variables y más pequeña la muestra). En todo caso, se podría intentar redefinir las variables para reducir la correlación (restar la media a veces funciona).
- Si tengo un R2 bajo? No es necesariamente un problema. R2 alto es mejor para predicción. Incluir más variables aumenta el R2. Pero incluir más variables sólo para incrementar R2 no es buena idea.

- Si queremos explicar variables transformadas? Supongamos que tenemos $\ln(w) = \alpha + \beta_1 d + u$, donde la variable dependiente es el logaritmo de los salarios, d es binaria y x es continua. Cómo se interpreta β_1? Considerando un modelo de Rubin, se puede ver que $\beta_1 = \ln(w_1) - \ln(w_0)$, donde w_1 es el salario con $d = 1$ y w_0 con $d = 0$. Es decir que el coeficiente β_1 se mide en puntos log, y es una aproximación al cambio porcentual. Es tradicional representar el cambio porcentual del salario con una transformación del coeficiente: como $\beta_1 = \ln(w_1 / w_0)$, el cambio porcentual de los salarios cuando $d = 1$ es

$$\frac{w_1 - w_0}{w_0} = \exp(\beta_1) - 1$$

- Si queremos utilizar combinaciones de variables? Supongamos que tenemos $y = \alpha + \beta_1 x + \beta_2 x^2 + u$. Cuál es el efecto de x? Depende de dónde evaluemos x: $\frac{\partial y}{\partial x} = \beta_1 + 2\beta_2 x$. Típicamente se evalúa en la media de x. Por ejemplo, en una regresión para ver los efectos de la variable edad

```
. regress y x_1 x_2 c.x_2#c.x_2 x_3
. margins, dydx(x_2) at(x_2=(20 45 60))
```

- Muestras estratificadas? Una muestra estratificada implica que cada observación puede representar distinta cantidad de unidades de un universo poblacional. Por eso, hace falta ponderar. En la práctica esto es ponderar los datos, teniendo cuidado de no ampliar el número de observaciones:
  `. regress y x [pw=pondera]`
- Ponderacion en STATA: algunas aclaraciones

pw=weight (pw: probability weights) ponderadores muestrales que representan la inversa de la probabilidad de participar en la muestra. STATA no cambia el N de la regresión

fw=weight (fw: frequency weights) representan el número de observaciones duplicadas. STATA considera que la cantidad de observaciones debe ampliarse a la suma total de los ponderadores ($N = \sum_i weight_i$) reduciendo el desvío de los coeficientes.

- Ver Angrist & Pischke, Cap 3.4.1
- Para una regresión típicamente utilizamos pw.
- Para datos agrupados se usa fw (Angrist & Pischke, Cap 3.1.2)

Regresión y causalidad

Un aspecto más difícil es el de analizar en qué medida el coeficiente de MCO me brinda informacion sobre la relación causal entre las variables de interés (por ejemplo, salarios y educación).
¿Qué relaciones u objetos son de interés? En general el objeto de interés es

$$\frac{\partial \mathbb{E}\left(y|X\right)}{\partial x_1}$$

(si x es una variable continua) es decir, cuánto aumenta el resultado cuando cambia un input, sólo uno, dejando constante todas las demás variables, incluso los inobservables (en el ejemplo, cuánto aumenta el salario cuando aumenta la educación).
Este efecto puede ser medido según las medias condicionadas o según una regresión.

Problemas

Los problemas para interpretar un resultado de una regresión como una relación o un efecto causal son múltiples, pero pueden ser considerados mayormente como un problema de endogeneidad

Supongamos que tenemos interés en estimar el efecto de x_1 (variable de interés)

Hace falta que se cumpla que la variable de interés sea exógena, es decir, que los inobservables sean independientes de la variable x_1, $\mathbb{E}(u|x_1) = 0$, o también $u \perp x_1$ (u y x_1 deben ser ortogonales).

Si no son ortogonales, entonces habrá un sesgo en β. Ejemplo:

- Si el modelo es

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u,$$

- Supongamos que no observamos x_2 (variable omitida), ahora x_2 será parte del error, e:

$$y = \beta_0^S + \beta_1^S x_1 + e,$$

entonces el $\beta_1^S = \frac{cov(y,x_1)}{v(x_1)}$. Reemplazando y por el modelo

$$\beta_1^S = \frac{cov(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, x_1)}{v(x_1)} = \quad (8)$$

$$= \beta_1 + \beta_2 \pi_{21} \quad (9)$$

donde π_{21} es el resultado de la siguiente regresión: $x_2 = \pi_0 + \pi_{21} x_1 + v$.

