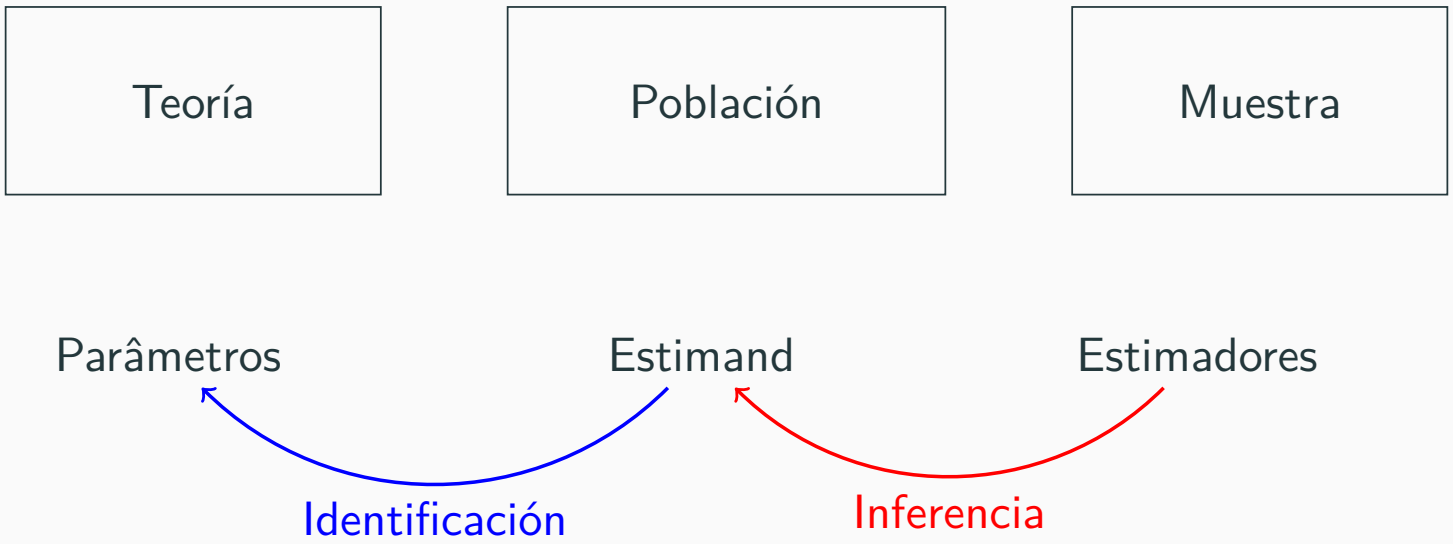
Repaso de Regresión

2025-02-06

Conceptos basicos



Conceptos basicos

La teoría económica postula relaciones sobre las variables. Normalmente un modelo contiene un *proceso generador de datos*, que depende de *parámetros*.

Identificación es el proceso de relacionar un parámetro teórico a alguna cantidad en *la población*.

La población es *fija*, pero trabajamos con una *muestra aleatória*.

Cuando aplicamos estimadores a datos de la muestra, relacionamos los resultados aleatórios a los *estimandos fijos*.

Esperanza condicional

Discutimos la esperanza condicional, pero para una variable continua, no es tan simple usarla.

En su conjunto de datos, puede ser que haya apenas una persona con ingresos iguales a $13,567.90. O tal vez nadie.

Una posibilidad es hacer como hacemos con histogramas y discretizar el espacio.

