Proyecto II
Métodos Estadísticos
para la Investigación en VIH/SIDA

Yanink Neried Caro Vega
Lizbeth Naranjo Albarrán

Facultad de Ciencias, UNAM

August 26, 2022

Índice

1 Regresión lineal simple

Un análisis de regresión es una técnica estadística para investigar y modelar la relación entre una variable (variable respuesta o dependiente) y otra u otras variables (variables explicativas o independientes).

El modelo de regresión lineal simple es un modelo con sólo una variable explicativa x (o independiente) que está relacionada con una variable respuesta y (o dependiente) a través de una línea recta. En la figura -1 se observa esta situación con el modelo ajustado a los datos. Sin embargo, es claro que pueden existir muchas rectas que ajusten estos datos. ¿Cómo garantizamos que obtenemos la mejor recta ajustada? ¿A qué se refiere con mejor ajuste?

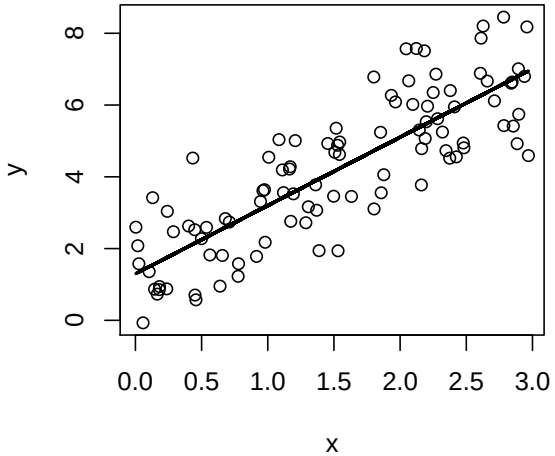

Figura 1: Modelo de regresión lineal simple.

El modelo de regresión lineal se puede escribir como:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X \qquad \text{(modelo determinista)} \\ Y &= \beta_0 + \beta_1 X + \varepsilon \qquad \text{(modelo estocástico poblacional)} \\ y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \text{(modelo estocástico muestral)} \end{aligned}$$

para $i = 1, \ldots, n$, y n el tamaño de la muestra,

- Y es la variable respuesta (o variable independiente).
- X es la variable explicativa (o variable dependiente).
- β_0 es una constante desconocida (intercepto).
- β_1 es una constante desconocida (pendiente).
- ε es un error aleatorio desconocido.

El modelo determinista plantea el modelo que proponemos ajustar. En el modelo estocástico poblacional el término de error ε se introduce porque al observar X y Y se pueden cometer errores de medición, además de que se pueden haber eliminado algunas variables que se consideraron sin importancia, las cuales causan fluctuaciones en la respuesta. En el modelo estocástico muestral se considera que se tiene una muestra aleatoria de tamaño n $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Los parámetros β_0 y β_1 generalmente son llamados coeficientes de regresión. Estos coeficientes tienen una simple y útil interpretación. El parámetro β_1 es la pendiente de la recta y se interpreta como el cambio en Y producido por el cambio en una unidad de X. Si el rango de los datos de X incluye $X = 0$, entonces el intercepto β_0 es el valor de la variable respuesta Y cuando $X = 0$.

Supuestos del modelo

Media cero: se puede suponer un error con media cero, es decir, $\mathbb{E}(\varepsilon) = 0$.

Varianza constante: la varianza de los errores sea constante, es decir, $\mathbb{V}\text{ar}(\varepsilon) = \sigma^2$, donde σ^2 es una constante desconocida y $\sigma^2 < \infty$.

Covarianza cero: los errores están no correlacionados, es decir, $\mathbb{C}\text{ov}(\varepsilon_i, \varepsilon_j) = 0, \forall i, j,\ i \neq j$.

Normalidad: los errores siguen una distribución Gaussiana o Normal, es decir, $\varepsilon \sim Normal$.

$$\varepsilon_i \sim \text{Normal}\left(0, \sigma^2\right) \qquad \text{v.a.i.i.d.}$$

El trabajo estadístico consistirá en, dada la información proporcionada por la muestra, estimar los parámetros del modelo β_0 y β_1, y la varianza desconocida σ^2, considerando los supuestos anteriores.

