Estadística Aplicada

Modelos de regresión

Clase 11

Ingeniería en ciberseguridad

La excelencia no se improvisa

Clase 11:

Analizar información contextual sobre hábitat, infraestructura y movilidad, mediante técnicas estadísticas descriptivas e inferenciales, para la adecuada toma de decisiones

Reto # 4

1. 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 y las que son consideradas variables de respuesta. Esto lo pudimos establecer a través de la función de correlación. En la clase 3 revisamos que, para asegurar esta relación, el levantamiento de datos debía seguir un proceso formal con el fin de garantizar que la data no estuviera sesgada. Por otro lado, en la clase 2, revisamos los conceptos de 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 esta ecuación (que en adelante llamaremos modelo), podríamos, en primer lugar, validar que se ajusta a las observaciones y, en segundo lugar, predecir nuevos valores de las variables de respuesta dentro de determinados intervalos de confianza.

Clase 11:

1. Tema. Modelos de regresión

1. Modelos de regresión lineal

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 estos modelos, es básicamente una ecuación, pero otros tipos de modelos podrían representarse mediante un sistema de ecuaciones. Un ejemplo de esto son los modelos epidemiológicos.

MODELOS LINEALES Y NO LINEALES

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

Figura 1

Fórmula de la correlación de Pearson

Creación de autor Alfonso Prado

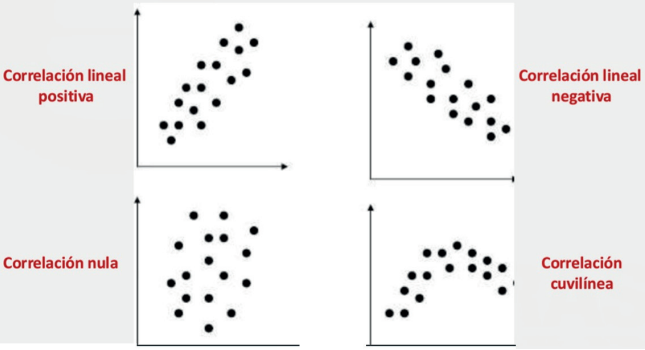
$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

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, sin decirnos en qué proporción lo hará o si esa proporción es constante. Estos son los conceptos de linealidad y homocedasticidad. Dependiendo de si la relación es lineal o no, utilizaremos distintos tipos de modelos. En esta clase, iniciaremos con 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 utilizáramos una variable que muestra una correlación nula o curvilínea, esta abonaría al incremento de la incertidumbre y, por ende, al error estándar de la estimación.

Figura 2

Relación lineal y no lineal

Creación de autor Alfonso Prado

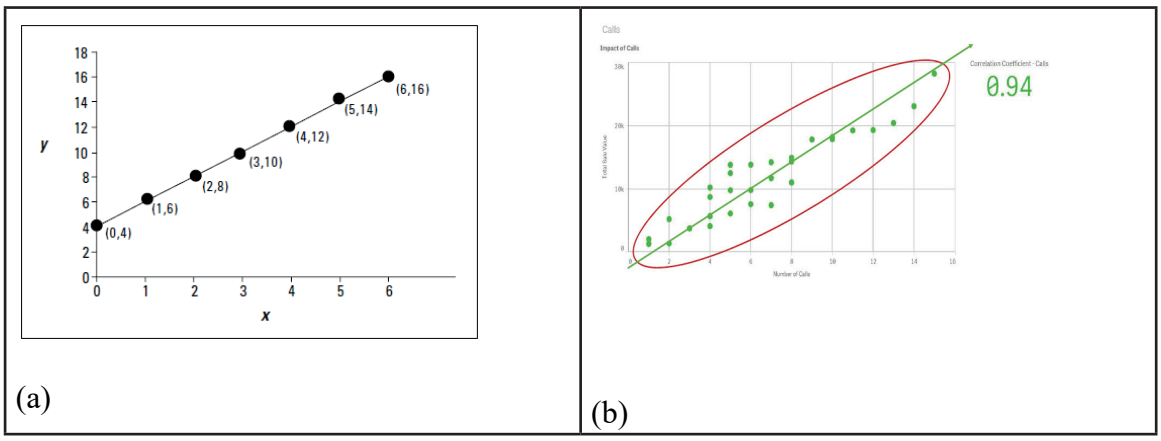


El problema es que los valores rara vez están alineados perfectamente, como se ve en la Figura 3a. Esto se debe a que las observaciones incluyen un error de muestreo. Lo que buscamos es que la recta se ubique más o menos por la media de las observaciones para un valor dado de X (Figura 3b).