Binned Scatter

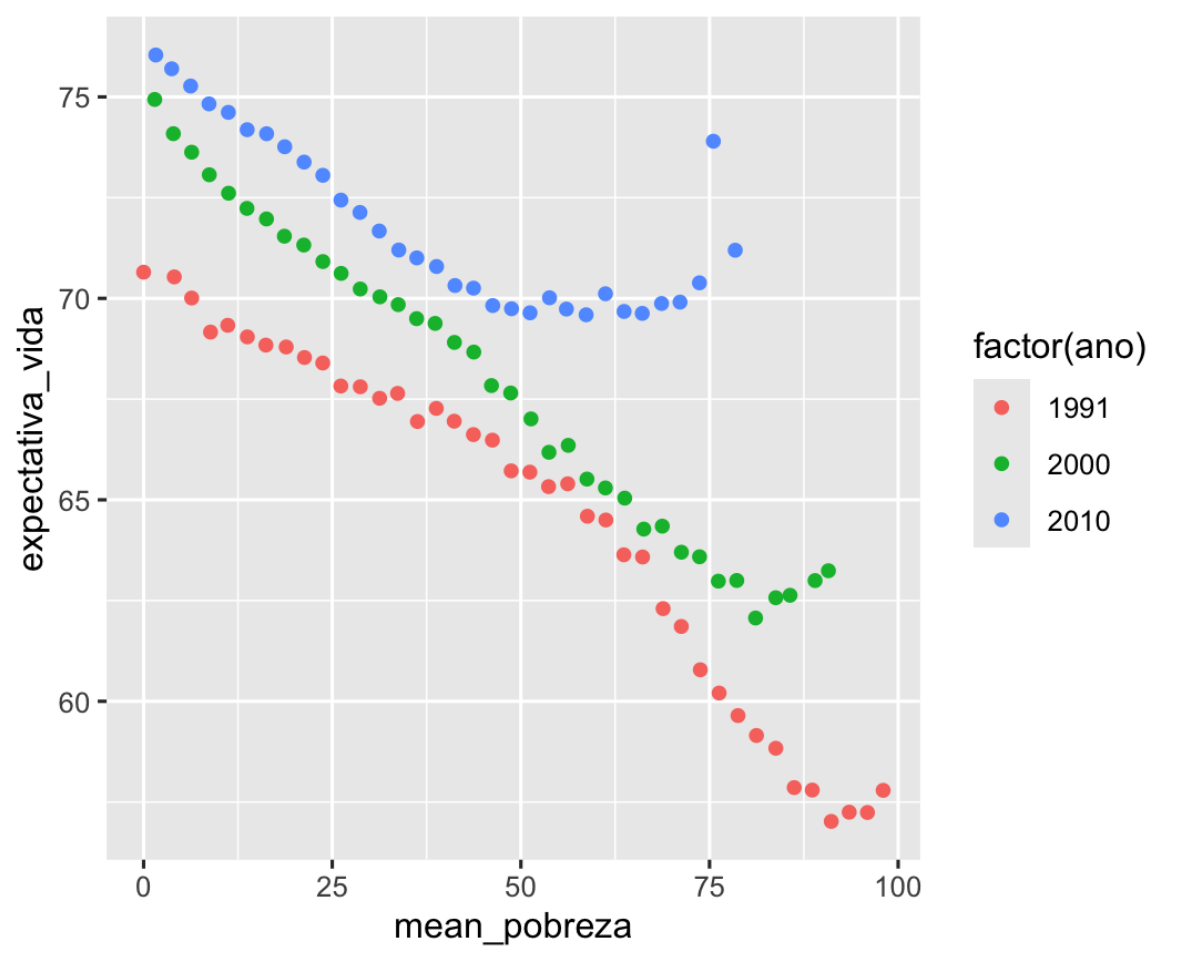


Figure 1: Ejemplo de binscatter

LOESS

Otra posibilidad es estimar la media para puntos próximos de cada valor de X. Una forma de hacerlo se llama LOESS (Locally Estimated Scatterplot Smoothing).

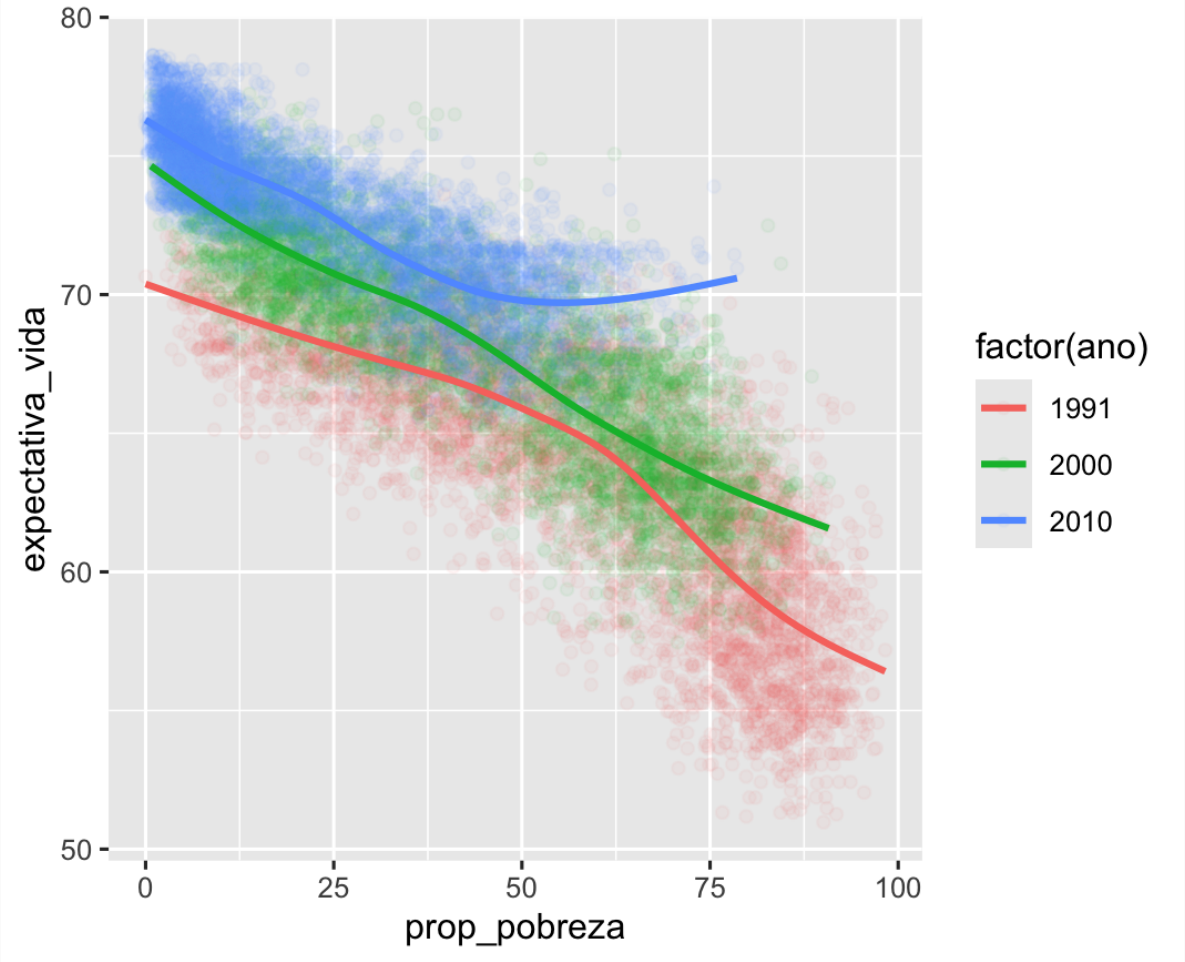


Figure 2: Ejemplo de LOESS

Esperanza condicional

La esperanza condicional tiene ciertas propiedades útiles.

