PUCE Virtual
EDUCACIÓN FLEXIBLE

Estadística 1 para administración

Modelos de regresión

Clase 11

La excelencia no se improvisa

INTRODUCCIÓN DE LA CLASE

En la clase 1 explicamos que, de todas las variables contenidas en un dataset, podríamos distinguir entre aquellas que son predictoras de otra variable que son consideradas variables de respuesta, esto lo pudimos establecer través de la función correlación. En la clase 3 revisamos además que para asegurar esta relación el levantamiento de datos debía seguir un proceso formal de tal manera de garantizar que la data no está sesgada. Por otro lado, en la clase 2 revisamos los conceptos distribución de probabilidades de los valores de una variable.

Una vez que descubrimos las relaciones entre variables, nos preguntamos si podemos establecer una ecuación que modele la relación entre dichas variables. Si pudiéramos encontrar dicha ecuación, (que en adelante la llamaremos modelo) podremos en primer lugar validar que se ajustan a las observaciones y en segundo lugar predecir nuevos valores de las variables de respuesta dentro de determinados intervalos de confianza.

Clase 11: Modelos de regresión

Un modelo no es más que la representación del funcionamiento de un proceso mediante una o más ecuaciones matemáticas que tratan de explicar cómo funciona dicho proceso. En la forma más simple, el modelo es una ecuación, pero otros tipos de modelos podrían representarse mediante un sistema de ecuaciones por ejemplo en el caso de modelos epidemiológicos.

DETERMINACIÓN DE LINEALIDAD Y HOMOCEDASTICIDAD

Determinación de Linealidad

Las funciones R de correlación cor() nos indican si existe una relación entre la variable predictora y la variable de respuesta a través del el cálculo del índice de correlación. Este puede ser calculado mediante métodos paramétricos y no paramétricos. Para el propósito de modelos preferirnos siempre los métodos paramétricos como la correlación de Pearson, por ser considerados más exactos.

Figura 1

Fórmula de la correlación de Pearson

Adaptación de la formula de correlación Szretter (2017)

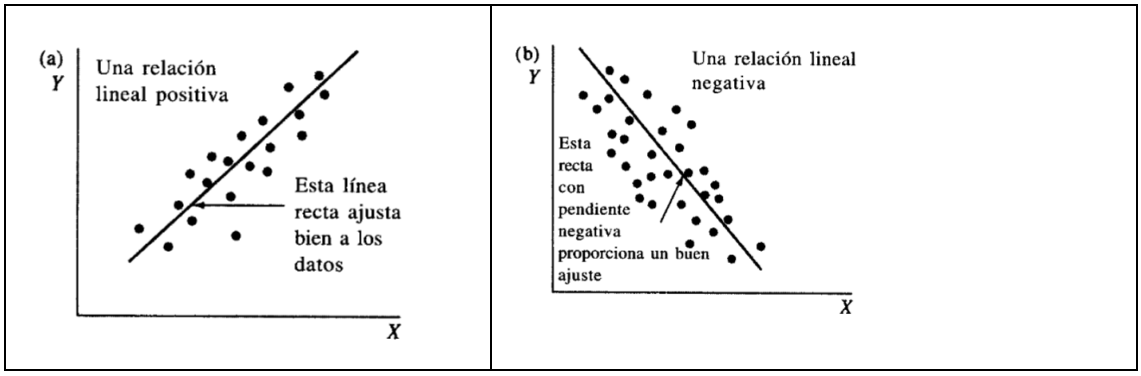
$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

