Tecnológico de Monterrey
Escuela de Ciencias Sociales
y Gobierno

Modelos Lineales

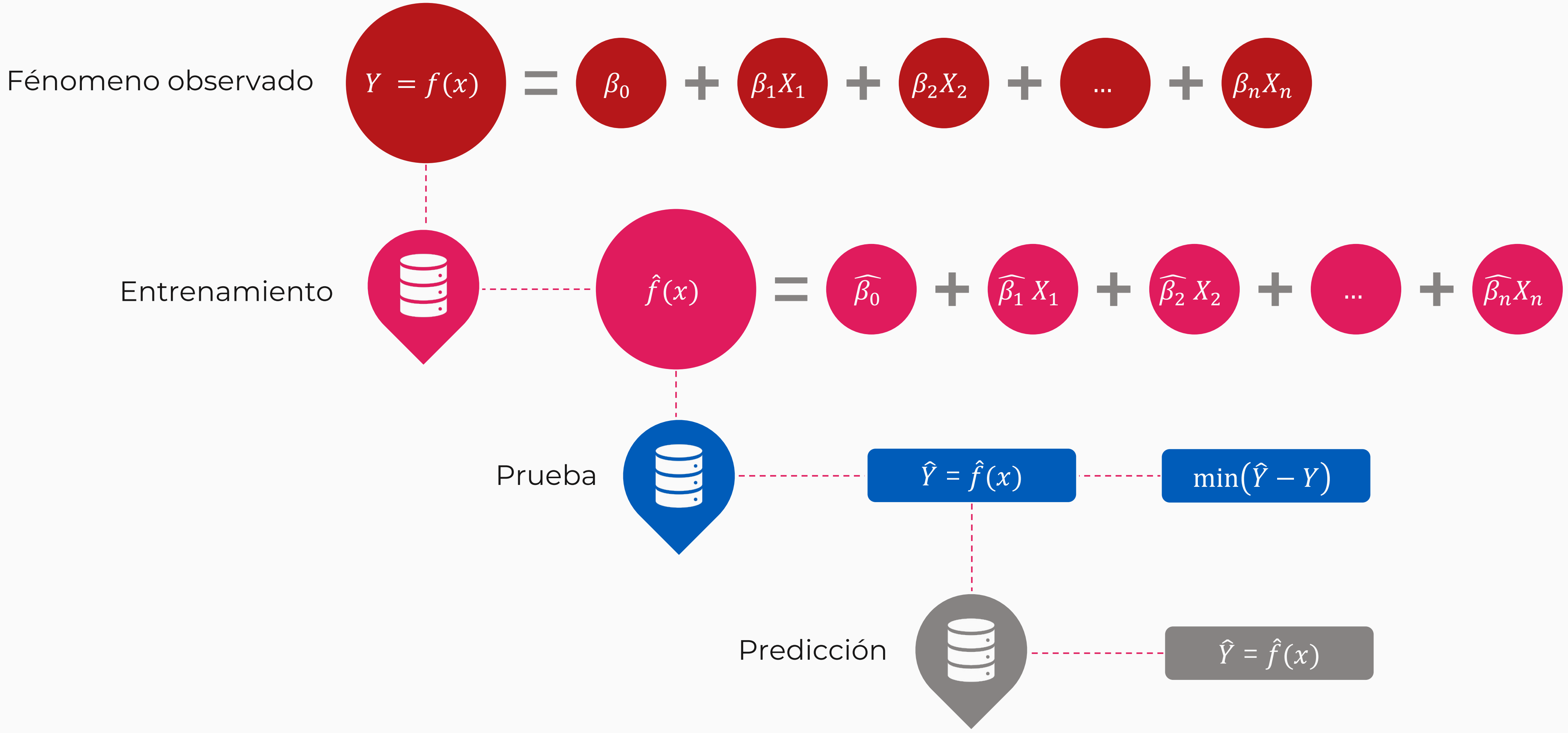
Mtro. René Rosado González
Director de Programa LTP

r.rosado@tec.mx

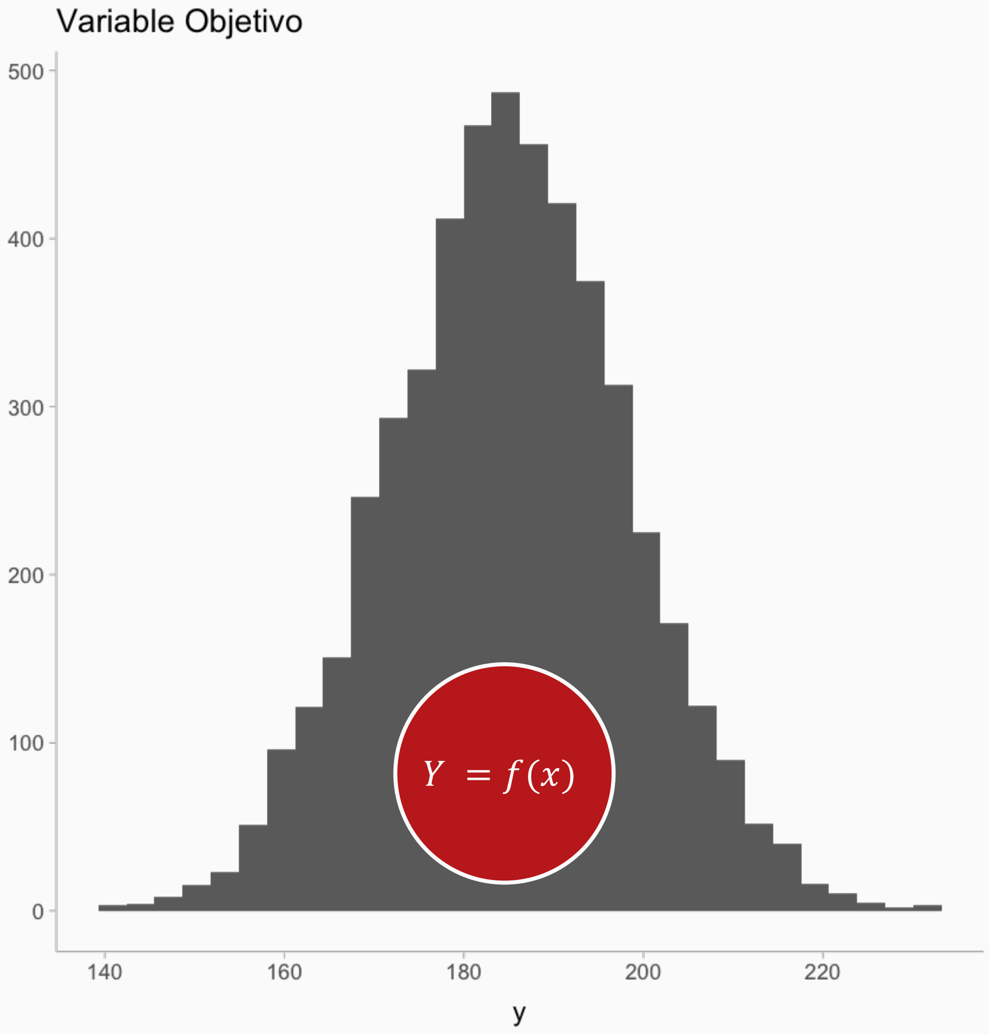
Modelos Lineales

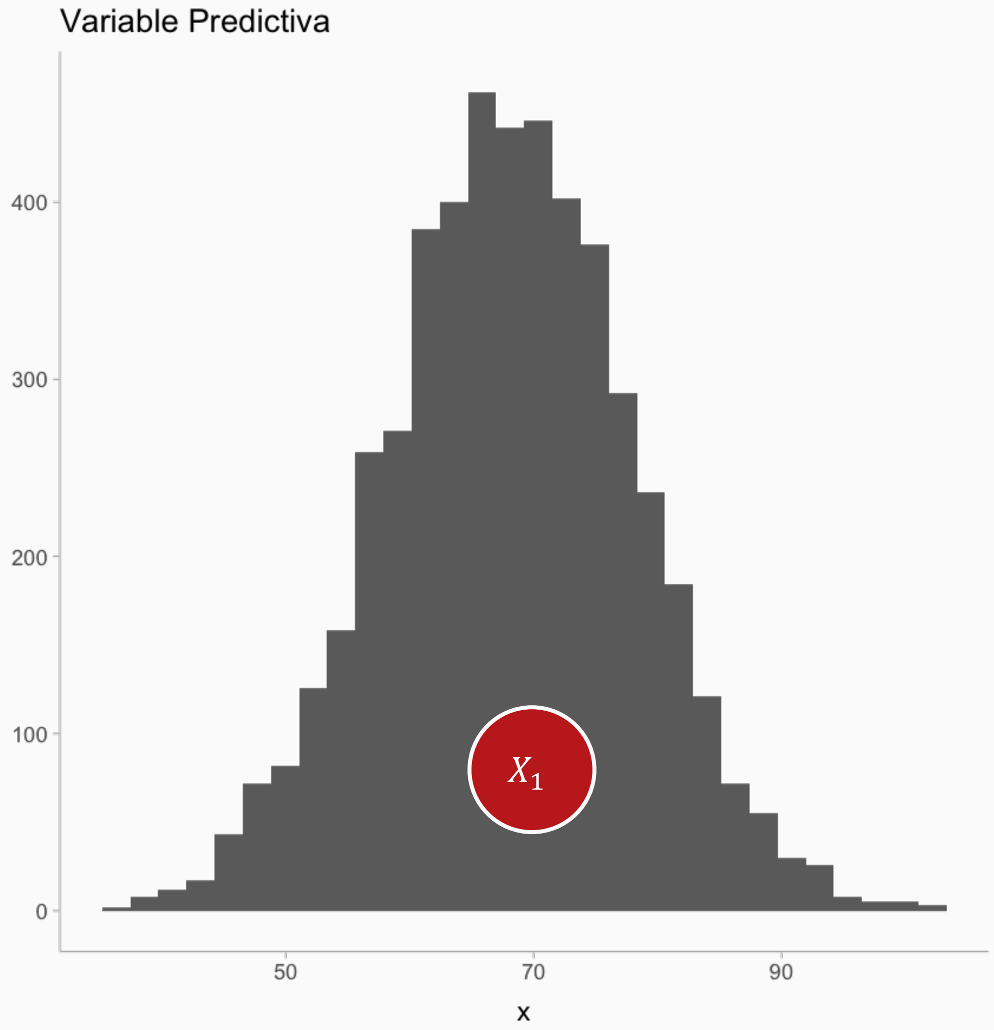
Fénomeno observado $Y = f(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$

Modelos Lineales

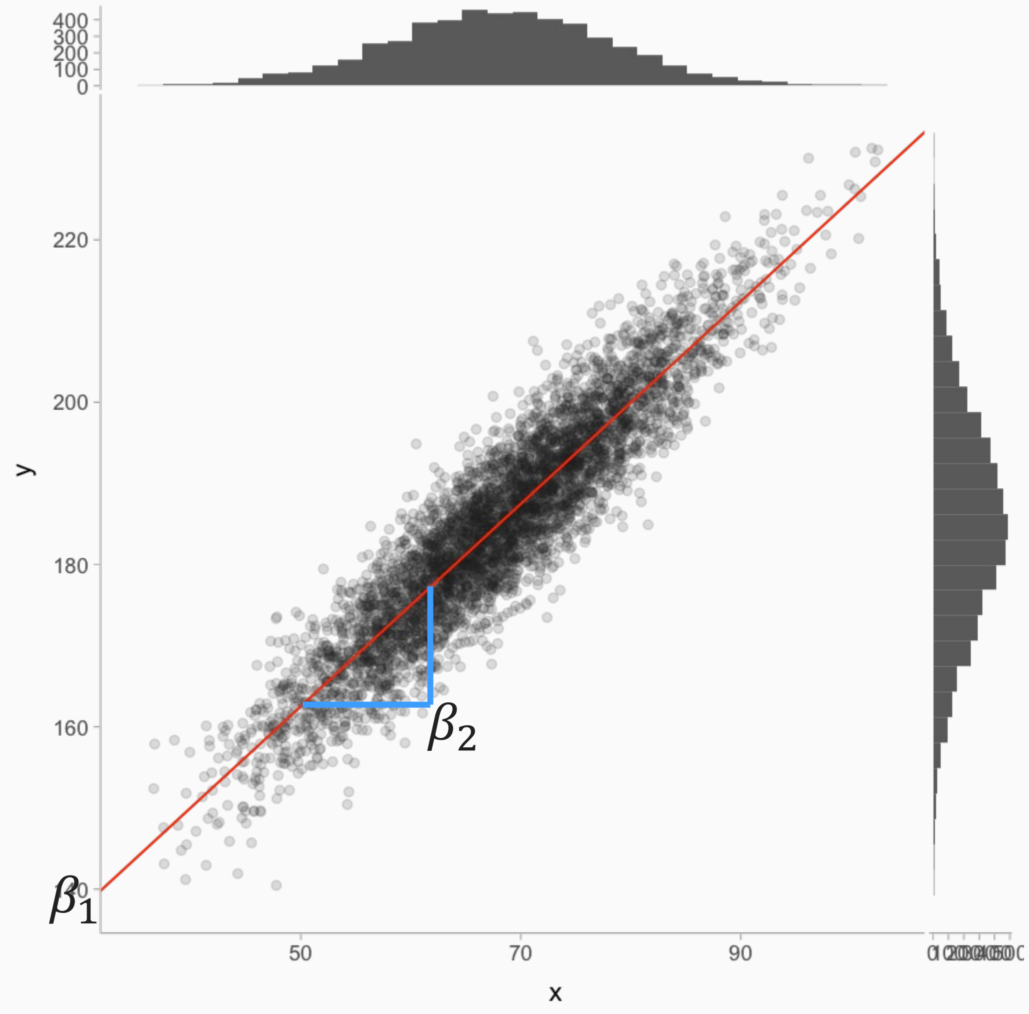


Regresión Lineal





Regresión Lineal Simple



$$y = x\beta$$

$$\begin{matrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{matrix} = \begin{matrix} \beta_1 & +\beta_2 x_{1,1} \\ \beta_1 & +\beta_2 x_{1,2} \\ \cdots & \cdots \\ \beta_1 & +\beta_2 x_{1,n} \end{matrix}$$

$$\hat{y} = x\hat{\beta} + \epsilon$$

$$\epsilon \sim N(0, \sigma_\epsilon^2 I)$$

- Homocedasticidad.
- Normalidad de los errores
- No correlación entre los errores