Para cualquier variables Y y X:

$$y = E[y|x] + u$$

Con $E[u|x] = 0$, y $cov(u, f(x)) = 0$.

Eso significa que cualquier variable puede descomponerse en una parte que es "explicada" por x, y otra que es "ortogonal" a x.

Esperanza condicional

Otra propiedad importante es que la esperanza condicional es el mejor predictor de Y, dentro de la clase de funciones de X.

Eso significa que, si quieres una predicción para el valor de Y, y la única información disponible es el valor de X, el mejor posible es $E[Y|X]$.

En este caso, "mejor predición" es la que minimiza la esperanza del cuadrado del error. Es decir:

$$E[Y|X] = \arg\min_{f(x)} E[Y - f(X)]^2$$

Regresión lineal

Mostrar la esperanza de Y para diferentes valores de X es muy útil.

Métodos como binscatter y LOESS son muy buenos para inspeccionar la relación, pero no son tan prácticos.

Una forma de resumir la información aún más es asumir que la relación entre Y y X tiene una forma específica. Por ejemplo, una línea. Esto lleva al modelo lineal:

$$y = \alpha + \beta x + u$$

Regresión lineal

Si la relación es lineal, podemos predecir el valor de Y para cualquier valor de X.

Sumarizamos la información en solamente dos números. Fácil de comprender.

Pero nos limitamos a una relación lineal. Si la relación es diferente, nuestros resultados serán inválidos.

También podemos omitir variación importante.

Regresión Poblacional

La regresión poblacional es la ecuación:

$$y = \alpha + \beta x + u$$

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

Podemos interpretar β como la inclinación de la línea. Para un aumento de 1 unidad de x, tenemos un aumento de β unidades de y.

Si X es ingresos, y Y es expectativa de vida en años, la unidad de medida de β es años por dólar.

Regresión Poblacional

$$y = \alpha + \beta x + u$$

Asumimos que $E[u] = 0$.

Asumimos que u_i es independiente en media de x:
$E[u|x] = E[u] = 0$.

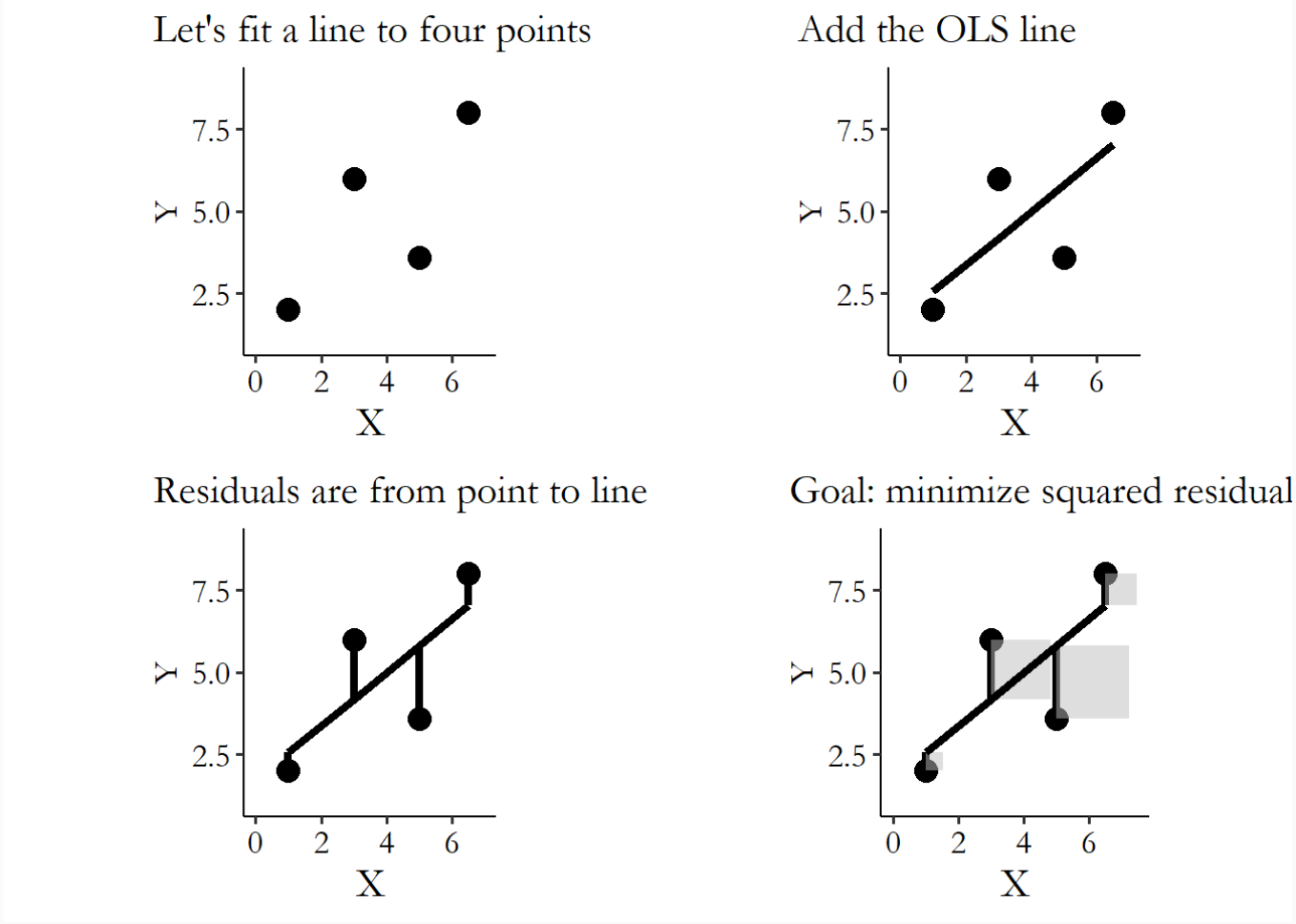
Con estas hipótesis, tenemos que:

