BITMONEY

HERRAMIENTA: REGRESION MULTIPLE

BITMONEY

Autores

M.I.C. Carlos Abraham Carballo Monsivais

I.S.C. Leticia Edith Trujillo Ballesteros

C.L.M.A. Sacbe García García

BITMONEY

Hackathon Blockchain 2020

Elaboración 2020, Primera edición

BITMONEY

Herramienta: Análisis de regresión lineal múltiple

Contenido

1. Contenido

BITMONEY

4 Análisis de regresión lineal múltiple

En esta unidad se estudiará el caso en el que más de una **variable independiente** puede influir en el comportamiento de la **variable dependiente**. Y Para describir la forma de la relación que liga a estas variables se utilizarán los llamados modelos de regresión múltiple.

Por ejemplo, supongamos que la variable x_1 define el precio anual del trigo, x_2 la cantidad de fertilizantes utilizada y la variable y, las hectáreas sembradas anualmente en una región; es posible estudiar el efecto del precio y de los fertilizantes en la producción de trigo.

4.1 Análisis de regresión lineal múltiple

Este modelo de regresión puede estudiarse como una extensión del modelo lineal simple en el que considerábamos una sola variable predictor x. Ahora vamos a considerar que la variable de respuesta y depende de varias variables x, conocidas por el investigador.

El modelo de regresión múltiple trata de estimar el efecto de las variables más importantes, mientras que englobando las demás (las no importantes o que desconoce) en el término que denominarnos error aleatorio.

Para simplificar, vamos a suponer que la variable y depende solamente de dos variables x_1 y x_2 , y que la relación que liga a las variables sigue siendo lineal.

La ecuación de regresión poblacional es la siguiente:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \quad (4.1.1)$$

Un ejemplo de esta situación de investigación podría ser que y fuera la cantidad de lluvia caída en una zona en particular, x_1 la humedad del ambiente, y x_2 la presión atmosférica. La variable de respuesta y depende de dos variables que llamaremos predictoras: x_1 y x_2.

Otro ejemplo podría ser que y fuera el salario pagado por una empresa, x_1 los años de antigüedad de los empleados y x_2 la calificación anual de cada uno de ellos.

Un modelo de regresión lineal poblacional de manera general se escribe de la siguiente manera:

$$\underbrace{y = \alpha}_{\text{Variable respuesta}} + \underbrace{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \varepsilon}_{\text{Variables explicativas}} \quad 4.1.2$$

Supuestos del Modelo

Los supuestos referidos a los ε son los mismos que estudiamos en el modelo de regresión lineal simple.

El ejemplo de enseguida se desarrolla a lo largo de esta lectura para comprender los conceptos que en su momento se esté tratando.

BITMONEY

hipertensión pediátrica. La relación entre la presión sistólica (SPB) y , el peso al nacimiento x_1 y edad en días x_2 se supone que es de la siguiente manera:

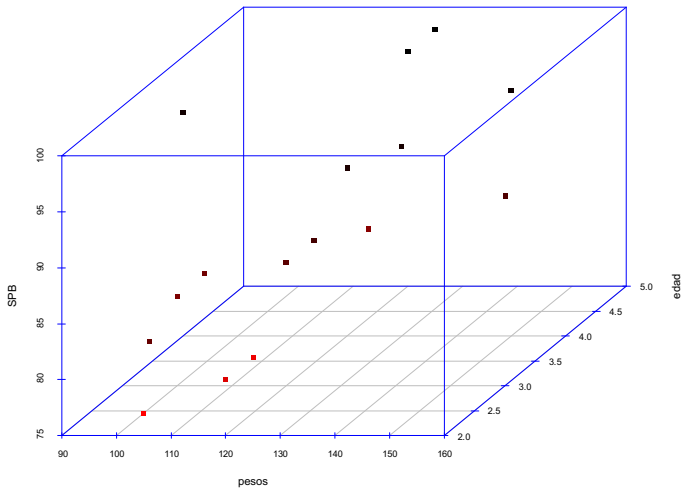
$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Suponer que SBP, peso al nacimiento y la edad son medidas en 16 infantes. Los datos se muestran en la tabla de continuación.

*en la plataforma hay un documento Excel denominado datos_hipertension que contiene los datos de la tabla de abajo.

infante	peso (oz)	edad$(dias)$	SPB$(mm\ Hg)$
1	135	3	89
2	120	4	90
3	100	3	83
4	105	2	77
5	130	4	92
6	125	5	98
7	125	2	82
8	105	3	85
9	120	5	96
10	90	4	95
11	120	2	80
12	95	3	79
13	120	3	86
14	150	4	97
15	160	3	92
16	125	3	88

El gráfico de dispersión de los datos se muestra enseguida.



BITMONEY

Del gráfico de dispersión anterior se puede ver que a valores más latos de pesos y edad la presión sistólica del infante es mayor.

