UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

Regresión y métodos predictivos

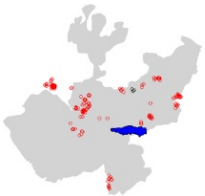
Alan Reyes-Figueroa

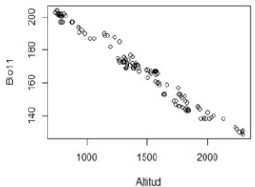
Introducción a la Ciencia de Datos (Aula 32) 13.mayo.2021

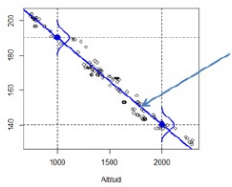
Regresión Lineal

Ejemplo motivacional:

Se mide la altura (X) y la temperatura promedio del trimestre más frío del año (Y), en varias localidades geográficas:







$$Promedio\ Bio11 = a_0 + a_1 Altitud$$

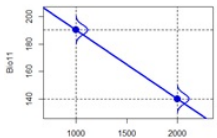
Hay una relación lineal entre Y y X. Matemáticamente, podemos escribir

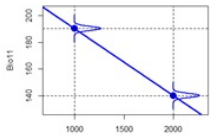
$$Y = b_0 + b_1 X + \varepsilon,$$

donde $b_0, b_1 \in \mathbb{R}$, y ε es una variable aleatoria que codifica la incertidumbre intrínseca del fenómeno de interés y nuestra ignorancia.

Regresión Lineal

Esta v.a. ε, típicamente satisface $\mathbb{E}(\varepsilon) = 0$. Si su varianza es $Var(\varepsilon) = \sigma$,





Efecto de la magnitud σ: entre mayor, más incertidumbre y más difícil predicir y estimar.

entonces σ codifica la variabilidad de esta incertidumbre.

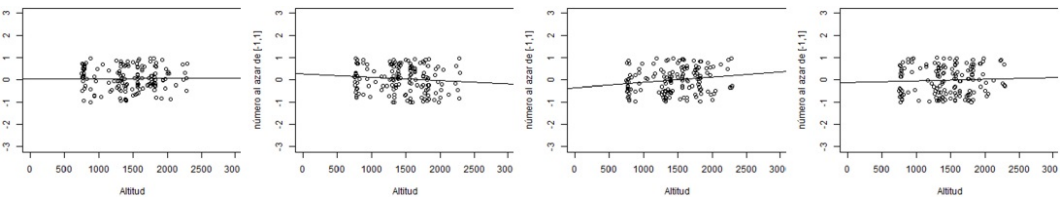
Dada una muestra $\{(X_i, Y_i)\}$, estimamos los parámetros vía máxima verosimilitud: Si $\theta = (b_0, b_1, \sigma)$, entonces

$$\mathcal{L}(\theta) = \mathbb{P}(\{(X_i, Y_i) = (\mathbf{x}_i, y_i)\}) = k\mathbb{P}(\{\varepsilon_i = y_i - b_0 - b_1\mathbf{x}_i\}).$$

Regresión

Muchas veces, estaremos sólo interesados en contestar preguntas como: ¿X afecta a Y?
En nuestro modelo esto equivalente a preguntar: ¿$b_1 \neq 0$?

Cuidado! Las estimaciones son variables aleatorias.
Ejemplo:
Supongamos que $Y = 0 + 0 \cdot X + \varepsilon$. Para diferentes muestras obtenemos:



Regresión Lineal

Tenemos que recurrir a pruebas de hipótesis.

A veces, tendremos interés especial en preguntas como: dado $X = \mathbf{x}$, ¿qué podemos decir sobre Y?
$Y \mid X = \mathbf{x}$ es una variable aleatoria. Podemos calcular su distribución y su promedio:

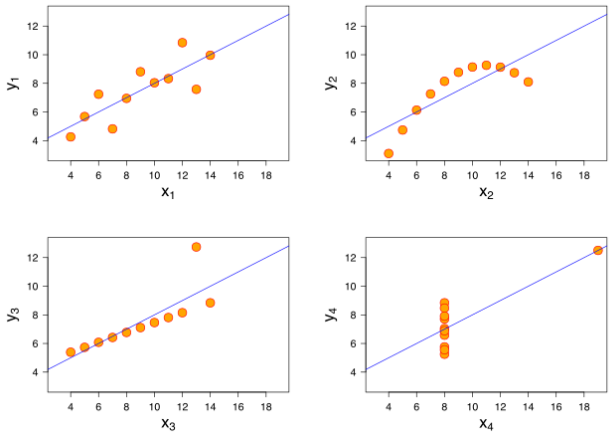
$$\boxed{\mathbb{E}(Y \mid X = \mathbf{x}) = b_0 + b_1\mathbf{x}.}$$