En resumen, la construcción del modelo de regresión lineal simple es

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

donde

$$\varepsilon \sim \text{Normal}\left(0, \sigma^2\right) \qquad \text{v.a.i.i.d.}$$

La variable X es controlada, y la variable Y es aleatoria. Por lo tanto, existe una distribución de probabilidad para Y en cada valor de X. La media y varianza son:

$$\begin{aligned} \mathbb{E}(Y|X) &= \beta_0 + \beta_1 X, \\ \mathbb{V}\text{ar}(Y|X) &= \sigma^2. \end{aligned}$$

```
n = 100 ; B0 = 2 ; B1 = 3 ; sigma2 = 1 ;   # valores de los parametros
set.seed(12345)                            # semilla
x = runif(n, 0,1)                          # variable explicativa
Eps = rnorm(n, mean=0,sd=sqrt(sigma2))     # errores
y = B0 + B1*x + Eps                        # variable respuesta
plot(x,y)                                  # grafica los puntos (x,y)
modelo <- lm(y ~ x)                        # regresion lineal simple
lines(x,modelo$fit, lwd=2)                 # grafica la linea ajustada
```

2 Estimación

Estimadores mínimos cuadrados

Los parámetros β_0 y β_1 son desconocidos y deben estimarse usando los datos muestrales. El método de mínimos cuadrados se usa para estimar β_0 y β_1. Estimaremos β_0 y β_1 de tal manera que la suma de los cuadrados de las diferencias entre las observaciones Y_i y la línea recta $\beta_0 + \beta_1 X_i$ sean mínimas.. De esta manera, los errores son

$$\varepsilon_i \;=\; Y_i - \beta_0 - \beta_1 X_i,$$

y se busca minimizar

$$\sum_{i=1}^{n} \varepsilon_i^2 \;=\; \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

La figura 2 ilustra la situación.

Entonces, la suma de los cuadrados de los errores es

$$S \;=\; S(\beta_0, \beta_1) \;=\; \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

Los estimadores mínimos cuadrados de β_0 y β_1, denotados por $\hat{\beta}_0$ y $\hat{\beta}_1$, deben satisfacer

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = 0, \qquad \left.\frac{\partial S}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = 0,$$

para encontrar el punto crítico, y después es necesario demostrar que éste es un mínimo.

Los estimados $\hat{\beta}_0$ y $\hat{\beta}_1$ son:

$$\begin{aligned} \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X}, \\ \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}}, \end{aligned}$$

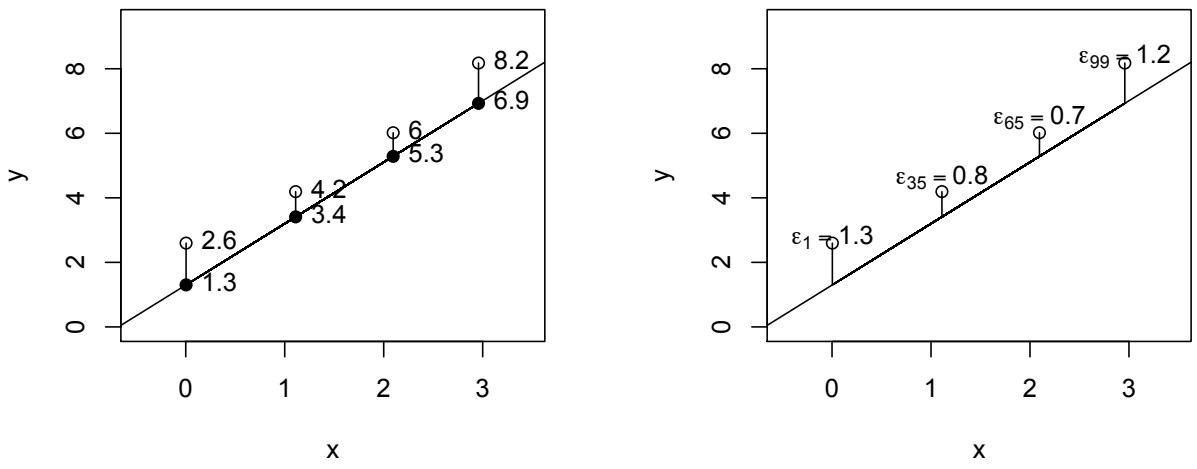

Figura 2: Errores del ajuste del modelo.

donde

$$\begin{aligned} S_{xy} &= \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n}(X_i - \bar{X})Y_i, \\ S_{xx} &= \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2, \end{aligned}$$

Por lo tanto $\hat{\beta}_0$ y $\hat{\beta}_1$ son los estimadores mínimos cuadrados de β_0 y β_1, respectivamente. El modelo de regresión ajustado es

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X.$$

Esta ecuación da una estimación puntual de la media de Y para un valor particular X.

