MCO: inferencia y más

Econometría I

Paula Pereda (ppereda@correo.um.edu.uy)
10 de setiembre de 2021

Regresión múltiple

Más variables explicativas

Pasamos de la **regresión lineal simple** (una variable de resultado y una variable explicativa)

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

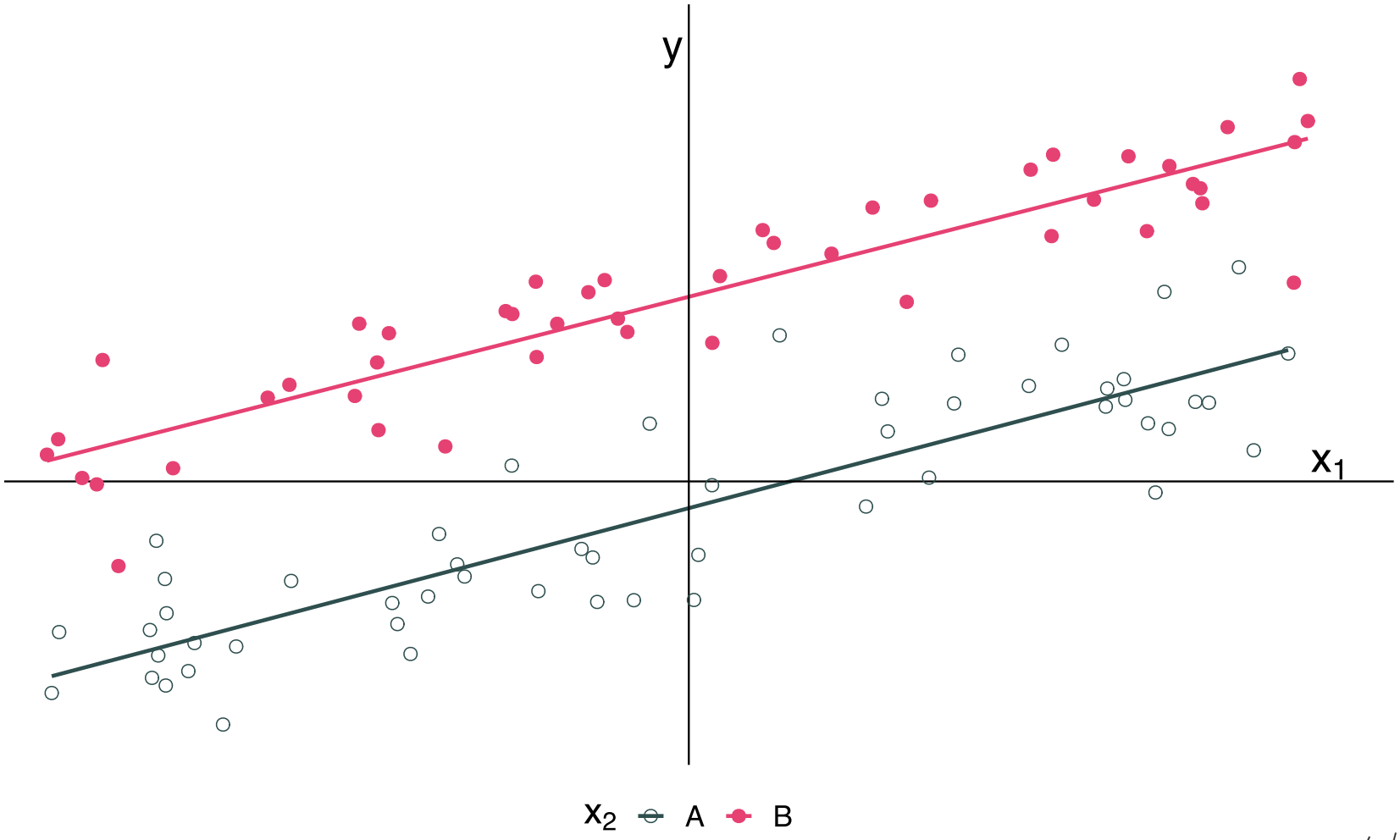
a la tierra de la **regresión lineal múltiple** (un variable de resultado y varios variables explicativas)

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$$

¿Por qué? Podemos explicar mejor la variación en y, mejorar las predicciones, evitar el sesgo de variables omitidas, ...

Regresión múltiple

Otra manera de pensar sobre esto:



Regresión múltiple

Mirar a nuestro estimador puede ayudar, también.