Ajuste por Mínimos Cuadrados Ordinarios

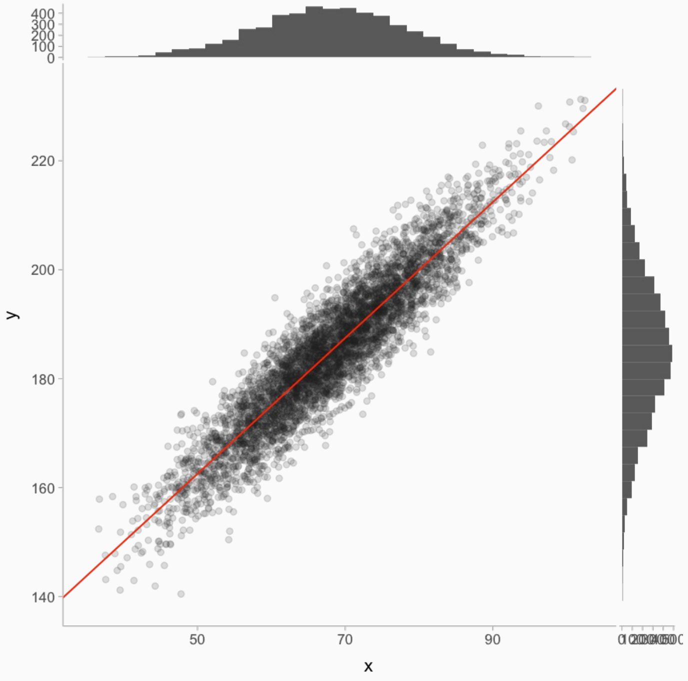
$$\min\left(f(y-\hat{y})\right) = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = S$$

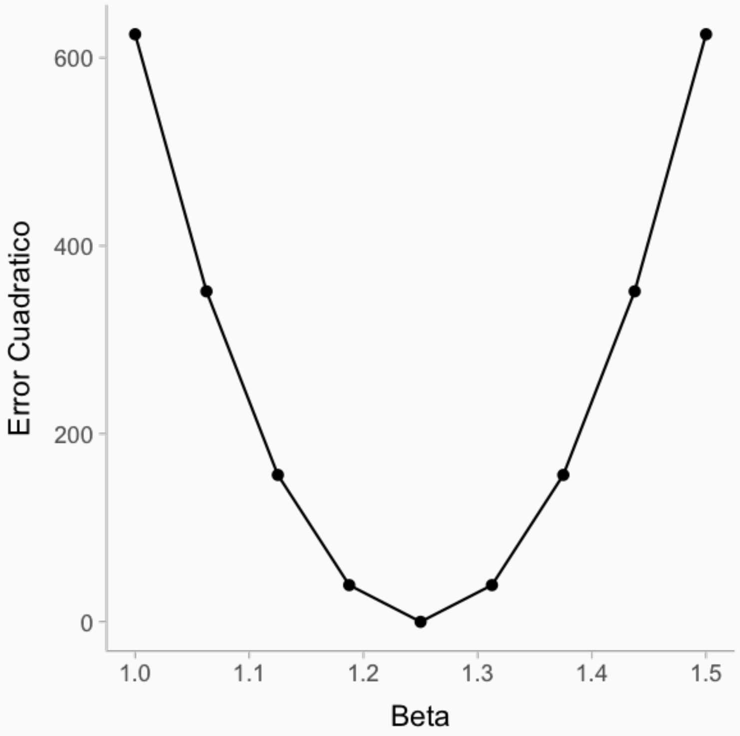
Condición de Primer Orden

$$\frac{\delta S}{\delta \beta_i} = 0$$

$$\beta_1^* = \bar{y} - b\bar{x}$$

$$\beta_2^* = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$





El Intercepto

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2$$

$$\frac{\partial S}{\partial \beta_1} = 2 \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i) = 0$$

$$\beta_1^* = \bar{y} - \beta_2 \bar{x}$$

La Pendiente

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2$$

$$\frac{\partial S}{\partial \beta_2} = 2\sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)\, x_i = 2\sum_{i=1}^{n} (y_i - (\bar{y} - \beta_2 \bar{x}) - \beta_2 x_i)\, x_i =$$

$$2\sum_{i=1}^{n} \big((y_i - \bar{y}) + \beta_2(\bar{x} - x_i)\big)\, x_i = 2\sum_{i=1}^{n} (y_i - \bar{y})\, x_i + \beta_2(\bar{x} - x_i)\, x_i = 0$$

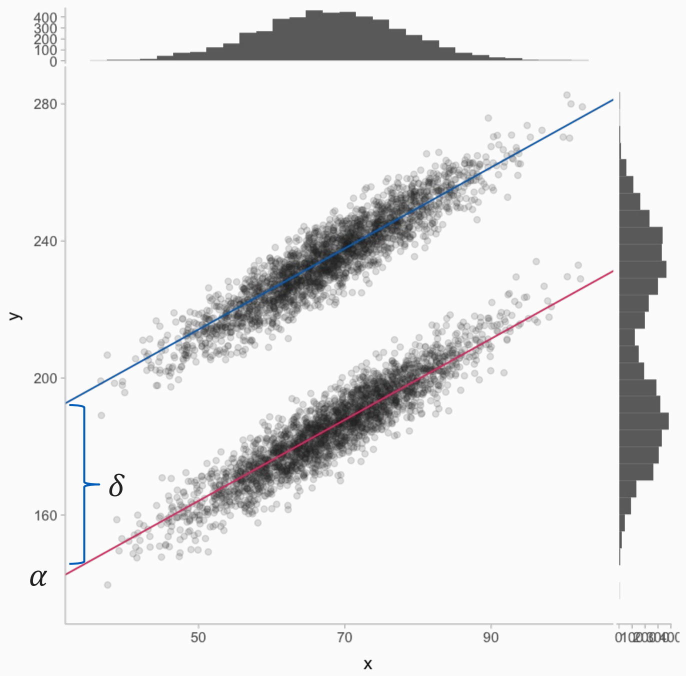
$$\beta_2^* = \frac{\sum_{i=1}^{n} (y - \bar{y})(\, x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \ \bar{x})^2} = \frac{cov(y, x)}{var(x)}$$

Un ejemplo

Modelos con interacciones

Diferentes interceptos

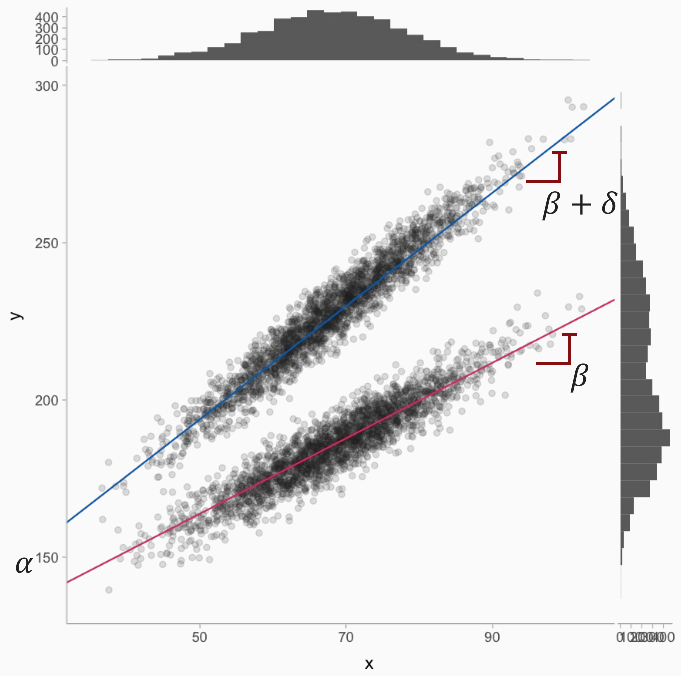
$$y = \alpha + \beta x + \delta i$$
$$= (\alpha + \delta i) + \beta x$$



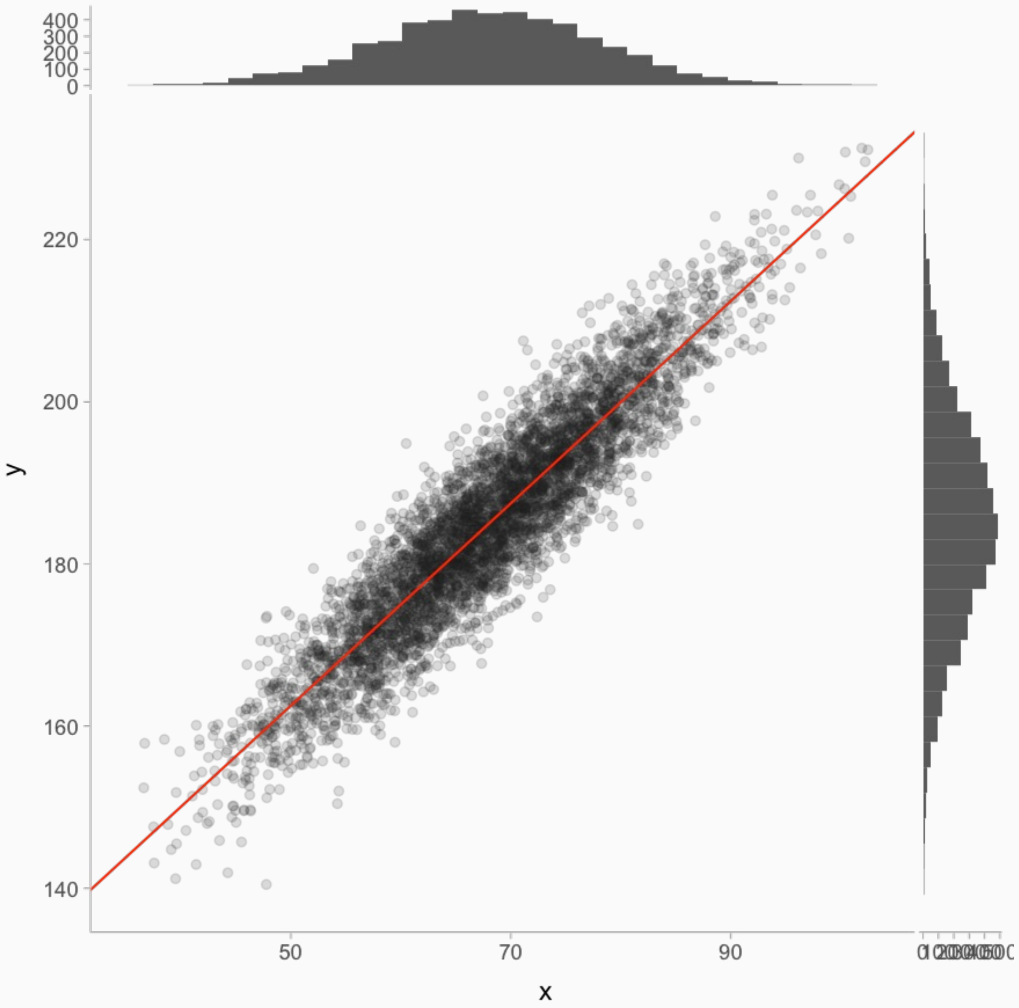
$$i\begin{cases}1 & si\ clase\ A\\ 0 & otra\ clase\end{cases}$$

Diferentes Pendientes

$$y = \alpha + \beta x + \delta i * x$$
$$= \alpha + (\beta + \delta i)x$$



Regresión Lineal Múltiple



$$y = x\beta$$

$$\begin{matrix} y_1 \\ y_2 \\ \dots \\ y_n \end{matrix} = \begin{matrix} \beta_1 & +\beta_2 x_{1,1} & \dots & +\beta_k x_{k,1} \\ \beta_1 & +\beta_2 x_{1,2} & \dots & +\beta_k x_{k,2} \\ \dots & \dots & \dots & \dots \\ \beta_1 & +\beta_2 x_{1,n} & \dots & +\beta_k x_{k,n} \end{matrix}$$

$$\hat{y} = x\hat{\beta} + \epsilon$$

$$\epsilon \sim N(0, \sigma_\epsilon^2 I)$$

- Homocedasticidad.
- Normalidad de los errores
- No correlación entre los errores