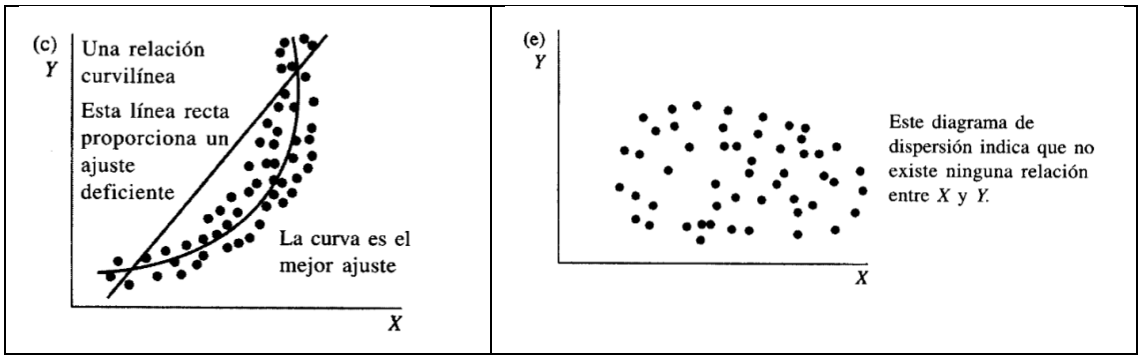
Es importante mencionar que el coeficiente de Pearson solo nos indica que, si la variable predictora sube o baja en valor, la variable de respuesta también subirá o bajará, pero no nos indica en qué proporción lo hará y si esa proporción es constante o no. Estos son los conceptos de linealidad y cedasticidad. Dependiendo de si la relación es lineal o no utilizaremos distintos tipos de modelos. En esta clase vamos a estudiar modelos lineales, por lo tanto, cuando hacemos el análisis de correlación, debemos encontrar ya sea una correlación positiva o negativa como se muestra en la figura 2, si decidiéramos utilizar una variable que muestra una correlación nula o curvilínea esta solo abonaría al incremento de la incertidumbre y por ende al error estándar de la estimación. La mejor forma de estimar la linealidad es a través de un diagrama de dispersion, donde el eje X contiene la variable predictora y el eje Y la varaible de respuesta.

Figura 2

Diagramas de dispersión con distinto tipo de relación. Solo la relaciones a y b serían aptas para modelar mediante una regresión simple

Tomado de Webster (2000) Capitulo 11





El problema es que los valores raramente están alineados perfectamente, como se ve en la figura 3a, esto es debido a que las observaciones incluyen un error de muestreo, lo que buscamos es que la recta se ubique más o menos por la media de las observaciones para un valor dado de X (Figura 3b)

Figura 3

Índice Pearson ideal (a) y realista (b)

Creación del autor Alfonso Prado

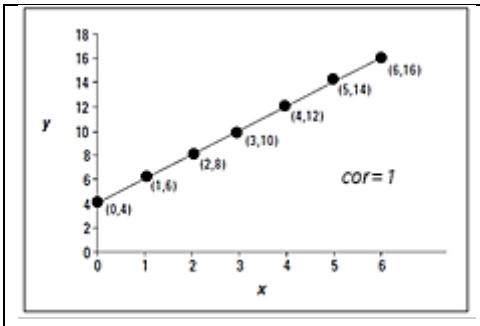


Figura 3a

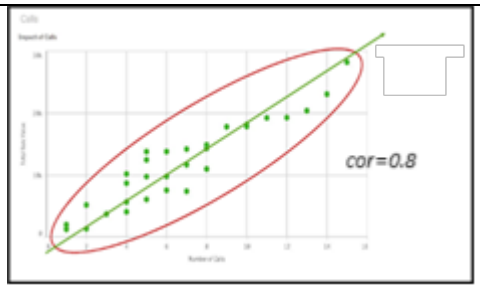


Figura 3b

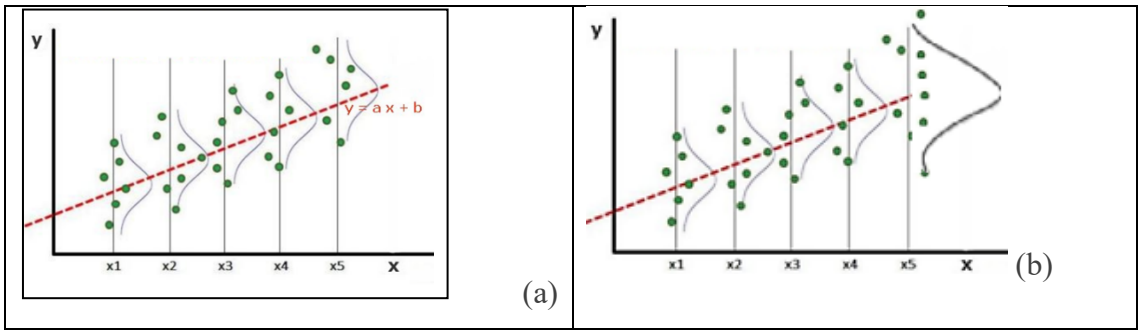
El coeficiente de correlación de Pearson generalmente está representado por una r para la muestra y la letra griega rho ρ para la población, puede asumir un valor entre -1 y 1 . Un valor de r=-1 indica una relacion negativa perfecta entre X y Y, todas las obserbaciones se alinean en una línea perfecta con pendiente negativa. Por el contrario si el r=1 indica una relación positiva perfecta a lo largo de una recta con pendiente positiva. Valores absolutos intermedios como por ejemplo .1<r<.3 indican una asociación débil (la variable no es válida para el modelo) , valores absolutos intermedios como por ejemplo .3<r<.5 indican una asociación mediana (posiblemente la variable es válida en conjunto con otras