4.2 Estimación de los parámetros del modelo

El modelo de regresión lineal para el ejemplo de **hipertensión pediátrica** empleando la ecuación 4.1.1 se expresa en forma matricial de la siguiente manera:

$$y = y_{16\times1} = \begin{bmatrix} 89 \\ 90 \\ 83 \\ \vdots \\ 97 \\ 92 \\ 88 \end{bmatrix} \quad X = X_{16\times3} = \begin{bmatrix} 1 & 135 & 3 \\ 1 & 120 & 4 \\ 1 & 100 & 3 \\ \vdots & \vdots & \vdots \\ 1 & 150 & 4 \\ 1 & 160 & 3 \\ 1 & 125 & 3 \end{bmatrix} \quad \varepsilon = \varepsilon_{16\times1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_{14} \\ \varepsilon_{15} \\ \varepsilon_{16} \end{bmatrix} \quad \beta = \beta_{3\times1} = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

x_1 *peso al nacimiento* y x_2 *edad en días*

$$\begin{bmatrix} 89 \\ 90 \\ 83 \\ \vdots \\ 97 \\ 92 \\ 88 \end{bmatrix} = \begin{bmatrix} 1 & 135 & 3 \\ 1 & 120 & 4 \\ 1 & 100 & 3 \\ \vdots & \vdots & \vdots \\ 1 & 150 & 4 \\ 1 & 160 & 3 \\ 1 & 125 & 3 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_{14} \\ \varepsilon_{15} \\ \varepsilon_{16} \end{bmatrix}$$

Para el primer renglón se tiene:

$$89 = 1 * \alpha + 135 * \beta_1 + 3 * \beta_2 + \varepsilon_1$$

De manera general

$$y = X_{16\times3} * \beta_{3\times1} + \varepsilon_{16\times1} \quad 4.2.1$$

La línea recta que relaciona la variable y con las variables regresoras, predictoras o independientes (x_1 y x_2) es mediante la siguiente ecuación:

$$\mu_{y/_{x_1\, x_2}} = X_{16\times3} * \beta_{3\times1}$$

La ecuación 4.2.1 debido a que únicamente contiene información de una muestra, entonces, se denomina modelo de regresión lineal muestral del ejemplo de **hipertensión pediátrica.**

En este caso, el interés consiste encontrar los valores $a, b_1\ y\ b_2$, que son los estimadores puntuales de los parámetros de $\alpha, \beta_1\ y\ \beta_2$, respectivamente, y como en el caso de regresión lineal simple, los valores encontrados de los estimadores minimizan la suma de cuadrados del valor observado y y el valor ajustado $\hat{y}$.

$$\min_{a,b_1,b_2} \sum_{i=1}^{16} (y_i - \hat{y}_i)^2 \quad 4.2.2$$

Solo que ahora

$$\mu_{y_i/_{x_{1i}\, x_{2i}}} = \hat{y}_i = \alpha + b_1 * x_{1i} + b_2 * x_{2i}$$

BITMONEY

La ecuación 4.2.2 se expresa de forma matricial de la siguiente forma:

$$\min(y-\hat{y})^T(y-\hat{y}) = \min_{\hat{\beta}}\left(Y-X\hat{\beta}\right)^T\left(y-X\hat{\beta}\right)$$

$$\hat{\beta}_{3\times1} = \begin{bmatrix} a \\ b_1 \\ b_2 \end{bmatrix}$$

Nota: Si X es una matriz, entonces X^T denota la matriz transpuesta de X.

Y la manera de minimizar la ecuación es derivar respecto al vector $\hat{\beta}$ e igualando a cero.

$$\frac{d\left(Y-X\hat{\beta}\right)^T\left(y-X\hat{\beta}\right)}{d\hat{\beta}} = 0 \quad 4.2.3$$

Encontrar la solución de la ecuación 4.2.3 se complica y posiblemente no sea comprendido por todos los participantes de este curso, por tal motivo enseguida únicamente se proporciona la solución de dicha ecuación, solución que los softwares estadísticos emplean para estimar los parámetros muestrales de la regresión.

$$\hat{\beta} = (X^TX)^{-1}X^Ty \quad 4.2.4$$

Nota: $(X^TX)^{-1}$ denota la inversa del producto X^TX.

Empleando el comando **regress** del software stata los valores de los parámetros muéstrales son los siguientes:

```
regress spb peso edad, beta
```

Tabla 1. Valores de los coeficientes de regresión.

	coeficientes	error estándar	t	P>|t|	[95% Conf. Intervalo]	
peso	0.126	0.034	3.66	0.003	0.051	0.199
edad	5.89	0.68	8.66	0.000	4.418	7.357
constante	53.45	4.53	11.79	0.000	43.66	63.241

$$\hat{\beta}_{3\times1} = \begin{bmatrix} a \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 53.45 \\ 0.126 \\ 5.89 \end{bmatrix}$$

De modo que la recta de regresión ajustada muestral es:

$$\hat{y}_i = 53.45 + 0.126 * x_1 + 5.89 * x_2 \quad 4.2.5$$

De la ecuación anterior se interpreta lo siguiente:

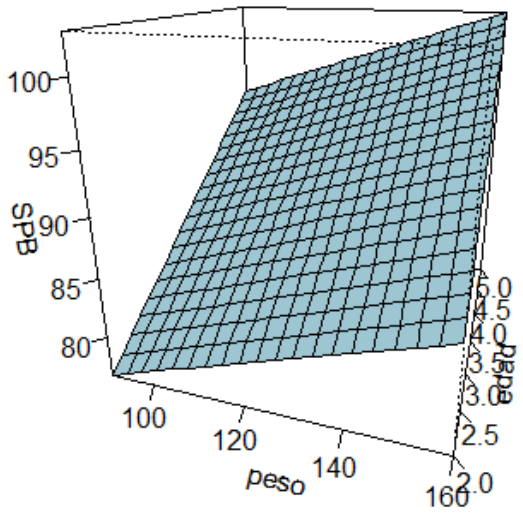
El peso y la edad se asocian de manera positiva con la presión sistólica, es decir, a mayor edad y pesos, mayor presión sistólica.