Para una regresión simple $y_i = \beta_0 + \beta_1 x_i + u_i$

$$
\begin{aligned}
\hat{\beta}_1 &= \\
&= \frac{\sum_i \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_i \left(x_i - \bar{x}\right)} \\
&= \frac{\sum_i \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)/(n-1)}{\sum_i \left(x_i - \bar{x}\right)/(n-1)} \\
&= \frac{\hat{\mathrm{Cov}}(x,\, y)}{\hat{\mathrm{Var}}(x)}
\end{aligned}
$$

Regresión múltiple

El estimador de regresión simple:

$$\hat{\beta}_1 = \frac{\hat{\text{Cov}}(x, y)}{\hat{\text{Var}}(x)}$$

Pasando a la regresión lineal múltiple, el estimador cambia ligeramente:

$$\hat{\beta}_1 = \frac{\hat{\text{Cov}}(\tilde{x}_1, y)}{\hat{\text{Var}}(\tilde{x}_1)}$$

donde $\tilde{x}_1$ es la variable *residualizada* x_1, la variación restante en x después de controlar las otras variables explicativas.

Regresión múltiple

Más formalmente, considere el modelo de regresión múltiple

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u_i$$

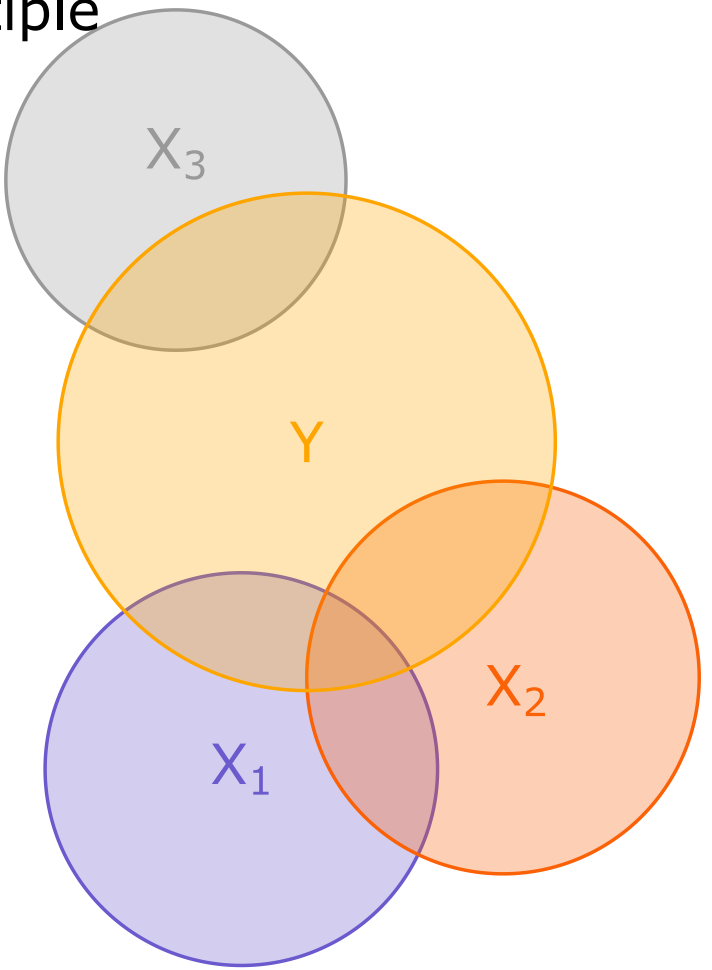
Nuestro x_1 residualizado (que llamamos $\tilde{x}_1$) proviene de hacer una regresión de x_1 en una intersección y todas las demás variables explicativas y recopilar los residuos, *es decir*,

$$\begin{aligned} \hat{x}_{1i} &= \hat{\gamma}_0 + \hat{\gamma}_2 \, x_{2i} + \hat{\gamma}_3 \, x_{3i} \\ \tilde{x}_{1i} &= x_{1i} - \hat{x}_{1i} \end{aligned}$$

lo que nos permite comprender mejor nuestro estimador de regresión múltiple MCO 🤓

$$\hat{\beta}_1 = \frac{\hat{\text{Cov}}(\tilde{x}_1, \, y)}{\hat{\text{Var}}(\tilde{x}_1)}$$

Regresión Múltiple



Ajuste del modelo

Las medidas de *bondad de ajuste* intentan analizar qué tan bien nuestro modelo describe (*ajusta*) los datos.

Medida común: R^2 [R-cuadrado] (*a.k.a.* coeficiente de determinación)

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i \left(y_i - \bar{y}\right)^2} = 1 - \frac{\sum_i \left(y_i - \hat{y}_i\right)^2}{\sum_i \left(y_i - \bar{y}\right)^2}$$

Observe a nuestro viejo amigo SCR: $\sum_i \left(y_i - \hat{y}_i\right)^2 = \sum_i e_i^2$.