La diferencia entre los valores observados Y_i y sus correspondientes valores ajustados $\hat{Y}_i$ se conoce como residuo o residual. El i-ésimo residual, para $i = 1, \ldots, n$, es

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i.$$

Los residuales juegan un papel importante para averiguar si el modelo de regresión ajustado es adecuado o no, y para detectar si se cumplen los supuestos del modelo.

Estimador de la varianza σ^2

El estimador de σ^2 se obtiene de la suma de cuadrados del error o de los residuales:

$$SC_{Error} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} \left[Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i \right) \right]^2 .$$

La suma de cuadrados de los residuales SC_{Error}, tiene $n-2$ grados de libertad, porque dos grados de libertad están asociados con las estimaciones de $\hat{\beta}_0$ y $\hat{\beta}_1$ al obtener $\hat{Y}_i$.

Se puede mostrar que $\mathbb{E}(SC_{Error}) = (n-2)\sigma^2$, por lo tanto, el estimador insesgado de σ^2 es

$$\hat{\sigma}^2 = \frac{SC_{Error}}{n-2} = MS_{Error},$$

donde MS_{Error} son los cuadrados medios de los residuales. Al estimador $\sqrt{\hat{\sigma}^2}$ algunas veces se le llama el *error estándar de la regresión.*

Estimación por máxima verosimilitud

Un supuesto adicional a los descritos en la sección 1 es la distribución Normal de los errores:

$$\varepsilon \sim \text{Normal}(0, \sigma^2),$$

por lo tanto se tiene que

$$Y \sim \text{Normal}(\beta_0 + \beta_1 X, \sigma^2).$$

La función de verosimilitud es

$$\begin{aligned} \mathcal{L} = \mathcal{L}(\beta_0, \beta_1, \sigma^2; \boldsymbol{Y}, \mathbf{X}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} (Y_i - \beta_0 - \beta_1 X_i)^2 \right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \right\}. \end{aligned}$$

Los estimadores máximo verosímiles son aquellos que maximizan la función de verosimilitud $\mathcal{L}$, o el logaritmo de la verosimilitud $\log \mathcal{L}$. Entonces, los estimadores máximo verosímiles $\tilde{\beta}_0$, $\tilde{\beta}_1$ y $\tilde{\sigma}^2$ de β_0, β_1 y σ^2, respectivamente, deben satisfacer el sistema de ecuaciones

$$\begin{aligned} \left. \frac{\partial \log \mathcal{L}}{\partial \beta_0} \right|_{\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\sigma}^2} &= 0, \\ \left. \frac{\partial \log \mathcal{L}}{\partial \beta_1} \right|_{\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\sigma}^2} &= 0, \\ \left. \frac{\partial \log \mathcal{L}}{\partial \sigma^2} \right|_{\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\sigma}^2} &= 0. \end{aligned}$$

La solución está dada por los estimadores:

$$\begin{aligned}\tilde{\beta}_0 &= \bar{Y} - \tilde{\beta}_1 \bar{X},\\ \tilde{\beta}_1 &= \frac{S_{xy}}{S_{xx}},\\ \tilde{\sigma}^2 &= \frac{\sum_{i=1}^{n}(Y_i - \tilde{\beta}_0 - \tilde{\beta}_1 X_i)^2}{n},\end{aligned}$$

Visto en términos de los estimadores mínimos cuadrados

$$\tilde{\beta}_0 = \hat{\beta}_0, \qquad \tilde{\beta}_1 = \hat{\beta}_1, \qquad \tilde{\sigma}^2 = \frac{n-2}{n}\hat{\sigma}^2.$$

En general, los estimadores máximo verosímiles tienen mejores propiedades estadísticas que los estimadores mínimos cuadrados. Los estimadores máximo verosímiles son insesgados o asintóticamente insesgados, y tienen varianza mínima cuando se comparan con todos los estimadores insesgados. Son estimadores consistentes (los estimadores difieren del parámetro verdadero por una cantidad pequeña cuando la muestra es grande) y son estadísticas suficientes (los estimadores contienen toda la información de la muestra original).

```
modelo <- lm(y ~ x)     ### modelo de regresion lineal

> summary(modelo)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-2.35764 -0.89720  0.04335  0.73949  2.40188

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2955     0.2253  10.190  < 2e-16 ***
x             2.7123     0.3813   7.113 1.87e-10 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.141 on 98 degrees of freedom
Multiple R-squared:  0.3405,Adjusted R-squared:  0.3338
F-statistic:  50.6 on 1 and 98 DF,  p-value: 1.875e-10
```

3 Intervalos de confianza de los parámetros

Para construir un intervalo de confianza *exacto* se requiere de una cantidad pivotal, que es una función de los datos y del parámetro de interés, y que tiene una distribución conocida que no depende de otros prámetros desconocidos.