$$E[y|x] = \alpha + \beta x$$

Mínimos Cuadrados Ordinarios

Regresión es una herramienta tremendamente útil.

Podemos estimar los parámetros de la regresión como la solución para un problema de minimizar los errores de previsión.



Mínimos Cuadrados Ordinarios

La solución para el problema de MCO es:

$$\hat{\beta} = \frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x})^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Note que:

- $\sum_i (x_i - \bar{x}) y_i$ es el análogo muestral de $cov(x, y)$,
- $\sum_i (x_i - \bar{x})^2$ el de $var(x)$.

Mínimos Cuadrados Ordinarios

Regresión es una herramienta tremendamente útil.

Una razón es que, si la verdadera esperanza Y condicional a X es lineal, entonces es idéntica a la línea de regresión.

Si la esperanza condicional **no es** lineal, entonces la línea de regresión es la mejor aproximación lineal a ella.

¡Recuerda!

La regresión poblacional es un objeto en la población. El estimador de MCO es un objeto construído con datos de una muestra.

- β es la inclinación de la esperanza condicional lineal, **fijo**. $\hat{\beta}$ es su estimador, **aleatório**.
- X es la variable aleatória. x_i es una observación en la muestra.
- u son los errores: la parte de Y que no se explica por X. $\hat{u}$ son los residuos, su contraparte muestral.

Propiedades algébricas de MCO

Por construcción, la suma de los residuos es siempre zero:

$$\sum \hat{u}_i = 0$$

También por construción, los residuos son no correlacionados con X:

$$\sum \hat{u}_i \hat{x}_i = 0$$

Unbiasedness

Decimos que el estimador de MCO es insesgado si su esperanza es el verdadero parámetro.

$$E[\hat{\beta}] = \beta$$

Eso es verdad bajo ciertas hipótesis:

1. El modelo está bien especificado (lineal en los parámetros).
2. Muestra aleatória.
3. $Var(X) \neq 0$
4. $E[u|x] = E[u] = 0$

Unbiasedness

Prueba:

$$\hat{\beta} = \frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x})^2}$$

$$\hat{\beta} = \frac{\sum_i (x_i - \bar{x})(\alpha + \beta x_i + u_i)}{\sum_i (x_i - \bar{x})^2}$$

$$\hat{\beta} = \frac{\sum_i \alpha(x_i - \bar{x}) + \beta(x_i - \bar{x})x_i + u_i(x_i - \bar{x}))}{\sum_i (x_i - \bar{x})^2}$$

$$\sum_i \alpha(x_i - \bar{x}) = \alpha \sum_i (x_i - \bar{x}) = \alpha n(\bar{x} - \bar{x}) = 0$$

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

$$\hat{\beta} = \beta + \frac{\sum_i (x_i - \bar{x}) u_i}{\sum_i (x_i - \bar{x})^2}$$

$$E[\hat{\beta}] = \beta + E[u_i x_i] = \beta + E[E[u_i | x_i] x_i] = \beta$$

Regresión Multipla

Cuando tenemos más de un regressor, digamos x_1 y x_2, el estimador de MCO es:

$$\beta_1 = \frac{\sum_i (\tilde{x}_{1i} - \overline{\tilde{x}_1}) y_i}{\sum_i (\tilde{x}_{1i} - \overline{\tilde{x}_1})^2}$$

Donde $\tilde{x}_{1i}$ son los residuos de una regression de x_1 en x_2.

Regresión Multipla

En una regresión múltipla, cambia la interpretación de β_1.

Ahora, β_1 es el aumento en y asociado con un aumento en x, *menteniendo constante* los otros términos de la regresión (x_2, etc).

Ejemplo

- Kilos de fertilizante por m^2 (X_1)
- Lluvia en mm (X_2)
- Rendimiento en kg por m^2 (Y)

$$Y = 30 + 0.4X_1 + 2X_2$$

Significa que una plantación con 1 kilo más de fertilizante tiene 0.4 kilos de rendimiento, manteniendo constante el nivel de lluvia.

Ejercicio

En la regresión:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

si $\beta_1 = 0$, significa que:

A) X_1 no tiene relación con Y en esos datos
B) X_1 no está correlacionado con Y controlando por X_2 y X_3
C) X_1 e Y tienen una correlación perfecta
D) X_1 debe ser removido del modelo porque no es relevante

Variancia de MCO

Podemos estimar la varianza de MCO así:

$$Var(\hat{\beta}) = \frac{\sum_i (x_i - \bar{x})^2 \hat{u}_i^2}{\sum_i (x_i - \bar{x})^4}$$

Si aceptamos la hipótesis de homocedasticidad, esto se simplifica a:

$$Var(\hat{\beta}) = \frac{\sum_i \hat{u}_i^2}{\sum_i (x_i - \bar{x})^2}$$

Normalmente se llama la primera ecuación de “errores robustos”.