R^2 literalmente nos dice la parte de la varianza en y que representan nuestros modelos actuales. Por lo tanto $0 \leq R^2 \leq 1$.

Regresión Múltiple

El problema: A medida que agregamos variables a nuestro modelo, el R^2 se incrementa *mecánicamente*.

Para ver este problema, podemos simular un conjunto de datos de 10.000 observaciones en y y 1.000 variables aleatorias x_k. **¡No hay relaciones entre y y x_k!**

Esquema de pseudocódigo de la simulación:

- Generamos 10.000 observaciones de y
- Generamos 10.000 observaciones sobre las variables x_1 hasta x_{1000}
- Regresiones
  - LM_1: Regresamos y en x_1; registro R^2
  - LM_2: Regresamos y en x_1 y x_2; registro R^2
  - LM_3: Regresamos y en x_1, x_2, y x_3; registro R^2
  - ...
  - LM_{1000}: Regresamos y en x_1, x_2, ..., x_{1000}; registro R^2

Regresión Múltiple

El problema: A medida que agregamos variables a nuestro modelo, el R^2 se incrementa *mecánicamente*.

Código de **R** para la simulación:

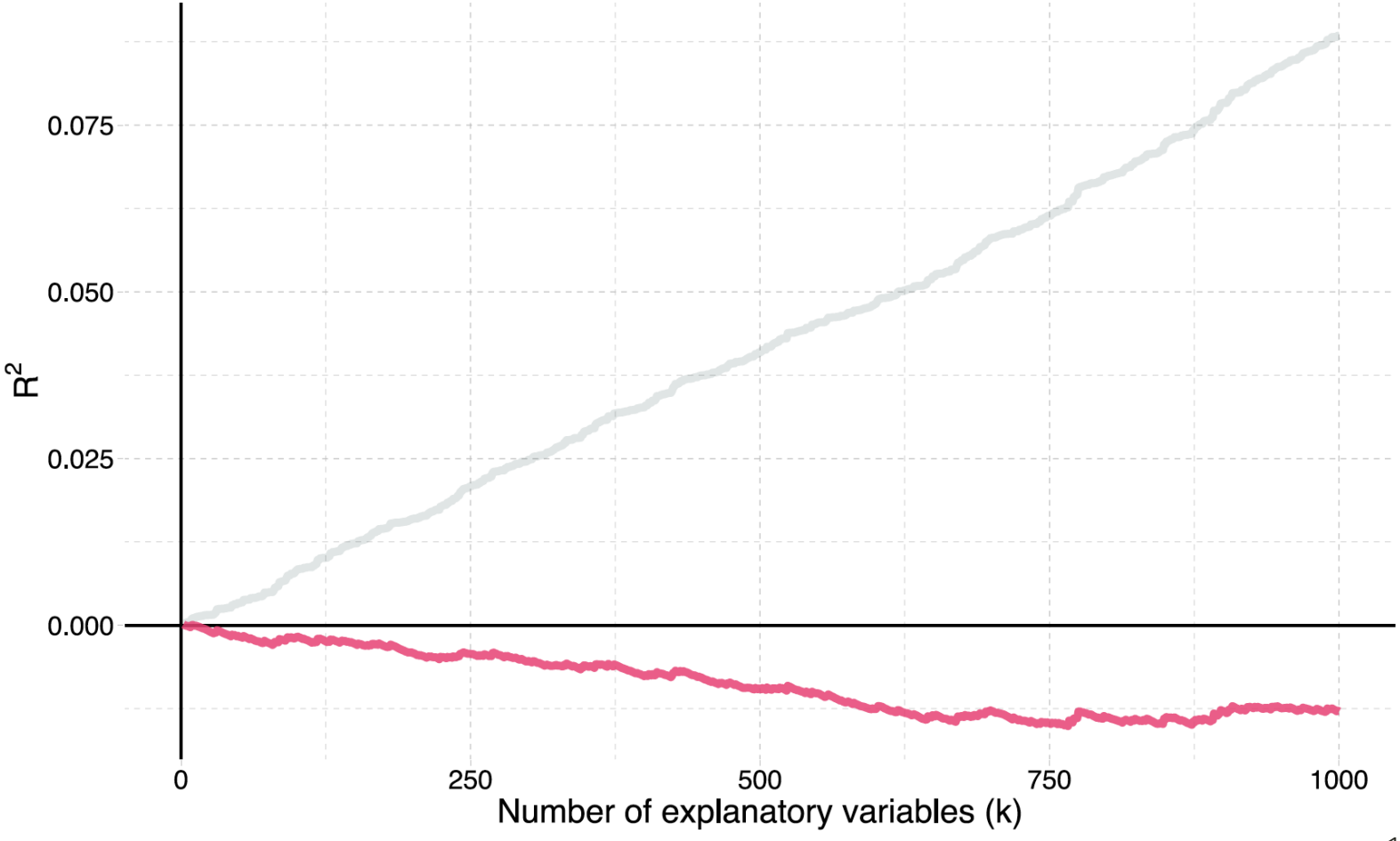
```
set.seed(1989)
y ← rnorm(1e4)
x ← matrix(data = rnorm(1e7), nrow = 1e4)
x %<>% cbind(matrix(data = 1, nrow = 1e4, ncol = 1), x)
r_df ← mclapply(X = 1:(1e3-1), mc.cores = 2, FUN = function(i) {
  tmp_reg ← lm(y ~ x[,1:(i+1)]) %>% summary()
  data.frame(
    k = i + 1,
    r2 = tmp_reg %$% r.squared,
    r2_adj = tmp_reg %$% adj.r.squared
  )
}) %>% bind_rows()
```