Si se mantiene constante el valor de la variable edad x_2, se fija en un valor, la presión sistólica aumenta en un razón de 0.126 ($^{mm\ Hg}/_{oz}$) por cada oz adicional. Por el contrario, si se mantiene

constante el valor del peso x_1, la presión sistólica aumenta en una razón de 5.89 $(^{mm\ Hg}/_{dia})$ por cada día adicional.

En este caso la ordenada al origen (constante) no tiene una interpretación exacta, por ejemplo: si un individuo presenta un peso de 0 oz y 0 días de nacido, no es correcto decir que tendrá una presión sistólica igual a 53.45.

Una ilustración visual de la ecuación 4.2.5 anterior es:



Calcular SPB promedio de un bebe cuyo peso al nacimiento fue 128 oz y 3 días de vida.

El SPB promedio del bebe estimado es:

$$53.45 + 0.126 * 128 + 5.89 * 3 \approx 87.20 mm\ Hg$$

En muchas ocasiones se tiene el interés por saber cuál de las dos variables predictoras influye más en el modelo de regresión con múltiples variables predictoras. Cuando esta es la situación, se emplea los coeficientes estándarizados. Si observa la ecuación 4.2.5 y si ve a los coeficientes parciales, las constantes que acompañan a las variables x_1 y x_2, la variable que más influye es la variable x_2 (edad). Sin embargo, como en los coeficientes parciales dependen de la unidad de medida de la variable a veces suelen ser engañosos. Para evitar este engaño se emplean los coeficientes estandarizados.

Empleando el comando el siguiente comando

```
regress spb peso edad, beta
```

La salida es la siguiente:

Tabla 2. Valores de coeficientes de regresión estandarizados.

	coeficientes	error estándar	t	P>|t|	Coef. estandarizados
peso	0.126	0.034	3.66	0.003	0.35
edad	5.89	0.68	8.66	0	0.83

constante	53.45	4.53	11.79	0

En la tabla anterior, la parte enmarcada de color rojo muestra los coeficientes de la regresión estandarizados. El coeficiente estandarizado de la variable edad es mayor, esto indica que dicha variable tiene mayor influencia en la ecuación de regresión que la variable peso.

Propiedades de los estimadores de mínimos cuadrados

No se demuestran las propiedades de los estimadores de un modelo de regresión lineal múltiple, pero se mencionan ya que serán útiles cuando se quiera hacer inferencias acerca de estos.

1. Los estimadores de a, b_1 y b_2 son insesgados.

$$E(a) = \alpha$$
$$E(b_1) = \beta_1$$
$$E(b_1) = \beta_2$$

Como los estimadores se contienen en un vector

$$\hat{\beta}_{3\times1} = \begin{bmatrix} a \\ b_1 \\ b_2 \end{bmatrix}$$

$$E(\hat{\beta}_{3\times1}) = E\left(\begin{bmatrix} a \\ b_1 \\ b_2 \end{bmatrix}\right) = \beta_{3\times1} = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

2. Es estimador $\hat{\beta}_{3\times1}$ es un estimador con mínima varianza.

$$cov(\hat{\beta}_{3\times1}) = \sigma^2(X^TX)^{-1}$$

$$X^TX = \begin{bmatrix} 16 & 1925 & 53 \\ 1925 & 236875 & 6405 \\ 53 & 6405 & 189 \end{bmatrix}$$

α	b_1	b_2

$$C = (X^TX)^{-1} = \begin{bmatrix} 3.34 & -0.021 & -0.20 \\ -0.021 & 0.00019 & -0.0004 \\ -0.20 & -0.0004 & 0.0752 \end{bmatrix}$$

Por consiguiente si $C = (X^TX)^{-1}$, la varianza de a, b_1 o b_2 es:

$$Var(a) = \sigma^2 C_{1,1}$$

Usando la matriz C de los datos que estamos trabajando se tiene:

$$Var(a) = \sigma^2 * 3.34$$

El valor estimado σ^2 aun no se muestra como estimarlo, esto se estudia más adelante, por el momento únicamente se comenta que $\hat{\sigma}^2 = 6.15$.

$$Var(a) = 6.15 * 3.34 = 20.54$$

BITMONEY

Y la desviación estándar es $\sqrt{Var(a)} = \sqrt{20.54} = 4.53$

Mientras que para las pendientes será:

$$Var(b_{j-1}) = \sigma^2 C_{j,j} \quad j = 2\ y\ 3$$

Usando la matriz C de los datos que estamos trabajando se tiene:

Varianza del coeficiente de regresión de la variable regresora peso:

$$Var(b_1) = \sigma^2 C_{2,2} = 6.15 * 0.00019 = 0.0011$$

La desviación estándar del coeficiente de regresión para la variable x_1 será

$$\sqrt{Var(b_1)} = \sqrt{0.0011} = 0.034$$

Debido a que realizar el cálculo de la matriz $(X^T X)^{-1}$ es muy tedioso, la varianza de los coeficientes se calculará empleando el software stata. En la tabla 1 los valores enmarcados de color rojos son los valores de la desviación estándar de los coeficientes de regresión. Los valores estándar serán útiles en la parte de pruebas de hipótesis.

