Regresión Lineal

En esta segunda parte de la materia nos detendremos a estudiar la *regresión*, un método de ajustes de curvas y datos utilizada en áreas de la informática y la matemática aplicada como el Machine Learning y el Análisis Numérico.

La regresión es el estudio de la dependencia entre diferentes aspectos de una situación o experimento. Esta herramienta matemática intenta determinar cúales de dichos aspectos pueden influenciar en el comportamiento del resto y de qué forma lo hacen. El análisis de regresión investiga la relación entre dos o más variables, con el fin de obtener información sobre una de ellas mediante el conocimiento de los valores de la otras y expresar los resultados de la forma más simple y concisa posible.

Si bien existen distintos tipos de regresión como la lineal, múltiple y la logística (utilizadas según la naturaleza de las variables en cuestión) enfocaremos nuestro estudio en la *regresión lineal*, que resulta ser la más comunmente utilizada y, además, el resto de métodos de regresión se construyen sobre el entendimiento del funcionamiento de la regresión lineal.

1 Regresión lineal

Comenzaremos el estudio de este modelo con un ejemplo sencillo...

Uno de los primeros usos de la regresión fue para estudiar la herencia de rasgos y características de generación en generación, como por ejemplo, la preservación de la altura entre madres e hijas.

Como en la mayoría de los problemas de regresión con una variable independiente (predictor) y una dependiente (respuesta), el análisis comienza con un gráfico. El siguiente representa un conjunto de 1375 alturas de madres de menos de 65 años y una de sus hijas mayores de 18 (Dataset: Heights.csv):

Con ayuda de un gráfico podemos obtener un primer acercamiento al caso de estudio e identificar información relevante a simple vista.

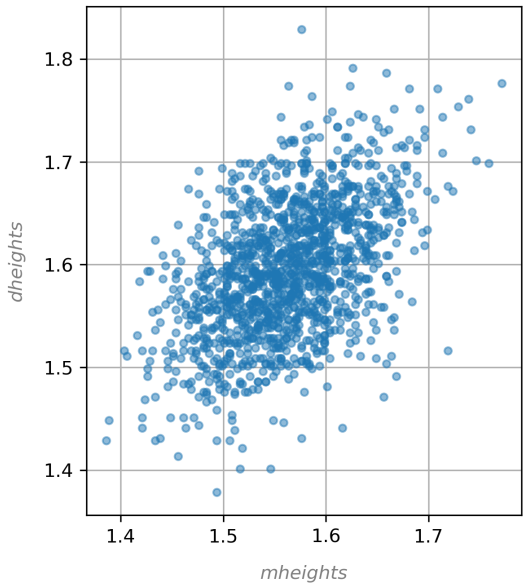


Figure 1: Gráfico de dispersión de las alturas (en metros) de las madres y las hijas

Dado que nuestro interés es determinar la herencia de altura de las madres (*mheights*) hacia las hijas (*dheights*), en la Figura 1 graficamos las alturas de las hijas contra las de las madres con estas últimas como variable independiente o predictora. La dispersión de los datos nos permite observar rápidamente que el rango de alturas aparenta ser el mismo tanto para las madres como para las hijas ([1.4 ; 1.8]).

En el gráfico, la dispersión de puntos aperenta ser mas o menos elíptica con su mayor eje "inclinado" hacia arriba. Si todas las hijas tuvieran la misma altura que sus madres, los puntos caerían exactamente sobre una línea recta de 45°. Pero claramente la distribución de los puntos no es del todo lineal ... Entonces, ¿cómo se distribuyen los puntos de *dheights* a medida que el valor de *mheights* cambia?

Para responder a la pregunta centrémosnos ahora en un aspecto importante de esta distribución de puntos, la *función del valor medio*, que definimos como

$$E(Y|X = x) = \textit{una función que depende del valor de } x$$

Esto se lee como "el valor esperado de la variable dependiente Y cuando la variable independiente X está fija en el valor x". Por ejemplo, en el caso de las alturas podemos pensar que la función de valor medio toma la forma de una línea recta, es decir,

$$E(dheight|mheight = x) = \beta_0 + \beta_1 x \tag{1}$$

Esta función en particular tiene dos parámetros, una ordenada al origen β_0 y una pendiente

β_1, los cuales son desconocidos (usualmente lo son). Asignándoles posibles valores podemos obtener la siguiente figura:

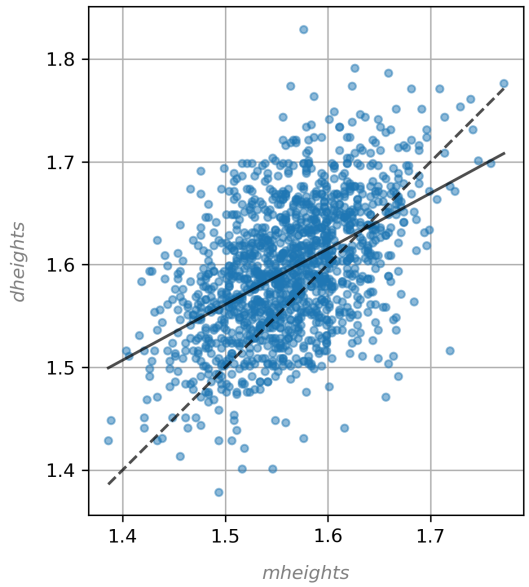


Figure 2: Los datos de alturas. La linea punteada es la función *E(dheight mheight) = mheight*, y la linea sólida es la estimada con mínimos cuadrados.

La Figura 2 muestra una recta punteada que parte de la función (1) con parámetros $\beta_0 = 0$ y $\beta_1 = 1$, es decir, una línea recta de 45°. Esta función sugiere que, en promedio, las hijas tienen la misma altura que las madres.

La segunda recta es la estimada utilizando el ***método de regresión lineal simple con mínimos cuadrados***. La recta de mínimos cuadrados tiene una pendiente positiva, significando que las madres altas tienden a tener hijas más altas que el promedio, pero más bajas que ellas, por ser la pendiente *menor* que 1. De forma similar, las madres bajas tienden a tener hijas bajas pero más altas que ellas.

Este último detalle curioso, peculiarmente, resulta ser el origen del término "*regresión*", debido a que los valores extremos de una generación (las madres) tienden a revertirse o *regresar* hacia la media poblacional en la siguiente generación (las hijas).

A continuación daremos un desarrollo teórico del método de regresión lineal *simple* y cómo éste puede explicar relaciones entre dos variables distintas.

1.1 Modelo de regresión lineal simple

Supongamos que nos convertimos en investigadores (de lo que más gusten) por un rato. Al realizar nuestros estudios nos encontraremos, en su mayoría, con situaciones o escenarios complejos, donde una variedad de factores cambian constantemente las propiedades o comportamientos de lo que estamos estudiando. ¿Por dónde comenzamos?

Lo primero que podríamos pensar es en dividir el problema en partes más simples con el fin de reducir su complejidad e intentar comprender cada una de estas partes por separado. Cada una de ellas se convierte en un nuevo problema a estudiar, pero esta vez más sencillo. Veamos un ejemplo:

Queremos comprender cómo varía el consumo de combustible a través de las distintas provincias del país. Este interrogante no parece tener una respuesta directa ya que podemos identificar una multiplicidad de factores que podrían impactar en el consumo de combustible . . . ¿La cantidad de autos influye en el consumo? ¿Cómo se ve modificado por los diferentes impuestos de las provincias? ¿Esto se ve infuenciado por los sueldos de los habitantes? ¿La cantidad de kilómetros recorridos son significativos?