El Modelo Matricial

$$
\begin{matrix} y_1 \\ y_2 \\ \dots \\ y_n \end{matrix} = \begin{matrix} \beta_1 & +\beta_2 x_{1,1} & \dots & +\beta_k x_{k,1} \\ \beta_1 & +\beta_2 x_{1,2} & \dots & +\beta_k x_{k,2} \\ \dots & \dots & \dots & \dots \\ \beta_1 & +\beta_2 x_{1,n} & \dots & +\beta_k x_{k,n} \end{matrix} = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{k,1} \\ 1 & x_{1,2} & \dots & x_{k,2} \\ \dots & \dots & \dots & \dots \\ 1 & x_{1,n} & \dots & x_{k,n} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_n \end{bmatrix}
$$

$$
\underline{y} = X\underline{\beta}
$$

El Modelo Matricial

$$\underline{y} - \underline{\hat{y}} = \underline{\epsilon} = \underline{y} - X\underline{\beta}$$

$$S = \underline{\epsilon}^T \underline{\epsilon} = \left(\underline{y} - X\underline{\beta}\right)^T \left(\underline{y} - X\underline{\beta}\right) = \left(\underline{y}^T - \underline{\beta}^T X^T\right)\left(\underline{y} - X\underline{\beta}\right) =$$

$$\underline{y}^T \underline{y} - \underline{y}^T X \underline{\beta} - X^T \underline{\beta}^T \underline{y} + \underline{\beta}^T X^T X \underline{\beta}$$

Condiciones de primer orden

$$\frac{\partial S}{\partial \underline{\beta}} = \begin{bmatrix} \frac{\partial S}{\partial \beta_1} \\ \vdots \\ \frac{\partial S}{\partial \beta_n} \end{bmatrix} = 0$$

El Modelo Matricial

$$\frac{\partial S}{\partial \underline{\beta}} = \frac{\left(\underline{y}^T \underline{y} - \underline{y}^T X \underline{\beta} \ - X^T \underline{\beta}^T \underline{y} + \underline{\beta}^T X^T X \underline{\beta}\right)}{\partial \underline{\beta}} = 0$$

$$= 0 - \frac{2\left(\underline{y}^T X \underline{\beta}\right)}{\partial \underline{\beta}} + \frac{\left(\underline{\beta}^T X^T X \underline{\beta}\right)}{\partial \underline{\beta}}$$

$$= -2\left(\underline{y}^T X\right) + 2(X^T X)\underline{\beta} = 0$$

$$\underline{\beta} = (X^T X)^{-1}(X^T y)$$

El Modelo Matricial (Bivariado)

$$\underline{\beta} = (X^T X)^{-1}(X^T y)$$

$$(X^T X) = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum_i^n x_i^2 \end{bmatrix}$$

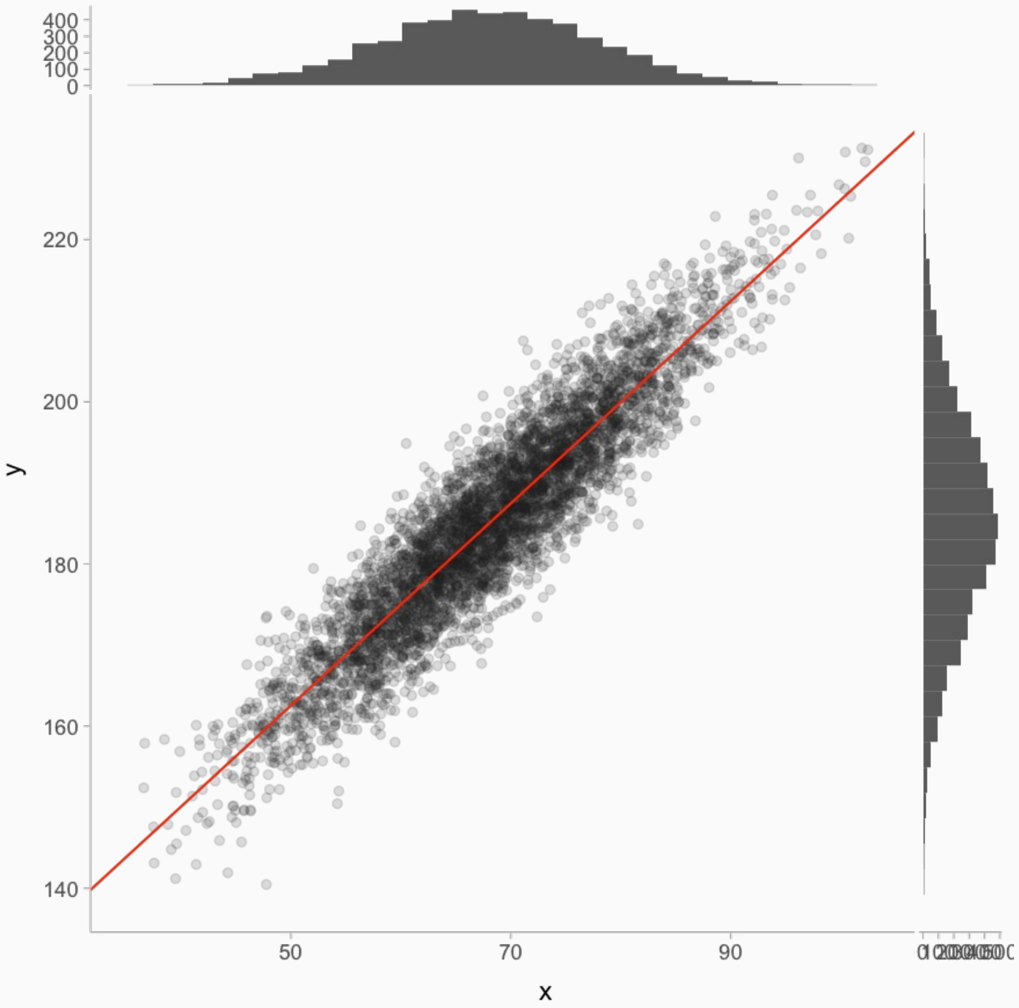
$$(X^T X)^{-1} = \begin{bmatrix} \sum_i^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix}$$

$$(X^T y) = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} n\bar{y} \\ \sum_i^n x_i y_i \end{bmatrix}$$

$$\underline{\beta} = \begin{bmatrix} \sum_i^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \begin{bmatrix} n\bar{y} \\ \sum_i^n x_i y_i \end{bmatrix} = \frac{\sum_i^n x_i y_i - n\overline{x\bar{y}}}{\sum_i^n x_i^2 - n\overline{x^2}} = \frac{cov(y,x)}{var(x)}$$

Un ejemplo

Regresión Lineal Simple



$$y = x\beta$$

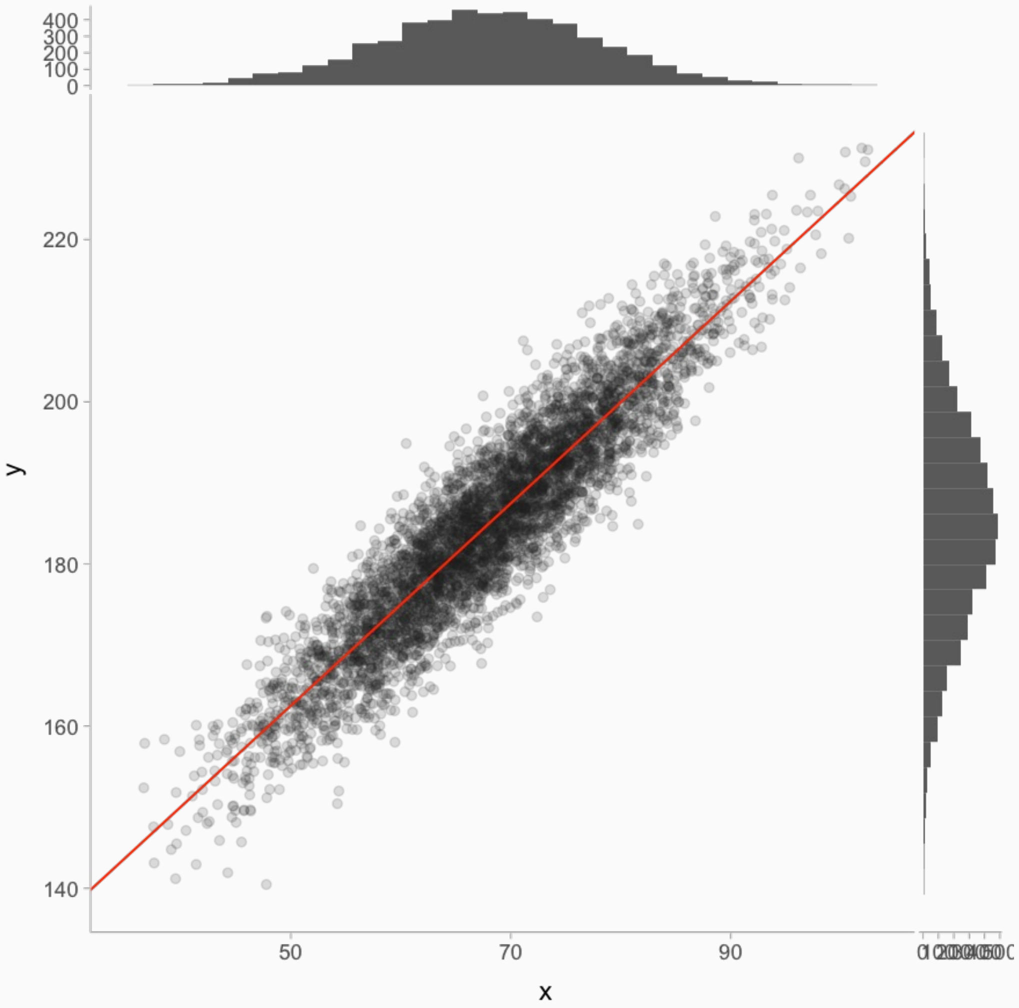
$$\begin{matrix} y_1 \\ y_2 \\ \dots \\ y_n \end{matrix} = \begin{matrix} \beta_1 & +\beta_2 x_{1,1} & \dots & +\beta_k x_{k,1} \\ \beta_1 & +\beta_2 x_{1,2} & \dots & +\beta_k x_{k,2} \\ \dots & \dots & \dots & \dots \\ \beta_1 & +\beta_2 x_{1,n} & \dots & +\beta_k x_{k,n} \end{matrix}$$

$$\hat{y} = x\hat{\beta} + \epsilon$$

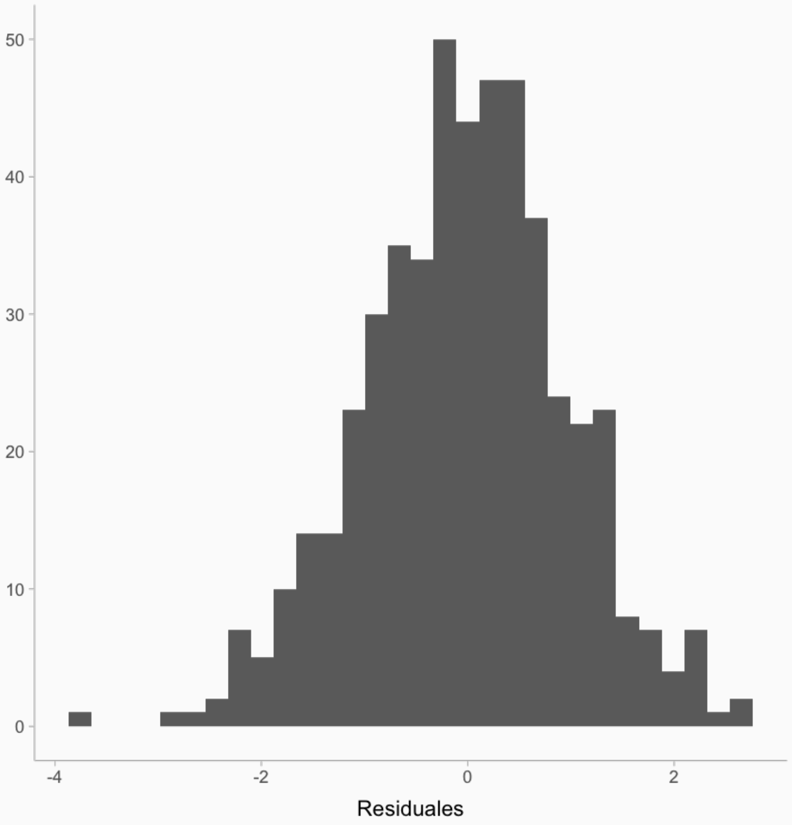
$$\epsilon \sim N(0, \sigma_\epsilon^2 I)$$