¿Cómo se estima σ^2 o error estándar de la regresión S_e^2?

Al igual en el modelo de regresión lineal simple, aquí, la varianza o error estándar de la regresión se estima como la división de la suma de cuadrados del error divida por los grados de libertad correspondiente a dicha suma de cuadrados.

$$SCe = y_{1\times16} y_{16\times1} - \hat{\beta}_{1\times3} X_{3\times16} = y^T y - \hat{\beta}^T X^T y$$

$$\hat{\sigma}^2 = S_e^2 = \frac{SCe}{16-3}$$

El cálculo del error estándar de la regresión empleando el software stata se lleva acabo con el mismo comando que se empleó para estimar los coeficientes de la regresión. La tabla de enseguida es la salida del software stata, y la parte de la tabla enmarcada de color negra es $SCe = 79.9$.

Los grados de libertad del error se cálculos de la siguiente forma:

n: Número de observaciones

p: Denota el número de coeficientes en el modelo de regresión, en este ejemplo, hay 3 y son: a, b_1 y b_2.

Por lo que los grados de libertad del error son: $16 - 3 = 13$

	Suma de cuadrados	Grados de libertad	MS
Regresión	591.03	2	295.51
Residual	79.9	13	6.15
Total	670.93	15	

El error estándar de la regresión es

$$\hat{\sigma}^2 = S_e^2 = \frac{79.9}{13} = 6.15$$

Con base a la tabla anterior S_e^2 es al también denominado cuadrado medio del error que más adelante se empleará para probar hipótesis de los coeficientes de regresión.

4.3 Prueba de hipótesis en el modelo de regresión lineal múltiple

La prueba de la significancia de la regresión es para determinar si hay una relación lineal entre la respuesta SPB y cualquiera de las variables regresoras peso (x_1) y edad (x_2).Enseguida se proporciona una prueba general o global de los coeficientes de las variables regresoras.

$$H_o: \beta_1 = \beta_2 = 0 \quad Vs \quad H_1: Alguna\ \beta_j \neq 0$$

EL rechazo de la hipótesis nula implica que al menos una de las regresoras x_1 y x_2 contribuye al modelo de regresión de forma significativa. Del mismo modo que como se hizo en la regresión lineal simple, la **suma de cuadrados total** se divide en dos **sumas de cuadrados**, que son: **suma de cuadrados de la regresión** y **suma de cuadrados del error**.

$$SCT = SCR + SCe$$

Donde:

$$SCT = y^T y - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$$

$$SCR = \hat{\beta}^T X^T y - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$$

$$SCe = y^T y - \hat{\beta}^T X^T y$$

	Suma de cuadrados	Grados de libertad	MS	$\boldsymbol{F_0}$
Regresión	SCR	k	$MS_R = {SCR}/{k}$	${MS_R}/{MS_e}$
Residual	SCe	$n-p$	$MS_e = {SCR}/{(n-p)}$	
Total	SCT	$n-1$		

k: Denota el número de variables regresoras

$p = k + 1$: Denota el número de coeficientes en el modelo de regresión.

Para probar la prueba de hipótesis establecida $H_o: \beta_1 = \beta_2 = 0$, se calcula el estadístico de prueba $\boldsymbol{F_0}$ y se rechaza H_o si

$$F_o > F_{\alpha,\ k\ ,n-p}$$

Otro modo para decir rechazar o no rechazar la hipótesis nula mediante el criterio del p valor.

Regresando con el caso de hipertensión pediátrica. Los cálculos no se realizarán a mano, se interpretarán las salidas proporcionadas por el software stata. Empleando el siguiente comando de stata

```
regress spb peso edad
```

	Suma de cuadrados	Grados de libertad	MS	F_0
Regresión	591.03	2	$591.03/2 = 295.52$	$295.52/6.15 = 48.05$
Residual	79.90	$16 - 3 = 13$	$79.90/(13) = 6.15$	
Total	670.94	$16 - 1 = 15$		

No. observaciones=16

$F(2,13)$= 48.05

P valor=0.0000

R-cuadrado=0.86

R-ajustado=0.8286

De acuerdo a la tabla del análisis de varianza se tiene que la recta de recta regresión es quien recoge la mayor variabilidad, esto debido a que la suma de cuadrados de la regresión es mucha mayor que la suma de cuadrado de los residuales. Esto se puede ver también en el valor del R-ajustado, el modelo de regresión explica el 82% de la variabilidad total.

