Correlación y Regresión Lineal

Abril 1 y 3, 2024

Prof. Sergio Béjar

Departamento de Estudios Políticos, CIDE

Objetivos Para Hoy

Usar correlación y regresión lineal para describir la relación entre dos variables.

Lo Que Hemos Construido

Todo lo que hemos estudiado hasta el momento nos sirve para construir una investigación cuantitativa "normal".

- Observamos la tendencia central y variación en nuestras variables.
- Hacemos inferencias sobre nuestras afirmaciones (i.e. hipótesis) de causa y efecto usando la lógica del muestreo aleatorio.

Si nuestro estadístico muestral es más que 1.96 errores estándar del parámetro poblacional, entonces tenemos mucha confianza (95%) rechazando el parámetro poblacional propuesto.

El Plan Para Hoy

Estudiaremos los siguientes temas.

1. **Análisis Correlacional.**
2. **Análisis de Regresión.**

Paquetes de R Que Usaremos

```
library(tidyverse) # para todo lo relacionado con nuestro workflow
library(stevemisc) # para formatear algunas cosas
```

Base de Datos que Usaremos

La base de datos que usaremos esta disponible en la página Github de la clase. Se llama `election.turnout.csv` y contiene datos sobre el porcentaje de votantes que participaron (i.e. turnout) en la elección presidencial de Estados Unidos en 2016.

Correlación

Pregunta: ¿El porcentaje de votantes a nivel estatal varia como consecuencia del nivel de educación estatal?

- Education: % de personas en el estado con preparatoria. (Datos estimados para 2015)
- Turnout: % de participación ciudadana en elección presidencial del 2016.

Podemos hacer una conclusion preliminar usando un **scatterplot**.

- Pero primero vamos a ver un poco nuestros datos.

Analizamos Un Poco los Datos

Estados menos educados en EEUU

```
datos %>% select(state, perhsed) %>%
  top_n(-5, perhsed) %>% arrange(perhsed)
```

```
##         state perhsed
## 1  California    81.8
## 2       Texas    81.9
## 3 Mississippi    82.3
## 4   Louisiana    83.4
## 5    Kentucky    84.2
## 6  New Mexico    84.2
```

Usando Otro Indicador de Educación...

Moraleja: Hay que tener cuidado con el indicador de educación que usamos...

```
datos %>% select(state, percoled) %>%
  top_n(-5, percoled) %>% arrange(percoled)
```

```
##           state percoled
## 1 West Virginia     19.2
## 2   Mississippi     20.7
## 3      Arkansas     21.1
## 4      Kentucky     22.3
## 5     Louisiana     22.5
```

Los Estados Más Educados

```
datos %>% select(state, perhsed) %>%
  top_n(5, perhsed) %>% arrange(-perhsed)
```

```
##           state perhsed
## 1       Montana    92.8
## 2     Minnesota    92.4
## 3 New Hampshire    92.3
## 4       Wyoming    92.3
## 5        Alaska    92.1
```

De Nuevo, Universidad (College) es Diferente...

```
datos %>% select(state, percoled) %>%
  top_n(5, percoled) %>% arrange(-percoled)
```

```
##                  state percoled
## 1 District of Columbia     54.6
## 2        Massachusetts     40.5
## 3             Colorado     38.1
## 4             Maryland     37.9
## 5          Connecticut     37.6
```

% de Participción (Turnout) en 2016...

```
datos %>% select(state, turnoutho) %>%
  top_n(5, turnoutho) %>% arrange(-turnoutho)
```

```
##           state turnoutho
## 1     Minnesota      74.2
## 2 New Hampshire      71.4
## 3         Maine      70.5
## 4      Colorado      70.1
## 5     Wisconsin      69.4
```