Figura 3

Índice Pearson ideal (a) y realista (b)

Creación de autor Alfonso Prado



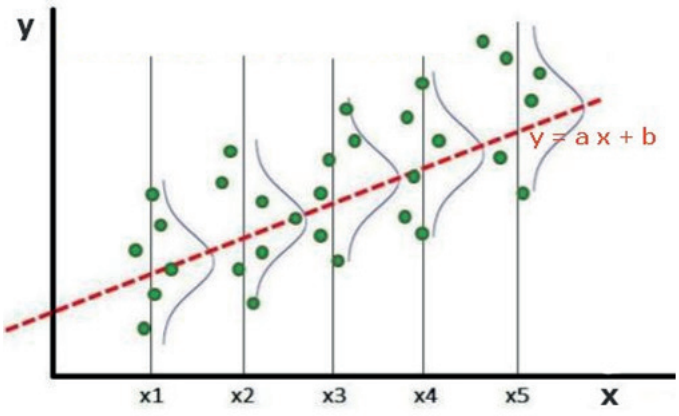
11.1) Determinación de linealidad y homocedasticidad

La homocedasticidad se refiere a una condición en la cual la varianza del residuo, o término de error, en un modelo de regresión es constante. Es decir, el término de error no varía mucho a medida que cambia el valor de la variable predictora. Si hay demasiada variación, diremos que su comportamiento es heterocedástico y, como consecuencia, el modelo no estará bien definido. La siguiente figura visualiza la varianza de la relación entre dos variables X y Y.

Figura 4

Concepto de cedasticidad

Creación de autor Alfonso Prado



Homocedasticidad y heterocedasticidad

Describe en concepto y pruebas

Enlace

Creación de modelos 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

Creación de autor Alfonso Prado

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

11.2) Mínimos cuadrados ordinarios MCO

De acuerdo con 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 con las siguientes fórmulas.

Figura 6

Fórmulas de cálculo de los coeficientes de regresión mediante Algoritmo MCO

Creación de autor Alfonso Prado

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

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

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

Figura 7

Función programática lm()