- Homocedasticidad.
- Normalidad
- No correlación

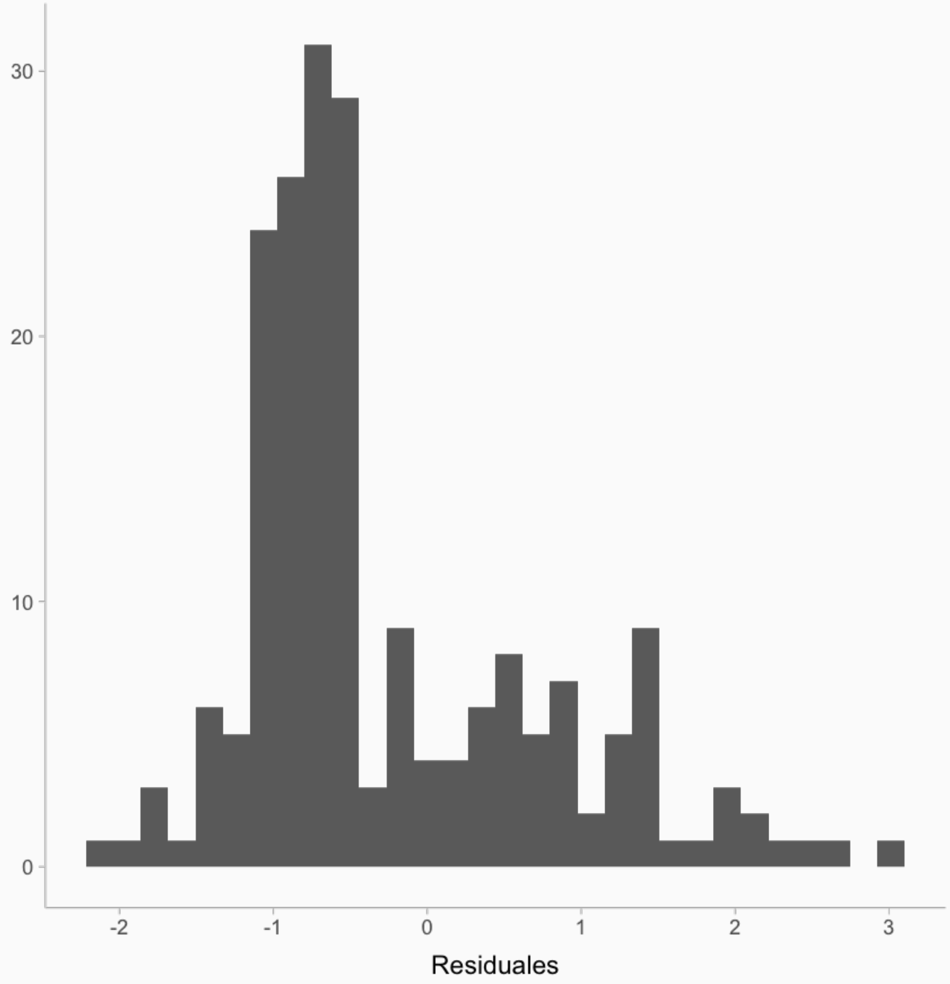
Regresión Lineal Gaussiana

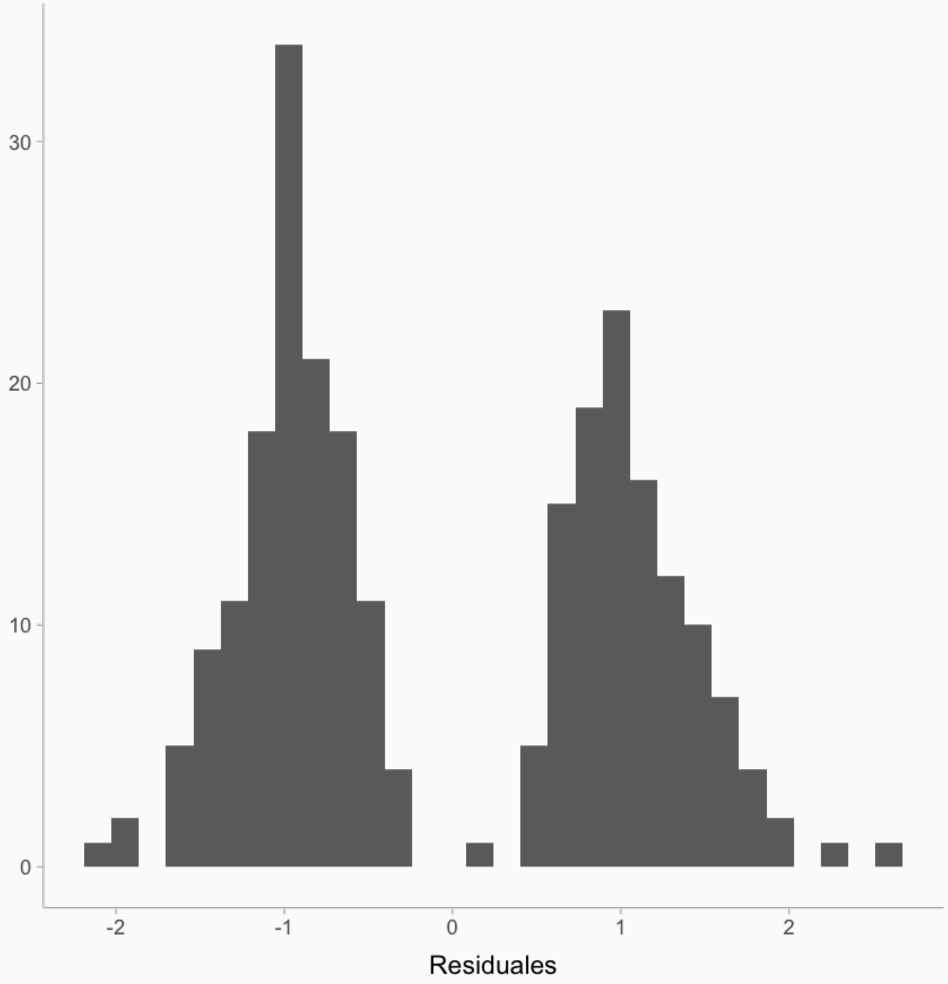


$$\hat{y} = x\hat{\beta} + \epsilon \qquad \epsilon \sim N(0, \sigma_\epsilon^2 I)$$



¡René, mis residuales no son normales!





“Normality is a fine ideal for those who have no imagination”

- Carl Jung

Modelos Lineales Generalizados

Fénomeno observado $Y = f(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$

Entrenamiento $\hat{f}(x) = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \widehat{\beta_2} X_2 + \ldots + \widehat{\beta_n} X_n$

Prueba $\hat{Y} = \hat{f}(x)$ $\min(\hat{Y} - Y)$

Predicción $\hat{Y} = \hat{f}(x)$

$$y = f(x)$$

Liga Canónica
Link fucntion

Modelos Lineales Generalizados

- Son un tipo de modelado estadístico formulado por John Nelder y Robert Wedderburn en 1972.

- Es un término general que abarca muchos otros modelos, lo que permite que la variable de respuesta y tenga una distribución de error diferente a la distribución normal.

- Los modelos que cubriremo en este curso son

    - regresión lineal,
    - regresión de Poisson.
    - regresión logística

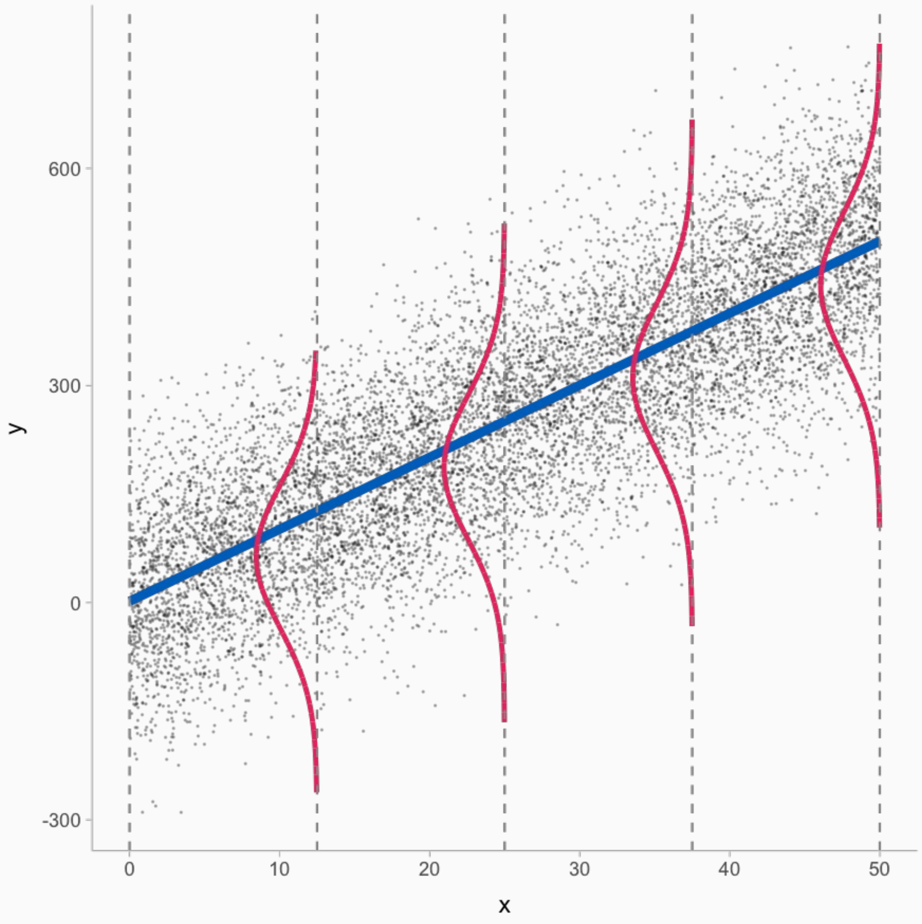
Modelos Lineales Generalizados

- Los datos deben son independientes y aleatorios.
- Cada variable aleatoria tiene la misma distribución de probabilidad.
- La variable de respuesta tiene una distribución de una familia exponencial (por ejemplo, binomial, Poisson, multinomial, normal).
- La variable de respuesta original no necesita tener una relación lineal con las variables independientes, pero la variable de respuesta transformada en la liga canónica depende linealmente de las variables independientes.

Preguntas no normales

- ¿El número de muertes de motociclistas en un año determinado está relacionado con las leyes estatales sobre uso de cascos?
- ¿La cantidad de inasistencias durante un año difiere para las escuelas públicas y privadas?
- ¿El número diario deingresos por asma a una Sala de Emergencias difiere según los índices de contaminación del aire?

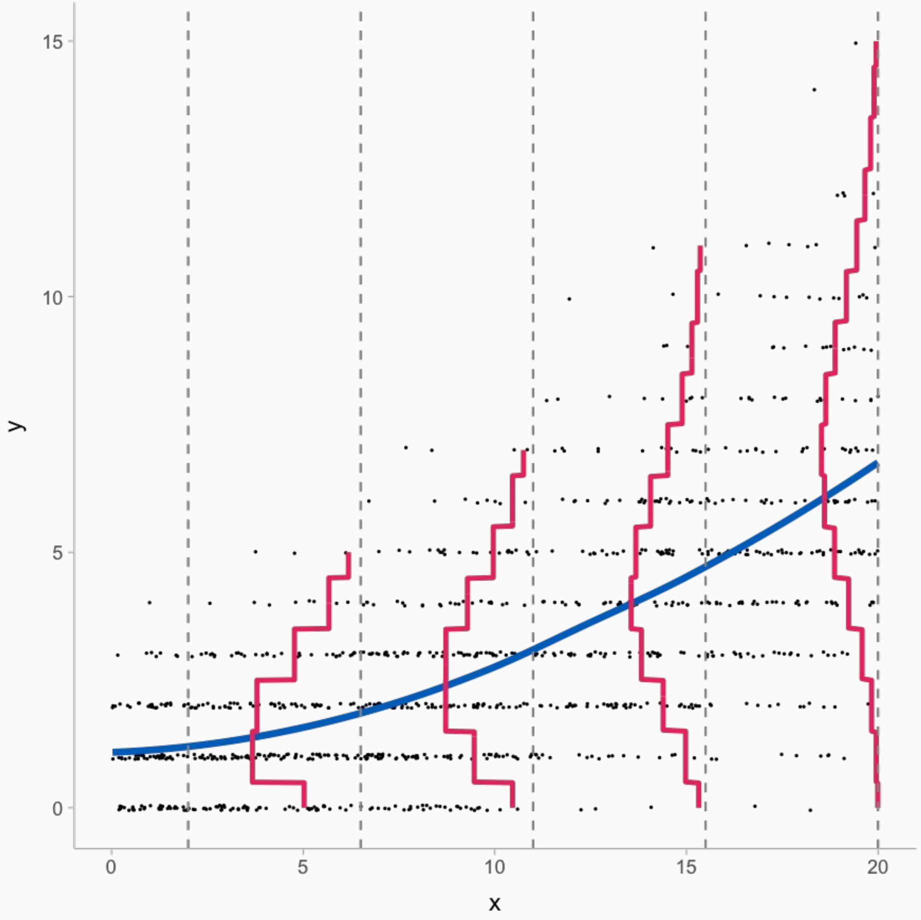
Regresión Gaussiana



Preguntas Poisson

- ¿El número de muertes de motociclistas en un año determinado está relacionado con las leyes estatales sobre uso de cascos?
- ¿La cantidad de inasistencias durante un año difiere para las escuelas públicas y privadas?
- ¿El número diario deingresos por asma a una Sala de Emergencias difiere según los índices de contaminación del aire?