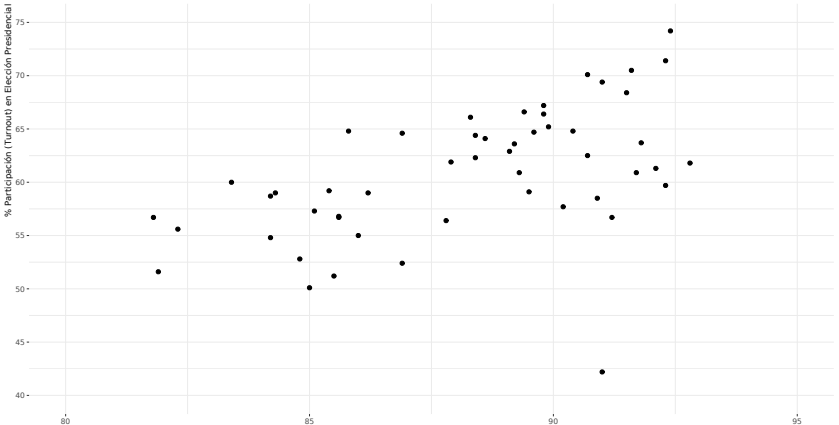
Los Estados con Menor Participación (Turnout)

```
datos %>% select(state, turnoutho) %>%
  top_n(-5, turnoutho) %>% arrange(turnoutho)
```

```
##           state turnoutho
## 1        Hawaii      42.2
## 2 West Virginia      50.1
## 3     Tennessee      51.2
## 4         Texas      51.6
## 5      Oklahoma      52.4
```

Scatterplot de Nivel de Educación Estatal y Turnout en la Elección de 2016

Los datos están dispersos consistente y positivamente. Hawaii es un claro outlier.



Correlación

La relación entre educación y turnout es identificable facilmente: es positiva.

- La relación no es perfecta, pero se ve bastante "fuerte".

¿Qué tan fuerte? El **coeficiente de correlación lineal de Pearson (r)** nos lo dirá.

Coeficiente de Correlación lineal de Pearson, *r*

$$\Sigma \frac{\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right)}{n-1}$$

. . . donde:

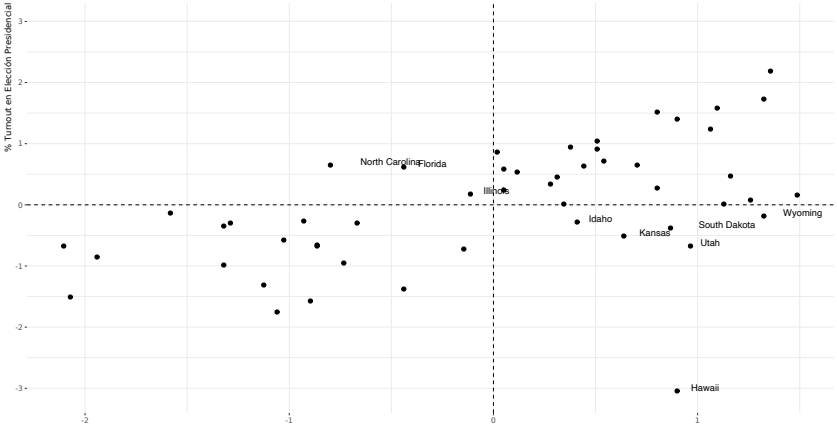
- x_i, y_i = observaciones individuales de x o y, respectivamente.
- $\overline{x}$, $\overline{y}$ = medias muestrales de x y y, respectivamente.
- s_x, s_y = desviación estándar muestral de x y y, respectivamente.
- n = número de observaciones en la muestra.

Propiedades de la *r* de Pearson

1. Es simétrica.
2. Está contenida entre -1 y 1.
3. Es estandarizada.

Scatterplot de Nivel de Educación Estatal y Turnout en la Elección Presidencial del 2016

Observaciones en los cuadrantes de correlación negativos indicados con nombre.



Datos: Elections Project, U.S. Census Bureau.

Educación y Turnout (Z Scores)

- Casos en cuadrante superior-derecho están por encima de medias de x y y.
- Casos en cuadrante inferior-izquierdo están por debajo de la media de x y y.
- Cuandrante superior-izquierdo e inferior-derecho son cuadrantes de correlación negativa.