variables) y valores absolutos intermedios como por ejemplo .5<r<.1 indican una asociación fuerte (la variable es pertinente para el modelo)

Determinación de Homocedasticidad y heterocedasticidad

Homocedasticidad se refiere a una condición en la cual la varianza de los datos es constante. Es decir, el término de error no varía mucho a medida que cambia el valor de la variable predictora. Cuando la varianza para un valor dado de X es distinta que otros valores de X entonces decimos que la data presenta heterocedasticidad. Como consecuencia el modelo no estará bien definido. La siguiente figura visualiza la varianza de la relación de dos variables X y Y .

Figura 4
Ejemplo de homocedasticidad (a) la varainza a lo largo de todos los valores es constante y Heterocedasticidad (b) la varianza de X5 es mucho mayor que en otros valores
Creción del autor Alfonso Prado



MODELO DE REGRESIÓN SIMPLE

En estos modelos tenemos solamente una variable predictora y por supuesto una variable de respuesta. La ecuación lineal estará definida por tres coeficientes:

- ✓ La intercepción (a) indica cual es el valor de la variable de respuesta cuando la predictora es igual a 0.
- ✓ La pendiente (b) indica cuánto cambia "Y" cuando "X" cambia en 1 unidad.
- ✓ El error ε o SE, conocido como error estándar de la regresión, representa la distancia promedio a la que caen los valores observados desde la línea de

regresión, le dice cuán incorrecto es el modelo de regresión en promedio usando las unidades de la variable de respuesta

Figura 5

Fórmula de la regresión lineal simple

Adaptado de Webster (2000) capítulo 11

$$Y = a + bX + \in$$

Donde:

Y es la variable de respuesta

X es la variable predictora

ε Es el error estándar

Mínimos cuadrados ordinarios MCO

De acuerdo a Webster (2000), "El propósito del análisis de regresión es determinar una recta que se ajuste a los datos muestrales mejor que cualquier otra recta". El algoritmo MCO producirá esta recta que se extiende por el centro del diagrama de dispersión. Este algoritmo calculará los valores de a y b de acuerdo a las siguientes fórmulas.

Figura 6

Fórmulas de cálculo de los coeficientes de regresión según el algoritmo MCO

Nota. Adaptación de la fórmula del algoritmo según webster(2000) capítulo 11

$$a = (\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy) / n(\Sigma x^2) - (\Sigma x)^2$$

$$b = n(\Sigma xy) - (\Sigma x)(\Sigma y) / n(\Sigma x^2) - (\Sigma x)^2$$

Para la creación del modelo R provee algunas funciones, la más simple es lm() cuya sintaxis se muestra en la siguiente figura.

Figura 7

Sintaxis de la función programática lm()

Adaptación de la fórmula para la resolución de un modelo según Rdocumentation (s.f.), https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/lm

```
lm(formula, data, subset, weights, na.action,
   method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
   singular.ok = TRUE, contrasts = NULL, offset, …)
```

La fórmula es el argumento más importante, misma que define cuál será la variable predictora y la variable de respuesta, y toma la forma de la siguiente expresión: respuesta ~ predictora. El signo "~" se lee como decir "se distribuye como" o "dependiendo de" cuando se ve en las funciones de regresión.

Coeficientes, valores ajustados y residuos

Como resultado de la ejecución de lm, éste devuelve un objeto que contiene varias propiedades, mediante las cuales podemos visualizar la relación entre las observaciones y los valores de la regresión y vamos a ubicar los siguientes conceptos:

Los coeficientes corresponden al cálculo hecho por el algoritmo MCO para estimar la regresión, ver figura 6.