Variancia de MCO

La maioría de los softwares calcula automaticamente la variancia y los errores estándar usando la hipotesis de homocedasticidad.

¡Pero eso no es una buena idea!

Si los errores son homocedasticos, no perdemos nada en usar errores robustos.

Si los errores no lo son, podemos estar calculando muy mal la variancia. Normalmente asumir homocedasticidad **subestima** los errores.

Variancia de MCO

Vamos analizar en más detalles:

$$Var(\hat{\beta}) = \frac{\sum_i \hat{u}_i^2}{\sum_i (x_i - \bar{x})^2}$$

Normalmente queremos minimizar la variância de β, para tener más seguridad en las estimativas.

Para eso podemos minimizar el numerador, o aumentar el denominador.

- Para disminuir los residuos, podemos incluir variables altamente explicativas.
- La variancia de X es dada por la población. Pero, incluir variables que están correlacionada con X disminuye su variancia residual.

Prueba de hipotesis

Queremos usar las estimaciones, derivadas de una muestra aleatória, para hacer inferencias sobre la población.

La prueba de hipótesis funciona así:

- 1: Se selecciona una hipótesis nula, H_0.
  - Por ejemplo: $\beta = 0$
- 2: Si el parámetro real en la población corresponde al que elegimos en la hipótesis nula, la estimativa $\hat{\beta}$ estará distribuida a su alrededor según una distribución normal.
- 3: Chequeamos si el $\hat{\beta}$ que obtuvimos es muy raro para la distribución del paso 2.
- 4: Si el $\hat{\beta}$ es muy raro, se rechaza la hipótesis.

Prueba de hipotesis

Más especificamente:

- Se elige una hipótesis nula: $\beta = \mu$
- La estadística de teste es: $z = \frac{\hat{\beta} - \mu}{\hat{se}}$
- Se compara la estadística de teste contra una distribución Normal.
- Si $|z| > 1.96$, rechazamos la hipótesis a un nível de significancia de 5%.

En terminos prácticos, utilizamos software estadístico para probar hipóteses.

Prueba de Hipótesis

- Cuando no podemos rechazar la hipótesis de que un parâmetro es cero, decimos que “no es estadísticamente signifiativo.”
- No significa que la relación no es importante en términos prácticos. Una muestra pequeña suele tener relaciones no estadisticamente significantes, porque es difícil rechazar hipótesis con pocos datos.

Prueba de Hipótesis

También podemos pruebar hipóteses sobre dos o más coeficientes.

- $\beta_1 = 0$ y $\beta_2 = 0$
- $3 * \beta_1 + 4 * \beta_2 - 2 * \beta_3 = 4$

En este caso, tenemos una **prueba F**. La idea es estimar (1) el modelo sin la restricción, (2) el modelo restricto, y (3) preguntar ¿qué tanto mejor es el modelo irrestricto?

- $SSR_r = \sum_i \hat{u}_{ir}^2$
- $SSR_u = \sum_i \hat{u}_{iu}^2$

$$F = \frac{(SSR_r - SSR_u)}{SSR_u} \cdot \frac{(n - k - 1)}{q}$$

Formas funcionales

Variables binarias o categóricas son muy comunes en ciencias sociales.

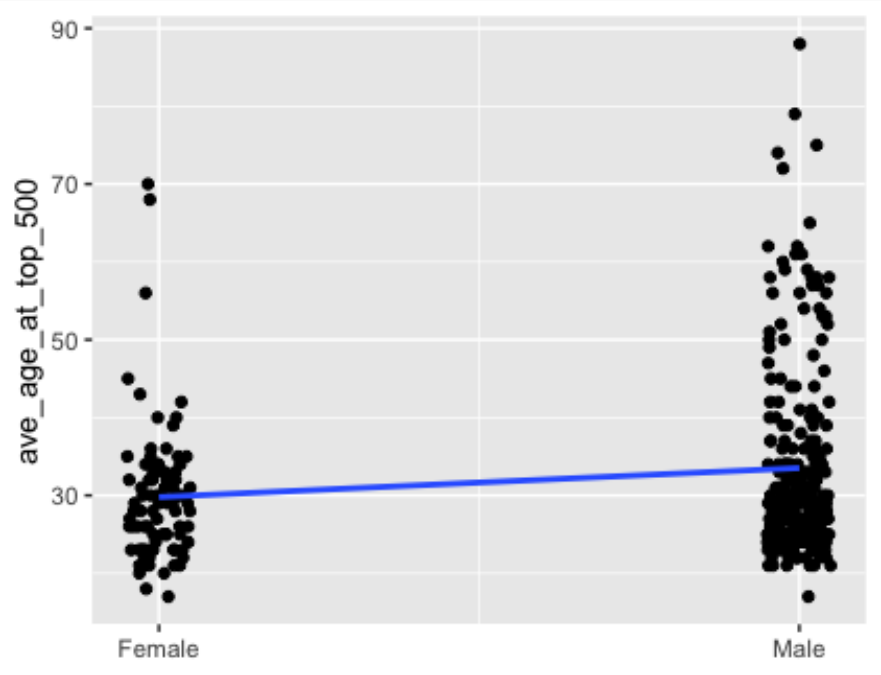
- Hombre o mujer,
- Casado o soltero,
- Raza,
- Religión,
- Tipo de ocupación.