Dicho esto, el coeficiente de correlación lineal r es $26.41369/50$, o .52.

- Podemos decir informalmente que hay una relación positiva fuerte entre las dos variables.

. . . Calculando en R

```
datos %>%
  mutate(z_perhsed = (perhsed - mean(perhsed))/sd(perhsed),
         z_turnoutho = (turnoutho - mean(turnoutho))/sd(turnoutho)) -> datos

with(datos, sum(z_perhsed*z_turnoutho)/(length(state)-1))
```

```
## [1] 0.5282739
```

```
with(datos, cor(perhsed,turnoutho))
```

```
## [1] 0.5282739
```

Nuestro Outlier, Hawaii...

```
datos %>%
  filter(state != "Hawaii") %>%
  summarize(cor = cor(perhsed, turnoutho))
```

```
##         cor
## 1 0.6540847
```

Regresión Lineal

El coeficiente de correlación tiene algunas características interesantes.

- Es otra herramienta analítica que puede ser usada como "primer paso".
- Útil para detectar **multicolinearidad**.
  - Esto es cuando dos variables independientes están muy correlacionadas y es difícil detectar el efecto parcial de cada una (lo veremos más adelante).

Pero es neutral en lo que es x y lo que es y.

- Es decir, no nos dice nada sobre la causa-efecto.

La regresión nos ayuda con eso.

Demistificando la Regresión

¿Les parece familiar?

$$y = mx + b$$

Demistificando la Regresión

Es la ecuación de una línea recta con pendiente e intercepto.

- b es el intercepto: el valor observado de y cuando $x = 0$.
- m es la pendiente, mide el cambio que hay y por cada cambio unitario en x.

Demistificando la Regresión

La ecuación pendiente-intercepto es, en esencia, la representación de una regresión lineal.

- Los estadísticos o econometristas statisticians preferieren la siguiente notación

$$y = a + b(x)$$

La b es el **coeficiente de regresión** que indica el cambio en y por cada cambio de unidad en x.

Un Ejemplo Simple

Supongamos que quiero explicar tu calificación en un examen (y) usando el número de horas que estudiaste para dicho examen (x).

Horas (x)	*Calificación (y)*
0	55
1	61
2	67
3	73
4	79
5	85
6	91
7	97

Tabla: Horas de estudio y calificación en examen.

Un Ejemplo Simple

En esta clase, el estudiante que estudió 0 hours sacó 55.

- El estudiante que estudió 1 hora obtuvo un 61.
- Quien estudió 2 obtuvo un 67.
- . . . y así sucesivamente. . .

Cada hora de estudio adicional produce un cambio de seis unidades en la calificación. Lo podemos denotar así:

$$y = a + b(x) = \text{Calificación} = 55 + 6(x)$$

Nótese que el intercepto de y es en 55.

Un Ejemplo Menos Simple

En realidad los datos nunca son tan simples. Compliquemos un poco...

Horas (x)	*Calificación (y)*	*Calificación Estimada ($\hat{y}$)*
0	53	55
0	57	
1	59	61
1	63	
2	65	67
2	69	
3	71	73
3	75	
4	77	79
4	81	
5	83	85
5	87	
6	89	91
6	93	
7	95	97
7	99	

Un Ejemplo Menos Simple

Complicando un poco los datos no cambia la linea de regresión.

- Notemos que la regresión promedia sobre diferencias.
- Una hora de estudio adicional, *en promedio*, corresponde a un incremento de seis unidades en la calificación.
- Hemos visto nuestras observaciones (y) y nuestros estimados ($\hat{y}$, o y-gorro).

Nuestra Línea de Regresión

La línea de regresión es:

$$\hat{y} = \hat{a} + \hat{b}(x) + e$$

. . . donde:

- $\hat{y}$, $\hat{a}$ y $\hat{b}$ son estimados de y, a, y b sobre los datos.
- e es el error.
  - El error contiene: error de muestreo aleatorio, y el error de predicción.