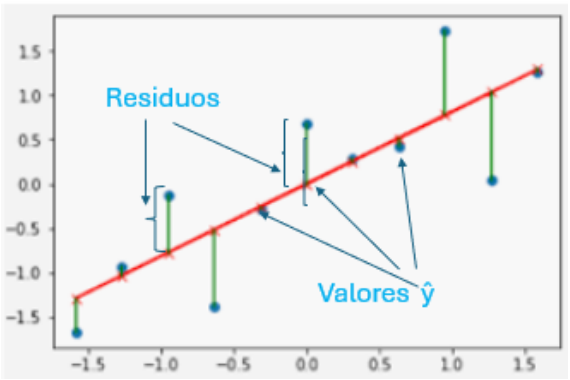
Valores ajustados, corresponden a los valores que la función de regresión estima para cada valor de las observaciones en base a lo indicado en la figura 5. Estos valores se los denota como $\hat{Y}$ (y-hat).

Los residuos corresponden a la diferencia entre $\hat{Y}$ y la observación (Y) y nos interesa que estos residuos sean lo más pequeño posible porque de eso depende la precisión del modelo.

Figura 8

Valores ajustados y residuos

Creación del auto Alfonso Prado



Coeficiente de determinación o R^2

Una vez creado el modelo podemos ejecutar la función summary, la cual nos entrega una serie de métricas que nos sirven para validar la bondad del modelo.

Figura 9
Métricas de la bondad del modelo
Creación de autor Alfonso Prado

```
Residual standard error: 2.437 on 1076 degrees of freedom
Multiple R-squared:  0.2513,    Adjusted R-squared:  0.2506
F-statistic: 361.2 on 1 and 1076 DF,  p-value: < 2.2e-16
```

Error estándar de los residuos: Este valor nos indica la varianza de los residuos, es decir, que tan dispersos están los valores de los residuos. Una dispersión muy grande significa que para muchas observaciones el residuo es alto y en general queremos que la varianza sea pequeña o moderada.

R^2: También llamado coeficiente de determinación, es una medida de qué tan bien se ajusta su modelo a los datos. Como habíamos visto en la fórmula de la regresión (figura 5) el valor $\hat{Y}$ está afectado por el ε o SE. R^2 indica que tanto de la variable de respuesta está dado por la variable predictora y que tanto por el error ε.
Valores más altos representan diferencias más pequeñas entre los datos observados y los valores ajustados.
En el mejor caso, los valores observados coinciden con los ajustados, la suma de cuadrados de los residuos (RSS) es 0 y $R^2 = 1$

Análisis de coeficientes
Si se ejecuta la función summary a un objeto de la clase lm nos retornará una cantidad de estadísticos que debemos analizar, la primera sección es la relacionada con los coeficientes, en la cual encontraremos no solo los coeficientes generados por el MCO, sino algunos estadísticos mencionados en la tabla 1
Tabla 1

Sección de Coeficientes

Creación del autor Alfonso Prado

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.88660    1.83235   18.49   <2e-16 ***
fheight      0.51409    0.02705   19.01   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

La primera columna contiene los nombres de las variables que son parte del modelo. El “intercept” corresponde al coeficiente a de la fórmula de la regresión y siempre estará presente con este nombre y a continuación estarán los nombres de las variables (“fheight” en el ejemplo) que han sido incluidas en el modelo.

La segunda columna contiene la estimación de los coeficientes, note la palabra estimación es decir que estos no son exactos y que los mismos pueden tener una variabilidad que figura en la tercera columna como “Error Estándar”.
Lo que nos interesa es que el error estándar sea lo más pequeño posible con respecto al coeficiente, esto es lo que refleja la cuarta columna t-value que es la relación entre el valor estimado y el error estándar.

La columna Pr>|t| corresponde a p-value de un t-test que prueba la hipótesis de si el valor del coeficiente podría ser igual a 0, y debe contrastarse con el nivel α que por defecto es .05 (95%). En otras palabras, nos permite validar si el coeficiente podría ser 0. Si la hipótesis es correcta indicaría que la variable no es estadísticamente significativa.

En conclusión, valoramos la pertenencia de las variables en base a los siguientes criterios:

- ✓ t-value alto indicará que el error en su cálculo es bajo
- ✓ p-value bajo indicará que el coeficiente no es cero.