Regresión Poisson



Regresión Poisson

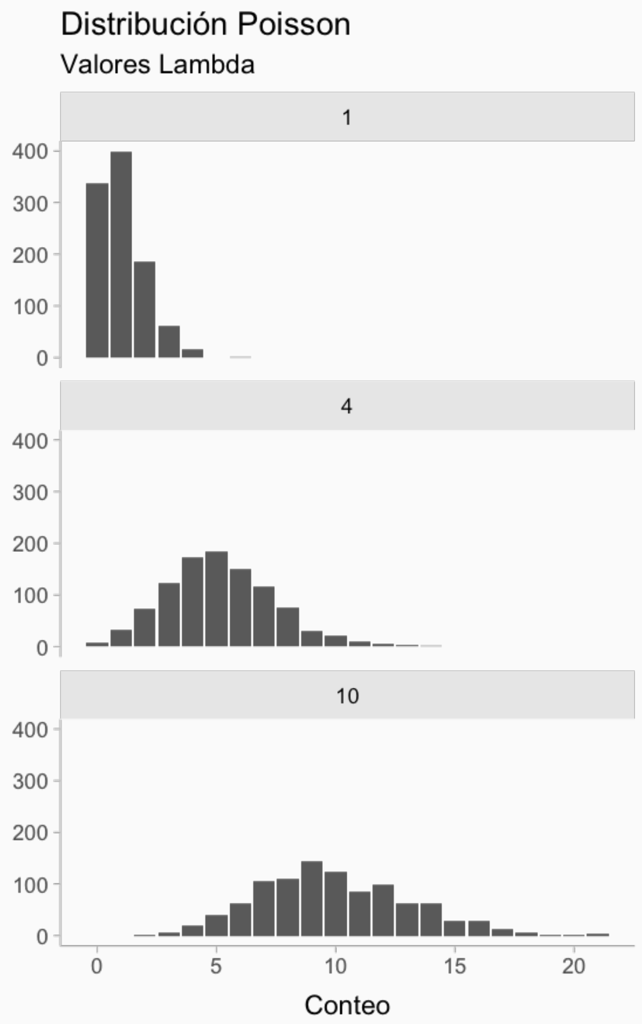
Ditribución Poisson:

$$\Pr(Y = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

s.a. $\lambda > 0$ y $\lambda = E(y) = Var(y)$

Liga canónica:

$$\lambda(x) = e^{\beta x} <=> \log(\lambda(x)) = \beta x$$



Máxima Versosimilitud

La verosimilitud es una función que nos dice qué tan probable es que observemos nuestros datos para un valor de parámetro dado.

Supongamos que temenos conteos observados de tamaño 4, 2, 8, 6 y 1:

$$Verosimilitud = \Pr(Y_1 = 4) * \Pr(Y_2 = 2) * \Pr(Y_3 = 8) * \Pr(Y_4 = 6) * \Pr(Y_5 = 1)$$

$$Verosimilitud = \frac{e^{-\lambda_1}\lambda_1^4}{4!} * \frac{e^{-\lambda_2}\lambda_2^2}{2!} * \frac{e^{-\lambda_3}\lambda_3^8}{8!} * \frac{e^{-\lambda_4}\lambda_4^6}{6!} * \frac{e^{-\lambda_5}\lambda_5^1}{1!} = \prod_i^N \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

$$-Logversoimilitud =$$

$$-logL \propto \lambda_1 - 4\log(\lambda_1) + \lambda_2 - 2\log(\lambda_2) + \lambda_3 - 8\log(\lambda_3) + \lambda_4 - 6\log(\lambda_4) + \lambda_5 - 1\log(\lambda_5)$$

$$logL \propto \sum_i^N -\lambda_i + y_i \log(\lambda_i) = \sum_i^n -e^{\beta_0+\beta_1 x_i} + y_i \log(\beta_0 + \beta_1 x_i)$$

Regresión Poisson

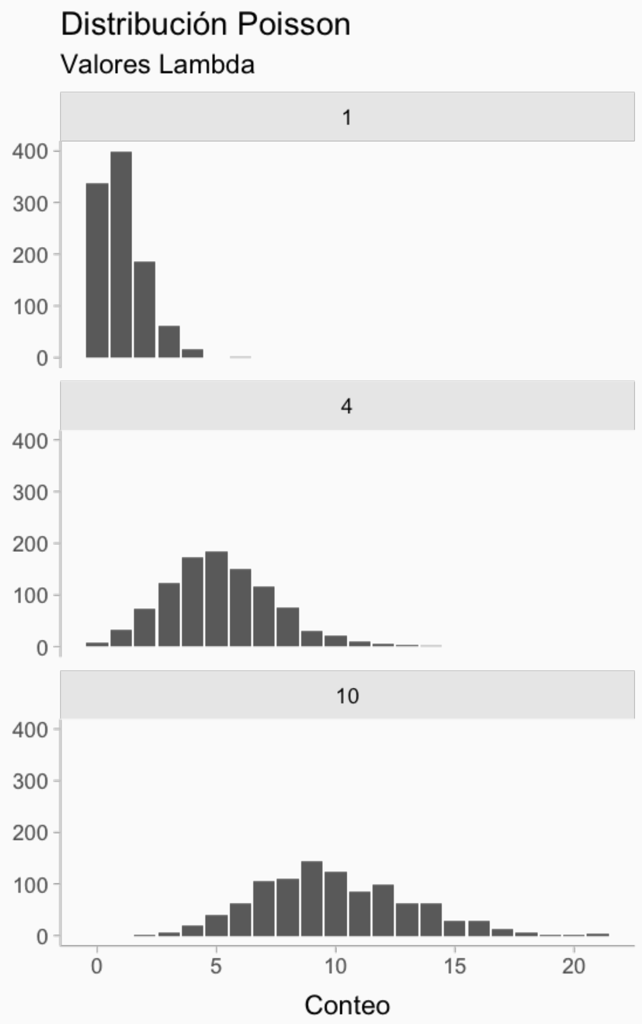
Ditribución Poisson:

$$\Pr(Y = k) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

s.a. $\lambda > 0$ y $\lambda = E(y) = Var(y)$

Liga canónica:

$$\lambda(x) = e^{\beta x} <=> \log(\lambda(x)) = \beta x$$



Regresión Poisson

Supuestos

1. Respuesta de Poisson

   La variable de respuesta es un conteo por unidad de tiempo o espacio, descrita por una distribución de Poisson.

2. Independencia

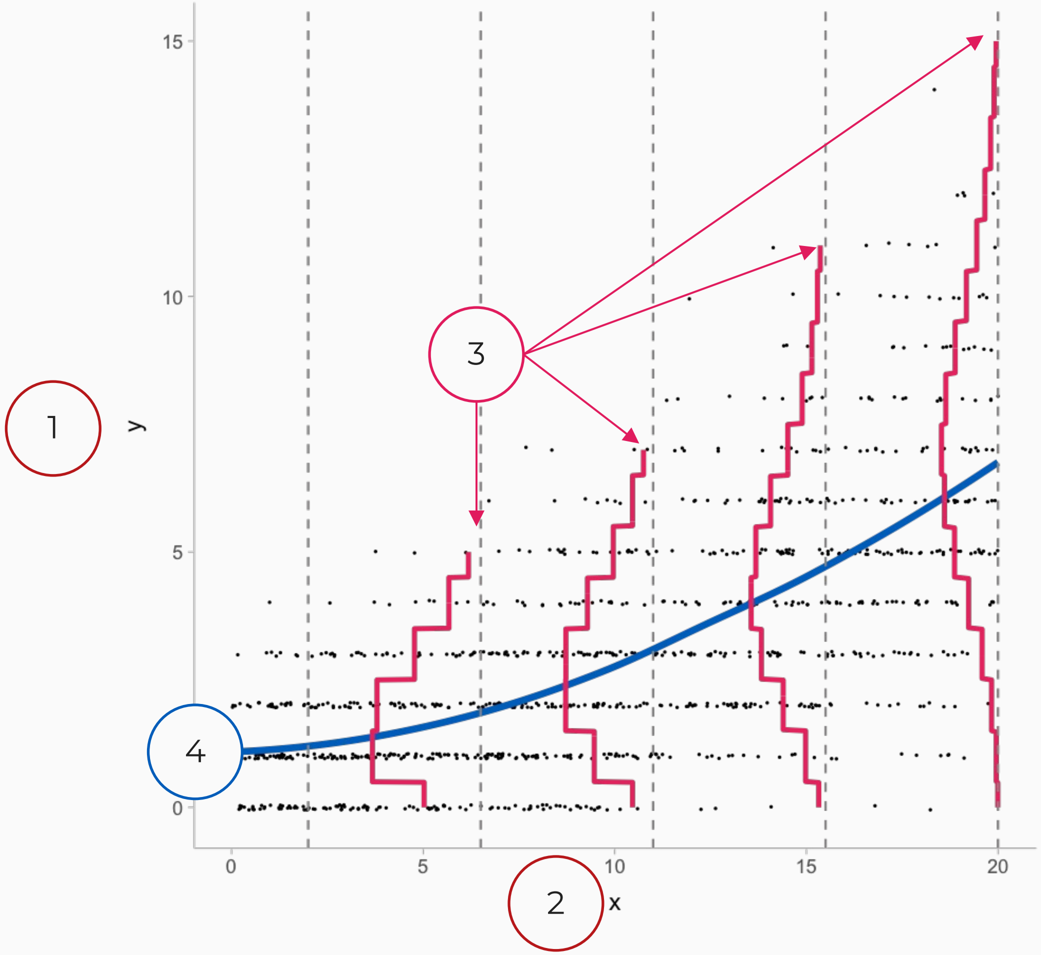
   Las observaciones deben ser independientes entre sí.

3. La media es igual a la varianza

   Por definición, la media de una variable aleatoria tipo Poisson debe ser igual a su varianza.

4. Linealidad

   El logaritmo de la tasa media, log(λ), debe ser una función lineal de x.



Regresión Poisson

Fénomeno observado $\log(E(y|x)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$

Entrenamiento $\log(E(y|x)) = \widehat{\beta_0} + \widehat{\beta_1}\, X_1 + \widehat{\beta_2}\, X_2 + \ldots + \widehat{\beta_n} X_n$

Prueba $\hat{Y} = \exp(\hat{f}(x))$ $\min(\hat{Y} - Y)$

Predicción $\hat{Y} = \exp(\hat{f}(x))$

$$y = \log(\lambda)$$

Liga Canónica
Link fucntion

Regresión Poisson

Interpretación

Ya que nuestro modelo tiene la siguiente forma:

$$\log(\lambda) = \beta_0 + \beta_1 x + \epsilon$$

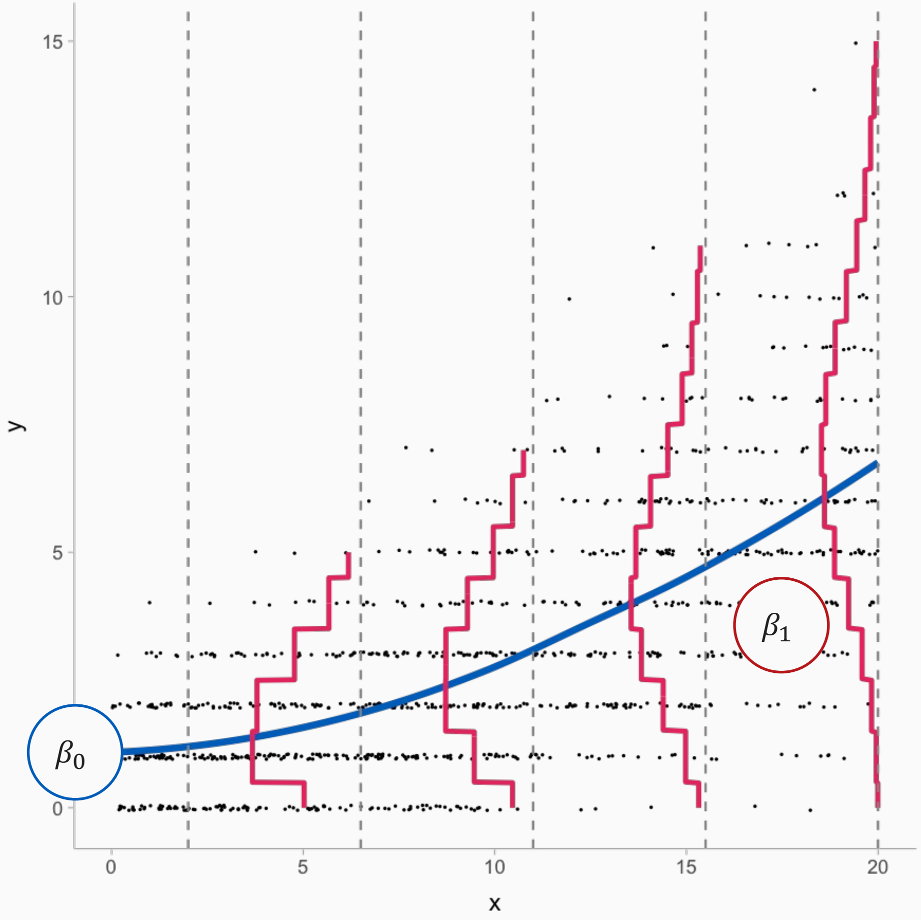
resulta más facil interpretarlo si lo regresamos a su forma original

$$\lambda = \mathrm{e}^{\beta_0 + \beta_1 x + \epsilon} = e^{\beta_0} e^{\beta_1 x} e^{\epsilon}$$

de esta forma:

e^{β_0}: es el conteo base cuando $x = 0$

e^{β_1}: es el factor multiplicativo en el que cambia el conteo ante un cambio en x (relación de tasa de incidencia)