Cada una de las preguntas que surgieron podría ser un problema en sí mismo. Es aquí donde intentaremos identificar los distintos factores que impactan en el sistema y cómo se relacionan entre ellos (si es que lo hacen).

En términos de la matemática, cuando queremos comprender la relación entre variables del mundo real, solemos modelizarlo mediante una *función*. En este capítulo veremos la función más simple que existe entre dos variables x e y, la lineal.

Como sabemos, una **función lineal** tiene por gráfica a una recta y una ecuación de la siguiente forma:

$$y = \beta_0 + \beta_1 x$$

donde β_0 es la **ordenada al origen** y β_1 la **pendiente**.

Este tipo de relaciones son totalmente determinísticas ya que, para un valor fijo de la **variable independiente** x, se *determina* un *único* valor para la **variable dependiente** y. Pero en la amplia mayoría de los casos las relaciones que debamos modelizar no serán determinísticas, es decir, si conocemos el valor de x no sabremos *con exactitud* el de y. Esta última variable es lo que conocemos como *variable alreatoria*, a la cual nombraremos Y.

Normalmente se realizan observaciones para varios escenarios de la variable independiente.

Sean $x_1, x_2, \ldots, x_n$ los valores de la variable independiente para las que se realizaran las observaciones y sean $_i$ e y_i, respectivamente, la variable aleatoria y el valor observado asociado con x_i. Entonces los datos se componen entonces por los n pares ordenados $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Veamos el siguiente ejemplo:

Ejemplo 1.1. (DATASET: Sight.csv)

Los problemas visuales asociados con la exposición de la luz de la pantalla de las computadoras se han vuelto un tanto comunes en estos últimos años. Algunos investigadores se han enfocado en la dirección vertical de la mirada fija como causa del cansancio e irritación del los ojos. Se sabe que esto está estrechamente relacionado con el área de la superficie ocular. Los datos representativos tomados por los investigadores se disponen en la siguiente tabla:

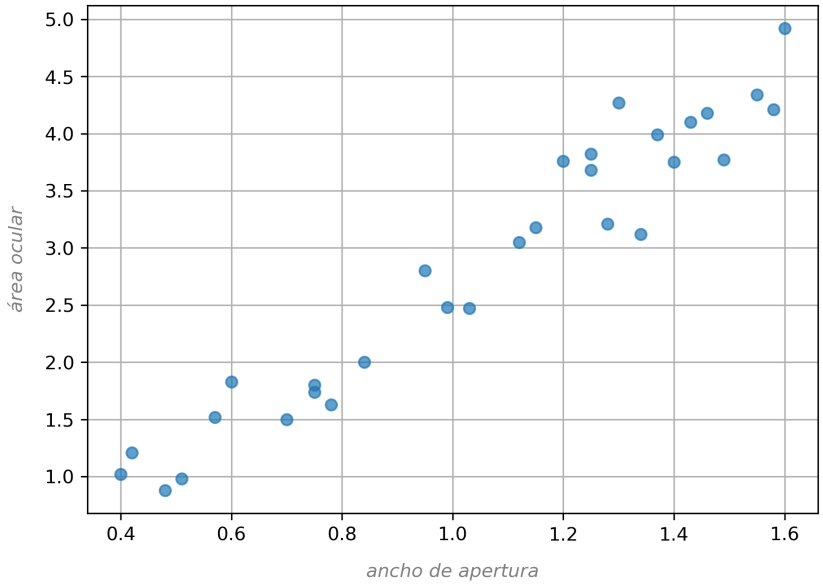
i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_i	0.4	0.42	0.48	0.51	0.57	0.60	0.70	0.75	0.75	0.78	0.84	0.95	0.99	1.03	1.12
y_i	1.02	1.21	0.88	0.98	1.52	1.83	1.50	1.80	1.74	1.63	2.0	2.8	2.48	2.47	3.05

i	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
x_i	1.15	1.20	1.25	1.25	1.28	1.30	1.34	1.37	1.40	1.43	1.46	1.49	1.55	1.58	1.60
y_i	3.18	3.76	3.68	3.82	3.21	4.27	3.12	3.99	3.75	4.10	4.18	3.77	4.34	4.21	4.92

y = *área ocular en* cm^2

x = *ancho horizontal de la apertura del ojo en cm*

Dados estos datos construiremos un gráfico de dispersión intentar comprender la relación entre las dos variables:



Como sabemos, una función no puede tomar dos valores distintos para un mismo valor fijo de x. Notemos que en la tabla $x_8 = x_9$ pero el valor de $y_8 = 1.80$ e $y_9 = 1.74$. Por lo tanto, esto nos indica que el valor de la variable y no está totalmente determinado por x.

De igual modo, es evidente que existe una relación lineal (aunque no perfecta) sustancial entre las dos variables. Pareciera que el valor de y podría ser pronosticado a partir de x encontrando una recta que esté razonablemente cerca de los puntos presentes en la gráfica.

Lo visto en el ejemplo nos mostró que el valor esperado o medio de la variable Y es una función lineal de x:

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

Pero para un valor fijo de x, Y *difiere* de su valor esperado en una cantidad aleatoria. En otras palabras, la i-ésima observación y_i usualmente no será igual al valor esperado $E(Y|X = x_i)$. Para dar cuenta de esta diferencia entre el valor observado y el dato esperado, se creó una cantidad llamada **término de error aleatorio** o **desviación aleatoria** e_i. De esta forma conseguimos la denominada **ecuación de modelo**:

Definición 1.2. Modelo de Regresión Lineal Simple

Existen parámetros β_0, β_1 y σ^2 de tal suerte que con cualquier valor fijo de la variable independiente x, la variable dependiente está relacionada con x por medio de la ***ecuación de modelo***

$$Y = \beta_0 + \beta_1 x + e$$

La cantidad e en la ecuación de modelo es una variable aleatoria, que se supone está normalmente distribuida con $E(e) = 0$ y $V(e) = \sigma^2$

Notemos que sin la presencia del error e, cualquier par observado (x, y) correspondería a un punto que cae exactamente sobre la recta $y = \beta_0 + \beta_1 x$, llamada **línea de regresión verdadera**.

Al incluir el término de error aleatorio es posible encontrar puntos (x, y) por encima de la línea de regresión verdadera (cuando $e > 0$) o por debajo (cuando $e < 0$). Esta variación en el valor del error se le atribuye precisamente al parámetro σ^2 que suele ser mayor a 0 y representa la varianza de la variable aleatoria e.

De esta forma, dados n puntos $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ observados, estos se dispersarán

en torno a la línea de regresión verdadera:

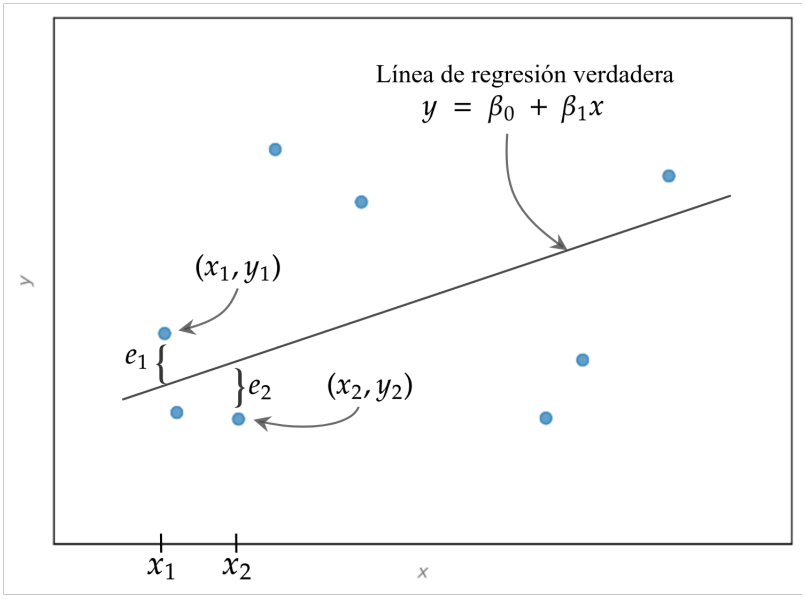


Figure 3: Puntos correspondientes a observaciones del modelo de regresión lineal

Los errores e_i dependen de parámetros desconocidos por lo que no son cantidades observables. Como se mencionó, son variables aleatorias que corresponden a la distancia vertical entre el punto y_i y la línea de regresión verdadera.