Por último, summary nos presenta unos códigos de significación del p-value, es una forma de indicar que tan importante es la variable y preferimos variables con “***” .
Una forma gráfica para entender la importancia de cada variable es mediante la función coefplot del paquete coefplot. Este gráfico nos presenta la probabilidad de que el coeficiente sea 0 mediante un diagrama de barras de error. Este diagrama contiene un

punto correspondiente al valor estimado de los coeficientes y dos barras que representa la variabilidad de ±1 σ (traza gruesa) y ±2 σ (traza delgada)

Valoramos una variable como significativa si las barras nunca cruzan la línea punteada de 0

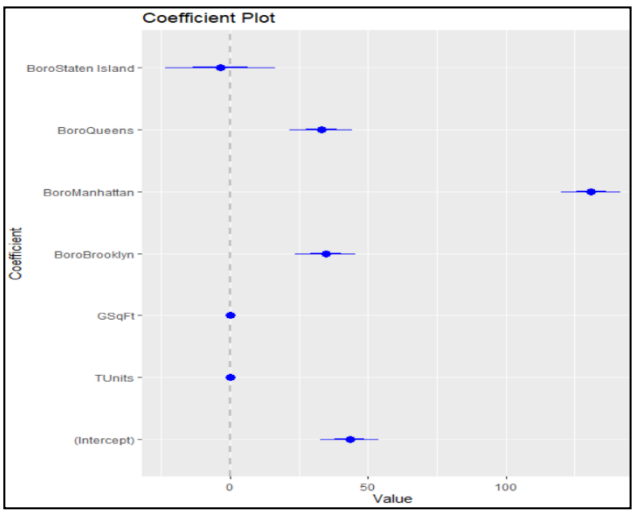


Figura 10

Coefplot de una regresión múltiple

Creación del autor Alfonso Prado

Análisis de residuos

Otra sección del summary del modelo que es de particular interés es la relacionada con los residuos.

Figura 11

Estadísticos de los Residuos

Creación del autor Alfonso Prado

```
Residuals:
    Min      1Q  Median      3Q     Max
-8.8772 -1.5144 -0.0079  1.6285  8.9685
```

En general nos interesan dos temas:

- ✓ Los residuos deben ser pequeños, residuos altos nos indican que el modelo no está bien definido
- ✓ Los residuos deben tener una distribución normal, si la variable de respuesta tiene una distribución normal esperaríamos que los residuos, que salen de la resta de los valores Y con respecto a la recta de regresión, también tengan una distribución normal.

Lo que el summary del modelo nos presenta es esta distribución en términos de cuartiles. La forma más fácil de validar si este supuesto se cumple es mediante un diagrama de caja,

según el cual la distancia del Q1 a la media debe ser similar a la distancia de la media al Q3. Ver figura 11a , por otro lado si el diagrama de caja se muestre como la figura 11b significaría que la predicción de ciertos valores arrojarán un error mayor que otros.