(la recta o modelo regresor estima el promedio de $Y \mid X = \mathbf{x}$). Si queremos minimizar $\mathbb{E}(Y - \widehat{Y}(\mathbf{x}))^2$ una propuesta es tomar

$$\widehat{Y}(\mathbf{x}) = \mathbb{E}(\widehat{Y \mid X = \mathbf{x}}) = \widehat{b}_0 + \widehat{b}_1\mathbf{x},$$

$\widehat{Y}$ es de nuevo es una v.a. Dependiendo de supuestos que podemos hacer sobre los datos, obtenemos información sobre la distribución de $\widehat{Y}(\mathbf{x})$.

Regresión Lineal



Cuarteto de Anscombe.

Regresión Lineal

El cuarteto de Ascome es una familia de varios conjuntos de datos con estadísticas similares, pero diferentes distribuciones (lo importante son las distribuciónes, más que los estadísticos que resumen).

Propiedad / Estadístico	Valor	Exactitud
Meadia de x	9	exacto
Varianza muestra de x	11	exacto
Media de y	7.5	$\pm$ 0.001
Varianza muestra de y	4.125	$\pm$ 0.003
Correlación entre x, y	0.816	$\pm$ 0.0001
Recta de regresión lineal	$y = 3.0 + 0.5x$	$\pm$ 0.001 y $\pm$ 0.0001, resp.
Coeficiente R^2	0.67	$\pm$ 0.001

Regresión Lineal (OLS)

Sea $\mathbf{X} = (X_1, \ldots, X_{d-1}) \in \mathbb{R}^{d-1}$ vector aleatorio, Y variable aleatoria. Dada una muestra $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, consideramos el modelo

$$Y = \beta_0 + \sum_{j=1}^{d-1} \beta_j X_j + \varepsilon, = \beta_0 + \beta^T X + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

donde $\beta^T = (\beta_1, \ldots, \beta_{d-1}) \in \mathbb{R}^{d-1}$.

Hagamos la inclusión de $\mathbb{R}^{d-1}$ a $\mathbb{R}^d$ por $\mathbf{x} \to (1, \mathbf{x})$, y definamos $\mathbb{X} = (\mathbf{x}_{jk}) \in \mathbb{R}^{n \times d}$ la matriz de datos, $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$, $\beta = (\beta_0, \beta_1, \ldots, \beta_{d-1}) \in \mathbb{R}^d$, y $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \in \mathbb{R}^n$. Entonces el modelo (1) se escribe como

$$\mathbf{y} = \mathbb{X}\beta + \varepsilon, \quad \text{con } \varepsilon \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 I). \tag{2}$$

Regresión Lineal

Como $\varepsilon = \mathbf{y} - \mathbb{X}\beta$, entonces la verosimilitud para ε es

$$\mathcal{L}(\varepsilon) = \mathbb{P}((\varepsilon_1, \ldots, \varepsilon_n) = \varepsilon) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{||\mathbf{y} - \mathbb{X}\beta||^2}{2\sigma^2}\right).$$

Obs! Maximizar esta verosimilitud equivale a minimizar la función de costo $\varepsilon = ||\mathbf{y} - \mathbb{X}\beta||^2$.

Solución: Si $J = ||\mathbf{y} - \mathbb{X}\beta||^2 = \langle \mathbf{y} - \mathbb{X}\beta, \mathbf{y} - \mathbb{X}\beta \rangle$, entonces

$$\nabla_\beta J = 2\langle -\mathbb{X}, \mathbf{y} - \mathbb{X}\beta \rangle = -2\mathbb{X}^T(\mathbf{y} - \mathbb{X}\beta) = -2\mathbb{X}^T\mathbf{y} + 2\mathbb{X}^T\mathbb{X}\beta = 0.$$

Obtenemos las **ecuaciones normales**

$$\boxed{(\mathbb{X}^T\mathbb{X})\beta = \mathbb{X}^T\mathbf{y}.} \quad \Rightarrow \quad \widehat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{y}.$$

Regresión Lineal

De las ecuaciones normales

$$\begin{aligned}\widehat{\beta} &= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{y} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T(\mathbb{X}\beta+\varepsilon) = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}\beta + (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\varepsilon \\ &= \beta + (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\varepsilon.\end{aligned}$$

Entonces, como una combinación lineal de variables normales es normal

$$\widehat{\beta} \sim \mathcal{N}(\beta, \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}),$$

y

$$\widehat{y}(\mathbf{x}) = \mathbf{x}^T\widehat{\beta} \sim \mathcal{N}(\mathbf{x}^T\beta, \mathbf{x}^T\sigma^2(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{x}).$$

Lo anterior permite hacer pruebas de hipótesis $H_0 : \beta_j = 0$, construir intervalos de confianza para β, o construir intervalos de predicción para $Y \mid X = \mathbf{x}$.