Regresión Múltiple

El problema: A medida que agregamos variables a nuestro modelo, el R^2 se incrementa *mecánicamente*.

Regresión Múltiple

Una solución: R^2 Ajustado



Regresión Múltiple

El problema: A medida que agregamos variables a nuestro modelo, el R^2 se incrementa *mecánicamente*.

Una solución: Penalizar por el número de variables, *por ejemplo*, R^2 ajustado:

$$\bar{R}^2 = 1 - \frac{\sum_i \left(y_i - \hat{y}_i\right)^2 / (n-k-1)}{\sum_i \left(y_i - \bar{y}\right)^2 / (n-1)}$$

Nota: El R^2 ajustado no necesita estar entre 0 y 1.

Incertidumbre e inferencia

Aprendiendo de nuestros errores

Como señaló nuestra simulación anterior, nuestro problema con la **incertidumbre** es que no sabemos si nuestra estimación muestral está *cerca* o *lejos* del parámetro poblacional desconocido. †

Sin embargo, no todo está perdido. Podemos usar los errores ($e_i = y_i - \hat{y}_i$) para tener una idea de qué tan bien nuestro modelo explica la variación observada en y .

Cuando nuestro modelo parece estar haciendo un "buen" trabajo, es posible que tengamos un poco más de confianza al usarlo para conocer la relación entre y y x.

Ahora solo tenemos que formalizar lo que realmente significa un "buen trabajo".

†: Excepto cuando corremos la simulación nosotros mismos, por eso nos gustan las simulaciones.

Aprendiendo de nuestros errores

En primer lugar, estimaremos la varianza de u_i (recordemos: $\mathbf{Var}(u_i) = \sigma^2$) usando nuestros errores al cuadrado, *es decir*,

$$s^2 = \frac{\sum_i e_i^2}{n - k}$$

donde k nos da el número de términos de constantes y variables que estimamos, *ejemplo*, $beta_0$ y β_1 darían $k = 2$.

s^2 es un estimador insesgado de σ^2.

Aprendiendo de nuestros errores

Luego mostramos que la varianza de $\hat{\beta}_1$ (para una regresión lineal simple) es

$$\mathbf{Var}\left(\hat{\beta}_1\right) = \frac{s^2}{\sum_i \left(x_i - \overline{x}\right)^2}$$

lo que muestra que la varianza de nuestro estimador de pendiente:

1. aumenta a medida que nuestras residuos se vuelven más ruidosas
2. disminuye a medida que aumenta la varianza de x

Aprendiendo de nuestros errores

Más comúnmente: El **error estándar** de $\hat{\beta}_1$

$$\hat{\mathrm{EE}}\left(\hat{\beta}_1\right) = \sqrt{\frac{s^2}{\sum_i \left(x_i - \overline{x}\right)^2}}$$

Recuerden: El error estándar de un estimador es la desviación estándar de la distribución del estimador.

Aprendiendo de nuestros errores

El error estándar en **R**'s `lm`, se ve así:

```
tidy(lm(y ~ x, pop_df))
```

```
> # A tibble: 2 x 5
>   term        estimate std.error statistic  p.value
>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
> 1 (Intercept)    2.53     0.422       6.00 3.38e- 8
> 2 x              0.567    0.0793      7.15 1.59e-10
```

Aprendiendo de nuestros errores

Usamos el error estándar de $\hat{\beta}_1$, junto con el propio $\hat{\beta}_1$, para aprender sobre el parámetro β_1.