Rango Intercuartilico

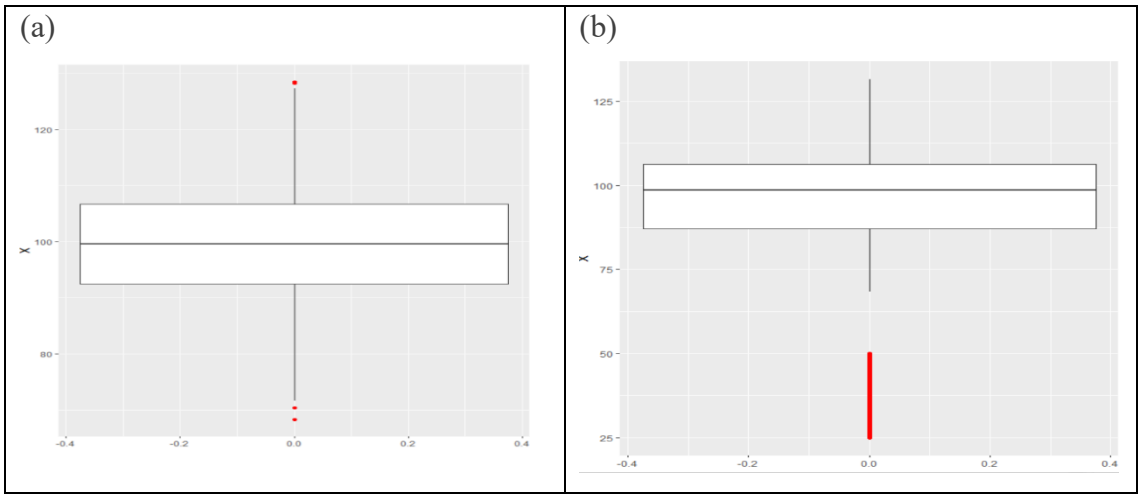
Definición y ejemplos

Enlace

Figura 12

Diagramas de Caja de una distribución de residuos normal (a) y sesgada(b)

Creación del autor Alfonso Prado



En este punto nos preguntamos ¿cuáles son las causas para que la distribución de los residuos no sea normal? Comúnmente esto tiene que ver con los valores atípicos (outliers en inglés) representados en el diagrama de caja por un color diferente (rojo en el ejemplo). En general, cualquier punto de datos que se encuentre fuera del rango intercuartilico (1.5 * IQR) se considera un valor atípico, donde, IQR se calcula como la distancia entre los valores del percentil 25 y 75 para la variable. Estos valores atípicos harán que el diagrama de caja se desvíe hacia cualquiera de sus extremos. El analista debería encontrar la razón por la cual se presentan estos “outliers”, siendo posible que los mismos sean errores de medida en cuyo caso el analista podría proceder a eliminar dichas observaciones.

Pero también es posible que los valores sean verdaderos en cuyo caso su eliminación afectará el modelo en el sentido de no poder predecir valores tan extremos. Muchos autores toman una opción intermedia que es conocida como la imputación que implica el reemplazo de valores atípicos por un valor máximo o mínimo.

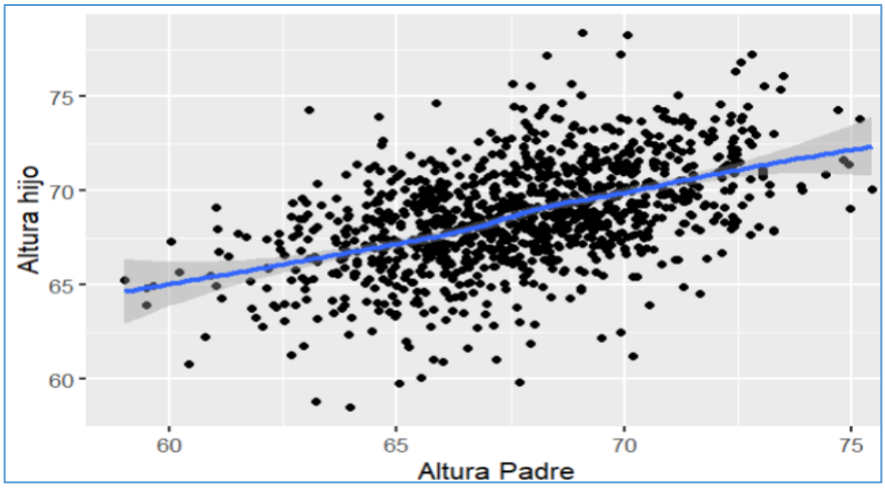
Veamos un ejemplo completo

Usaremos el dataset father.son que es parte del paquete UsingR, este dataset contiene 1078 observaciones de altura de padres con respecto a la altura de sus hijo. El propósito es demostrar que se puede obtener un modelo que permita predecir la altura del hijo (variable de respuesta) en función de la altura del padres (variable predictora)