Regresión Lineal

Luego de estimar $\widehat{\beta}$, los valores predichos por el modelos de regresión son

$$\widehat{\mathbf{y}} = \mathbb{X}\widehat{\beta} = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{y} = P\mathbf{y}. \tag{3}$$

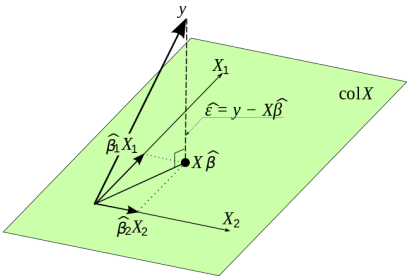
Aquí, $P = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$ es una matriz de proyección: es la matriz de proyección sobre el espacio V generado por las columnas de $\mathbb{X}$. Esta matriz La matriz $M = I - P$ es la matriz de proyección en el espacio ortogonal a V. Ambas matrices P y M son simétricas e idempotentes (lo que significa que $P^2 = P$ y $M^2 = M$), y satisfacen

$$P\mathbb{X} = \mathbb{X} \quad \text{y} \quad M\mathbb{X} = \mathbf{o}.$$

La matriz M crea los residuos de la regresión:

$$\widehat{\varepsilon} = \mathbf{y} - \widehat{\mathbf{y}} = \mathbf{y} - \mathbb{X}\widehat{\beta} = \mathbf{y} - P\mathbf{y} = (I - P)\mathbf{y} = M\mathbf{y} = M(\mathbb{X}\beta + \varepsilon) = M\mathbb{X}\beta + M\varepsilon = M\varepsilon.$$

Regresión Lineal



Proyecciones $P\mathbb{X}$ y $M\mathbb{X}$.

Usando estos residuales, podemos estimar el valor de σ^2 mediante la estadística χ^2:

$$
\begin{aligned}
s^2 &= \frac{\widehat{\varepsilon}^T\widehat{\varepsilon}}{n-d} = \frac{(M\mathbf{y})^T(M\mathbf{y})}{n-d} = \frac{\mathbf{y}^T M^T M\mathbf{y}}{n-d} \\
&= \frac{\mathbf{y}^T M\mathbf{y}}{n-d} = \frac{S(\widehat{\beta})}{n-d}, \\
\widehat{\sigma}^2 &= \frac{n-d}{n} s^2.
\end{aligned}
$$

El denominador $n - d$ es el número de **grados de libertad**. s^2 es el estimador OLS para σ^2, mientras que $\widehat{\sigma}^2$ es el estimador máximo verosímil de σ^2. (son similares, pero el primero es insesgado, mientras que el segundo no). s es llamado el **error estándar** de la regresión.

Regresión Lineal

Bondad de ajuste: Evaluamos el ajuste de la regresión OLS comparando cuánto se puede reducir la variación inicial en la muestra, al hacer la regresión sobre X.

El **coeficiente de determinación** R^2 se define como la relación entre la varianza "explicada" y la varianza "total" de la variable dependiente Y, en los casos en que la suma de cuadrados de la regresión es igual a la suma de cuadrados de los residuos:

$$R^2 = \frac{\sum_i (\widehat{y}_i - \overline{y})^2}{\sum_i (y_i - \overline{y})^2} = \frac{\mathbf{y}^T P^T J P \mathbf{y}}{\mathbf{y}^T J \mathbf{y}} = 1 - \frac{\mathbf{y}^T M \mathbf{y}}{\mathbf{y}^T J \mathbf{y}} = 1 - \frac{RSS}{TSS},$$

donde TSS es la suma total de cuadrados de la variable dependiente, $J = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$.