Obtención del Coeficiente de Regresión

¿Cómo se obtiene el coeficiente de regresión en datos más complejos?

- Empezamos con el **error de predicción**, formalmente: $y_i - \hat{y}$.
- Los elevamos al cuadrado. Esto es: $(y_i - \hat{y})^2$
  - La suma de los errores al cuadrado es igual a cero.

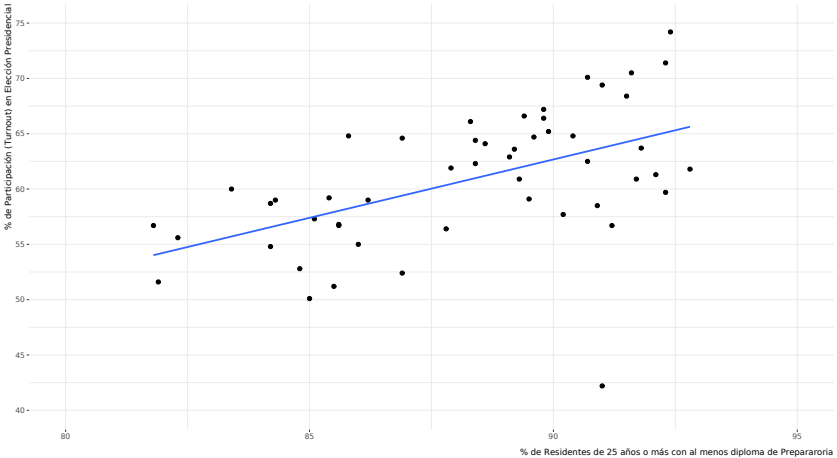
El coeficiente de regresión que resulta *minimiza* la suma de diferencias al cuadrado ($(y_i - \hat{y})^2$).

- En otras palabras: regresión de "mínimos cuadrados ordinarios" (OLS - Ordinary Least Squares).

El siguiente gráfico nos da una representación de esto usando el ejemplo de educación y turnout estatal.

Educación y Turnout en la Elección del 2016 (EEUU)

La línea que minimiza la suma de los errores al cuadrado se hace a través de esos puntos.



Error Estándar del Coeficiente de Regresión

Cada parámetro (o variable) en el modelo de regresión viene con un error estándar.
- Que estima con que precisión el modelo estima el valor desconocido del coeficiente(s).

El procedimiento para estimar los errores estándar no es tan sencillo.

- Necesitamos la diagonal de raíces cuadradas de la matriz de varianza-covarianza.
- Y para ello requerimos álgebra matricial, que sale de nuestros objetivos en esta clase.

En R lo podemos obtener con facilidad.

Paquete `lm` en `R`

```
summary(M1 <- lm(turnoutho ~ perhsed, data=datos))
```

```
##
## Call:
## lm(formula = turnoutho ~ perhsed, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.529  -3.510   1.176   3.676   8.994
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -32.3027    21.3948  -1.510    0.138
## perhsed       1.0553     0.2423   4.355 6.77e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.247 on 49 degrees of freedom
## Multiple R-squared:  0.2791, Adjusted R-squared:  0.2644
## F-statistic: 18.97 on 1 and 49 DF,  p-value: 6.765e-05
```

Lo que Vemos con `summary`

Lo más importante:

- "Estimaciones": intercepto de y, y coeficientes de regresión.
- Errores Estándar: un estimado de la variabilidad alrededor del coeficiente estimado.
- Pruebas estadísticas (estadístico t, valor de p): los usaremos para hacer inferencias.
- R^2s: mide que tan bien nuestro modelo se ajusta a los datos.

Otros estadísticos (menos importantes):

- Estadístico F: "significancia total" del modelo.
- Error estándar residual: error estándar de los residuos.
  - Usado para calcular los errores estándar junto a la matriz varianza-covarianza.
- Distribución de los residuales: nos da un resúmen del rango de los residuales.