Si los errores se distribuyen en forma normal e independiente, es decir $\varepsilon_i \sim \text{Normal}(0, \sigma^2)$ v.a.i.i.d. para $i = 1, \ldots, n$, entonces los intervalos del $100(1-\alpha)\%$ de confianza para β_0 y β_1 son:

$$\begin{aligned} \beta_0 &\in \left(\hat{\beta}_0 - t_{1-\alpha/2,n-2} se(\hat{\beta}_0)\ ,\ \hat{\beta}_0 + t_{1-\alpha/2,n-2} se(\hat{\beta}_0)\right), \\ \beta_1 &\in \left(\hat{\beta}_1 - t_{1-\alpha/2,n-2} se(\hat{\beta}_1)\ ,\ \hat{\beta}_1 + t_{1-\alpha/2,n-2} se(\hat{\beta}_1)\right). \end{aligned}$$

El intervalo del $100(1-\alpha)\%$ de confianza para σ^2 es:

$$\sigma^2 \in \left(\frac{(n-2)\hat{\sigma}^2}{\chi^2_{1-\alpha/2,n-2}}\ ,\ \frac{(n-2)\hat{\sigma}^2}{\chi^2_{\alpha/2,n-2}}\right).$$

La interpretación es la usual, por ejemplo, para β_1: *Si se tomaran muestras repetidas del mismo tamaño y se calcularan los intervalos del* 95% *de confianza de la pendiente de la muestra, entonces el* 95% *de los intervalos contendría el verdadero valor de* β_1.

```
> confint(modelo, parm=c("(Intercept)","x"), level=0.95)
                2.5 %   97.5 %
(Intercept) 1.848451 2.742525
x           1.955566 3.468948
```

4 Pruebas de hipótesis

En el modelo de regresión lineal simple, las pruebas de hipótesis se realizan para verificar si los parámetros involucrados toman algún valor particular propuesto por el investigador.

Pruebas de hipótesis de los parámetros

De manera general, la prueba de hipótesis es

$$H_0 : \beta_j = b_j \qquad vs. \qquad H_1 : \begin{cases} \beta_j \neq b_j \\ \beta_j > b_j \\ \beta_j < b_j \end{cases}$$

donde b_j es un valor fijo, para $j = 0, 1$.

Aunque las pruebas de hipótesis pueden realizarse sobre cualquier valor b_j, existen valores que revisten un especial interés. Para β_0, puede ser de interés verificar si la recta de regresión pasa por el origen, es decir que $\beta_0 = 0$, por lo que es relevante plantear la hipótesis

$$H_0 : \beta_0 = 0 \qquad vs. \qquad H_1 : \beta_0 \neq 0.$$

Para β_1, la prueba más relevante es determinar si este parámetro es o no distinto de cero, es decir,

$$H_0 : \beta_1 = 0 \qquad vs. \qquad H_1 : \beta_1 \neq 0,$$

ya que, de no rechazar $H_0 : \beta_1 = 0$, o equivalentemente, de aceptar $H_0 : \beta_1 = 0$, implicaría que:

- no existe cambio en Y por unidad de cambio en X (las dos variables son independientes, en el sentido de que los cambios en X no generan cambios en Y);
- X no sirve para explicar a Y;
- no hay relación lineal entre X y Y.

Por el contrario, si se rechaza $H_0 : \beta_1 = 0$, equivalentemente, si acepta $H_1 : \beta_1 \neq 0$, implicaría que:

- cambios en X provocan cambios en Y;
- X es importante para explicar a Y;
- existe relación lineal entre X y Y.

Pruebas de hipótesis para β_0

Para β_0 se tiene que

$$T \;=\; \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{\sigma}^2(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}})}} \;=\; \frac{\hat{\beta}_0 - \beta_0}{se(\hat{\beta}_0)} \;\sim\; t_{(n-2)},$$

y bajo $H_0 : \beta_0 = b_0$ entonces

$$T \;=\; \frac{\hat{\beta}_0 - b_0}{\sqrt{\hat{\sigma}^2(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}})}} \;=\; \frac{\hat{\beta}_0 - b_0}{se(\hat{\beta}_0)} \;\sim\; t_{(n-2)},$$

donde al valor $\sqrt{\hat{\sigma}^2(\frac{1}{n}+\frac{\bar{X}^2}{S_{xx}})}$ se le conoce como error estándar de $\hat{\beta}_0$ y se denota como $se(\hat{\beta}_0)$.