Después de derivar la distribución de $\hat{\beta}_1$, [†] tenemos dos opciones (relacionadas) para la inferencia estadística formal (aprendizaje) sobre nuestro parámetro desconocido β_1:

- **Intervalos de confianza:** Utilizar la estimación y su error estándar para crear un intervalo que, cuando se repite, generalmente [††] contendrá el parámetro verdadero.
- **Pruebas de hipótesis:** Determinan si existe evidencia estadísticamente significativa para rechazar un valor o rango de valores hipotéticos.

†: *Pista:* es normal con la media y la varianza que hemos derivado/discutido anteriormente)

††: *Ejemplo*, los intervalos de confianza del 95% construidos de manera similar contendrán el parámetro verdadero el 95% del tiempo.

Intervalos de confianza

Construimos intervalos de confianza nivel $(1-\alpha)$ para β_1

$$\hat{\beta}_1 \pm t_{\alpha/2,\text{gl}}\ \hat{\text{EE}}\left(\hat{\beta}_1\right)$$

$t_{\alpha/2,\text{gl}}$ denota el $\alpha/2$ cuantial de una distribución t con $n-k$ grados de libertad.

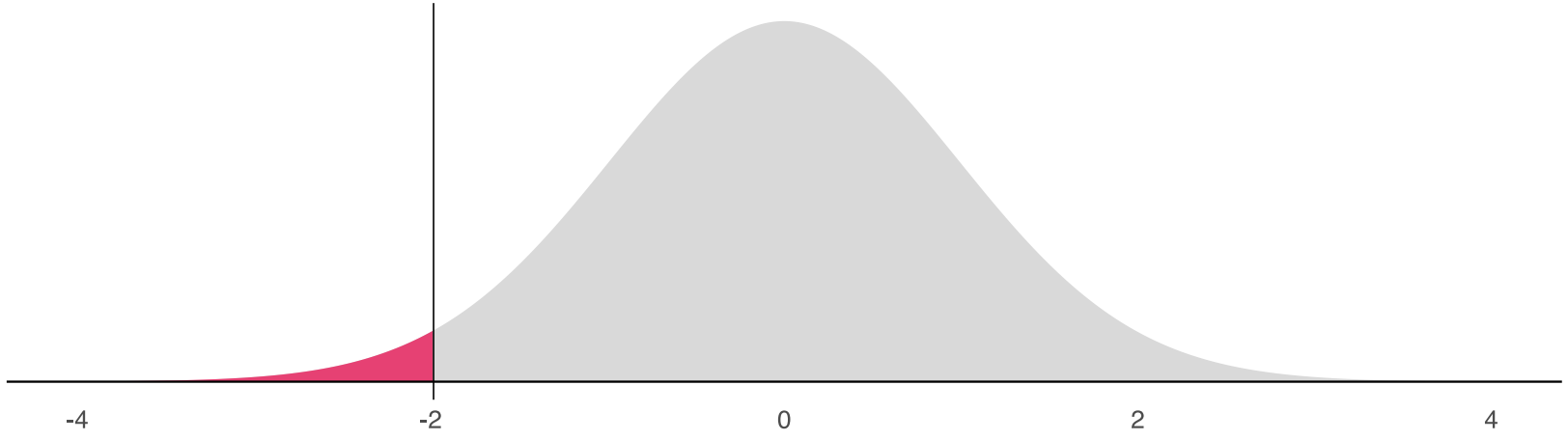
Incertidumbre e inferencia

Intervalos de confianza

Construimos intervalos de confianza nivel $(1-\alpha)$ para β_1

$$\hat{\beta}_1 \pm t_{\alpha/2,\text{gl}}\ \hat{\text{EE}}\left(\hat{\beta}_1\right)$$

Por ejemplo, 100 obs., dos coeficientes (*es decir*, $\hat{\beta}_0$ y $\hat{\beta}_1$ $\textbf{implica}\ k=2$) y $\alpha = 0.05$ (para un intervalo de confianza del 95%) nos da $t_{0.025,\,98} = -1.98$



Intervalos de confianza

Construimos intervalos de confianza nivel $(1-\alpha)$ para β_1

$$\hat{\beta}_1 \pm t_{\alpha/2,\text{gl}} \, \hat{\text{EE}}\left(\hat{\beta}_1\right)$$

Ejemplo:

```
lm(y ~ x, data = pop_df) %>% tidy()
```

```
> # A tibble: 2 x 5
>   term        estimate std.error statistic  p.value
>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
> 1 (Intercept)    2.53      0.422       6.00 3.38e- 8
> 2 x              0.567     0.0793      7.15 1.59e-10
```

Nuestro intervalo de confianza del 95% es $0.567 \pm 1.98 \times 0.0793 = [0.410, \, 0.724]$

Intervalos de confianza

Así que tenemos un intervalo de confianza para β_1, *i.e.*, $[0.410,\ 0.724]$.