Podemos trabajar con esas variables en regresiones utilizando variables *dummy*, o sea, variables que tienen valor 1 para un cierto grupo, y 0 para otros.

Variable Binaria

Ejemplo: Entre álbumes musicales en el top 500 de la revista Rolling Stone, ¿los hombres son más viejos que las mujeres cuando producen sus discos?

$$Edad_i = 29.8 + 3.7 Hombre_i + u_i$$



Variable Binaria

- En este caso, como 1 = Hombre, y 0 = Mujer, decimos que Mujer es la "referencia".
- La interpretación de β es simplemente la diferencia de promedio entre hombres y mujeres.
- En una regresión simple, el intercepto es el promedio para las mujeres.

Variables Categoricas

¿Qué tipo de disco es más popular?

- Podemos regredir la mayor posición de cada disco contra su "tipo".

	(1)	(2)	(3)
(Intercept)	52.900	164.263	
	(2.939)	(10.773)	
typeLive	-4.900	-116.263	48.000
	(12.610)	(16.323)	(12.262)
typeCompilation	111.363		164.263
	(11.167)		(10.773)
typeStudio		-111.363	52.900
		(11.167)	(2.939)
Num.Obs.	665	665	665

Variables Categóricas

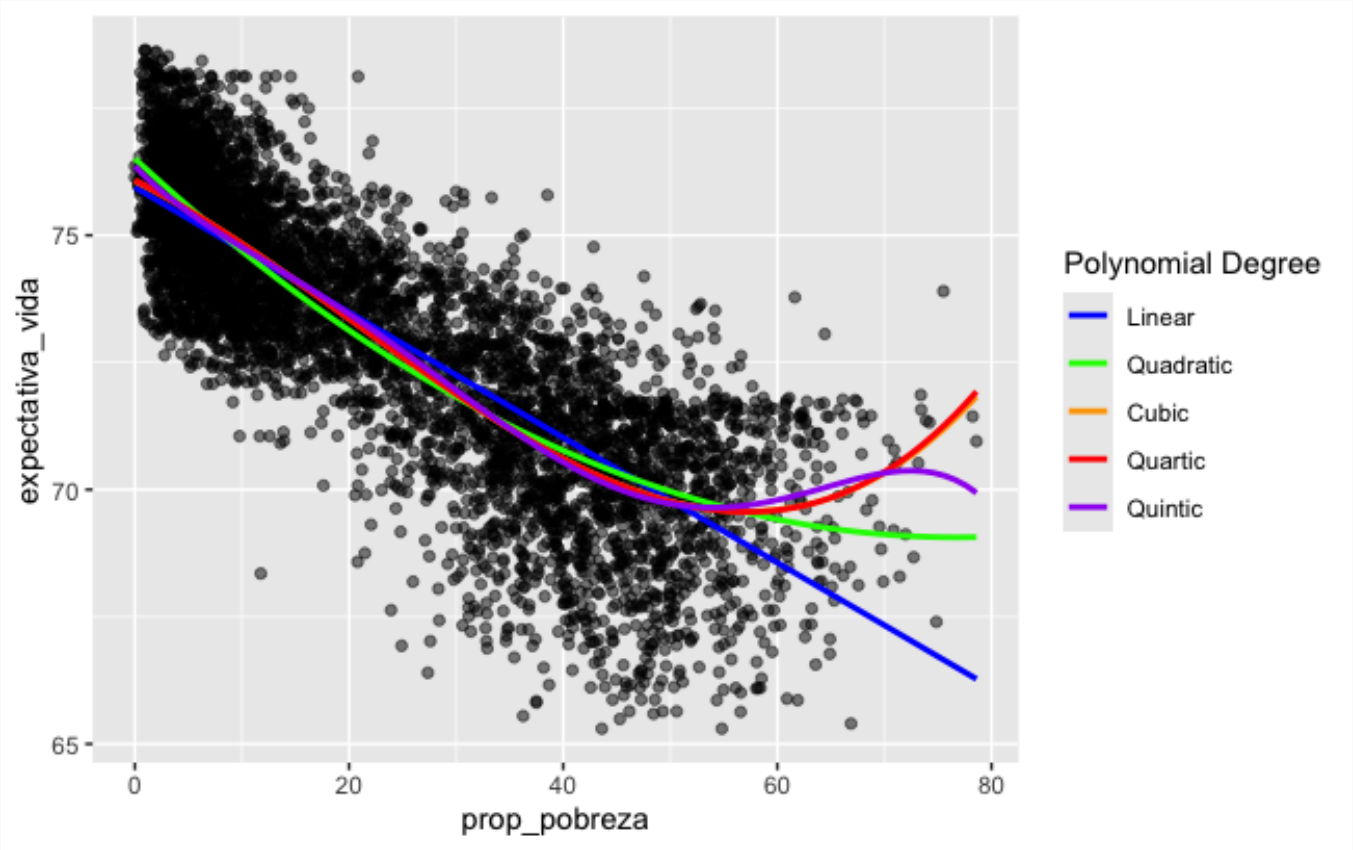
- Para variables categóricas, tenemos que elegir una categoría como referencia.
- Se interpreta cada coeficiente como la diferencia entre su categoría y la referencia.

