Regresión Logística

Mayo 8, 2024

Prof. Sergio Béjar

Departamento de Estudios Políticos, CIDE

Plan para Hoy

Discutir regresión logística (logit) para variables binarias.

Yendo Más lejos en Estadística Aplicada

Ya tenemos herramientas para responder a nuestras propias preguntas en CP/RI.

- Sabemos como manipular datos.
- Creemos que la variación en y se puede atribuir a la variación en x.
- Después de controlar por explicaciones alternativas (z), nuestra regresión lineal produce el efecto parcial de x en y.

La regresión lineal (OLS) nos da la línea que mejor se adapta a los datos.

- Para producir esa linea se minimiza la suma al cuadrado de diferencias al cuadrado (de aquí viene: OLS).

OLS

OLS tiene muchas propiedades.

- Mejor estimador lineal sin sesgo (BLUE - Best Linear Unbiased Estimator).
- Es fácil de ejecutar y de interpretar.

Sería una *lástima* que algo pasara con alguno de nuestros supuestos.

El Problema de las Variables Dependientes Binarias

Uno de los mayores problemas con los que nos podemos encontrar tiene que ver con la variable dependiente.

- OLS asume que la VD se distribuye normalmente.

Pero en muchas ocasiones nos vamos a encontrar con VDs que son binarias.

- Candidato gana/pierde.
- Ciudadano votó/no votó.
- Programa exitoso/no exitoso.
- Guerra sucedió/no sucedió.

Muchos fenómenos políticos/sociales se miden con variables binarias ("esta"/"no esta").

Implicaciones cuando Supuestos de OLS son Violados

1. Tus errores serán **heteroesquedásticos.**
2. Tus $\hat{y}$s no harán mucho sentido.

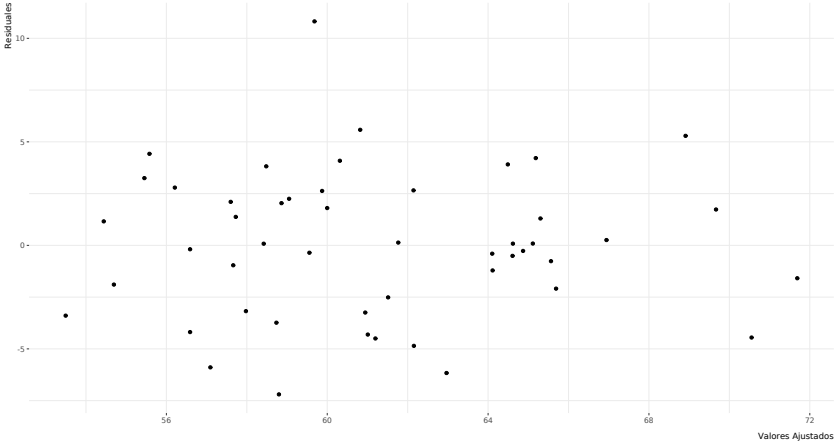
La Historia de Dos Regresiones

```r
# ¿Cuál fue el % de turnout?
# Vamos a omitir Hawaii y DC porque son outliers potenciales.
M1 <- lm(turnoutho ~ percoled + ss,
         data=subset(turnout_EU, state %nin% c("Hawaii", "District of Columbia")))

# ¿Ganó Trump (1) o no (0)?
M2 <- lm(trumpw ~ percoled + ss,
         data=subset(turnout_EU, state %nin% c("Hawaii", "District of Columbia")))
```

Un Gráfico de Residuales-Ajustados Debe Verse Así.

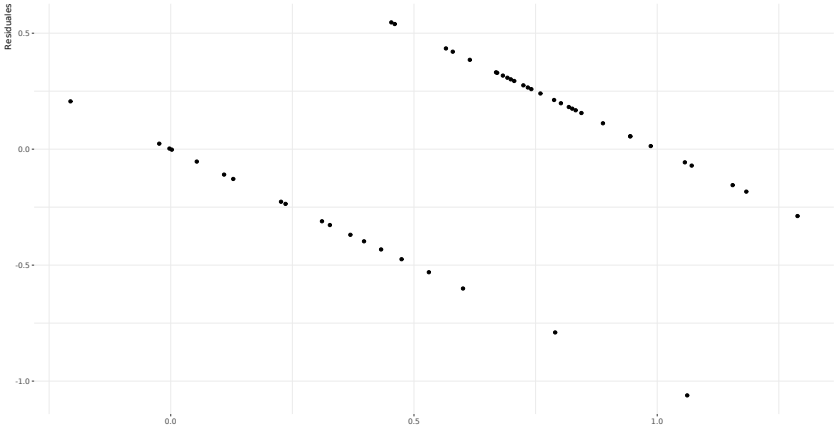
La variación entre lo que estimamos (fit) y el error que resulta de ello (residuales) es normal.