El valor del R ajustado es una medida que se emplea en regresión lineal múltiple para cuantificar el porcentaje de la varianza explicada por el modelo de regresión. Para el caso de regresión lineal simple se utilizó R cuadrada.

Empleando el criterio del p valor, con un nivel de significancia del $\alpha = 0.05$ se rechaza la hipótesis nula, esto debido a que el P valor (0.0000) es mucha más pequeño que 0.05.

¡Ojo!

La tabla del análisis de varianza proporciona un análisis de los coeficientes de regresión de manera conjunta, en otras palabras, nos permite determinar si la regresión tiene significancia, pero no dice explícitamente cuales coeficientes son los significativos. En este caso del ejemplo que hemos venido desarrollando, el análisis de varianza lo que nos dice es que al menos uno de los coeficientes de regresión es distinto de cero, y al menos quiere decir que puede que sólo x_1, puede que sólo x_2 o puede que ambos. Pero, para saber cuáles son los significativos se emplea una prueba de hipótesis parcial de los coeficientes de regresión, esto se lleva a cabo empleando una prueba de hipótesis empleando el estadístico t.

Pruebas de coeficientes individuales de los coeficientes de regresión

Una vez determinado que al menos uno de las variables regresaras es importante, la pregunta lógica es ¿cuál de ellos sirven?

La hipótesis para probar la significancia de cualquier coeficiente individual de regresión, por ejemplo, el coeficiente de regresión de la variable peso:

$$H_0: \beta_1 = 0 \quad Vs \quad H_0: \beta_1 \neq 0$$

BITMONEY

Si no se rechaza la hipótesis nula quiere decir que se puede eliminar la variable regresora x_1 del modelo. El estadístico de prueba es

$$t_0 = \frac{b_1}{\sqrt{\hat{\sigma}^2 C_{2,2}}}$$

Entonces, se rechaza la hipótesis nula $H_0: \beta_1 = 0$ si

$$|t_0| = t_{\alpha/2,\ n-p}$$

De igual manera se puede emplear el concepto del p valor para decidir rechazar la hipótesis nula.

$$t_0 = \frac{b_1}{\sqrt{\hat{\sigma}^2 C_{2,2}}} = \frac{0.126}{\sqrt{6.15 * 0.00019}} = 3.68$$

$$P(t_{16-3} > 3.68 | H_0) = P(t_{13} > 3.68 | H_0) \approx 0.003$$

Nuevamente, los cálculos se desarrollan en el software stata y empleando el comando regress e interpretando las salidas.

```
regress spb peso edad
```

La salida tras ejecutar el código anterior es:

	coeficientes	error estándar	t	P>|t|	[95% Conf. Intervalo]	
peso	0.126	0.034	3.66	0.003	0.051	0.199
edad	5.89	0.68	8.66	0.000	4.418	7.357
constante	53.45	4.53	11.79	0.000	43.66	63.241

En la tabla anterior, el cuadro enmarcado de color azul se muestra el p valor de cada uno de los coeficientes de las variables regresoras. Si se propone un nivel de significancia $\alpha = 0.05$, entonces se rechazan cada una de las siguientes hipótesis nulas.

$$H_0: \alpha = 0 \quad Vs \quad H_0: \alpha \neq 0$$

$$H_0: \beta_1 = 0 \quad Vs \quad H_0: \beta_1 \neq 0$$

$$H_0: \beta_2 = 0 \quad Vs \quad H_0: \beta_2 \neq 0$$

Por otra parte, también se puede estimar mediante intervalos de confianza los coeficientes de regresión. En la tabla anterior, la parte remarcada de color rojo ilustra la estimación de coeficientes de regresión mediante un intervalo de confianza. Como puede ver, ningún intervalo contiene al cero, entonces los coeficientes son distintos de cero.

Estimación de valores medio

Anteriormente se planteó la siguiente pregunta:

> *Calcular el valor de SPB medio para un bebe cuyo peso al nacimiento fue 128 oz y 3 días de vida*.

El SPB promedio del bebe estimado fue:

BITMONEY

$$53.45 + 0.126 * 128 + 5.89 * 3 \approx 87.20 mm\ Hg \quad 4.3.1$$

En regresión lineal simple se estableció la manera de estimar la respuesta media de la variable respuesta mediante un intervalo de confianza, ahora se muestra como hacer lo mismo, pero en una regresión lineal múltiple.