Transformaciones de variables

Regressiones "lineales" pueden ser más flexibles que simplemente líneas rectas.

- Primeramente, podemos incluir polinomios.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + u_i$$



Transformaciones de variables

¿Cómo se interpretan regresiones con polinomios?

En general un coeficiente se interpreta como el efecto de una variable manteniendo las otras constantes.

¡Pero no es posible aumentar X^2 mantenendo X constante!

Para obtener una interpretación en términos de efecto marginal, es necesario usar una derivada.

Ejemplo

$$E[Y] = 2 + 7X - 2X^2$$

El efecto marginal de X en Y es: $\frac{\partial E[Y]}{\partial X} = 7 - 4X$

Transformaciones de variables

La segunda transformación importante es el logaritmo.

En vez de regredir Y en X, podemos estimar:

$$Y_i = \alpha + \beta \ln(X_i) + u_i$$

Razones para considerar el logaritmo:

- 1: Si tenemos una distribución muy asimétrica, con valores mucho más grandes que el promedio (p.e. ingresos), la transformación log la hace más "bien comportada" con menos valores extremos.
- 2: Ciertas variables naturalmente tienen relaciones multiplicativas, o exponenciales. El log las hace más aproximadamente lineales.

Transformaciones de variables

Interpretamos como cambios *relativos* en la variable con log.

$$\log(Sueldos_i) = 1 + 0.08 \cdot Educación_i + \varepsilon_i$$

- 1 año extra de educación aumenta los sueldos en un 8%.

$$Puntage_i = 40 + 20 \cdot \log(gastos\_escolares_i) + \varepsilon_i$$

- Un aumento de 1% en los gastos resulta en 0.2 puntos más.

$$\ln(GDP_c) = 2 + 0.35 \ln(Capital_c) + \varepsilon_c$$

- Un país con 1% más capital tendrá 0.35% más de producto.

Transformaciones de variables

Otras transformaciones frecuentemente usadas:

- El logaritmo no está bien definido para valores cero ($log(0) = -\inf$). Una alternativa es usar $log(1 + x)$.
  - Sin embargo, no es recomendable por diversas razones.
- Una alternativa es la raiz cuadrada.
  - No tiene la interpretación simple del logaritmo.
- Otra es la transformación *seno hiperbolico inverso*: $\ln(x + \sqrt{x^2 + 1})$.
  - Muy similar a log para valores altos, pero tiene problemas similares al $log(1 + x)$.
- Finalmente, la *winsorización* consiste en transformar los valores arriba del X% superiores en el valor del percentil X%.

Interaciones

Algunas veces, la relación entre dos variables depende de una tercera.

- El efecto del precio de la becina en cuántos km una persona se mueve por semana depende de la variable *tiene auto*.
- Una misma política puede tener efectos distintos para mujeres y hombres, o personas con diferentes niveles de ingresos.

Para captar este tipo de relación, usamos *interacciones* entre variables.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i \cdot Z_i + \epsilon_i$$

Interaciones

Una interacción entre variables binarias se puede interpretar como una diferencia adicional para un grupo.

Ejemplo

$$Ingresos_i = \beta_0 + \beta_1 Mujer_i + \beta_2 EdSupr_i + \beta_3 Mujer_i \cdot EdSupr_i + \varepsilon_i$$

Ingresos promedios de:

- Hombres sin educación superior: β_0
- Mujeres sin educación superior: $\beta_0 + \beta_1$
- Hombres con educación superior: $\beta_0 + \beta_2$
- Mujeres con educación superior: $\beta_0 + \beta_1 + \beta_2 + \beta_3$

Interaciones

Una interacción entre una variable continua y una binaria se puede interpretar como diferentes inclinaciones para cada grupo.

Ejemplo

$$Ingresos_i = \beta_0 + \beta_1 Mujer_i + \beta_2 Educ_i + \beta_3 Mujer_i \cdot Educ_i + \varepsilon_i$$

Un año de educación aumenta los ingresos promedios de:

- Hombres en β_2
- Mujeres en $\beta_2 + \beta_3$

Interaciones

Una interacción entre variables continuas se puede interpretar como una variación en el efecto marginal.

Ejemplo

$$Ingresos_i = \beta_0 + \beta_1 Exper_i + \beta_2 Educ_i + \beta_3 Exper_i \cdot Educ_i + \varepsilon_i$$

Un año de educación aumenta los ingresos promedios en:

$$\beta_2 + \beta_3 Exper_i$$

Es decir, entre dos personas sin ninguna experiencia, una con 10 años de educacion y la otra con 13, la diferencia es $3\beta_2$. Si tuvieran 5 años de experiencia, la diferencia seria $3 \cdot (\beta_2 + 5 \cdot \beta_3)$.

Regresion no lineal

Podemos usar regresion lineal de manera muy flexible para modelar relaciones no lineales. Pero hay limites. En ciertas situaciones, es más indicado usar modelos no lineales.

El problema más frecuente es el de modelar una variable binaria:

- ¿Un cliente quedará inactivo?
- ¿Un alumno va terminar el curso?
- ¿Un paciente va se recuperar?

En este caso, si usamos regresión lineal, le llamamos *modelo de probabilidad lineal*. Podemos interpretar las predicciones de la regresión lineal como probabilidades.

Regresion no lineal

El modelo de probabilidad lineal tiene algunas ventajas:

- Simple interpretación.
- Aproxima la función de esperanza condicional.