Modelo linear simple donde turnout es DV y educación universitaria y estado pivotal son vars. explicativas.

Un Gráfico de Residuales-Ajustados NO Debe Verse Así

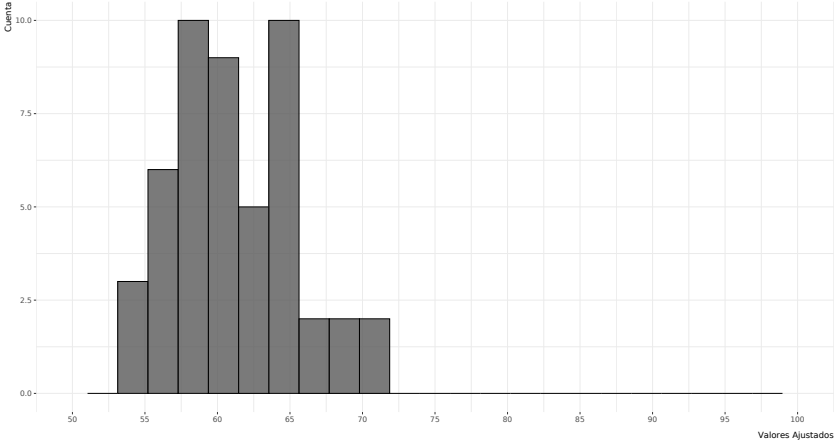
Si ves que hay patrones claros como en este gráfico, OLS no es (con alta probabilidad) el modelo que quieres.



Modelo linear simple donde Gana Trump o no es DV y educación universitaria y estado pivotal son vars. explicativas..

Los Valores Estimados por el Modelo Lineal Deben ser Plausibles

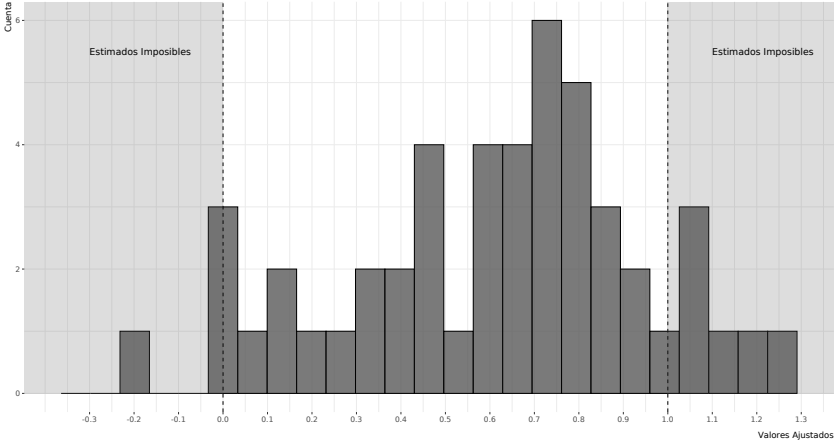
En este ejemplo lo son. Turnout en todos los estados está entre 50s bajos y 70s medios, que es lo que estamos estimando.



Hacer un histograma para una variable continua no es lo mejor, pero lo hago para efectos ilustrativos.

Los Valores Estimados por el Modelo Lineal deben ser Plausibles

En este ejemplo no lo son. La probabilidad debe estar delimitada entre 0 y 1.