El caso del niño cuyo peso al nacer fue 128 oz y tiene 3 días de edad. Esta información se puede contener en el siguiente vector:

$$X_0 = \begin{bmatrix} 1 \\ 128 \\ 3 \end{bmatrix}$$

El valor ajustado expresado en la ecuación 4.3.1 se puede expresar en forma matricial de la siguiente forma:

$$\hat{y}_0 = X_0^T \hat{\beta} \quad 4.3.2$$

$$\hat{\beta} = \begin{bmatrix} 53.45 \\ 0.1256 \\ 5.89 \end{bmatrix}$$

Y precisamente la ecuación 4.3.2 es un estimador insesgado de $E(\hat{y}_0|X_0)$ y la varianza de $\hat{y}_0|X_0$ es

$$Var(\hat{y}_0) = \hat{\sigma}^2 X_0^T (X^T X)^{-1} X_0$$

Por consiguiente un intervalo de confianza de $100 * (1 - \alpha)$ por ciento de la respuesta media en el punto $x_1 = 1280\ oz$ y $x_2 = 3\ dias$ es

$$\hat{y}_0 - t_{\alpha/2,\ n-k-1}\sqrt{\hat{\sigma}^2 X_0^T (X^T X)^{-1} X_0} \leq E(\hat{y}_0|X_0) \leq \hat{y}_0 - t_{1-\alpha/2,\ n-k-1}\sqrt{\hat{\sigma}^2 X_0^T (X^T X)^{-1} X_0} \quad 4.3.3$$

Empleando el software stata y ejecutando el comando **margins** se obtiene el intervalo de confianza

```
margins, at (peso= (128) edad= (3))
```

	coeficientes	error estándar	t	P>|t|	[95% Conf. Intervalo]	
constante	87.188	.714	121.7	0.000	85.6437	88.7323

En la tabla anterior con un marco de color verde se señala la estimación puntual de la respuesta media, mientras que en el marco de color morado se señala la estimación mediante un intervalo de confianza.

Para un modelo de regresión lineal con k variables el vector de los valores de las variables regresoras es:

$$X_0 = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}$$

Para encontrar el intervalo de confianza para el caso del vector anterior también se aplica la ecuación 4.3.3.

BITMONEY

Modelo lineal con una interacción

El modelo que se ha planteado hasta el momento es de la siguiente forma:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Sin embargo, se puede establecer el siguiente modelo de regresión lineal:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 * x_2 + \varepsilon$$

Un modelo de regresión lineal será lineal siempre y cuando sea lineal en sus parámetros. Ejemplos:

$$y = \alpha + \beta_1 x_1^2 + \beta_2 x_2 + \varepsilon \quad es\ un\ modelo\ de\ regresión\ lineal$$

Los siguientes no corresponden a modelos de regresión lineal simple:

$$y = \alpha + \beta_1 x_1 + \frac{\beta_1}{\beta_2} x_2 + \varepsilon$$

$$y = \alpha + \beta_1 x_1 + \beta_2^2 x_2 + \varepsilon$$

Regresando al ejemplo que hemos venido desarrollando un modelo con interacción se tiene:

$$\hat{y}_i = a + b_1 x_1 + b_2 x_2 + b_3 x_1 * x_2$$

```
regress spb peso edad
```

	Suma de cuadrados	Grados de libertad	MS
Regresión	618.18	3	206.06
Residual	52.75	12	4.39
Total	670.94	15	

No. observaciones=16

$F(2,13)$=46.87

P valor=0.0000

R-cuadrado=0.9241

R-ajustado=0.9017

	coeficientes	error estándar	t	P>|t|	[95% Conf. Intervalo]	
peso	0.55	0.173	3.17	0.008	0.1727	0.9297
edad	21.28	6.223	3.42	0.005	7.7271	34.8474
Peso*edad	-0.128	0.051	-2.49	0.029	-0.24061	-0.0157
Constante	2.55	20.837	0.12	0.905	-42.85	47.953

Si vemos el R ajustado de este último modelo es más alto que el modelo que no tiene interacción, además de que resultan significativos todos los coeficientes con un nivel significancia del 0.05. Por lo que también este último es también se podría considerar.

BITMONEY

Posibles problemas en regresión lineal múltiple

A continuación, vamos a tratar los problemas más importantes que se pueden presentar cuando analizamos modelos de regresión múltiples.

Los dos puntos que se refiere a continuación no es un problema de violación de los supuestos; simplemente es una situación que puede ocasionar problemas en las inferencias en el modelo de regresión.

1) Multicolinealidad

Este problema se suscita cuando las variables predictoras están muy correlacionadas entre sí. Esta situación impide que se puedan medir aisladamente los efectos de cada una de ellas y su contribución en la ecuación de regresión como predictora de la variable de respuesta y.

En estos casos, los estimadores presentan grandes varianzas y, a menudo, ocultan contribuciones importantes de las variables predictoras. Por este motivo, se debe tener mucho cuidado al elegir las variables predictoras y no agregar variables en la ecuación por el solo hecho de que se han medido.

La multicolinealidad puede identificarse estudiando las correlaciones entre las variables predictoras por medio del coeficiente de correlación lineal y calculando un estadístico denominado VIF (**factor de inflación de la varianza**)

Correlaciones entre las variables.

La correlación entre las variables predictoras, en este caso, la variable peso y edad es de 0.10, de manera que no hay indicios que haga suponer que el modelo $\hat{y}_i = 53.45 + 0.126 * x_1 + 5.89 * x_2$ tenga problemas de multicolinealidad.

En stata el comando para encontrar la correlación entre variables es

```
cor peso edad
```

Estimación del VIF

El VIF es una medida comúnmente empleada en estadística para saber si el modelo de regresión lineal múltiple padece de multicolinealidad. Valores pequeños del VIF, valores menores a 10, no indican multicolinealidad, por el contrario, valores del VIF mayores a 10 indica problema de multicolinelidad entre las variables.