Con respecto a los errores hacemos dos suposiciones importantes. Primero, asumimos que están normalmente distribuidos con $E(e_i) = 0$ y $V(e_i) = \sigma^2$, es decir, el valor esperado del error e_i será 0 y tendrá desviación estándar σ. La segunda suposición es que todos los errores son independientes, lo que significa que el valor del error para un caso observado no afecta en el valor de otro.

Entonces, teniendo en cuenta estas suposiciones y que el valor de la variable aleatoria Y está afectado por el del *error aleatorio* e, podemos decir que éste determina las propiedades de Y. Veamos que en general

$$Y = \beta_0 + \beta_1 x + e$$

Luego, para un valor fijo de x, Y es una variable aleatoria tal que:

$$\begin{aligned} E(Y|X = x) = &\quad E(\beta_0 + \beta_1 x + e) \overset{(1)}{=} \beta_0 + \beta_1 x + E(e) \overset{(2)}{=} \beta_0 + \beta_1 x \\ V(Y|X = x) = &\quad V(\beta_0 + \beta_1 x + e) \overset{(1)}{=} V(e) \overset{(3)}{=} \sigma^2 \end{aligned}$$

(1) Los términos β_0 y $\beta_1 x$ son constantes ya que el valor de x está fijo

(2) $E(e) = 0$

(3) $V(e) = \sigma^2$

Notemos que el *valor medio* de Y, en lugar de Y misma, es una función lineal de x. La línea de regresión verdadera $y = \beta_0 + \beta_1 x$ es por consiguiente la **línea de valores medios**. La pendiente β_1 de la línea de regresión verdadera se interpreta como el cambio esperado de Y cuando el valor de x incrementa en una unidad. La segunda relación manifiesta que la cantidad de variabilidad en la distribución de valores Y es la misma con cada valor diferente de x, es decir, se mantiene constante (homogeneidad de varianza).

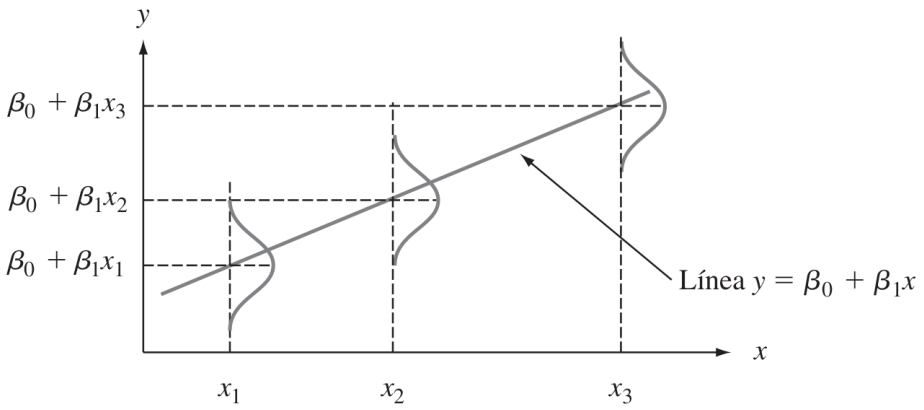


Figure 4: Distribución de Y con diferentes valores de x

Por último, con x fija, como Y es la suma de una constante $\beta_0 + \beta_1 x$ y una variable aleatoria normalmente distribuída e, Y "hereda" esta distribución. Además, debido a la independencia entre los valores de e_i, los datos observados y_i también serán independientes.

1.2 Método de mínimos cuadrados

A partir de esta sección se supondrá que las variables x e y están relacionadas de acuerdo con el modelo regresión lineal simple. Entonces, recordando lo visto, cuando se fija el valor de x_i, el *valor esperado* de la variable aleatoria Y_i resulta ser una función lineal de la variable x de la forma

$$E(Y_i|X = x_i) = \beta_0 + \beta_1 x_i$$

Observando esta ecuación podemos preguntarnos ¿De dónde sacamos β_0 y β_1? ¿Cuáles son sus valores? Un investigador casi nunca conocerá los valores de estos parámetros, ya que de hacerlo no existiría el modelo que estamos estudiando en estos momentos (regresión lienal). Conoceríamos la linea de regresión verdadera y no más habría trabajo que hacer.

Pero no estamos con las manos vacías. Tenemos a nuestra disposición un conjunto de datos compuesto por n pares observados $(x_1, y_1), (x_n, y_n), \ldots, (x_n, y_n)$ con los cuales podemos *estimar*

los valores de los parámtros β_0 y β_1. Ahora, intentemos comprender intuitivamente cómo vamos a estimarlos.

De acuerdo con el modelo de regresión lineal simple, los puntos observados estarán distribuidos en torno a la línea de regresión verdadera de una manera aleatoria. Para ilustrarlo elijamos 2 valores, en primer instancia, para cada uno de los parámetros β_0 y β_1 para **ajustar** la recta a los datos del Ejemplo 1.1.

Eligiendo primeramente $\beta_0 = -0.4$ y $\beta_1 = 3.08$ obtenemos una función $f(x)$ graficada en color naranja en la Figura 5. Por otro lado, tomando $\beta_0 = 1$ y $\beta_1 = 0.5$ tenemos la función $g(x)$ en verde:

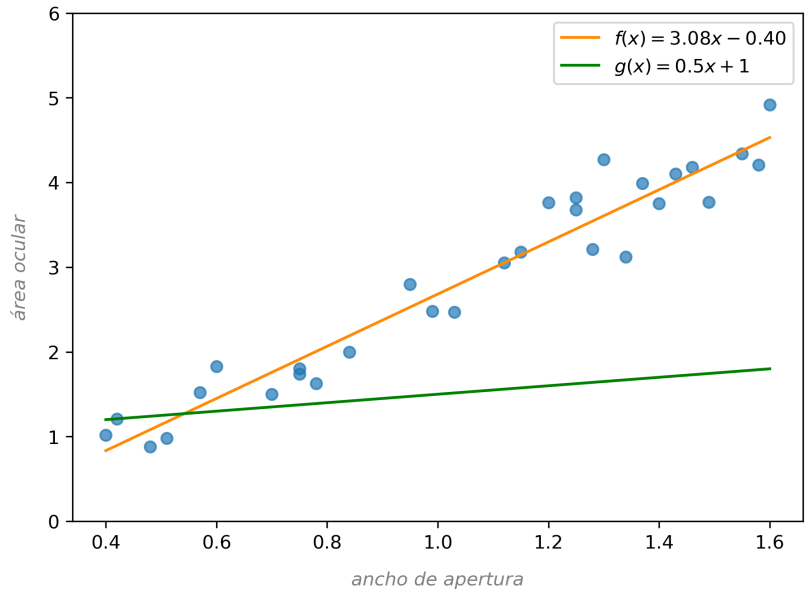


Figure 5: Ajuste de los datos del archivo Sight.csv

Ahora surge la pregunta, ¿cuál se **ajusta** mejor a los datos? En otras palabras, queremos identificar cuál de las funciones explica mejor la relación existente entre las variables (*area* y *ancho* en este caso).