¿Qué Estimados Están Fuera de los Límites?

```
turnout_EU %>%
  filter(state %nin% c("Hawaii", "District of Columbia")) %>%
  mutate(fitted = fitted(M2)) %>%
  filter(fitted > 1 | fitted < 0) %>%
  select(state, trumpw, percoled, ss, fitted)
```

```
##           state trumpw percoled ss       fitted
## 1      Arkansas      1     21.1  0  1.155048142
## 2   Connecticut      0     37.6  0 -0.002637214
## 3      Kentucky      1     22.3  0  1.070852844
## 4     Louisiana      1     22.5  0  1.056820294
## 5      Maryland      0     37.9  0 -0.023686039
## 6 Massachusetts      0     40.5  0 -0.206109186
## 7   Mississippi      1     20.7  0  1.183113242
## 8        Nevada      0     23.0  1  1.061609405
## 9 West Virginia      1     19.2  0  1.288357365
```

Limitantes con OLS

Substantivamente, los coeficientes de regresión nos llevan a inferencias erróneas.

- Recuerda: Los coeficientes OLS asumen efectos lineales y constantes de x en y.
- Cuando solo tenemos 0s y 1s, los efectos lineales no son muy intuitivos.

Regresión Logística

Abordaremos el problema de variables dependientes binarias con la **regresión logística**.

- Nos dará el efecto de un cambio en una unidad de x en la *probabilidad logarítmica natural de y.*

Vamos a empezar viendo que significa la "probabilidad logarítmica natural de y".

Odds (Momios)

La palabra **momios** es muy utilizada en el mundo de las apuestas deportivas.

- Y está muy relacionada con la probabilidad.

Dada cierta probabilidad de que un evento p ocurra, los momios del evento son iguales a:

$$\text{Momios} = \frac{p}{1-p}$$

¿Alguna vez escuchaste algo como "los momios son 4 a 1 a favor" del caballo 8?

- Traducción: por cada 5 intentos, esperamos que el caballo 8 gane en 4 ocasiones, en promedio.

Educación y Probabilidad de Votar (Datos Hipotéticos)

Vamos a ver esto con unos datos hipotéticos.

Table 1: Educación a nivel individual y Votación (Datos Hipotéticos)

Vote	**0: Low**	**1: Mid-Low**	**2: Middle**	**3: Mid-High**	**4: High**
No	94	80	50	20	6
Yes	6	20	50	80	94
Total	*100*	*100*	*100*	*100*	*100*

- El objetivo aquí es desmenuzar momios y logits en forma accesible.

Educación y Probabilidad de Votar

Evidentemente hay una relación positiva.

- i.e. gente más educada tiene una probabilidad más alta de votar.

Pero también estamos viendo la probabilidad de no-linealidad en VDs discretas.

- El efecto de 0 a 1 en x es un cambio de .14 en la probabilidad de votar.
- De 1 a 2 en x: cambio de .30.
- De 2 a 3 en x: cambio de .30 de nuevo.
- De 3 a 4 en x: cambio de .14.

Podemos pensar en algo análogo a un punto de inflexión.

Visualizando los Momios

Table 2: Probabilidad/Momios de Educación y Votar (Datos Hipotéticos)

Educación	**p(votar)**	**Momios de Votar**
0: Bajo	.06	$.06/.94 = .06$
1: Medio-Bajo	.20	$.20/.80 = .25$
2: Medio	.50	$.50/.50 = 1$
3: Medio-Alto	.80	$.80/.20 = 4$
4: Alto	.94	$.94/.06 = 16$

Visualizando Momios

La columna derecha, momios de votar, convierte probabilidades a momios.

- e.g. $\frac{p}{1-p}$ cuando $x = 0 = \frac{.06}{.94} = .06382979$.
- Una vez que llegamos a la categoria "Medio", los momios son enteros.
  - Cuando los momios son igual a 1, esperamos un votante por cada no-votante.

Relación de Momios (Odds Ratio)

Una forma en la que podemos ver como x afecta a y es utilizando la **relación de momios (odds ratio)**.

Momios y Relación de Momios

Veamos a la tabla.

- Momios de votar en la categoria de educación baja: .06.
- Momios de votar en la categoria de educación media-baja: .25.

Los momios de votar para la categoria media-baja son más de cuatro veces los momios de votar en la categoría de educación baja.

- $\frac{.25}{.06} = 4.1\overline{6}$
- Hagan esto para todos los otros valores y la relación de momios va a ser 4 siempre.

$$\text{Odds ratio} = \frac{1}{.25} = \frac{4}{1} = \frac{16}{4} = 4$$

Cambio Porcentual en Momios

También podemos calcular el **cambio porcentual en momios**.