Las reglas de decisión para las distintas hipótesis alternativas son las siguientes:

- Para $H_1 : \beta_0 \neq b_0$, se rechaza H_0 al nivel de significancia α si y sólo si $|T| > t_{(n-2),1-\alpha/2}$, donde $t_{(n-2),1-\alpha/2}$ es el cuantil $1-\alpha/2$ de una distribución $t_{(n-2)}$.
- Para $H_1 : \beta_0 > b_0$, se rechaza H_0 al nivel de significancia α si y sólo si $T > t_{(n-2),1-\alpha}$, donde $t_{(n-2),1-\alpha}$ es el cuantil $1-\alpha$ de una distribución $t_{(n-2)}$.
- Para $H_1 : \beta_0 < b_0$, se rechaza H_0 al nivel de significancia α si y sólo si $T < t_{(n-2),\alpha}$, donde $t_{(n-2),\alpha}$ es el cuantil α de una distribución $t_{(n-2)}$.

Pruebas de hipótesis para β_1

De manera similar, para β_1 se tiene que

$$T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 \frac{1}{S_{xx}}}} = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \sim t_{(n-2)},$$

y bajo $H_0 : \beta_1 = b_1$ entonces

$$T = \frac{\hat{\beta}_1 - b_1}{\sqrt{\hat{\sigma}^2 \frac{1}{S_{xx}}}} = \frac{\hat{\beta}_1 - b_1}{se(\hat{\beta}_1)} \sim t_{(n-2)},$$

donde al valor $\sqrt{\hat{\sigma}^2 \frac{1}{S_{xx}}}$ se le conoce como error estándar de $\hat{\beta}_1$ y se denota como $se(\hat{\beta}_1)$.

Las reglas de decisión para las distintas hipótesis alternativas son las siguientes:

- Para $H_1 : \beta_1 \neq b_1$, se rechaza H_0 al nivel de significancia α si y sólo si $|T| > t_{(n-2),1-\alpha/2}$.
- Para $H_1 : \beta_1 > b_1$, se rechaza H_0 al nivel de significancia α si y sólo si $T > t_{(n-2),1-\alpha}$.
- Para $H_1 : \beta_1 < b_1$, se rechaza H_0 al nivel de significancia α si y sólo si $T < t_{(n-2),\alpha}$.

Prueba de significancia de la regresión

Se verá el desarrollo formal para las hipótesis

$$H_0 : \beta_1 = 0 \qquad vs. \qquad H_1 : \beta_1 \neq 0.$$

Estas hipótesis se relacionan con la *significancia de la regresión*. El no rechazar $H_0 : \beta_1 = 0$, es decir aceptar $H_0 : \beta_1 = 0$ implica que no hay relación lineal entre X y Y. Eso puede implicar varias cosas:

- No existe cambio en Y por unidad de cambio en X (los cambios en X no generan cambios en Y).
- X no sirve para explicar Y.
- No hay relación lineal entre X y Y.

Se rechaza H_0 al nivel de significancia α si $F_0 > \mathrm{F}^{1-\alpha}_{(1,n-2)}$, donde $\mathrm{F}^{1-\alpha}_{(1,n-2)}$ es el cuantil $1-\alpha$ de una distribución $\mathrm{F}_{(1,n-2)}$,

$$F_0 \;=\; \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2)} \;\sim\; \mathrm{F}_{(1,n-2)},$$

5 Análisis de varianza (ANOVA)

Se puede usar un análisis da varianza apropiado para la prueba de significancia de la regresión. El análisis de varianza se basa en particionar el total de la variabilidad de la variable respuesta Y:

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,\\ SC_{Total} &= SC_{Reg} + SC_{Error},\end{aligned}$$

- SC_{Total}: suma de cuadrados total de las observaciones, es la variabilidad total de las observaciones;
- SC_{Reg}: suma de cuadrados debida a la regresión o suma de cuadrados del modelo, es la variabilidad entre la línea de regresión y los datos;
- SC_{Error}: suma de cuadrados del error, también conocida como suma de cuadrados de los residuales o residuos, es la variabilidad entre los residuales.