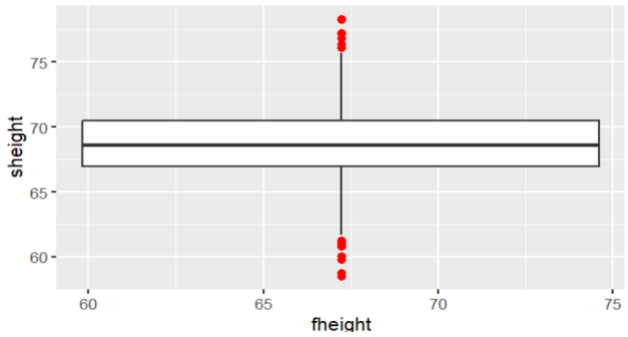
```
#Vamos a validar que existe una relación entre las variables
cor(father.son)
```

```
            fheight   sheight
fheight 1.0000000 0.5013383
sheight 0.5013383 1.0000000
```

```
#Vemos que la correlación es positiva pero no fuerte, lo cual puede ser debido a que la
relación no es completamente lineal.
#Podemos validad la linealidad mediante un diagrama de dispersión
ggplot (data=father.son , aes(x=fheight , y= sheight)) +
  geom_point()+
  geom_smooth(method=loess)+
  xlab("Altura Padre")+
  ylab("Altura hijo")
```




```
#En segundo lugar debemos validar si la data presenta datos atípicos (outliers), lo
haremos con un diagrama de caja
ggplot (data=father.son , aes(x=fheight,y=sheight)) +
  geom_boxplot(outlier.colour="red", outlier.shape=16,
               outlier.size=2, notch=FALSE)
```



#Creamos nuestro modelo

padre_hijoLM <- lm(sheight ~fheight , data= father_son_sin_outliers)

Revisando el modelo

summary(padre_hijoLM)

```
Call:
lm(formula = sheight ~ fheight, data = father_son_sin_outliers)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3517 -1.4722 -0.0741  1.6019  7.7910

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.69140    1.72826   21.23   <2e-16 ***
fheight      0.47300    0.02551   18.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.262 on 1057 degrees of freedom
Multiple R-squared:  0.2454,    Adjusted R-squared:  0.2447
F-statistic: 343.7 on 1 and 1057 DF,  p-value: < 2.2e-16
```

#Entoces el modelo sería

sheight=36.91 + 0.473*fheigth

#Pero recuerde que estos coeficientes no son exactos, en realidad debemos calcular cual es el intervalo de confianza para cada coeficiente

confint(padre_hijoLM, level=0.95)

```
                2.5 %     97.5 %
(Intercept) 33.3001794 40.0826200
fheight      0.4229353  0.5230576
```

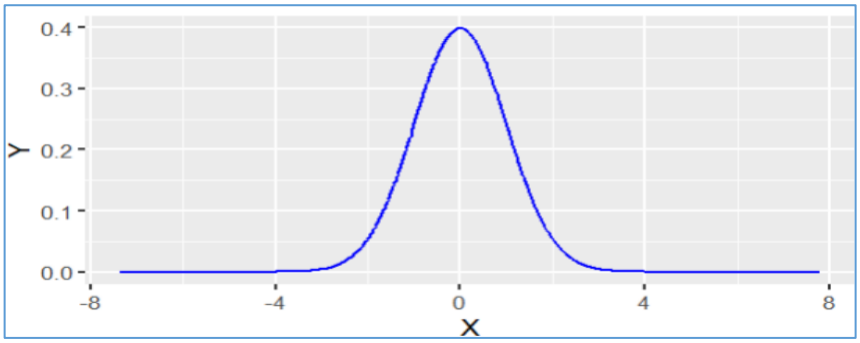
#Podemos validar si los residuos son normales

residuos <- padre_hijoLM$residuals

resdf <- data.frame(X=residuos, Y=dnorm(residuos))

ggplot (data=resdf , aes(x=X, y=Y)) +

```
geom_line(color="BLUE"
```




```
#Concluimos que los residuos son normales
#Obtenemos  un R2 de 0.24, lo cual indica que no es muy bueno
#Y error estándar de los residuos de 0.226
```

Referencias citadas en la Clase 11

Webster A., (2000), Estadística aplicada a los negocios, Irwin Professional Publishing

Schmuller J.(2022), Statistical Analysis with Excel For Dummies, Wiley

Teck K, (2022) , Practical t-test Power Analysis with R, Practical Assessment, Research & Evaluation , Volumen 27, No 18

Berman H.G., "*F Distribution* , descargado de https://stattrek.com/probability-distributions/f-distribution

Rodrigo J. (2016), *Análisis de varianza para comparar múltiples medias*, https://cienciadedatos.net/documentos/19_anova

Szretter, M. (2017), Apunte de Regresión Lineal, https://mate.dm.uba.ar/~meszre/apunte_regresion_lineal_szretter.pdf

PUCE Virtual

La excelencia no se improvisa

síguenos