Cambio Porcentual en Momios

Vamos a considerar de nuevo los momios de votar en las dos categorias más bajas.

- Calcular el incremento de unidad (aquí: .25 - .06 = .19).
- Dividimos eso entre los momios del valor más bajo (aquí: .06)
- Esto nos da un valor de $3.1\overline{6}$.
- Multiplicamos eso por 100 para obtener el cambio porcentual

Si hacemos esto para todos los otros valores, obtendremos valores de 3 (i.e. 300%).

Logits (Natural Logged Odds de *y*)

Hemos visto que cada unidad de cambio en x no produce un cambio consistente en y.

- Pero, el efecto del cambio en la relación de momios y cambio porcentual es consistente.
- El siguiente paso es hacer la transformación logarítmica natural de los momios, o **logit**.

Transformación Logarítmica Natural

El término clave aquí es transformación *natural* logarítmica.
En cálculo, el logaritmo natural con base e es común.

Pregunta

$$f(x) = (1 + \frac{1}{x})^x$$

¿Qué pasa en esta fórmula cuando x se va hacia el infinito?

Transformación Logarítmica Natural

Cuando x tiende al infinito, el exponente tiene también al infinito.

- Sin embargo, el denominador también.

Esto significa que estaremos tomando el exponencial de infinito para un valor cercano a 1, que resulta básicamente en 1.

- Bernoulli descubrió que el límite está entre 2 y 3.

Leonhard Euler propuso que la respuesta es e (un número irracional) y lo podemos denotar como $e = 2.7182818284$, aproximadamente.

Transformación Logarítmica Natural

Tomamo el logaritmo natural de todos los momios de y y lo agregamos a nuestra table.

Table 3: Probabilidad, Momios, y Logits de Educación y Votar (Datos Hipotéticos)

Educación	p(votar)	Momios de Votar	Momios Logged
0: Low	.06	$.06/.94 = .06$	-2.8
1: Mid-low	.20	$.20/.80 = .25$	-1.4
2: Middle	.50	$.50/.50 = 1$	0
3: Mid-high	.80	$.80/.20 = 4$	1.4
4: High	.94	$.94/.06 = 16$	2.8

Regresión Logística

Nuestra y no nada más es 0s y 1s, sino funciones logit aplicadas a los momios de 0s y 1s para todos los valores de x. Formalmente:

$$\text{Momios logged de } y = \hat{a} + \hat{b}(x)$$

¿Cómo se vería esto en el ejemplo simple que estamos viendo?

Regresión Logística

$$\text{Momios logged de votar} = -2.8 + 1.4(x)$$

Recordemos que:

- $\hat{a}$ es el estimado de los momios logged de y cuando $x = 0$ (entonces: -2.8)
- 1.4 es el $\hat{b}$ que observamos en la columna derecha de la tabla.

Interpretación de Regresión Logística

Una unidad de incremento en x resulta en un incremento de 1.4 en los momios logged de y.

- Aunque esto es una interpretación correcta, no es muy intuitiva.

¿Cómo obtenemos una interpretación más digestible/substantiva?

Interpretación de Regresión Logística

"Exponenciamos" el coeficiente de nuestra regresión.

$$\text{Exp}(\hat{b}) = \text{Exp}(1.4) = e^{1.4} = 4$$

¿Se ve familiar ese 4?

Interpretación de Regresión Logística

Es la relación de momios.

- Recuerda: tu coeficiente de regresión es el estimado del tamaño del efecto de una unidad a otra (más alta) en todo el rango de x.

Interpretación de Regresión Logística

También podemos obtener el cambio porcentual en los momios.

$$\text{Cambio porcentual en los momios de } y = 100 * (\text{Exp}(\hat{b}) - 1)$$

Con estos datos, obtenemos un resultado de 300. Cada incremento de unidad en x (en este caso: educación) incrementa los momios de votar en un 300 por ciento.

Interpretando Regresión Logística

También podemos obtener probabilidades (por ejemplo: cuando $x = 0$)

$$\text{Probabilidad} = \frac{\text{Momios}}{1 + \text{Momios}} = \frac{e^{-2.8}}{1 + e^{-2.8}} = .06$$

...o en R

```
exp(-2.8)/(1 + exp(-2.8))
```

```
## [1] 0.05732418
```

```
plogis(-2.8)  # hagan esto mejor
```

```
## [1] 0.05732418
```

Ahora Con Datos Reales

¿Cómo se vería esto con datos reales? Vamos a hacer un ejemplo con datos del General Social Survey, olas 2016 y 2018.