Problemas: endogeneidad

Este problema surge por varias razones (vinculadas entre sí):

- Endogeneidad. Las variables inobservables están correlacionadas con la variable de interés.
- Variables omitidas. Hay variables que no incluimos en la regresión y que están correlacionadas con x_1.
- Error de medida. La variable x_1 se mide con error y en ese caso el coeficiente estará sesgado (será en valor absoluto menor)
- Determinación conjunta o simultaneidad. Si una variable (como un precio) resulta de un equilibrio entre oferta y demanda es erróneo considerar una regresión de cantidad consumida y precio como una regresión de la demanda.

Regresión lineal. Cómo?

Regresión y predicción en STATA:

```
. reg y x1 x2
. predict yhat,xb
```

El coeficiente de $x2$ puede recuperarse también de la siguiente manera

```
. reg x2 x1
. predict x2bar, residual
. reg y x2bar
```

El error estándar de $x2$ puede recuperarse de la siguiente manera

```
. reg y x1 x2
. predict uuu, residuals
. sum uuu
. local sduu = r(sd)
. sum x2bar
. local sdx = r(sd)
. disp ‘sduu’/(‘sdx’*sqrt(_N))
```

Regresión lineal. Ejemplo en EPH

Variable de ingresos laborales en EPH: p21

```
. gen lnw = log(p21)
```

Otras variables: educación: nivel_ed, sexo: ch04, edad: ch06

```
. gen sexo = ch04 - 1
. gen edad = ch06
. gen edad2 = edad*edad
. replace nivel_ed=0 if nivel_ed==7 // la variable nivel_ed=7 es
sin educación; se reemplaza por cero
. xi i.nivel_ed, prefix(_E) // convierte la variable nive_ed (de
0 a 6) en 6 dummies con prefijo _E
```

Antes de la regresión: miramos posible multicolinealidad (correlate); miramos si las relaciones son lineales o no (con lowess)

```
. correlate lnw edad* sexo if lnw!=.
. lowess lnw edad, xlabel(15(5)70) ylabel(6(1)11)
lineopts(lwidth(vthick))
```

Regresión

```
. reg lnw sexo edad edad2 _E* [pw=pondera]
```

Efectos de la variable edad

```
. reg lnw sexo edad c.edad#c.edad _E* [pw=pondera]
. margins, dydx(edad) at(edad = (20 40 60))
```

```
. reg lnw sexo edad edad2  E* [pw=pondera]
(sum of wgt is    4.0
```

Nro de observaciones válidas en la regresión

```
Linear regression                               Number of obs =      2996
                                                F(  9,  2986) =     81.07
                                                Prob > F      =    0.0000
                                                R-squared     =    0.2263
                                                Root MSE      =      .746
```

Resultado del test H0 todos los coefs = 0

El modelo explica un 22% de lnw

Variable dependiente (*y*)

Es $s^2 = \frac{\sum_i \widehat{u_i}^2}{N-k}$

lnw	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf. Interval]	
sexo	-.5506008	.0321852	-17.11	0.000	-.6137082	-.4874933
edad	.0574672	.0[illegible]		0.000	.0384075	.0765269
edad2	-.0006273	[illegible]		0.000	-.0008528	-.0004018
_Enivel_ed_1	-.3266775	.2[illegible]		0.204	-.8310796	.1777247
_Enivel_ed_2	-.1909995	.2452562	-0.78	0.436	-.6718877	.2898887
_Enivel_ed_3	-.058713	.2451035	-0.24	0.811	-.5393018	.4218758
_Enivel_ed_4	.1803964	.2442404	0.74	0.460	-.2985001	.6592929
_Enivel_ed_5	.3578336	.2460562	1.45	0.146	-.1246241	.8402913
_Enivel_ed_6	.7403672	.243663	02		.2626014	1.218133
_cons	8.010337	.3111657	25.74	0.000	7.400216	8.620458

Las mujeres tienen un salario (*w*) menor: -42%=exp(-0.55)-1

Error estándar

Variables explicativas (*X*)

Constante

Media de *y* cuando *X=0*

Intervalo de confianza

Resultado del test de dos colas $\beta_k = 0$

Referencias

Angrist & Pischke, "Mostly Harmless Econometrics: An empiricit's Companion", Cap 3