Y... ¿qué significa?

Informalmente: El intervalo de confianza nos da una región (intervalo) en la que podemos depositar algo de confianza (confianza) para contener el parámetro.

-

Más formalmente: Si muestrea repetidamente de nuestra población y construye intervalos de confianza para cada una de estas muestras, $(1 - \alpha)$ por ciento de nuestros intervalos (*ejemplo*, 95%) contendrá el parámetro de población *en algún lugar del intervalo*.

De nuevo a nuestra simulación...

Intervalos de confianza

Extrajimos 10.000 muestras (cada una de tamaño $n = 30$) de nuestra población y estimamos nuestro modelo de regresión para cada una de estas simulaciones:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$$

(repetido 10.000 veces)

Ahora, estimemos intervalos de confianza del 95% para cada uno de estos intervalos...

Intervalos de confianza

De nuestras simulaciones previas: 97.6% el 95% de los intervalos de confianza contienen el verdadero valor de parámetro de β_1.

Testeo de hipotesis

En muchas aplicaciones, queremos saber más que una estimación puntual o un rango de valores. Queremos saber qué dice nuestra evidencia estadística sobre las teorías existentes.

Queremos probar hipótesis planteadas por funcionarios, políticos, economistas, científicos, amigos, vecinos raros, *etcétera*.

Ejemplos

- ¿El aumento de la presencia policial **reduce la delincuencia**?
- ¿Construir un muro gigante **reduce el crimen**?
- ¿El cierre de un gobierno **afecta negativamente a la economía**?
- ¿El cannabis legal **reduce la conducción en estado de ebriedad** o **reduce el uso de opiáceos**?
- ¿Los estándares de calidad del aire **aumentan la salud** y/o **reducen el empleo**?

Testeo de hipotesis

La prueba de hipótesis se basa en resultados e intuición muy similares.

Si bien la incertidumbre ciertamente existe, aún podemos construir pruebas estadísticas *confiables* (rechazando o no rechazando una hipótesis planteada).

MCO prueba *t* Nuestra hipótesis (nula) establece que β_1 es igual a un valor c, *por ejemplo*, H_0) $\beta_1 = 0$

Bajo los supuestos del modelo lineal clásico (MLC),

$$t_{\text{estadístico}} = \frac{\hat{\beta}_1 - c}{\hat{\text{EE}}\left(\hat{\beta}_1\right)}$$

sigue una distribución t con $n - k$ grados de libertad.

Testeo de hipotesis

Para una prueba **de dos caras** de nivel $\alpha $, rechazamos la hipótesis nula (y concluimos con la hipótesis alternativa) cuando

$$\left|t_{\text{estadístico}}\right| > \left|t_{1-\alpha/2,\, df}\right|$$

lo que significa que nuestra **estadística de prueba es más extrema que el valor crítico**.

Alternativamente, podemos calcular el **valor p** que acompaña a nuestra estadística de prueba, que efectivamente nos da la probabilidad de ver nuestra estadística de prueba *o una estadística de prueba más extrema* si la hipótesis nula fuera cierta.

Valores p muy pequeños (generalmente < 0.05) significan que sería improbable ver nuestros resultados si la hipotesis nula fuera realmente cierta; tendemos a rechazar el valor nulo para valores p por debajo de 0.05.

Testeo de hipotesis

`R` y `Stata` testean por defecto contra el valor cero.

```
lm(y ~ x, data = pop_df) %>% tidy()
```

```
> # A tibble: 2 x 5
>   term        estimate std.error statistic  p.value
>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
> 1 (Intercept)    2.53      0.422      6.00 3.38e- 8
> 2 x              0.567     0.0793     7.15 1.59e-10
```

H_0) $\beta_1 = 0$ *vs.* H_1) $\beta_1 \neq 0$

¿Qué observamos del valor-p?

El valor-$p < 0.05$.

$t_{\text{stat}} = 7.15$ y $t_{0.975,\, 28} = 2.05$ lo que implica que el valor-$p < 0.05$.

Entonces, **rechazamos H_0**.

Testeo de hipotesis

¡De vuelta a nuestra simulación! Veamos qué está haciendo realmente nuestra estadística de t.

En esta situación, podemos conocer (y hacer cumplir) la hipótesis nula, ya que generamos los datos.

Para cada una de las 10.000 muestras, calcularemos la estadística t, y luego podremos ver cuántas estadísticas de t exceden nuestro valor crítico (2.05, como encima).

La respuesta debería ser aproximadamente el 5 por ciento, nuestro nivel α.