Es bastante evidente que la función $g(x)$ no resulta una estimación razonable de la línea verdadera ya que los datos se encuentran muy lejos de esta recta. Esto nos lleva a afirmar que $f(x)$ resulta ser un mejor ajuste para este conjunto de puntos. Pero... ¿qué acabamos de medir para hacer dicha afirmación? Esta pregunta es la motivación del **método de mínimos cuadrados**.

Observemos que lo que hicimos fue identificar cuál recta quedaba más *cerca* de los datos, por lo que utilizamos una noción de *distancia* entre cada punto y la recta estimada. Tal es así

que el método de mínimos cuadrados determina que la recta de *mejor ajuste* es aquella tal que las distancias verticales de los puntos observados hasta la recta sean las mínimas posibles. La medida que tomará el método es la suma de los cuadrados de estas desviaciones.

Definición 1.3. Método de Mínimos Cuadrados

La desviación vertical del punto (x_i, y_i) *con respecto a la línea* $y = \beta_0 + \beta_1 x$ *es*

$$\textit{la altura del punto - altura de la línea} = y_i - (\beta_0 + \beta_1 x)$$

La suma de las desviaciones verticales al cuadrado de los puntos $(x_1, y_1), \ldots, (x_n, y_n)$ *a la línea es entonces*

$$f(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

Las estimaciones puntuales de β_0 *y* β_1*, denotadas por* $\hat{\beta}_0$ *y* $\hat{\beta}_1$ *llamadas* ***estimaciones de mínimos cuadrados****, son aquellos valores que reducen al mínimo a* $f(b_0, b_1)$*. Es decir,* $\hat{\beta}_0$ *y* $\hat{\beta}_1$ *son tales que*

$$f(\hat{\beta}_0, \hat{\beta}_1) \leq f(b_0, b_1) \qquad \forall\ b_0, b_1$$

La ***línea de regresión estimada*** *o* ***línea de mínimos cuadrados*** *es entonces la línea cuya ecuación es*

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

La siguiente figura toma el mismo ejemplo de la Figura 5, y muestra gráficamente el significado de las desviaciones verticales de cada punto observado con respecto a la recta en cuestión:

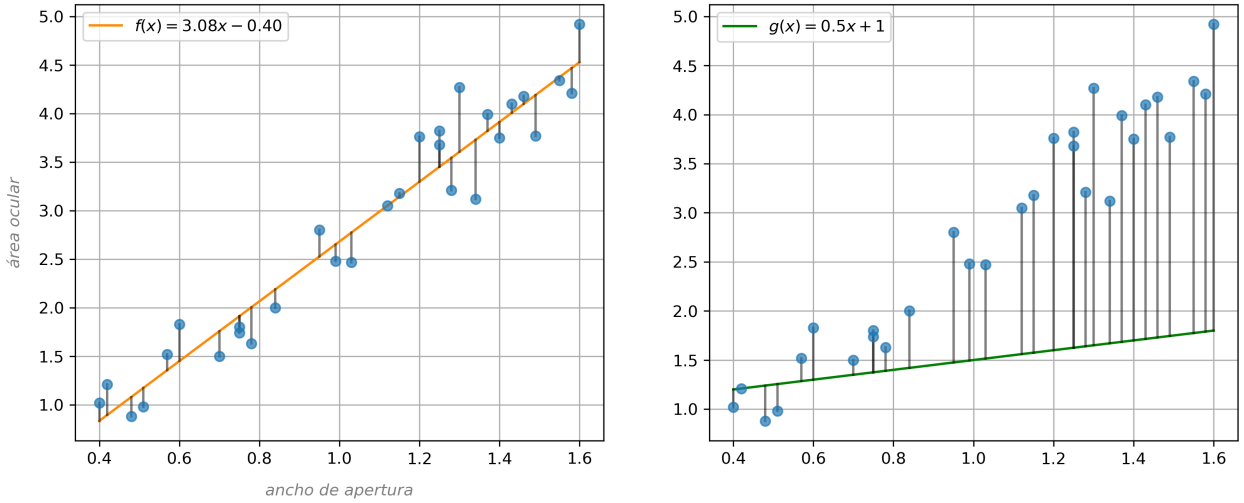


Figure 6: Desviaciones verticales de los puntos observados con respecto a las rectas

Luego, el método de mínimos cuadrados toma cada una de dichas desviaciones, las eleva al cuadrado y las suma. Aquella recta que minimice la suma será la que se considerará la *recta de ajuste*. Claramente éste no es el caso de $g(x)$, ubicada en el gráfico de la derecha de la figura 6, cuya suma de desviaciones parece ser muy grande.

Para minimizar $f(b_0, b_1)$ haremos uso de los contenidos vistos en la primera parte de la materia, donde aprendimos a calular máximos y mínimos de funciones de varias variables mediante derivadas parciales. De esta forma, si calculamos las derivadas parciales de $f(b_0, b_1)$ con respecto a b_0 y b_1, y luego las igualamos a 0, podremos hallar los valores minimizantes de estos parámetros:

$$\frac{\partial f(b_0, b_1)}{\partial b_0} = \sum_{i=1}^{n} -2(y_i - b_0 - b_1 x_i) = 0$$

$$\frac{\partial f(b_0, b_1)}{\partial b_1} = \sum_{i=1}^{n} -2x_i(y_i - b_0 - b_1 x_i) = 0$$

Realizando algunas operaciones algebráicas y reordenando términos se obtiene el siguiente sistema de ecuaciones, llamadas **ecuaciones normales**:

$$nb_0 + b_1 \left(\sum_{i=1}^{n} x_i \right) = \sum_{i=1}^{n} y_i$$

$$b_0 \left(\sum_{i=1}^{n} x_i \right) + b_1 \left(\sum_{i=1}^{n} x_i^2 \right) = \sum_{i=1}^{n} x_i y_i$$

La estimación de mínimos cuadrados de la pendiente β_1 de la recta de regresión verdadera será entonces

$$\hat{\beta}_1 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}$$

Otra forma de expresar a S_{xy} y S_{xy} es:

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\sum x_i \sum y_i}{n} \qquad S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum x_i)^2}{n}$$

La estimación de mínimos cuadrados de la ordenada al origen β_0 es:

$$\hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \overline{y} - \hat{\beta}_1 \overline{x}$$

1.3 Estimación de σ^2

El parámetro σ^2 determina la cantidad de variabilidad en el modelo de regresión. Un valor grande de σ^2 conducirá a (x_i, y_i) observados que están bastante dispersos en torno a la línea de regresión verdadera, mientras que σ^2 sea pequeña los puntos observados tenderán a quedar cerca de la línea verdadera.

Este parámetro podrá ser estimado una vez que hayamos calculado el modelo de regresión lineal, es decir la recta de la forma $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Si reemplazamos sucesivamente los valores x_i en la ecuación, obtendremos los denominados **valores ajustados** $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$.

Definición 1.4. *Las desviaciones verticales entre los valores observados y valores ajustados*

$$\hat{e}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \qquad i = 1, \ldots, n$$

recibirán el nombre de ***residuos***

Los residuos son cantidades que determinan directamente a la variabilidad del modelo. Notemos que representan los errores que se generan al ajustar una recta a puntos que no siguen una relación totalmente determinística. La Figura 6 muestra esta noción de residuos. A pesar de que no lo mencionamos en su entonces, la función $f(x)$ del gráfico de la izquierda de esta figura fue calculada mediante el método de mínimos cuadrados, por lo que resulta ser la recta que

minimiza las magnitudes de los errores (o residuos).