Criterios de Información

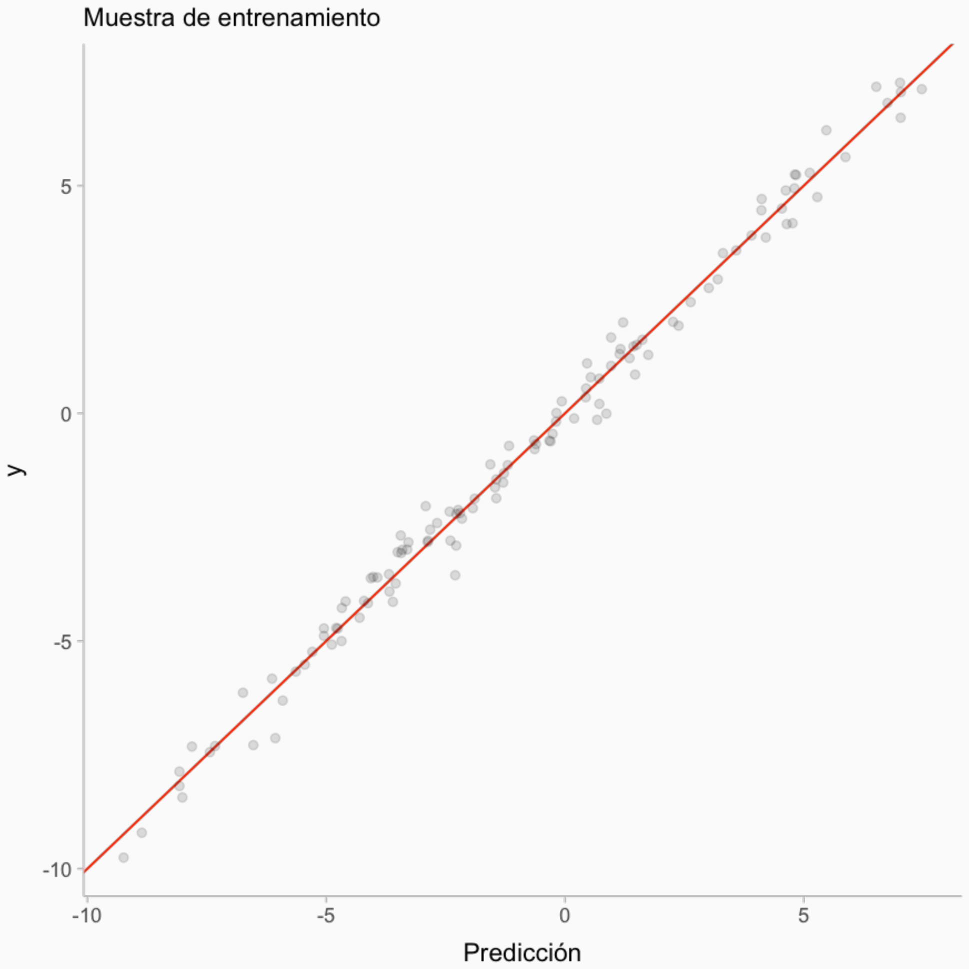
Para la selección de nuestro λ utilizaremos los criterios de Akaike (AIC) y Schwarz (BIC). Ambos introducen una penalización en función del número de parámetros en el modelo, evitando así un sobre ajuste excesivo derivado de introducir muchas variables.

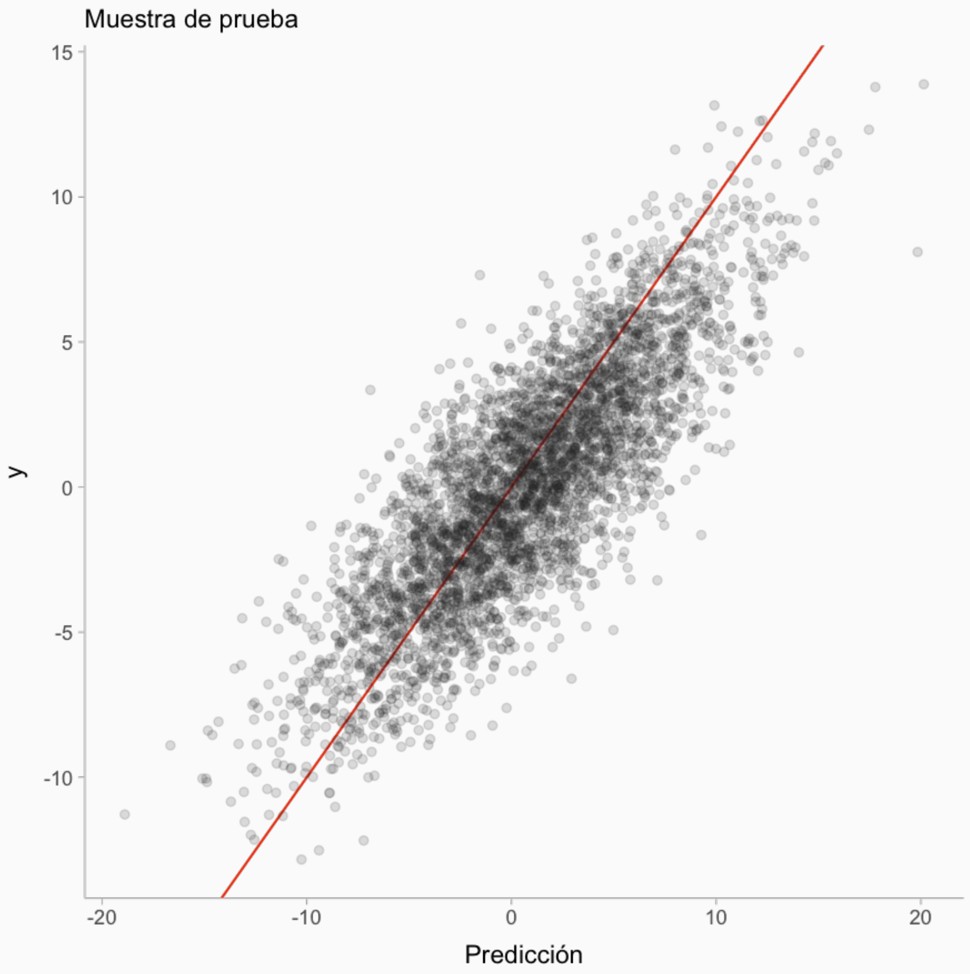
$$AIC = n * \log(\epsilon^2) + 2 * Grados\ de\ Libertad$$

$$BIC = n * \log(\epsilon^2) + 2 * Grados\ de\ Libertad * \log(n)$$

Un ejemplo

Varianza de los Parámetros





Regularización

Cuando los coeficientes presentan problemas de varianaza podemos imponer restricciones en sus rangos

$$\min\big(f(\hat{y} - y)\big) + \lambda \sum_{j=1}^{p} h(\beta_j)$$

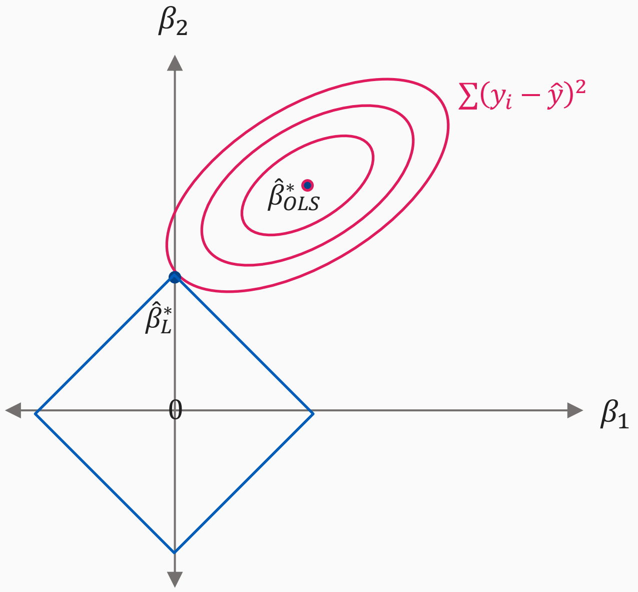
Definimos un valor de $\lambda > 0$ para mejorar nuestro modelo en función de su capacidad predictiva:

- Incrementar el valor de λ implica aumentar el sesgo de nuestro modelo pero reducir su varianza.
- Dado lo anterior, se asume que los regresores pueden ser escalados (estandarizados a z-scores)

Regularización LASSO

Regularización L1

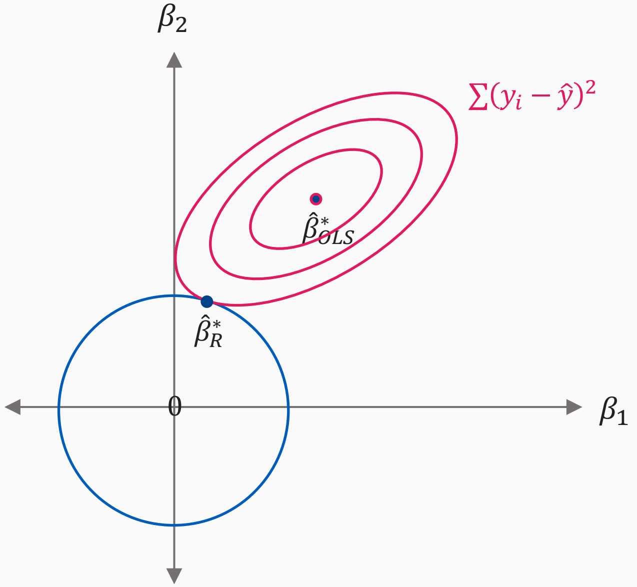
$$L_{LASSO}(\hat{\beta}) = L_1(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2 + \lambda\sum_{j=1}^{p}|\hat{\beta}_j| = \left\|y_i - x_i\hat{\beta}\right\|^2 + \lambda\left|\hat{\beta}\right|$$



Regularización RIDGE

Regularización L2

$$L_{Ridge}(\hat{\beta}) = L_2(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2 + \lambda\sum_{j=1}^{p}\hat{\beta}_j^2 = \left|\left|y_i - x_i\hat{\beta}\right|\right|^2 + \lambda\left|\left|\hat{\beta}\right|\right|^2$$



$$\hat{\beta}_R{}^* = (x^T x + \lambda I)^{-1} x^T y$$

Regularización RIDGE Heterocedastica

En lugar de resolver el problema de heteroscedasticidad igualando las varianzas de todos los predictores a través de escalarlos, podemos usarlos como pesos en el proceso de estimación.

$$L_{RH}(\hat{\beta}) = L_{2H}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2 + \lambda\sum_{j=1}^{p} w_j\hat{\beta}_j^2 = \left|\left|y_i - x_i\hat{\beta}\right|\right|^2 + \lambda\left|\left|w_j\hat{\beta}\right|\right|^2$$

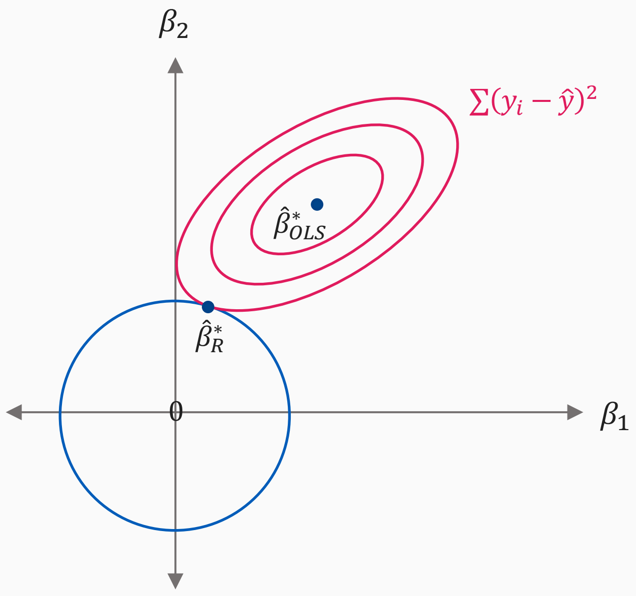
Simplemente hacemos una regregresión univariada para cada uno de los predictors. A partir de ello, extraemos la varianza de cada coeficiente $\hat{\sigma}_{\beta}^2$ y la utilizamos como ponderador.

$$\frac{1}{\hat{\sigma}_{\beta}^2} = w$$

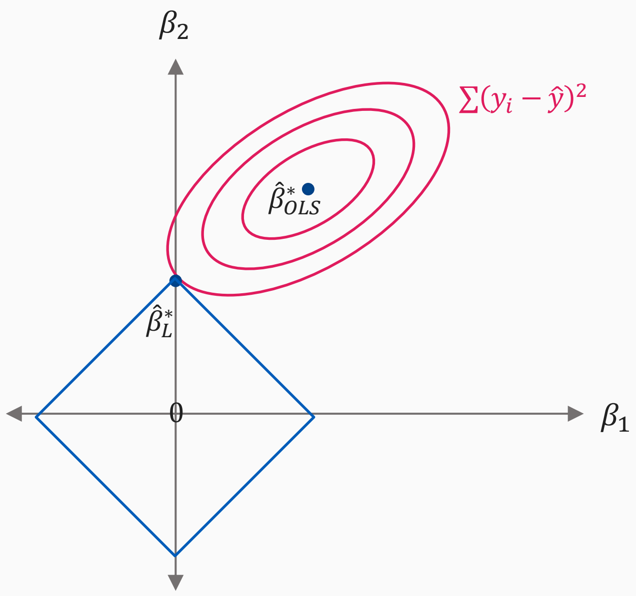
Los coeficientes con varianzas más altas (mayor incertidumbre) son por lo tanto más penalizados

RIDGE vs LASSO

- No restringido al origen.
- Permite atender multicolinealidad y heterocedasticidad encogiendo los coeficientes.