- y: ¿Pueden las mujeres obtener un aborto legal por alguna razón?

Table 4: ¿Pueden las mujeres obtener un aborto legal por alguna razón? (GSS, 2016-2018)

Respuesta	No. de Observaciones	Porcentaje
No	1733	51.98%
Sí	1601	48.02%

Variables Explicativas

- Mujer (1 = mujer)
- Efectos Fijos para Raza (blancos [omitidos], afro-americanos, otros)
- Hispano (1 = sí)
- Nivel Educativo (años en escuela [0:20])
- ID Partidario (muy D a muy R [0:6])
- Actividad Religiosa (nunca a varias veces al día [0:10])

```r
gss_abortion %>%
  filter(year >= 2016) %>%
  mutate(race = fct_relevel(race, "White"),
         relactiv = relactiv - 1,
         female = ifelse(sex == "Female", 1, 0)) -> Data

M3 <- glm(abany ~ female + factor(race) + hispanic + educ +
            pid + relactiv, data=Data,
          family=binomial(link="logit"))

modelsummary(list("¿Se puede abortar por alguna razón?" = M3), output="latex",
             title = "Actitudes sobre Aborto en GSS (2016-2018)",
             stars = TRUE, gof_omit = "IC|F|Log.|R2$",
             coef_map = c("female" = "Mujer",
                          "factor(race)Black" = "Raza = Afro-Americano",
                          "factor(race)Other" = "Raza = Otra",
                          "hispanic" = "Hispano",
                          "educ" = "Años de Educación",
                          "pid7" = "ID Partidario (D to R)",
                          "relactiv" = "Actividad Religiosa",
                          "(Intercept)" = "Intercepto"),
             align = "lc")
```

Table 5: Actitudes sobre Aborto en GSS (2016-2018)

	¿Se puede abortar por alguna razón?
Mujer	-0.087
	(0.078)
Raza = Afro-Americano	-0.388***
	(0.111)
Raza = Otra	-0.181
	(0.143)
Hispano	-0.354**
	(0.125)
Años de Educación	0.153***
	(0.014)
Actividad Religiosa	-0.196***
	(0.018)
Intercepto	-0.895***
	(0.218)
Num.Obs.	3175
RMSE	0.46

- $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Interpretación de Tabla 5

- No hay diferencia significativa entre mujeres y hombres.
- Los afro-americanos son menos propensos a pensar que las mujeres deberían poder abortar por alguna razón que los blancos.
- No hay diferencua entre otras razas (i.e. Asiáticos) y blancos.
- Los Hispanos son menos propensos a estar a favor del aborto por cualquier razón.
- La educación tiene un efecto positivo.
- Ser Republicano y religioso disminuye la propensidad a contestar "sí".

Extrayendo Algunas Cantidades de Interés.

Digamos que nos interesa entender el efecto de ser Hispano.

- Exp(-0.354) = $e^{-0.354} = .701$. Esto es la relación de momios (Odds Ratio).
- 100*($e^{-0.354}$ - 1) = -29.81%. El cambio porcentual en los momios de estar a favor del aborto por cualquier razón.

Ahora veamos el efecto de un incremento en educación.

- Relación de momios (Odds ratio): $e^{.153} = 1.165$.
- Probabilidad de cambio en momios: 100*($e^{.153}$ - 1) = 16.53%

El intercepto también da información útil ($\hat{a}$ = -.895).

- Nos dice los momios logged de contestar "sí" para un individuo que es blanco, no-hispano, muy demócrata y que nunca va a servicios religiosos.
- La probabilidad de que esa persona conteste "sí" es: .290

Conclusión

Las variables dependientes binarias violan supuestos de OLS y producen estimados engañosos.

- Por eso utilizamos regresión logística.
- El proceso de inferencia es el mismo, pero los coeficientes comunican cosas algo diferentes.
- Es la misma regresión, solo que sobre una VD transformada.

La compu hace el trabajo duro por nosotros, pero es importante saber que es lo que la compu está haciendo en este caso.

Table of Contents