Regresión Lineal con términos generales

Bajo el modelo de regresión lineal, podemos incluir modelos más generales que (1)

$$Y = \beta_0 + \sum_{j=1}^{d-1} \beta_j X_j + \varepsilon, = b_0 + \beta^T X + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Por ejemplo, podemos construir nuevas variables aleatorias, que sean potencias o productos de las variables X_j ya incluidas, o funciones no lineales de éstas.
En el caso en que incluimos potencias, tenemos un **modelo de regresión polinomial**. Por ejemplo, para el caso de $d-1$ variables

$$Y = \beta_0 + \sum_{j=1}^{d-1} \beta_j X_j + \sum_{j=1}^{d-1} \gamma_j X_j^2 + \sum_{j \neq k} \alpha_{jk} X_j X_k + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2). \tag{4}$$

Regresión Lineal con términos generales

Similarmente, podemos incluir funciones no lineales de las variables X_j, digamos $f_\ell(X_j)$, pero consideradas como una combinación lineal. Este modelo se llama comunmente un modelo de **regresión lineal con términos no lineales**.

$$Y = \beta_0 + \sum_{j=1}^{d-1} \beta_j X_j + \sum_{\ell} \gamma_\ell f_\ell(X_1, X_2, \ldots, X_{d-1}) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2). \tag{5}$$

Obs! Cualquiera de los dos casos anteriores se trata de forma similar como la regresión lineal ordinaria, tomando en consideración que al introducir términos $X_{jk} = X_j X_k$ o términos no lineales $f_\ell(X)$, podemos estar introduciendo también correlaciones entre estas nuevas variables.

Supuestos Básicos

El modelo clásico de regresión lineal (OLS), asume varios supuestos:

- Muestra finita $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, n finito.
- **Exogeneidad**: $\mathbb{E}(\varepsilon \mid \mathbb{X}) = 0$. En consecuencia $\mathbb{E}(\varepsilon) = 0$ y $\mathbb{E}(\mathbf{x}^T\varepsilon) = 0$.
- **Independencia lineal**: Los regresores en $\mathbb{X}$ son l.i., esto significa que $\mathbb{P}(\text{rank}\,\mathbb{X} = d) = 1$.
- **Segundos momentos finitos**: La matrix $Q_{xx} = \frac{1}{n}\mathbb{X}^T\mathbb{X}$ es finita y positiva semidefinida.
- **Errores esféricos**: $Var(\varepsilon) = \sigma^2 I$. Es común separar este supusto en dos partes:
  - **Homoscedasticidad**: $\mathbb{E}(\varepsilon_j^2 \mid \mathbb{X}) = \sigma^2$, $\forall j$.
  - **No autocorrelación**: $\mathbb{E}(\varepsilon_j\varepsilon_k \mid \mathbb{X}) = 0$, $\forall j \neq k$.
- **Normalidad**: $\mathbf{E} \mid \mathbb{X} \sim \mathcal{N}(0, \sigma^2 I)$.

Supuestos Básicos

En algunas aplicaciones, se asume que los datos son independientes e idénticamente distribuidos (i.i.d.). Esto significa que las observaciones son tomadas de una muestra aleatoria que cumple todos los supuestos anteriores.

Este enfoque permite establecer resultados asintóticos (cuando el tamaño de la muestra $n \to \infty$). Que se entiende como una posibilidad teórica de añadir observaciones independientes.

Los supuestos en este caso son:

- **Observaciones i.i.d**: $(\mathbf{x}_i, y_i)$ es indep. de y tiene la misma distribución que $(\mathbf{x}_j, y_j)$.
- **No multicolinealidad**: $Q_{xx} = \frac{1}{n}\mathbf{x}_i\mathbf{x}_i^T$ es definida positiva, $\forall i$.
- **Exogeneidad**: $\mathbb{E}(\varepsilon_i \mid \mathbf{x}_i) = 0$, $\forall i$.
- **Homoscedasticidad**: $\mathbf{v}(\varepsilon_i \mid \mathbf{x}_i) = \sigma^2$, $\forall i$.

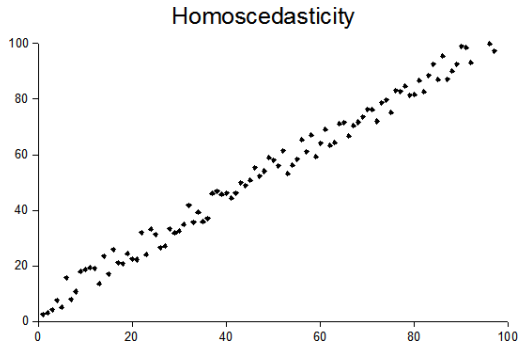
Supuestos Básicos

Finalmente, en el caso de series de tiempo, se asume que el proceso estocástico $\{(\mathbf{x}_i, y_i)\}$ es estacionario y ergódico. Si $\{(\mathbf{x}_i, y_i)\}$ no es estacionario, los resultados del modelo OLS son espúrios, a menos que $\{(\mathbf{x}_i, y_i)\}$ sea co-integrado.