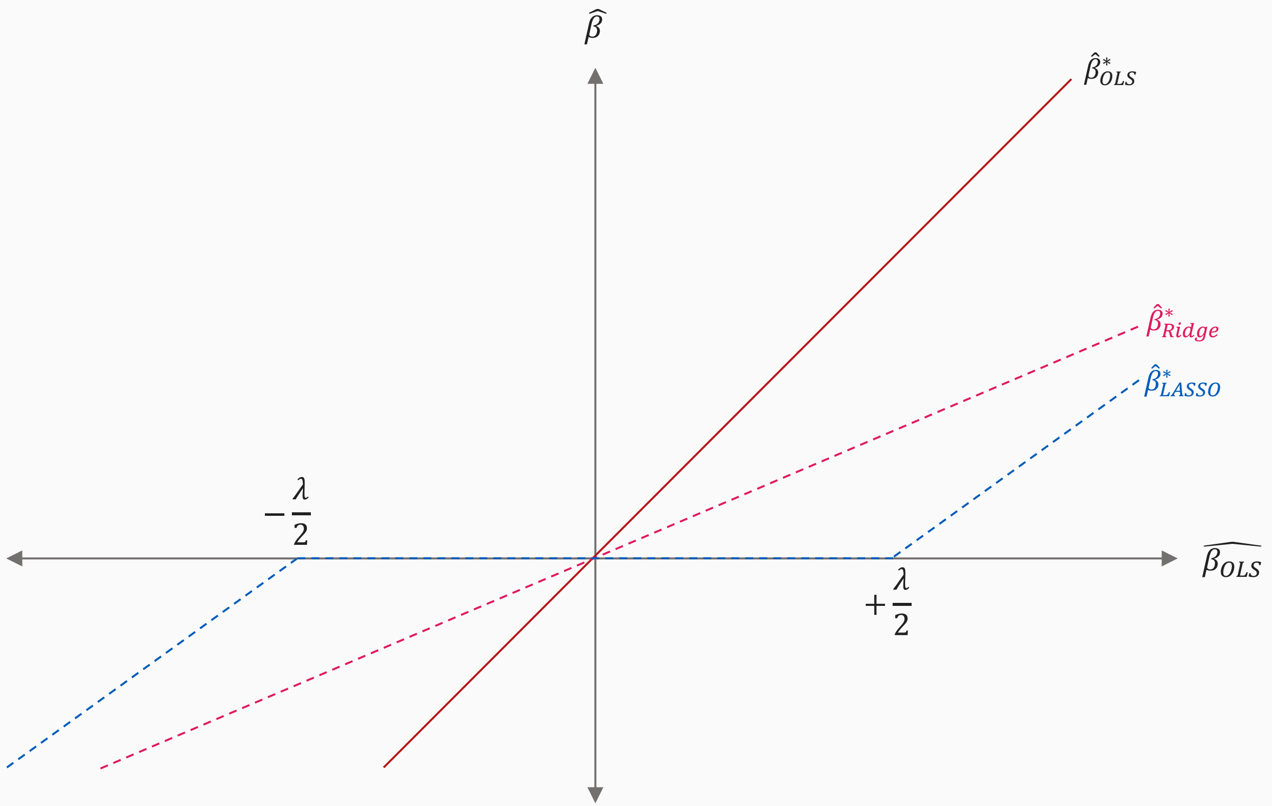


- Permite excluir variables del modelo.
- Reduce la varianza al desechar variables ruidosas o redundantes.



Resumen

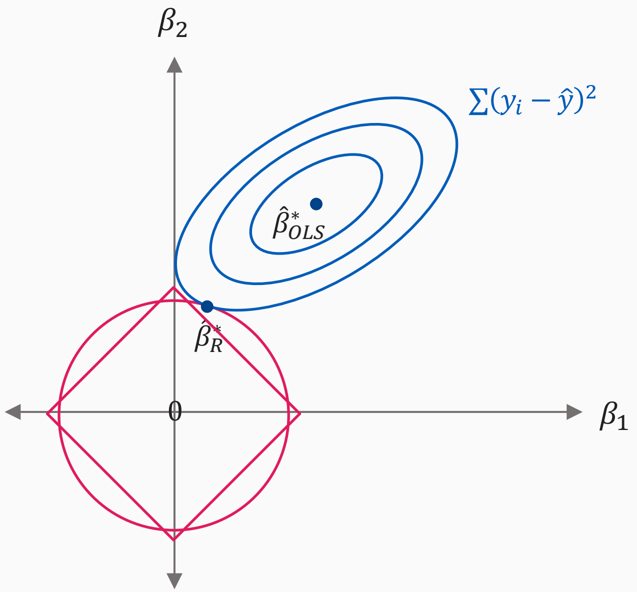
Estimadores Regularizados vs. Estimadores OLS



Tecnológico de Monterrey
Escuela de Ciencias Sociales
y Gobierno

Regularización Mixta

$$L_{Mix}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2 + \lambda\left(\alpha\sum_{j=1}^{p}\hat{\beta}_j^2 + (1-\alpha)\sum_{j=1}^{p}|\hat{\beta}_j|\right) = \left\|y_i - x_i\hat{\beta}\right\|^2 + \lambda\left(\alpha\left\|\hat{\beta}\right\|^2 + (1-\alpha)|\hat{\beta}|\right)$$

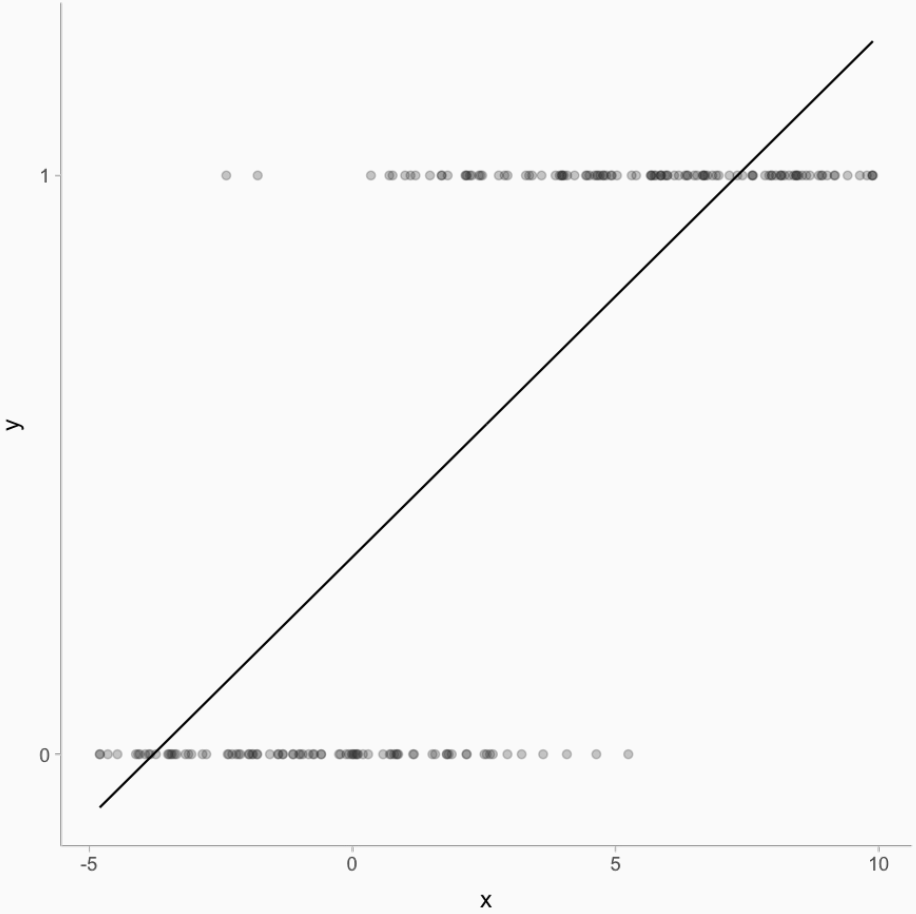


Un ejemplo

Preguntas no normales

- ¿Con qué probabilidad un partido puede ganar cierto distrito electoral?
- ¿Cuál es la probabilidad de que una persona reincida en un acto delectivo?
- ¿Cómo podemos detectar actos de corrupción o fraudes?

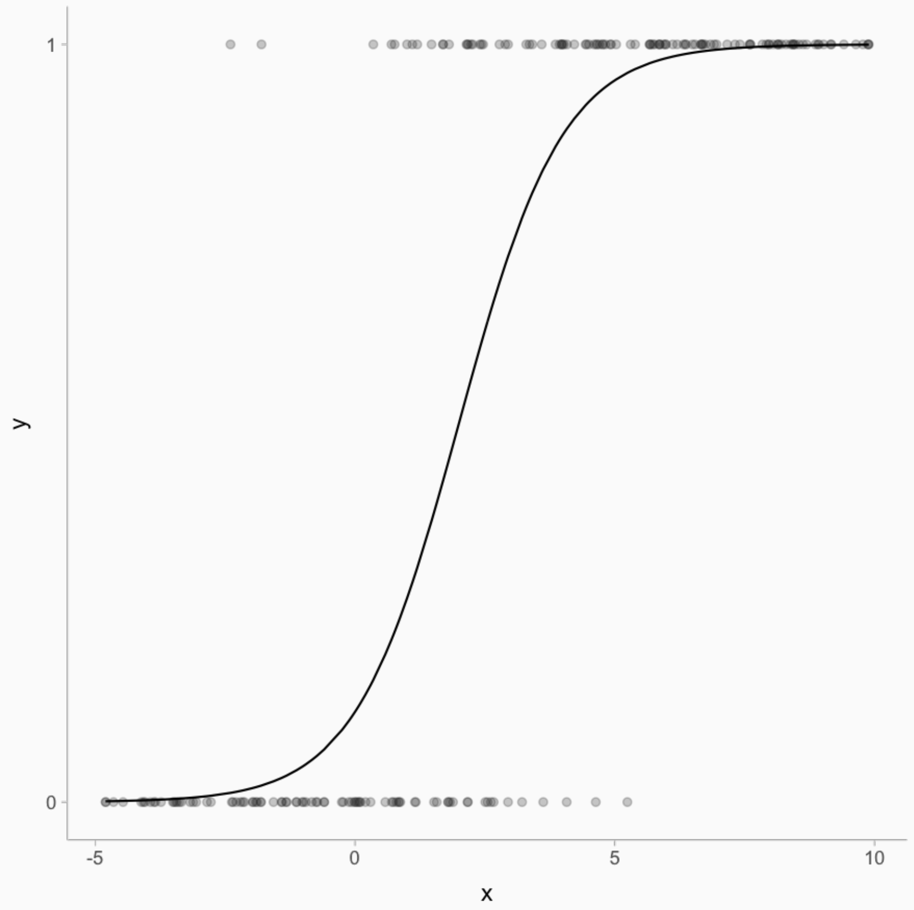
Regresión Gaussiana



Preguntas Categóricas

- ¿Con qué probabilidad un partido puede ganar cierto distrito electoral?
- ¿Cuál es la probabilidad de que una persona reincida en un acto delectivo?
- ¿Cómo podemos detectar actos de corrupción o fraudes?

Regresión Lógistica



Modelos Lineales Generalizados

Fénomeno observado

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Entrenamiento

$$\log\left(\frac{p(\hat{x})}{1-p(\hat{x})}\right) = \widehat{\beta_0} + \widehat{\beta_1}\, X_1 + \widehat{\beta_2}\, X_2 + \ldots + \widehat{\beta_n} X_n$$

Prueba

$\hat{Y} = \hat{f}(x)$ — $\min\left(\hat{Y} - Y\right)$

Predicción

$\hat{Y} = \hat{f}(x)$

$$y = \log\left(\frac{p(x)}{1-p(x)}\right)$$

Liga Logítica
Logistic fucntion

Liga Logística

Podemos definir la probalidad de $y = 1$ como

$$P(Y = 1 \mid X = x) = \frac{e^{x\beta}}{1 + e^{x\beta}}$$

y su complemento como:

$$P(Y = 0 \mid X = x) = 1 - P(Y = 1 \mid X = x) = \frac{1}{1 + e^{x\beta}}$$

Si calculamos la razón de probabilidades obtenemos algo así

$$\frac{P(Y = 1)}{P(Y = 0)} = \frac{p(x)}{1 - p(x)} = e^{\beta x}$$

La log-razón de probabilidades es lineal en x. A esta función se le conoce como liga logística

$$y = \log\left(\frac{p(x)}{1 - p(x)}\right) = \beta x$$

Regresión Logística

Supuestos

1. Respuesta categórica

   La variable respuesta es catagórica o suma de respuestas catagórica.