Los grados de libertad asociados a las sumas de cuadrados del análisis de varianza son los siguientes:

- $SC_{Total} \to n-1$: porque uno de los grados de libertad se pierde como resultado de las desviaciones de los datos;
- $SC_{Reg} \to 1$: porque está totalmente determinada por un sólo parámetro, $\hat{\beta}_1$;
- $SC_{Error} \to n-2$: porque se establecen dos condiciones paras las desviaciones entre los valores observados y los ajustados al estimar los parámetros $\hat{\beta}_0$ y $\hat{\beta}_1$.

Se puede usar la estadística de prueba F del análisis de varianza para probar las hipótesis $H_0 : \beta_1 = 0$ *vs.* $H_1 : \beta_1 \neq 0$. Usando las propiedades de la distribución Normal se obtiene una cantidad pivotal de la siguiente manera.

El análisis de varianza (ANOVA) se resumen como se muestra en la Tabla 1. Note que la tabla resumen la ANOVA resume la suma de cuadrados (SC), los grados de libertad ($g.l.$). Note que los cuadrados medios del error (CM_{Error}) en realidad es igual a la varianza estimada $\hat{\sigma}^2$. La tabla ANOVA es, en general, un resumen la prueba de hipótesis de significancia de la regresión, es decir $H_0 : \beta_1 = 0$ v.s. $H_1 : \beta_1 \neq 0$, y también es común mostrar su $p-value$, que se calcula como $\mathbb{P}[F > F^{observada}]$.

Análisis de Varianza (ANOVA) $H_0 : \beta_1 = 0$ *vs.* $H_1 : \beta_1 \neq 0$				
Fuente de Variación	Grados de Libertad	Suma de Cuadrados	Cuadrados Medios	F
Regresión	1	$SC_{Reg} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	$CM_{Reg} = SC_{Reg}$	
Error	$n-2$	$SC_{Error} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$CM_{Error} = \frac{SC_{Error}}{n-2}$	$F = \frac{CM_{Reg}}{CM_{Error}}$
Total	$n-1$	$SC_{Tot} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$		

Table 1: Tabla de análisis de varianza para regresión lineal simple.

```
> anova(modelo)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1  65.863  65.863  50.596 1.875e-10 ***
Residuals 98 127.572   1.302
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

6 Coeficiente de determinación

Una vez que el modelo se ha ajustado y que se considera adecuado dentro del área de aplicación, es necesario verificar qué tan bien ajustado está. Una de las medidas que permiten determinar la bondad del ajuste del modelo se llama *coeficiente de determinación*, denotado como R^2, y se define como

$$R^2 \;=\; \frac{SC_{Reg}}{SC_{Total}} \;=\; \frac{S_{yy} - SC_{Error}}{S_{yy}} \;=\; 1 - \frac{SC_{Error}}{S_{yy}} \;=\; 1 - \frac{SC_{Error}}{SC_{Total}},$$

además se tiene que $0 \leq R^2 \leq 1$, ya que $0 \leq SC_{Error} \leq SC_{Total}$.

El coeficiente de determinación se interpreta como el porcentaje o proporción de la varianza total de la respuesta (las Y's) que está explicado por el modelo de regresión (o los regresores X's). Si mucha variabilidad de la respuesta es explicada por los regresores, entonces SC_{Error} será pequeña comparada con SC_{Total}, lo que implica que SC_{Reg} será grande respecto a SC_{Total}, entonces $R^2 \approx 1$.

En la práctica, un valor de $R^2 < 1$ no implica que el modelo no sirva, o que se tenga una mala estimación, se requiere revisar varios elementos y supuestos del modelo. En general lo que sucede con R^2 es que al incluir variables explicativas el R^2 aumenta, sin embargo, no significa que sean 'buenas' variables para explicar la variable respuesta ni que con esto se obtenga un mejor modelo.

7 Predicción

Intervalos de confianza de la respuesta media

Una aplicación importante de un modelo de regresión es estimar la respuesta media $\mathbb{E}(Y)$, para determinado valor de la variable regresora x, es decir, digamos $\mathbb{E}(Y|x_0)$, donde x_0 es cualquier valor de la variable regresora dentro del rango de los datos originales de x que se usaron para ajustar el modelo. Un estimador insesgado de $\mathbb{E}(Y|x_0)$ se determina a partir del modelo ajustado como sigue:

$$\begin{aligned} \mathbb{E}(Y|x_0) \;&=\; \mu_{Y|x_0} \;=\; \beta_0 + \beta_1 x_0, \\ \widehat{\mathbb{E}(Y|x_0)} \;&=\; \hat{\mu}_{Y|x_0} \;=\; \hat{\beta}_0 + \hat{\beta}_1 x_0. \end{aligned}$$

Por lo tanto, un intervalo del $(1-\alpha)100\%$ de confianza para $\mathbb{E}(Y|x_0)$ es

$$\mathbb{E}(Y|x_0) \;\in\; \left(\hat{\mu}_{Y|x_0} - t_{1-\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\tfrac{1}{n} + \tfrac{(x_0-\bar{x})^2}{Sxx}\right]} \;,\; \hat{\mu}_{Y|x_0} + t_{1-\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\tfrac{1}{n} + \tfrac{(x_0-\bar{x})^2}{Sxx}\right]} \right).$$

Note que el ancho del intervalo de confianza para $\mathbb{E}(Y|x_0)$ es una función de x_0, alcanza su mínimo cuando $x_0 = \bar{x}$, y crece a medida que aumenta $|x_0 - \bar{x}|$. Esto significa que se espera que las mejores estimaciones de Y sean aquellas con valores de x cercanos al centro de los datos, y que la precisión de esas estimaciones se reduzca al moverse hacia la frontera del espacio de x.