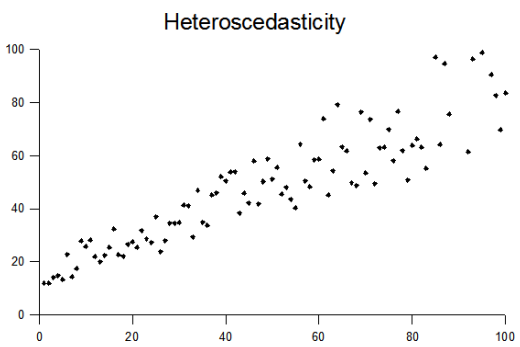
Los supuestos en este caso son:

- **Estacionariedad y ergodicidad**: $\{(\mathbf{x}_i, y_i)\}$ es estacionario y ergódico.
- **Regresores predeterminados**: $\mathbb{E}(\mathbf{x}_i \, \varepsilon_i) = 0, \forall i$.
- **No multicolinealidad**: $Q_{xx} = \mathbb{E}(\mathbf{x}_i \mathbf{x}_i^T)$ es de rango completo y positiva definida, $\forall i$.
- **Martingala con momentos finitos**: $\{\mathbf{x}_i \, \varepsilon_i\}$ es una martingala, con una matriz de segundos momentos $Q_{xx\varepsilon^2} = \mathbb{E}(\varepsilon_i^2 \, \mathbf{x}_i \mathbf{x}_i^T)$ finita.

Supuestos Básicos

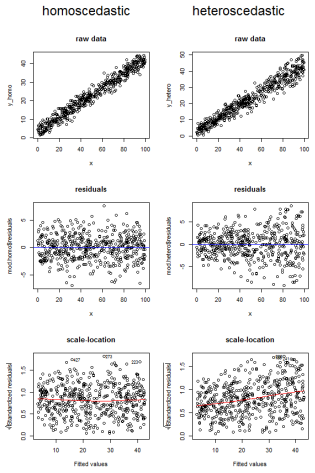


(a) Homoscedasticidad



(b) Heteroscedasticidad.

Supuestos Básicos



UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

Prueba de Hipótesis

Para determinar si $\beta_i =$ o ó no, se hace una prueba de hipótesis.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.988
Model:	OLS	Adj. R-squared:	0.988
Method:	Least Squares	F-statistic:	3959.
Date:	Thu, 13 May 2021	Prob (F-statistic):	1.05e-93
Time:	15:31:54	Log-Likelihood:	-142.71
No. Observations:	100	AIC:	291.4
Df Residuals:	97	BIC:	299.2
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.6302	0.301	5.415	0.000	1.033	2.228
x1	-0.1378	0.139	-0.991	0.324	-0.414	0.138
x2	3.1519	0.135	23.409	0.000	2.885	3.419

Omnibus:	1.426	Durbin-Watson:	2.235
Prob(Omnibus):	0.490	Jarque-Bera (JB):	1.075
Skew:	0.249	Prob(JB):	0.584
Kurtosis:	3.099	Cond. No.	24.6

Prueba de Hipótesis

Cada vez que se hace una regresión, queremos evaluar la hipótesis de si $\beta_i = 0$ (β_i el parámetro i del modelo). Esto se hace mediante una prueba de hipótesis sobre β_i. Consideramos la prueba $H_0 : \beta_i = 0$, $H_1 : \beta_i \neq 0$. Si se falla en rechazar H_0, esto significa que β_i no es significativo (y puede no usarse en el modelo). Rechazar significa que β_1 es significativo, y la variable X_i afecta a la variable Y.

La prueba se hace calculando la estadística $t = \frac{\widehat{\beta}_i}{se(\widehat{\beta}_i)}$, con $se(\widehat{\beta}_i)$ el error estándar del estimador. Para efectuar la prueba, derivamos el p-valor para dicha $\widehat{\beta}_i$.

- Calculamos el estadístico t para el punto medio $\mathbb{E}(\beta_i) = 0$ (de acuerdo a la hipótesis nula).
- Calculamos el p-valor de la probabilidad acumulada para el estadístico t, según la distribucipon t, con $n - d$ grados de libertad, y nivel de significancia α.
- Decidimos rechazar/no rechazar H_0 en función del p-valor y α:

$$\text{rechazamos } H_0 \iff p\text{-valor} < \alpha.$$