Incertidumbre e inferencia

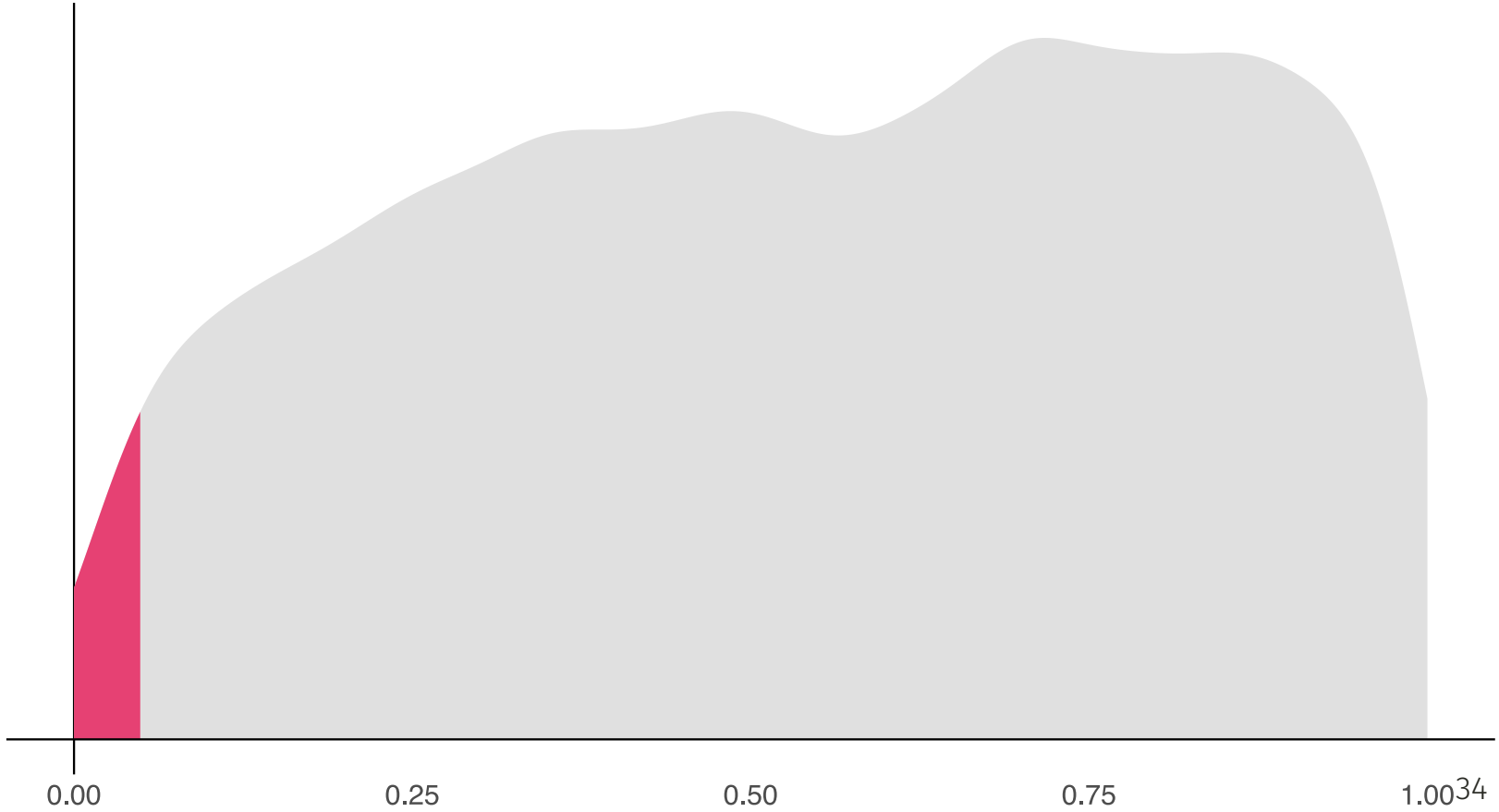
En nuestra simulación, el porcentaje 2.4 de nuestras estadísticas de t rechaza la hipótesis nula.

La distribución de nuestras estadísticas de t (sombreando las regiones de rechazo).

Incertidumbre e inferencia

En consecuencia, 2.4 % de nuestros p-valores rechazan la hipótesis nula.

La distribución de nuestros valores p (sombreando los valores p por debajo de 0,05).