```
> x2 = data.frame(x=c(0.2,0.4,0.6,0.8))
> predict(modelo, newdata=x2, interval="confidence", level=0.95)
       fit      lwr      upr
1 2.837939 2.512245 3.163633
2 3.380390 3.139315 3.621466
3 3.922842 3.686274 4.159410
4 4.465293 4.149670 4.780917
```

Intervalos de predicción

Una aplicación importante del modelo de regresión es predecir nuevas observaciones Y que correspondan a un nivel especificado de x. Si x_0 es el valor de interés, entonces

$$\hat{y}_0 \quad = \quad \hat{\beta}_0 + \hat{\beta}_1 x_0$$

es el estimador puntual del nuevo valor de la respuesta y_0.

El intervalo de confianza para la respuesta media de $x = x_0$ es inadecuado para este problema, porque es un estimador del intervalo para la media de Y, y no es un estimador sobre futuras observaciones. El intervalo de predicción del $(1-\alpha)100\%$ de confianza para una observación futura en x_0 es

$$y_0 \quad \in \quad \left(\hat{y}_0 - t_{1-\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]} \;,\; \hat{y}_0 + t_{1-\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]} \right).$$

Este intervalo de predicción es de ancho mínimo en $x_0 = \bar{x}$, y se ensancha a medida que aumenta $|x_0 - \bar{x}|$.