Teniendo esto en consideración, definiremos la estimación de la variabilidad σ^2 de la siguiente forma:

Definición 1.5. *La* ***suma de cuadrados de error o residuales*** *denotada por SCE, es*

$$SCE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

Luego, el estimador insesgado de σ^2 *es*

$$\hat{\sigma}^2 = \frac{SCE}{n-2}$$

Observemos que para calcular la suma de cuadrados residuales debemos realizar operaciones aritméticas tediosas y evaluar la ecuacion de la recta para cada valor de x_i, por lo que a continuación presentamos una forma alternativa para dicha cantidad:

$$SCE = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$$

donde

$$S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2$$

1.4 Coeficiente de determinación

En esta sección presentaremos una métrica con la que podremos medir cúan bueno fue el trabajo realizado por nuestro modelo de regresión lineal, es decir, nos indicará cúan bien se ajustó la recta a los datos observados.

Como ya sabemos, la varianza de una variable aleatoria nos determina el grado de dispersión de los valores con respecto a su media o valor esperado. Si tomamos los datos del Ejemplo 1.1 podemos interpretar esta variabilidad de los datos al graficar los puntos junto con la recta horizontal $f(x) = \overline{y}$

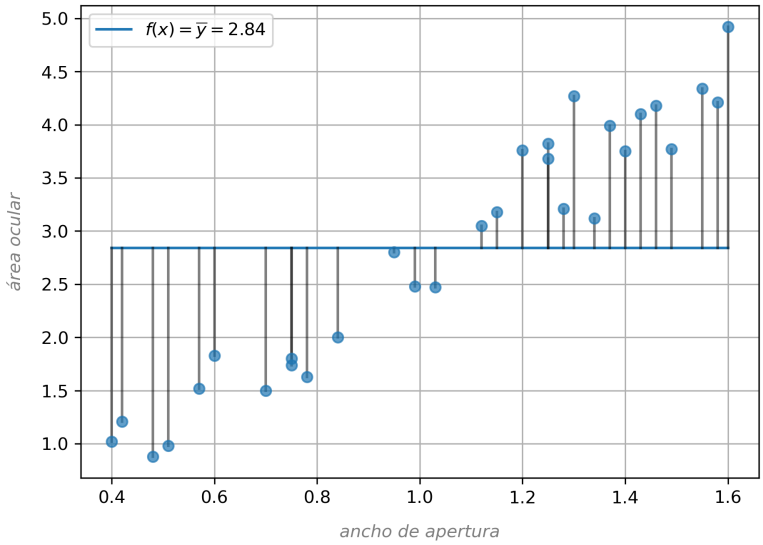


Figure 7: Desviaciones verticales de los puntos observados con respecto a la media muestral

Si tomamos cada uno de los desvios de esta figura, los elevamos al cuadrado y los sumamos obtenemos la denominada **suma total de cuadrados**. Simbólicamente

$$\sum_{i=1}^{30}(y_i - \overline{y})^2$$

que no es otra cosa que la cantidad S_{yy} ya mencionada. Esta describe la variabilidad **total** observada (ignorando a la variable pronosticadora x) de la variable y.

Cuando agregamos la variable independiente x podemos obtener la suma de los errores al cuadrado SCE que describe la variabilidad **no explicada** por el modelo de regresión. Veamos a qué nos referimos con variación explicada o no explicada por el modelo.

La explicaión dada por un modelo lineal de la variación de una variable puede ser interpretada con la siguiente pregunta: *"Si trazo las desviaciones de los puntos observados a la recta horizontal de la media y luego reemplazo dichos puntos por la recta de ajuste y vuelvo a trazar las desviaciones, ¿éstas últimas desviaciones se parecen a las primeras?"*. Si lo hacen, decimos que el modelo *explica* gran parte o la totalidad de la variación de la variable. En caso contrario el modelo *no explica* la variabilidad. Por ejemplo, la siguiente es una figura con un modelo que explica la totalidad de la variación de y:

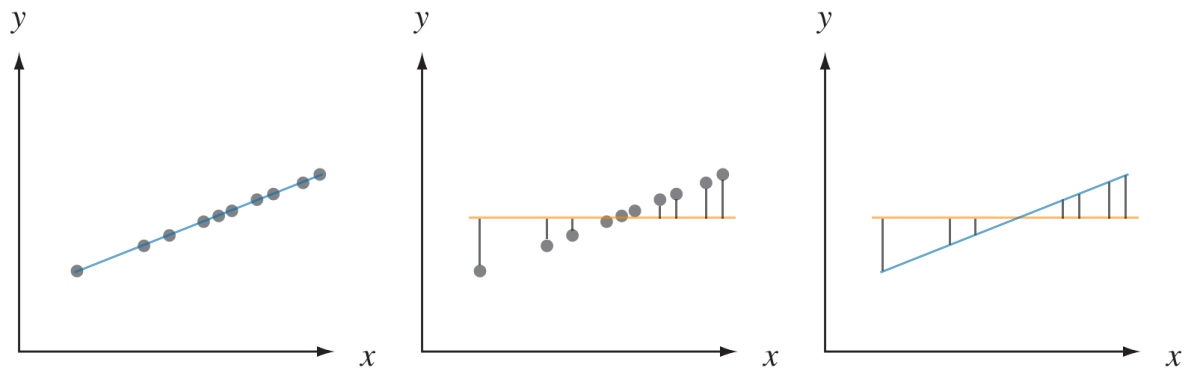
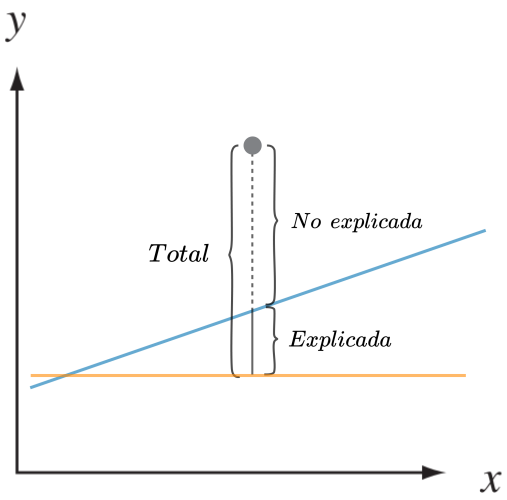


Figure 8: Modelo que explica la totalidad de la variación.

Entonces, dada una recta estimada, se define lo siguiente con respecto a la variabilidad:



Si observamos los casos de la Figura 9 podemos observar que en el gráfico *(a)* todos los puntos caen en una línea recta. De esta forma podemos decir que toda la variación (100%) de la variable y puede ser atribuida al hecho de que x e y están linealmente relacionadas en combinación con la variación de x. Es decir, el modelo de regresión lineal ajustaría a la perfección los datos. Por otro lado, los puntos de la gráfica *(b)* no caen exactamente en una línea pero sí tienen una tendencia lineal, por lo que si trazáramos la recta de mínimos cuadrados, las desviaciones de los puntos a esta recta serían pequeñas. Es razonable conluir en este caso que gran parte de la variación de y observada puede ser atribuida a la relación lineal aproximada. Finalmente, cuando la gráfica es como la figura *(c)*, existe una variación sustancial en torno a la recta estimada con respecto a la variación total de y, así que el modelo no explica la variación de y relacionando y con x.