2. Independencia

   Las observaciones deben ser independientes entre sí.

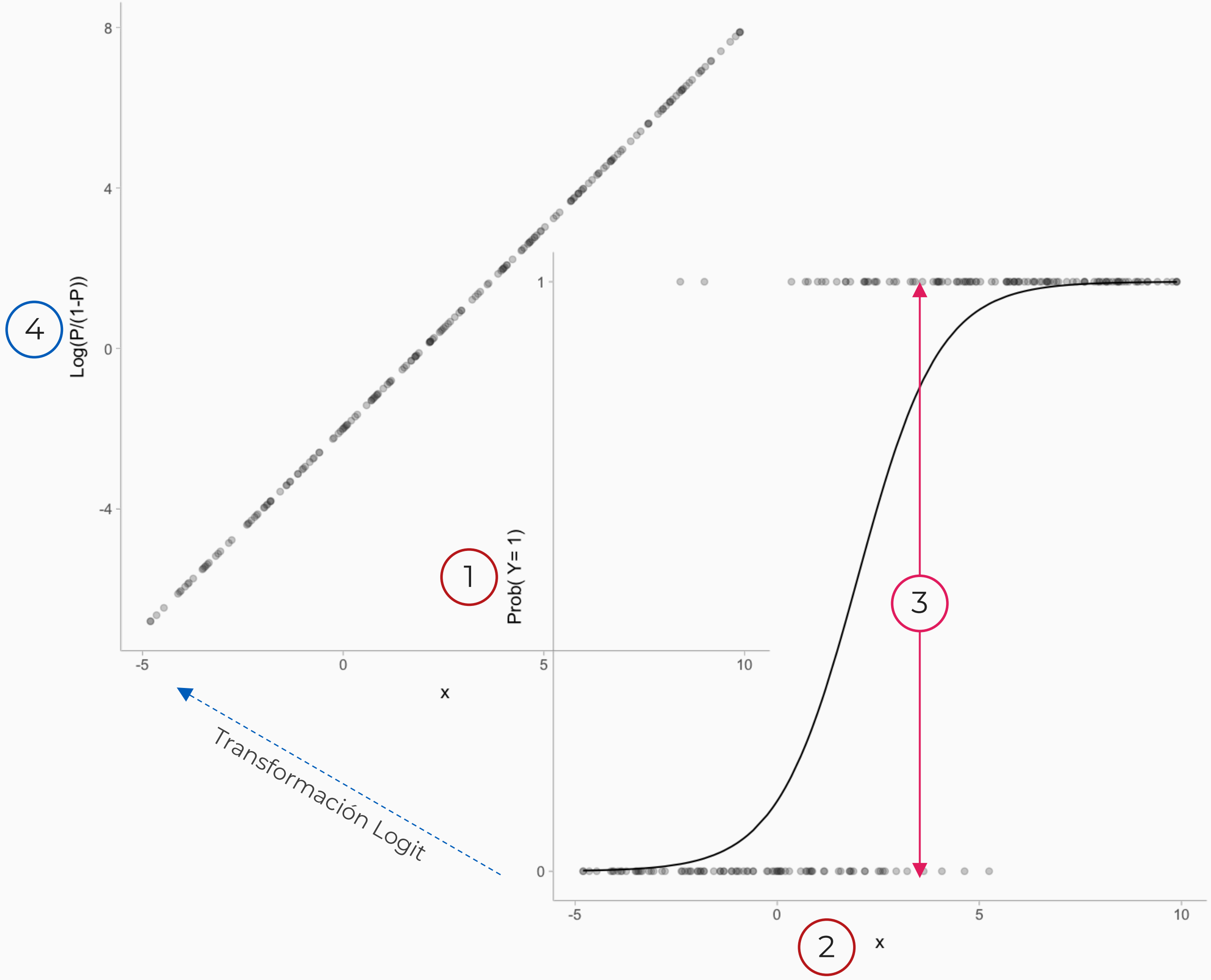
3. Varianza estructurada

   Por definición la varianza de una variable binomial está dada por $np(1-p)$

4. Linealidad

   La log-razón de probabilidades es lineal en x.

$$y = \log\left(\frac{p(x)}{1-p(x)}\right) = \beta x$$



Un ejemplo

Matriz de Confusión

Clasificado

Observados		Positivo	Negativo
	Positivo	Verdaderos Positivos (VP)	Falsos Negativos (FN) (Error tipo II)
	Negativo	Falsos Positivos (FP) (Error tipo I)	Verdaderos Negativos (VN)

Métricas para Clasificación

Exactitud

- $\frac{VP+VN}{VP+VN+FP+FN}$

Precisión

- $\frac{VP}{VP+FP}$

Sensibilidad

(Tasa de Verdaderos Positivos)

- $\frac{VP}{VP+FN}$

Tasa de Falsos Positivos

(Error Tipo I)

- $\frac{FP}{VP+FP}$

Especificidad

(Tasa de Verdaderos Negativos)

- $\frac{VN}{VN+FP}$

Tasa de Falsos Negativos

(Error Tipo II)

- $\frac{FN}{VN+FN}$

Puntaje F-Beta

$$F_{\beta} = (1+\beta^2) * \frac{precisión * sensibilidad}{(\beta^2 * precisión) + sensibilidad)}$$

Al elegir beta en su puntaje F-beta, cuanto más nos importe la sensibilidad sobre la precisión, debemos elegir una beta más alta.

$$F_1 = 2 * \frac{precisión * sensibilidad}{precisión + sensibilidad}$$

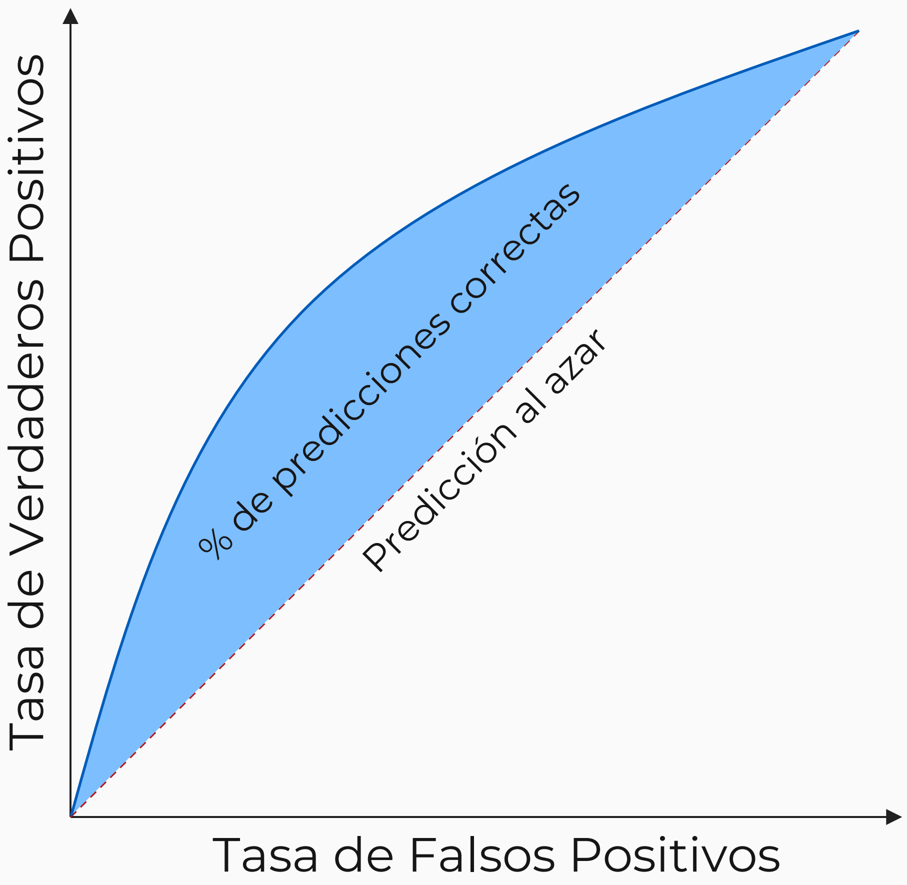
Por ejemplo, con la puntuación F1, nos preocupamos por igual por la sensibilidad y la precisión.

$$F_2 = 5 * \frac{precisión * sensibilidad}{(4 * precisión) + sensibilidad}$$

Con la puntuación F2, la sensibilidad es el doble de importante para nosotros.

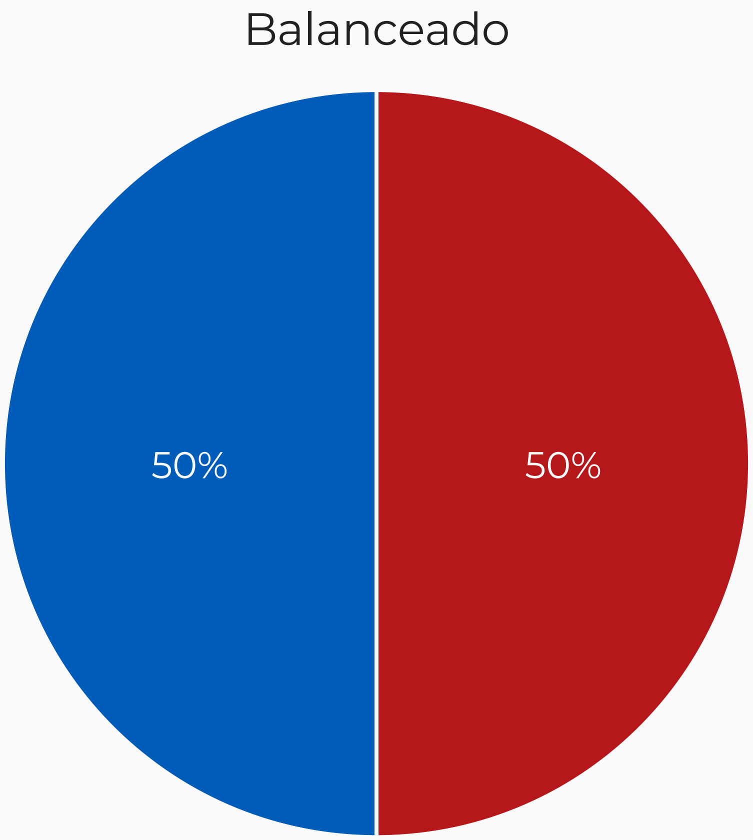
$$\dots$$

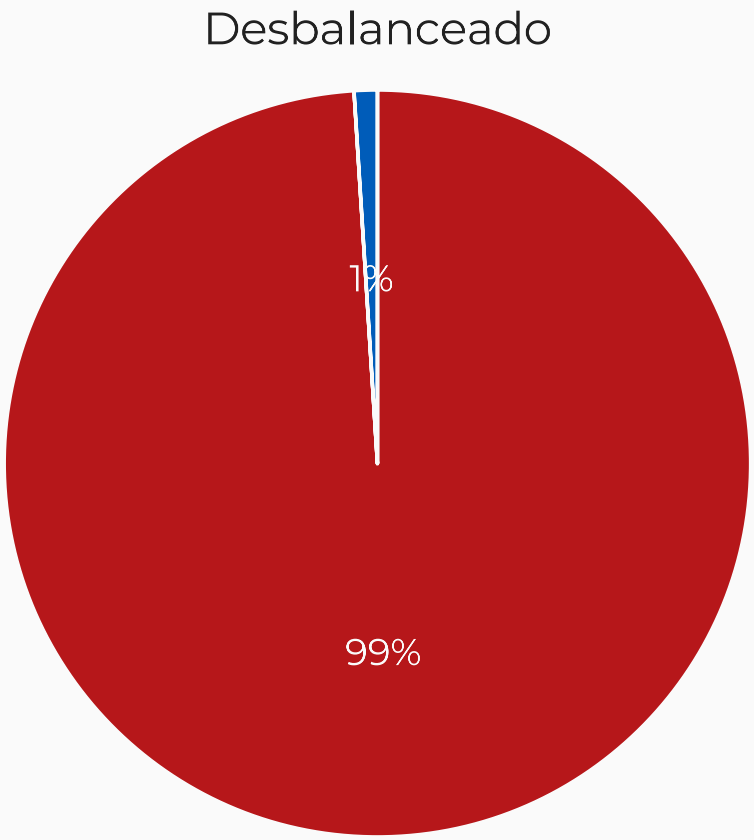
Area Bajo la Curva Característica Operativa del Receptor



Tecnológico de Monterrey
Escuela de Ciencias Sociales
y Gobierno

Tipos de Paneles





Un ejemplo