Aplicaciones en R

Ejercicio C3.1 (Wooldridge)

Un problema de interés para los funcionarios de salud (y para otros) es determinar los efectos que el fumar durante el embarazo tiene sobre la salud infantil. Una medida de la salud infantil es el peso al nacer; un peso demasiado bajo puede ubicar al niño en riesgo de contraer varias enfermedades. Ya que es probable que otros factores que afectan el peso al nacer estén correlacionados con fumar, deben considerarse. Por ejemplo, un nivel de ingresos más alto en general da como resultado el acceso a mejores cuidados prenatales y a una mejor nutrición de la madre. Una ecuación que reconoce estos factores es

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 faminc + u.$$

i) ¿Cual es el signo más probable para β_2?

ii) ¿Cree que *cigs* y *faminc* estén correlacionados? Explique por qué la correlación puede ser positiva o negativa.

iii) Ahora, calcule la ecuación con y sin *faminc* utilizando los datos del archivo BWGHT. RAW. Dé los resultados en forma de ecuación incluyendo el tamaño de la muestra y la *R*-cuadrada. Explique sus resultados enfocándose en si el añadir *faminc* modifica de manera sustancial el efecto esperado de *cigs* sobre *bwght*.

Ejercicio C3.2 (Wooldridge)

Utilice los datos del archivo HPRICE1.RAW para estimar el modelo

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u,$$

donde *price* es el precio de casas dado en miles de dólares.

i) Escriba los resultados en forma de ecuación.

ii) ¿Cual es el incremento en precio estimado para una casa con una habitación (*bdrms*) más, manteniendo constante la superficie en pies cuadrados (*sqrft*)?

iii) ¿Cual es el incremento en precio estimado para una casa con una habitación adicional de 140 pies cuadrados? Compare esto con su respuesta al inciso (ii).

iv) ¿Qué porcentaje de la variación en el precio se explica por la extensión en pies cuadrados y el número de habitaciones?

v) La primera casa en la muestra tiene $sqrft = 2{,}438$ y $bdrms = 4$. Determine el precio de venta estimado para esta casa con la línea de regresión de MCO.

vi) El precio de venta de la primera casa en la muestra fue \$300,000 (así que $price = 300$). Determine el residual para esta casa. ¿Sugiere esto que el comprador pagó de más o de menos por la casa?

Ejercicio C3.4 (Wooldridge)

Para este ejercicio, utilice los datos del archivo ATTEND.RAW.

i) Obtenga los valores mínimo, máximo y promedio para las variables *atndrte*, *priGPA*, y *ACT* (porcentaje de asistencia a clases, calificación promedio general acumulada, calificación en el examen de admisión a la universidad, respectivamente).

ii) Estime el modelo

$$atndrte = \beta_0 + \beta_1 priGPA + \beta_2 ACT + u,$$

y escriba los resultados en forma de ecuación. Interprete el intercepto. ¿Tiene un significado útil?

iii) Analice los coeficientes de pendiente estimados. ¿Hay alguna sorpresa?

iv) ¿Cuál es el *atndrte* si $priGPA = 3.65$ y $ACT = 20$? ¿Qué piensa de este resultado? ¿Hay alumnos en la muestra con estos valores de las variables explicativas?

v) Si el alumno A tiene $priGPA = 3.1$ y $ACT = 21$ y el alumno B tiene $priGPA = 2.1$ y $ACT = 26$, ¿cuál es la diferencia predicha en sus tasas de asistencia?

Se dispone de una muestra de 30 observaciones de datos que representan salario y experiencia laboral de economistas de Montevideo en 2013 y 2014. Las variables son: y = salario (en miles USD); x = experiencia posterior a recibirse (años).

Se tiene el siguiente modelo:

$$\mathrm{y_i} = \beta_1 + \beta_2 \mathrm{x_i} + \varepsilon_\mathrm{i}$$

Se sabe que:

$$\sum_{i=1}^{30} y_i^2 = 65692,27 \; ; \sum_{i=1}^{30} y_i = 1365,1 \; ; \sum_{i=1}^{30} y_i x_i = 26046,9$$

$$\sum_{i=1}^{30} x_i^2 = 12230 \; ; \; \sum_{i=1}^{30} x_i = 550$$

$$\mathrm{SCR} = 3089,95 \; ; \mathrm{SCT} = 3574,67$$

1.1 Estime por mínimos cuadrados ordinarios los parámetros de dicho modelo.

1.2 Calcule la matriz de varianzas y covarianzas de los coeficientes estimados.

1.3 Interprete los resultados obtenidos y analice su significancia estadística.

1.4 Calcular el R2 del modelo. ¿Qué puede decir al respecto?

1.5 ¿Qué relación debería encontrarse entre el vector de residuos y las variables explicativas?

Variable	coefficient	Std.Error	t-statistic	Prob
C	-11,21972	**0,908694**	-12,34708	0.0000
LCAP	0,245249	**0,0236266**	10,38020	0,0000
LTRA	**1,467**	0,078350	18,72435	0,0000