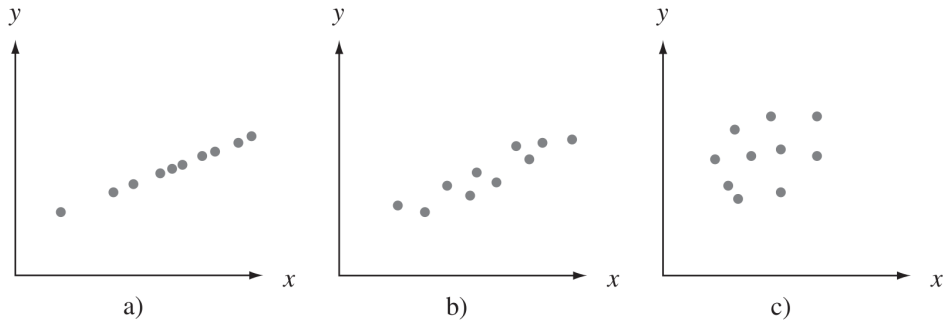


Figure 9: Utilización del modelo para explicar la variación de y : a) datos con los cuales toda la variación es explicada; b) datos con los cuales la mayor parte de la variación es explicada; c) datos con los cuales poca variación es explicada.

Una vez comprendida esta idea de explicación de la variabilidad, podemos proceder a construir una métrica que mida la calidad del ajuste que obtenemos con el método de mínimos cuadrados. Tomaremos las siguientes 3 cantidades:

- **Suma total de cuadrados**

$$STC = S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

- **Suma de cuadrados de error**

$$SCE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- **Suma de cuadrados debido a la regresión**

$$SCR = \sum_{i=1}^{n} (\hat{y} - \overline{y})^2$$

La variación total de la variable y denotada por STC podemos escribirla como la suma de la variabilidad **explicada** SCR y la **no explicada** SCE. Luego

$$STC = SCR + SCE$$

Si movemos un poco los términos podemos obtener

$$SCR = STC - SCE$$

Finalmente, dividiendo a ambos miembros por la variabilidad total (STC) obtenemos la proporción o porcentaje de explicación de la variación por parte del modelo:

$$\frac{SCR}{STC} = 1 - \frac{SCE}{STC}$$

Esta proporción es de tal significancia que recibe su propio nombre y será la métrica que buscábamos.

Definición 1.6. *El **coeficiente de determinación**, denotado por R^2, está dado por*

$$R^2 = 1 - \frac{SCE}{STC}$$

Se interpreta como la proporción de variación y observada que puede ser explicada por el modelo de regresión lineal simple (atribuida a una relación lineal aproximada entre y y x).

Debemos destacar que la suma de desviaciones al cuadrado con respecto a la línea de mínimos cuadrados (SCE) es más pequeña que la suma de desviaciones al cuadrado con respecto a *cualquier* otra línea. Esto se debe a que dicha suma se minimizó mediante el método de mínimos cuadrados. De esta forma, siempre $SCE < STC$ a menos que la recta de mínimos cuadrados sea la propia recta horizontal de la media, en cuyo caso $SCE = STC$.

La razón (SCE/STC) representa la proporción de variabilidad total que no puede ser explicada por el modelo de regresión lineal, siendo un valor entre 0 y 1. Luego $1 - (SCE/STC)$ también será un número entre 0 y 1.

Mientras más alto es el valor de R^2, mejor será nuestro modelo de regrsión lineal simple al explicar la variación de y.

1.5 Inferencias estadísticas

En esta sección comenzaremos presentando algunas propiedades sobre los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$, donde simplemente nos centraremos en los resultados y no en el camino con el que se arriba a estos. Aquellos interesados pueden profundizar en este tema con bibliografía mas extensa que excede a los contenidos de la materia.

Mencionemos algunas propiedades sobre los estimadores puntuales $\hat{\beta}_0$ y $\hat{\beta}_1$:

- Tanto $\hat{\beta}_0$ como $\hat{\beta}_1$ son estimadores *insesgados*, es decir:

$$E(\hat{\beta}_0) = \beta_0 \qquad E(\hat{\beta}_1) = \beta_1$$

- Las varianzas de los estimadores son de la forma

$$V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} \qquad V(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right)$$

- Tanto $\hat{\beta}_0$ como $\hat{\beta}_1$ estan normalmente distribuidas con las varianzas y valores medios mencionados en los puntos anteriores.

1.5.1 Inferencias sobre β_1

Muchos procedimientos inferenciales previamente discutidos en Matemática 3 se basaron en estandarizar un estimador restando primero su valor esperado y luego dividiéndolo entre su desviación estándar estimada. Un resultado similar en este caso permite dar más inferencias sobre β_1.

Teorema 1.7. *La suposición del modelo de regresión lineal simple implica que la variable estándar (o estadístico)*

$$T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}}$$

tiene una distribución t Student con $n-2$ grados de libertad.

Intervalo de confianza para β_1

Partiendo con el enunciado de probabilidad

$$P\left(-t_{\frac{\alpha}{2},n-2} < \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} < t_{\frac{\alpha}{2},n-2} \right) = 1 - \alpha$$

y manipulando las desigualdades para aislar β_1, da la forma del intervalo de confianza

Definición 1.8. *Un* ***intervalo de confianza de*** $100(1-\alpha)\%$ ***para la pendiente*** β_1 *de la línea de regresión verdadera es:*

$$\left[\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}\ ;\ \hat{\beta}_1 + t_{\frac{\alpha}{2},n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}\right]$$

Este intervalo tiene la misma forma que muchos de los vistos en Matemática 3. Está centrado en la estimación puntual del parámtero y el ancho depende del nivel de confianza deseado y de la cantidad de variabilidad del estimador $\hat{\beta}_1$

Tests de Hipótesis sobre β_1

Como se vio en los preliminares, todo test o prueba de hipótesis comienza fijando la *hipótesis nula*, que representa aquella aseveración que "favoreceremos" hasta que se demuestre que la hipótesis alternativa es la correcta.

En este caso deseamos probar la hipótesis de que la pendiente β_1 es igual a una constante, por ejemplo θ. Entonces supongamos las hipótesis:

$$H_0:\ \beta_1 = \theta \qquad\qquad H_a:\ \beta_1 \neq \theta$$

Luego, el siguiente paso en una prueba es la elección del estadístico de prueba. Utilizaremos uno similar al presentado en el Teorema 1.7:

$$T = \frac{\hat{\beta}_1 - \theta}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}}$$

El cual, bajo la hipótesis nula $\beta_1 = \theta$ tiene distribución student con $n-2$ grados de libertad.

Por último hay que fijar la regla de decisión:

$$Regla: \begin{cases} \text{rechazar} \quad H_0 \quad si \quad |T| > t_{\frac{\sigma}{2},\ n-2} \\ \text{aceptar} \quad H_0 \quad si \quad |T| \leq t_{\frac{\sigma}{2},\ n-2} \end{cases}$$

- Si $H_a : \beta_1 > \theta$ se rechaza $H_0 : \beta_0 = \theta$ si $T > t_{\alpha,n-2}$
- Si $H_a : \beta_1 < \theta$ se rechaza $H_0 : \beta_0 = \theta$ si $T < -t_{\alpha,n-2}$

Destaquemos el caso especial cuando tomamos $H_0 : \ \beta_1 = 0$ contra $H_a : \ \beta_1 \neq 0$ ¿Qué significa esto en nuestro modelo de regresión lineal?

Notemos que aceptar $H_0 : \ \beta_1 = 0$ implica aceptar que la pendiente de la recta de regresión verdadera es nula, lo que es equivalent a concluir que no existe ninguna relación lineal entre las variables x e Y. Por otro lado, si se rechaza esta hipótesis nula, significaría que x tiene cierta importancia al explicar la variabilidad de Y. También puede indicar que le modelo lineal es adecuado, o que, aunque exista efecto lineal, pueden obtenerse mejores resultados agregando términos polinomiales de mayor grado en x.