- $VIF \leq 1$ las variables regresoras no están correlacionados
- $1 < VIF \leq 5$ las variables regresoras están moderadamente correlacionados
- $FIV\ está\ entre\ 5\ y\ 10$ Las variables regresoras están altamente correlacionadas

Haciendo uso del software stata se verifica si el siguiente modelo de regresión tiene problemas de multicolinealidad

$$\hat{y}_i = 53.45 + 0.126 * x_1 + 5.89 * x_2$$

Ejecutando el código **vif** se tiene:

```
vif
```

Variable	VIF
edad	1.01
peso	1.01

El vif es casi 1 por lo que el modelo no tiene problemas de multicolinealidad.

2) Error de especificación

Se comete un error de especificación cuando se establece una dependencia errónea de la variable de respuesta con las variables predictoras.

Este problema ocurre cuando omitimos variables predictoras importantes, introducimos variables innecesarias o suponemos que existe una relación lineal cuando en realidad la relación es curvilínea.

Cuando olvidamos incluir en el modelo variables importantes, la consecuencia suele ser la obtención de estimadores sesgados y varianzas de los estimadores más grandes.

Cuando incluimos variables innecesarias, ya vimos que se puede producir un efecto de multicolinealidad si éstas variables están muy correlacionadas entre sí. Suponer una relación lineal cuando no lo es, afecta mucho a la predicción de la variable de respuesta sobre todo si la misma se debe realizar fuera de su rango de variación.

4.4 Selección de variables

En muchas situaciones se dispone de un conjunto grande de posibles variables regresoras, una primera pregunta es saber si todas las variables deben estar en el modelo de regresión y, en caso negativo, se quiere saber qué variables deben entrar y qué variables no deben estar en el modelo de regresión.

Intuitivamente parece bueno introducir en el modelo todas las variables regresoras significativas (según el contraste individual de la t) al ajustar el modelo con todas las variables posibles. Pero este procedimiento no es adecuado porque en la varianza del modelo influye el número de variables. Además, puede haber problemas de multicolinealidad cuando hay muchas variables regresoras.

Para responder a estas preguntas se dispone de diferentes procedimientos estadísticos. Bajo la hipótesis de que la relación entre las variables regresoras y la variable respuesta es lineal existen procedimientos "paso a paso" (o stepwise) que permiten elegir el subconjunto de variables regresoras que deben estar en el modelo. Estos algoritmos se presentan en esta sección.

•"**Eliminación progresiva" ("Backward Stepwise Regression").** Este procedimiento parte del modelo de regresión con todas las variables regresoras y en cada etapa se elimina la variable menos influyente según el contraste individual de la t (o de la F) hasta una cierta regla de parada. El procedimiento de eliminación progresiva tiene los inconvenientes de necesitar mucha capacidad de cálculo si k es grande y llevar a problemas de multicolinealidad si las variables están relacionadas. Tiene la ventaja de no eliminar variables significativas.

En stata este tipo de selección de variable se realiza ejecutando el siguiente código.

```
stepwise, pr(0.05):regress spb peso edad
```

La salida es la siguiente:

BITMONEY

Se inició con todas las variables regresoras (peso y edad) y aquellas que tuvieran un p valor mayor a 0.05 se sacaron del modelo.

	Suma de cuadrados	Grados de libertad	MS
Regresión	591.03	2	$^{591.03}/_{2} = 295.52$
Residual	79.90	$16 - 3 - 1 = 13$	$^{79.90}/_{(13)} = 6.15$
Total	670.94	$16 - 1 = 15$	

	coeficientes	error estándar	t	P>|t|	[95% Conf. Intervalo]	
peso	0.126	0.034	3.66	0.003	0.051	0.199
edad	5.89	0.68	8.66	0.000	4.418	7.357
constante	53.45	4.53	11.79	0.000	43.66	63.241

Todas tienen bastante influencia en la regresión, motivo por el cual ninguna se eliminó.

•**"Introducción progresiva" ("Fordward Stepwise Regression").** Este algoritmo funciona de forma inversa que el anterior, parte del modelo sin ninguna variable regresora y en cada etapa se introduce la más significativa hasta una cierta regla de parada. El procedimiento de introducción progresiva tiene la ventaja respecto al anterior de necesitar menos cálculo, pero presenta dos graves inconvenientes, el primero, que pueden aparecer errores de especificación porque las variables introducidas permanecen en el modelo, aunque el algoritmo en pasos sucesivos introduzca nuevas variables que aportan la información de las primeras. Este algoritmo también falla si el contraste conjunto es significativo pero los individuales no lo son, ya que no introduce variables regresoras.

En stata este tipo de selección de variable se realiza ejecutando el siguiente código.

```
stepwise, pe(0.05):regress spb peso edad
```

La salida es la misma que proporciona **backward regression**

En este caso se fueron incorporando variables que tuvieran un p valor menor a 0.05. Por lo tanto, las dos variables son importantes para el modelo de regresión.

4.5 Verificación de los supuestos

Los supuestos del modelo de regresión lineal múltiple son exactamente los mismo para que el del modelo de regresión lineal simple.