También en R pero con Fórmulas

```
X <- model.matrix(M1) # Intercepto + perhsed

# Suma de cuadrado residuales
sigma2 <- sum((datos$turnoutho - fitted(M1))^2) / (nrow(X) - ncol(X))

sqrt(sigma2) # error estándar residual
```

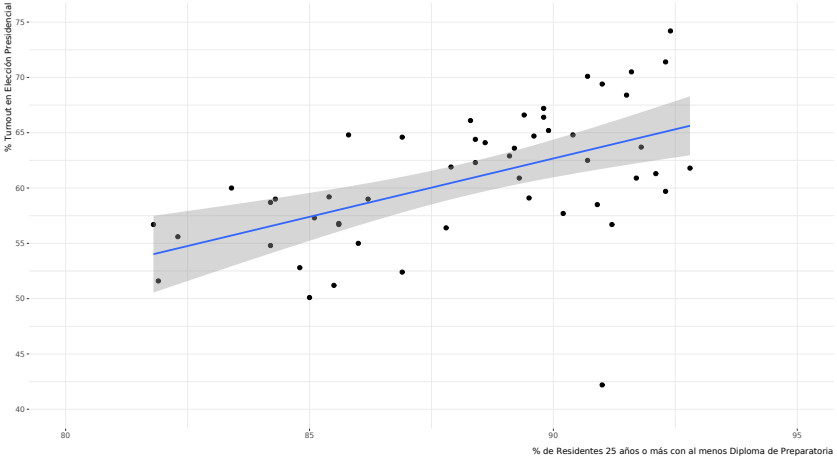
```
## [1] 5.246687
```

```
sqrt(diag(solve(crossprod(X))) * sigma2)
```

```
## (Intercept)     perhsed
##   21.394761    0.242304
```

Educación y Turnout en la Elección del 2016 (EEUU)

La línea que minimiza la la suma de los errores al cuadrado se hace a través de esos puntos.



Regresión: Educación y Turnout

Esta sería nuestra línea de regresión:

$$\hat{y} = -32.30 + 1.05(x)$$

Interpretamos esto de la siguiente forma:

- Para el estado en donde nadie se gradúa de preparatoria, el turnout será de -32.30%.
  - *Un poco extraño el resultado, pero es porque no centramos las variables...*
- Por cada incremento unitario en el porcentaje de personas que se gradúan de prepa, el turnout se incrementará en apoximadamente 1.05%.

Inferencia en Regresión

¿Qué podemos decir sobre b-gorro ($\hat{b} = 1.05$?)

- Si usamos una muestra diferente, ¿observaríamos algo muy diferente?
- ¿Cómo lo podemos saber?

Inferencia en Regresión

Esto lo hemos visto antes. ¿Se acuerdan de los valores Z?

$$Z = \frac{\overline{x} - \mu}{s.e.}$$

Inferencia en Regresión

Hacemos lo mismo, pero usando una distribución t Student.

$$t = \frac{\hat{b} - \beta}{s.e.}$$

$\hat{b}$ es el coeficiente de regresión. ¿Qué es β?

Inferencia en Regresión

β es **cero**.

- Estamos probando si el coeficiente de regresión es un artefacto del proceso de muestreo.
- Estamos probando que hay (o no) relación entre x y y.

Inferencia en Regresión

Esto simplifica las cosas.

$$t = \frac{\hat{b}}{s.e.}$$

Inferencia en Regresión

En el ejemplo de educación y turnout, es lo siguiente.

$$t = \frac{1.05}{.24} = 4.35$$

El coeficiente de regresión está a más de cuatro errores estándar de cero.

- La probabilidad de observarlo si β fuera realmente cero es de .000067.
- Por lo tanto, inferimos que nuestro coeficiente es estadísticamente signiificativo con 95% de confianza.

Conclusión

- El coeficiente de correlación no ayuda a saber si existe relación entre dos variables. Pero hay otras herramientas para hacer una inferencia más adecuada.
- Una de esas herramientas en la regresión lineal simple, a la que regresaremos la próxima semana.

Table of Contents