Pero hay problemas importantes:

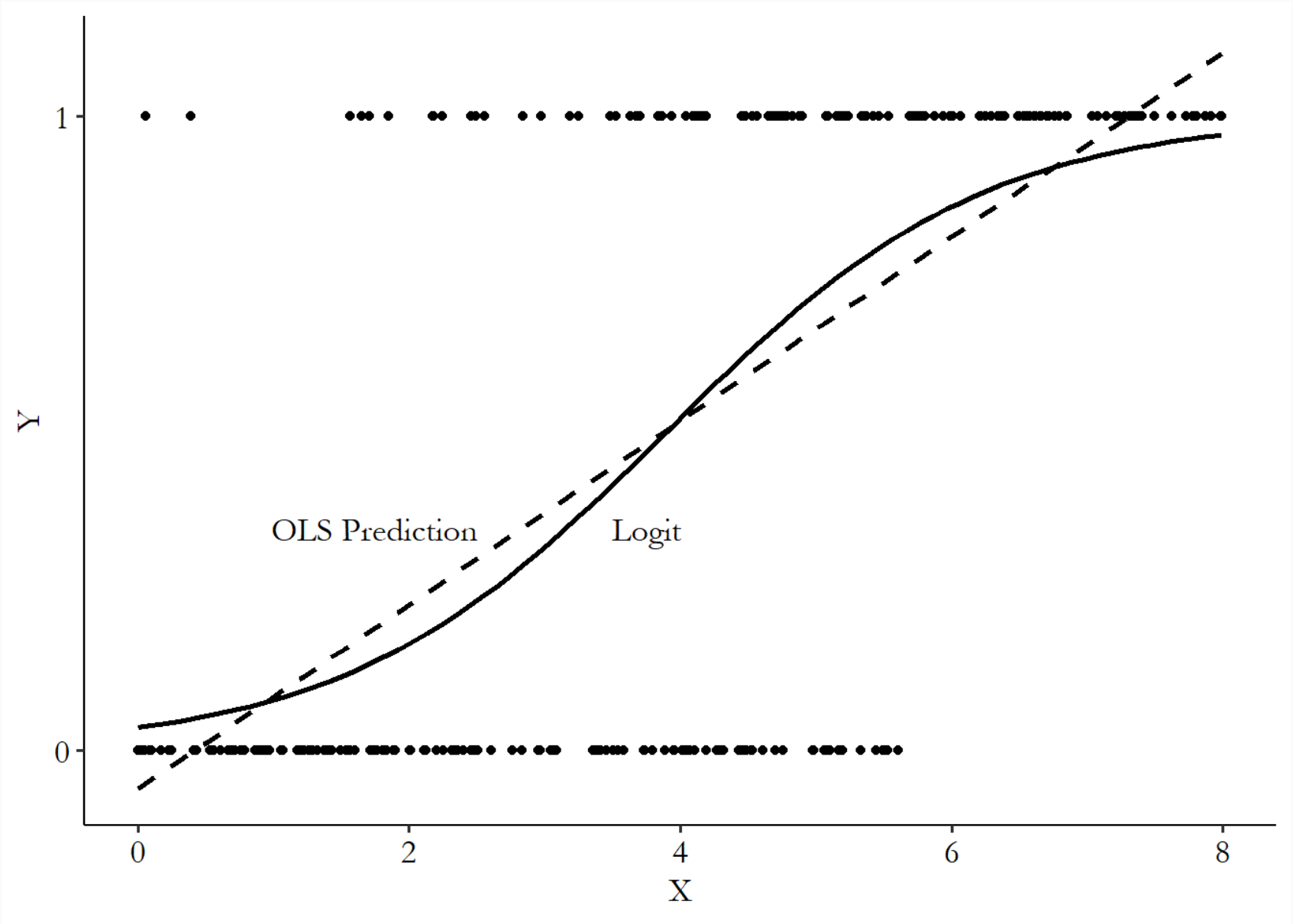
- Puede generar predicciones mayores que 1 o menores que 0.
- El modelo no puede ser válido para todos los valores de X.

En este caso, normalmente se usa un modelo Logit o Probit, que se puede escribir como:

$$E[Y] = F(\beta_0 + \beta_1 X)$$

Con F siendo una función que tiene valores entre 0 y 1. Para el Probit, $F(x) = \Phi(x)$ (la función de densidad acumulada de la distribucion normal), y para el Logit, $F(x) = \frac{e^x}{1+e^x}$.

Regresión no lineal



Regresion no lineal

La interpretación de los resultados no es sencilla. Por si mismos, los coeficientes no tienen una interpretación.

La mejor manera de presentar los resultados es de calcular *efectos marginales*. Es decir, podemos calcular como la probabilidad predicta cambia cuando aumentamos X en 1 unidad.

Pero en modelos no lineales, los efectos marginales dependen del valor de X. El efecto marginal en un Probit es mayor cuando la probabilidad es proxima de 50%, y menor cuando es proxima de 0 o 1.

Regresion no lineal

Dos formas principales de presentar los efectos marginales.

- Podemos calcular el efecto marginal para un individuo con valores promedios de X.
- O podemos calcular el efecto marginal para cada individuo en la muestra, y tomar el promedio de esos efectos.

Regresión no lineal

Otros tipos de regresión no lineal frecuentemente usadas:

- Para datos de conteo: Poisson, Negative Binomial
- Para variables categoricas: Multinomial Logit
- Para variables censuradas: Tobit