Aquí también se hace uso del residual estandarizado cuya ecuación es:

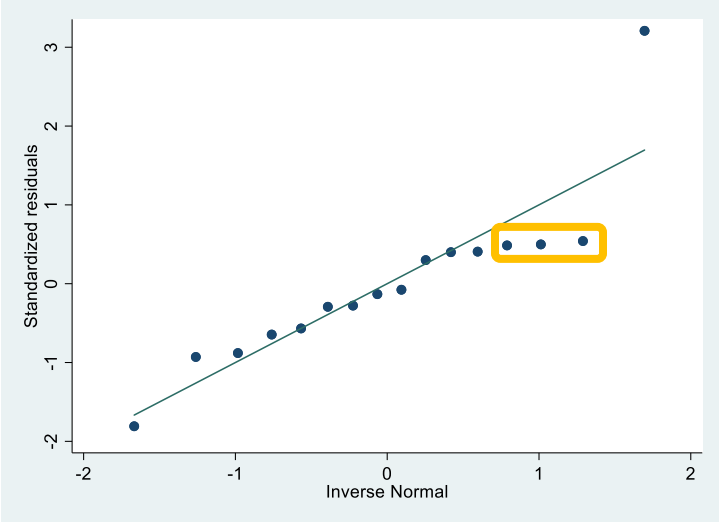
$$d_i = \frac{e_i}{\sqrt{MS_e}}$$

$$e_i = y_i - \hat{y}_i \quad i = 1, 2, 3, \dots, n$$

Analizando los residuales se comprueba que:

a) si la distribución de e es aproximadamente Normal.
b) si su variabilidad es constante.
c) si existen observaciones atípicas (outliers).
d) si los residuales o errores aleatorios se distribuyen independientemente.

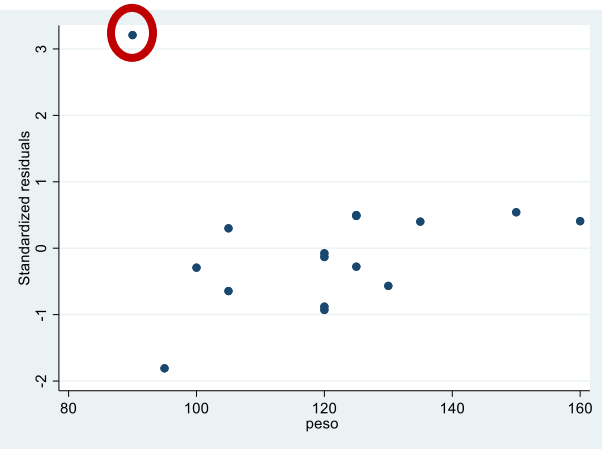
Supuesto de normalidad de los errores

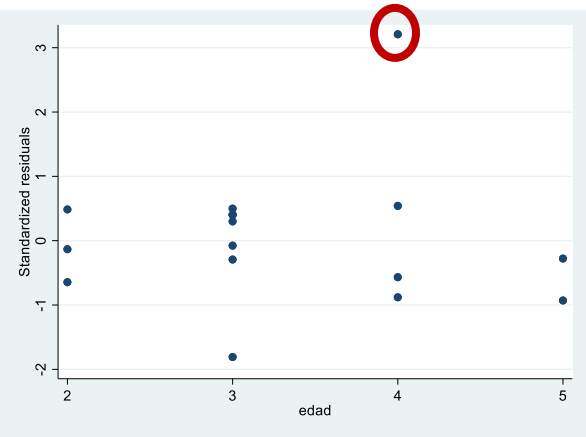


Con base al gráfico de probabilidad se alcanza a apreciar que existe un valor outlier, además, los residuales pegados al lado derecho de la gráfica parecen desviarse se la línea recta. En este caso diremos que los residuales tienen una dudosa distribución normal.

BITMONEY

Varianza constante





Con base a los gráficos anteriores se aprecia una varianza constante en de los residuales a lo largo de las variables predictoras. Sin embargo, es fácil apreciar un valor outlier que se distingue del resto de residuales.

Resumen de los residuales

Variable	Obs	Mean	Std. Dev.	Min	Max
res_est	16	0.0144176	1.074425	-1.808236	3.208419

De los resúmenes de los residuales, se tienen valores de residuales estandarizados mayor a 3, por lo que se tienen valores atípicos u outlier.

Residuales no correlacionados.

El valor del estadístico Durbin y Watson es: 2.214182

Como el valor del estadístico de Durbin y Watson está muy cercano a 2, entonces no hay evidencia de autocorrelación de los residuales.

Codigo para corroborar los supuestos:

```
%establece la regresión lineal múltiple

regress spb peso edad

%define los residuales estandarizados

predict res_est, rstandard

%establece el gráfico de probabilidad normal de los residuales estandarizados

%qnorm res_est

%Gráfico de los residuales estandarizados con cada una de las variables regresoras

twoway(scatter res_est peso)

twoway(scatter res_est edad)
```

BITMONEY

```
%resumen de los residuales estandarizados

sum res_est

%estadístico durbin y watson

gen time=_n

tsset time

estat dwatson
```