1.5.2 Inferencias sobre β_0

Intervalo de confianza para β_0

Intervalos de confianza de nivel $(1-\alpha)$ se deducen de manera análoga a lo visto para β_1. En este caso utilizamos el estadístico

$$T = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)}}$$

El cual tiene distribución student con $n-2$ grados de libertad. Luego, el intervalo es:

$$\left[\hat{\beta}_0 - t_{\frac{\alpha}{2},n-2}\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)} \ ; \hat{\beta}_0 + t_{\frac{\alpha}{2},n-2}\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)} \ \right]$$

Tests de Hipótesis sobre β_0

De manera similar a lo visto sobre β_1, se puede realizar tests de hipótesis sobre β_0. Específicamente si tenemos las hipótesis

$$H_0 : \ \beta_0 = \gamma \qquad\qquad H_a : \ \beta_0 \neq \gamma$$

y el estadístico de prueba es

$$T = \frac{\hat{\beta}_0 - \gamma}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right)}}$$

que bajo $H_0 : \ \beta_0 = \gamma$ tiene distribución student con $n-2$ grados de libertad. Finalmente, la regla de decisión de la prueba es igual a la de β_1:

$$Regla : \begin{cases} \text{rechazar} \quad H_0 \quad si \quad |T| > t_{\frac{\sigma}{2},\ n-2} \\ \text{aceptar} \quad H_0 \quad si \quad |T| \leq t_{\frac{\sigma}{2},\ n-2} \end{cases}$$

- Si $H_a : \beta_0 > \gamma$ se rechaza $H_0 : \beta_0 = \gamma$ si $T > t_{\alpha,n-2}$
- Si $H_a : \beta_0 < \gamma$ se rechaza $H_0 : \beta_0 = \gamma$ si $T < -t_{\alpha,n-2}$

1.5.3 Intervalo de confianza para la respuesta media

Así como vimos en el inicio de la Sección 1.2, dado un punto, el valor esperado de la variable aleatoria Y_i para un valor fijo de x era:

$$E(Y_i|X = x_i) = \beta_0 + \beta_1 x_i$$

Luego, si utilizamos el método de mínimos cuadrados para hallar $\hat{\beta}_1$ y $\hat{\beta}_0$ y fijamos un valor de x, x^*, entonces $\hat{\beta}_0 + \hat{\beta}_1 x^*$ puede ser considerada como una estimación puntual del valor medio de Y o una predicción del valor Y (cuando $x = x^*$)

Como $\hat{\beta}_1$ y $\hat{\beta}_0$ varían según la muestra aleatoria utilizada (son estadísticos, es decir, variables aleatorias), luego se desprende que $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$ cambia de valor de muestra en muestra para un mismo x^* fijo, por lo que también es un estadístico.

Del mismo modo que los intervalos de confianza para β_1 y β_0 estaban basados en propiedades de la distribución de muestreo de β_1 y β_0, un intervalo de confianza para un valor y medio en regresión está basado en propiedades de la distribución de muestreo del estadístico $\hat{\beta}_0 + \hat{\beta}_1 x^*$.

Sea $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$, donde x^* es un valor fijo de x. Entonces

- La esperanza de $\hat{Y}$ es

$$E(\hat{Y}) = E(\hat{\beta}_0 + \hat{\beta}_1 x^*) = \beta_0 + \beta_1 x^*$$

De forma tal que $\hat{\beta}_0 + \hat{\beta}_1 x^*$ es un estimador puntual insesgdo de $\beta_0 + \beta_1 x^*$.

- La varianza de $\hat{Y}$ es

$$\hat{Y} = V(\hat{\beta}_0 + \hat{\beta}_1 x^*) = \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}} \right)$$

y la desviación estandar es la raíz cuadrada de esta expresión.

- $\hat{Y}$ tiene distribución normal con esperanza y varianza anteriores.

Así como los procedimientos inferenciales para β_1 y β_0 se basaron en la variable T obtenida estandarizándolos, una variable T estandarizando $\hat{\beta}_0 + \hat{\beta}_1 x^*$ conduce a un intervalo como el que veremos a continuación.

Teorema 1.9. *La variable aleatoria*

$$T = \frac{\hat{\beta}_0 + \hat{\beta}_1 x^* - (\beta_0 + \beta_1 x^*)}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}} \right)}}$$

tiene distribución student con $n - 2$ grados de libertad.

Definición 1.10. *Un **intervalo de confianza de** $100(1-\alpha)\%$ **para** $\beta_0 + \beta_1 x^*$ de la línea de regresión verdadera es:*

$$\left[\hat{\beta}_0 + \hat{\beta}_1 x^* - t_{\frac{\sigma}{2}, n-2} \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}} \right)} \; ; \; \hat{\beta}_0 + \hat{\beta}_1 x^* + t_{\frac{\sigma}{2}, n-2} \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}} \right)} \right]$$

Observaciones:

- Notar que el ancho del intervalo de confianza depende de x^*. El ancho será minimo cuando $x^* = \overline{x}$ y crece a medida que $|x^* - \overline{x}|$ aumenta.

- Al repetir los calculos anteriores para varios valores diferentes de x^* pueden obtenerse intervalos de confianza para cada valor correspondiente de $\beta_0 + \beta_1 x^*$

1.5.4 Intervalo de predicción para valores futuros de Y

Como hemos visto, un intervalo de confianza incluye los valores posibles de un *parámetro* de la población, cuyo valor es fijo pero desconocido. En cambio, un valor futuro de Y no es un parámetro sino una variable aleatoria. Es por esto que se hace referencia a un intervalo de valores factibles para un valor de Y como un **intervalo de predicción** en lugar intervalo de confianza.

Notemos que el *error de estimación* de Y es $(\beta_0 + \beta_1 x^*) - (\hat{\beta}_0 + \hat{\beta}_1 x^*)$, una diferencia entre una cantidad fija (pero desconocida) y una variable aleatoria.

Por otro lado, el *error de predicción* es $Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)$, que resulta ser una diferencia entre dos variables aleatorias. Existe por lo tanto más incertidumbre en la predicción que en la estimación, de forma tal que un intervalo de predicción será más ancho que uno de confianza.

Ahora, como hemos visto, $E(Y^*) = \beta_0 + \beta_1 x^*$ y $E(\hat{Y}^*) = E(\hat{\beta}_0 + \hat{\beta}_1 x^*) = \beta_0 + \beta_1 x^*$, por lo tanto

$$E(Y^* - \hat{Y}^*) = 0$$

Es decir, el error esperado en la predicción de Y es 0. Por otro lado

$$V(Y^* - \hat{Y}^*) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\right)$$

$Y^* - \hat{Y}^*$ tiene distribución normal con varianza y esperanza anteriores.

Teorema 1.11. *La variable aleatoria*

$$T = \frac{Y^* - \hat{Y}^*}{\sqrt{\sigma^2 \left(1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\right)}}$$

tiene distribución student con $n - 2$ grados de libertad.

Por el argumento usual llegamos a:

Definición 1.12. *Un* ***intervalo de predicción del*** $100(1-\alpha)\%$ ***para*** Y^* *es:*

$$\left[\hat{Y}^* - t_{\frac{\sigma}{2},n-2}\sqrt{\sigma^2\left(1+\frac{1}{n}+\frac{(x^*-\overline{x})^2}{S_{xx}}\right)}\,;\hat{Y}^* + t_{\frac{\sigma}{2},n-2}\sqrt{\sigma^2\left(1+\frac{1}{n}+\frac{(x^*-\overline{x})^2}{S_{xx}}\right)}\,\right]$$

La interpretación del nivel de predicción $100(1-\alpha)\%$ es idéntica al de los niveles de confianza. Si se utiliza repetidamente, a la larga los intervalos resultantes en realidad contendrán los valores y observados el $100(1-\alpha)\%$ del tiempo.