```
> x2 = data.frame(x=c(0.2,0.4,0.6,0.8))
> predict(modelo, newdata=x2, interval="prediction", level=0.95)
       fit         lwr      upr
1 2.837939  0.55046802 5.125410
2 3.380390  1.10342663 5.657354
3 3.922842  1.64635085 6.199333
4 4.465293  2.17923446 6.751352
```

El intervalo de predicción en x_0 siempre es más ancho que el intervalo de confianza para la respuesta media en x_0, porque el intervalo de predicción depende tanto del error del modelo ajustado como del error asociado con observaciones futuras. La figura 3 muestra las bandas formadas por los intervalos de confianza para la respuesta media y los intervalos de predicción para una nueva observación.

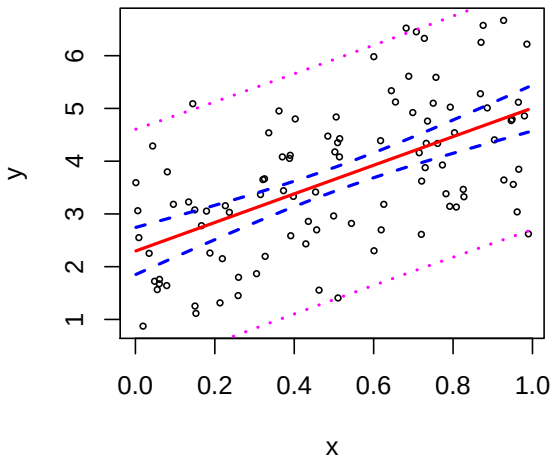

Figura 3: Intervalos de confianza para la respuesta media $\mathbb{E}(Y|X)$ e intervalos de predicción para una nueva observación Y^*.

8 Regresión lineal múltiple

La mayoría de los fenómenos reales son multicausales, por esta razón, un modelo de regresión más acorde a estudios reales es el modelo de regresión lineal múltiple, que es la generalización del modelo simple. En este caso supondremos que la variable respuesta Y puede explicarse a través de un conjunto de k covariables $X_1, \ldots, X_k$.

El modelo de regresión lineal múltiple se escribe como:

$$Y_i \quad = \quad \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i,$$

para $i = 1, \ldots, n$, donde para cada i-ésimo individuo

- Y_i es la variable respuesta,
- X_{ij} es el valor observado de la covariable j,
- ε_i es el término de error.

La figura 4 muestra un conjunto de datos para $k = 2$ covariables.

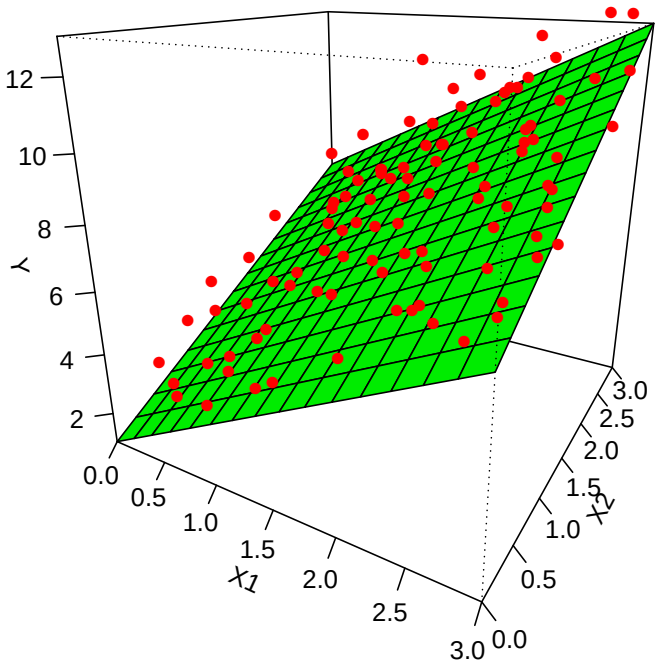

Figura 4: Modelo de regresión lineal múltiple ($k = 2$).

El modelo en términos matriciales

La notación matricial del modelo de regresión lineal múltiple es:

$$\boldsymbol{Y} \quad = \quad \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

donde

$$\boldsymbol{Y}_{n\times 1} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \mathbf{X}_{n\times p} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix},$$

$$\boldsymbol{\beta}_{p\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \qquad \boldsymbol{\varepsilon}_{n\times 1} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

y se tiene que $p = k + 1$, así el vector de parámetros de regresión $\boldsymbol{\beta}$ es de dimensión p, siento $p = k + 1$ cuando se considera el parámetro de intercepción, $p = k$ cuando no se considera. Además, se usan los mismos supuestos que en regresión lineal simple para el término de error

$$\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}, \qquad \mathbb{V}\text{ar}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_{n\times n},$$

es decir, los errores tienen media cero, varianza constante y covarianza cero.

Además, se asume que las Y's están medidas en escala continua (la escala de medición de las variables predictoras da lugar a distintos modelos, ver de modelos lineales generalizados).

Los supuestos del modelo implican que

$$\mathbb{E}(\boldsymbol{Y}) = \mathbf{X}\boldsymbol{\beta}, \qquad \mathbb{V}\text{ar}(\boldsymbol{Y}) = \mathbb{C}\text{ov}(\boldsymbol{Y}) = \sigma^2 \mathbf{I}_{n\times n},$$

Representación estructural

La forma estructural del modelo de regresión lineal es:

$$\underbrace{\boldsymbol{Y}}_{\text{Datos (respuesta)}} = \underbrace{\mathbf{X}\boldsymbol{\beta}}_{\substack{\text{Estructura sistemática} \\ \text{(forma funcional,} \\ \text{propuesta por el investigador)}}} + \underbrace{\boldsymbol{\varepsilon}}_{\text{Variación aleatoria}}$$

Esta forma estructural es, en general, la forma básica de cualquier modelo estadístico. Obsérvese lo importante que resulta la opinión de un experto del área de aplicación (el investigador) para definir la estructura de el modelo.

Bibliografía

Cohen, Y. and J. Y. Cohen (2008). *Statistics and Data with R: An applied approach through examples*. John Wiley & Sons.

Frees, E. W. (2009). *Regression Modeling with Actuarial and Financial Applications*. Cambridge University Press.

Montgomery, D. C., E. A. Peck, and G. G. Vining (2007). *Introduction to Linear Regression Analysis* (4th ed.). John Wiley & Sons.

Rencher, A. C. and G. B. Schaalje (2008). *Linear Models in Statistics* (2nd ed.). New Jersey: John Wiley & Sons.

Weisberg, S. (2005). *Applied Linear Regression*. John Wiley & Sons.