Creación de autor Alfonso Prado

```
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 "~" 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, este 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.

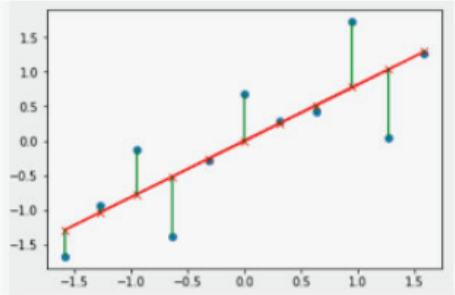
Valores ajustados, corresponden a los valores que la función de regresión estima para cada valor de las observaciones en base a lo indicado en la figura 5. Estos valores se los denota como $\hat{Y}$ (y-hat).

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 de autor Alfonso Prado



Análisis de coeficientes

Al ejecutar la función summary sobre un objeto de la clase lm, nos retornará una cantidad de estadísticos que debemos analizar. La primera sección está relacionada con los coeficientes, donde encontraremos no solo los coeficientes generados por el MCO, sino también algunos estadísticos mencionados en la tabla 1.

Tabla 1

Sección de Coeficientes

Creación de 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. A continuación, estarán los nombres de las variables (por ejemplo, “fheight”) que han sido incluidas en el modelo.

La segunda columna contiene la estimación de los coeficientes; nótese la palabra estimación, es decir, que estos no son exactos y que los mismos pueden tener un intervalo de confianza, 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 se refleja en 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. Esta debe contrastarse con el nivel α, que por defecto es .05 (95%). En otras palabras, nos permite validar si el coeficiente pudiera 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.

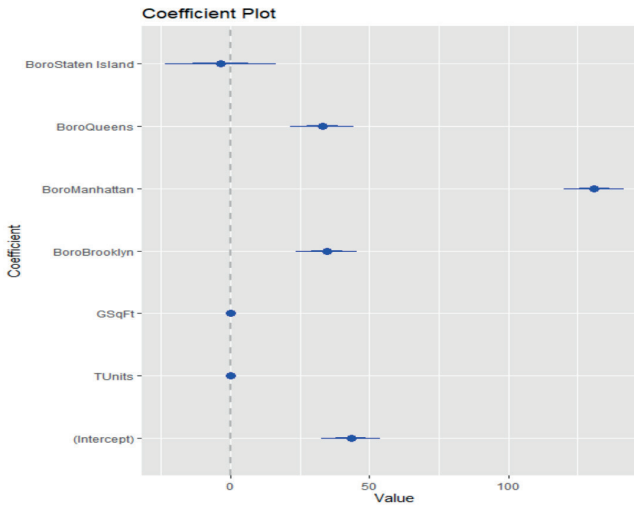
Por último, summary nos presenta unos códigos de significación del p-value, es una forma de indicar que tan importante es la variable y preferimos variables con "***" .

Una forma gráfica de 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 representan 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 9

Coefplot de una regresión

Creación de autor Alfonso Prado



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

Figura 10

Estadísticos de los Residuos

Creación de 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:

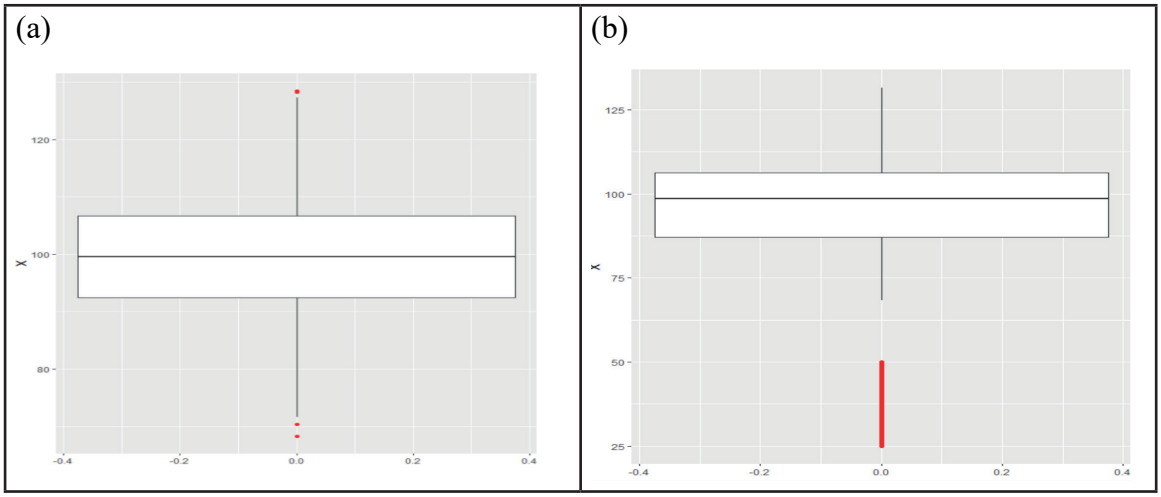
- ✓ Los residuos deben ser pequeños, residuos altos nos indican que el modelo no está bien definido
- ✓ Los residuos deben tener una distribución normal, si la variable de respuesta tiene una distribución normal esperaríamos que los residuos, que salen de la resta de los valores Y con respecto a la recta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

Figura 11

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

Creación de 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 intercuartílico (ejemplo: 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

debe 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 eliminar dichas observaciones.

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

Imputación

Describe el concepto y usos

imputación

11.3) Coeficiente de determinación

Otra sección importante que nos entrega el summary es la relacionada con las métricas de la bondad del modelo.

Figura 12

Métrica 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, qué 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 y R^2 cuadrado ajustado: 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 $R^2 = 1$

El valor ajustado contiene además una penalización en función de cantidad de variables que contiene y es una medida apropiada para medir un modelo de regresión múltiple.

Veamos un ejemplo completo

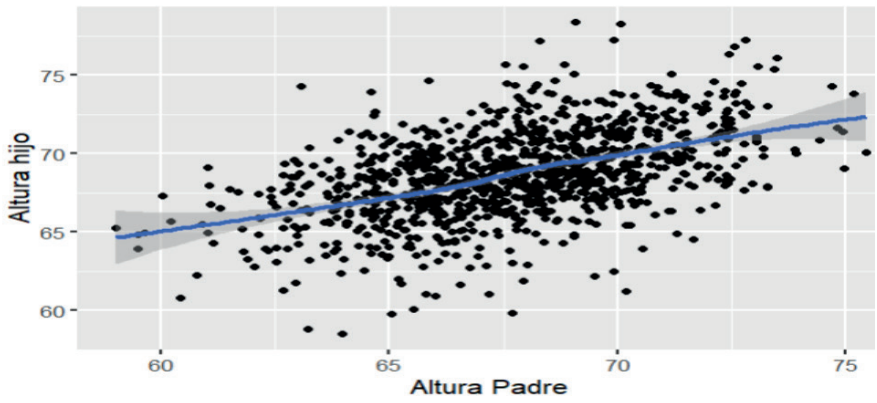
Usaremos el dataset father.son que es parte del paquete UsingR, este dataset contiene 1078 observa-

ciones 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)

```
#Cargamos el dataset

require(UsingR)

data(father.son)

names(father.son)

#En primer lugar analizamos la data

#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")
```



Referencias citadas en la Clase 11.

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

Bruce A. et al., (2017) Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python, O'Reilly Media, Inc.

Definición de los términos citados en la Clase 11.

INTERVALO DE CONFIANZA	Conocido también como límites de confianza. Es un rango de valores en el cual se encontraría el valor del parámetro, con una probabilidad determinada. Generalmente se construye intervalos de confianza con 95% de probabilidad
ANALISIS DE RESIDUOS	En el contexto de la regresión lineal, llamamos **residuos** a las diferencias entre los valores de la variable dependiente observados y los valores que predecimos a partir de nuestra recta de regresión.

Profundización Clase 11.

Modelos de regresión simple

Describe ejemplos

Estadistica-clase11-v1

PUCE Virtual

La excelencia no se improvisa

síguenos