1.6 Ejemplo integrador

A continuación veamos un ejemplo donde integramos la mayoría de los conceptos vistos a lo largo de este apunte:

Ejemplo 1.13. (DATASET: Nails.csv)
En un experimento para investigar la relación entre el diámetro de un clavo (x) *y su fuerza de retirada dinal* (y) *se colocaron clavos de forma anular enhebrados en madera de abeto de Douglas, y después se midieron sus fuerzas de retirada en* N/mm*. Se obtuvieron los resultados siguientes para 10 diámetros diferentes (en mm):*

i	1	2	3	4	5	6	7	8	9	10
x_i	2.25	2.87	3.05	3.43	3.68	3.76	3.76	4.5	4.5	5.26
y_i	54.74	59.01	72.92	50.85	54.99	60.56	69.08	77.03	69.97	90.7

1. *Obtenga y grafique la recta de mínimos cuadrados que se ajusta a los datos tabulados.*
2. *Calcule la estimación de la varianza y el coeficiente de determinación. De una interpretación esta última métrica.*
3. *Determine un intervalo de confianza del 95% para la media de la fuerza de retirada de clavos de 4mm de diámetro*
4. *Determine un intervalo de predicción del 95% para la media de la fuerza de retirada de clavos de 4mm de diámetro*
5. *¿Puede concluir que la media de la fuerza de retu¿irada de clavos de 4mm de diámetro es 60 N/mm con un nivel de significancia de 0,05?*

Resolución

Realizando las cuentas a partir de los datos de la tabla, obtenemos los resúmenes estadísticos:

$$S_{xx} = 6.287 \qquad S_{yy} = 1360.86625 \qquad S_{xy} = 69.29655 \qquad \overline{x} = 3.733 \qquad \overline{y} = 65.985 \qquad n = 10$$

1. Recordando lo visto en el método de mínimos cuadrados, sabemos que

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

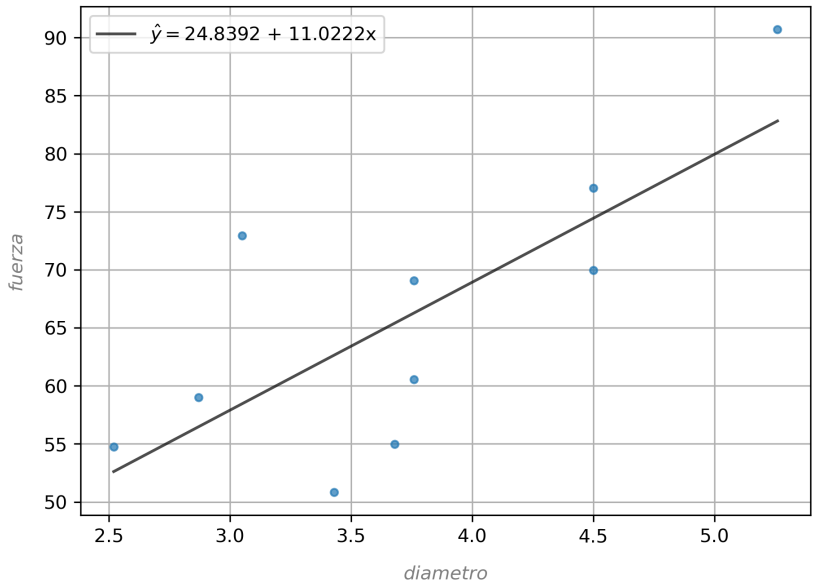
Tomando los datos estadísticos obtenidos podemos calcular la recta estimada de mínimos cuadrados:

$$\hat{\beta}_1 = 11.022179 \qquad \hat{\beta}_0 = 24.8392$$

$$\Downarrow$$

$$\hat{y} = 24.8392 + 11.022179\ x$$

A continuación grafiquemos el conjunto de datos junto con la recta de mínimos cuadrados:



2. Revisando la teoría podemos recordar que la estimación de la varianza está dada por

$$\hat{\sigma}^2 = \frac{SCE}{n-2} = \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right) \frac{1}{n-2}$$

$$\Downarrow$$

$$\hat{\sigma}^2 = \left(1360.86625 - \frac{(69.29655)^2}{6.287} \right) \frac{1}{8} = \frac{597.06726}{8} = 74.6334$$

Por otro lado, para hallar el coeficiente de determinación debemos tener en cuenta la suma total de cuadrados (STC) y la suma de cuadrados de error (SCE), ya que R^2 es de la forma:

$$R^2 = \frac{SCR}{STC} = 1 - \frac{SCE}{S_{yy}}$$
$$\Downarrow$$
$$R^2 = 1 - \frac{597.06726}{1360.86625} = 0.561259$$

Notemos que el coeficiente de determinación se encuentra relativamente en el medio entre 0 y 1. Como sabemos, nosotros buscamos un modelo cuyo coeficiente de determinación se acerque lo más posible a 1, de esta forma indicaría que la variabilidad total de la muestra es explicada por el modelo lineal.

Interpretando el valor de R^2, podemos decir que aproximadamente el 56.13% de la variabilidad de la fuerza de retirada de los clavos puede ser explicada su el diámetro. Por sí solo el parámetro del diámetro (x) no logra determinar por completo la variable y. Caso contrario, los puntos caerían todos en la recta de regresión estimada, haciendo la suma de errores cuadrados 0. Esta estimación bastante pobre se puede deber a la falta de datos con los que se trabajan y/o por su naturaleza. Podría ser el caso de que sigan otro tipo de relación que no sea lineal.

3. Al buscar un intervalo del 95%, debemos utilizar un $\alpha = 0.05$:

$$(1-\alpha) = 0.95 \qquad \Longleftrightarrow \qquad \alpha = 1 - 0.95 = 0.05$$

Tomando lo que vimos en la Definición 1.10, definiendo al valor fijo $x^* = 4$ y calculando $t_{\frac{\alpha}{2},n-2} = t_{0.025,8} = 2.306$, tenemos que el intervalo es:

$$\left[68.9279 - 2.306\sqrt{74.6334\left(\frac{1}{10} + \frac{(4-3.733)^2}{6.287}\right)}\ ;\ 68.9279 + 2.306\sqrt{74.6334\left(\frac{1}{10} + \frac{(4-3.733)^2}{6.287}\right)}\right]$$
$$\Downarrow$$
$$[\,62.281\ ;\ 75.575\,]$$

4. Ahora queremos calcular el intervalo de predicción para futuras observaciones. Recordar que lo que siempre estimamos con la recta de regrsión de minimos cuadrados son los valores medios de las variables aleatorias Y_i. Nuevamente utilizaremos:

$$\alpha = 0.05 \qquad n = 10 \qquad x^* = 4 \qquad t_{0.025,8} = 2.306$$

Entonces tenemos que el intervalo es:

$$\left[68.9279-2.306\sqrt{74.6334\left(1+\frac{1}{10}+\frac{(4-3.733)^2}{6.287}\right)}\ ;\ 68.9279+2.306\sqrt{74.6334\left(1+\frac{1}{10}+\frac{(4-3.733)^2}{6.287}\right)}\right]$$

$$\Downarrow$$

$$[\,47.925\ ;\ 89.928\,]$$

5. No se puede concluir que la media de la fuerza de retirada de clavos de 4mm de diámetro es de 60 N/mm con nivel de significancia de 0.05 (o nivel de confianza de 0.95) ya que:

$$60 \notin [\,62.281\ ;\ 75.575\,]$$