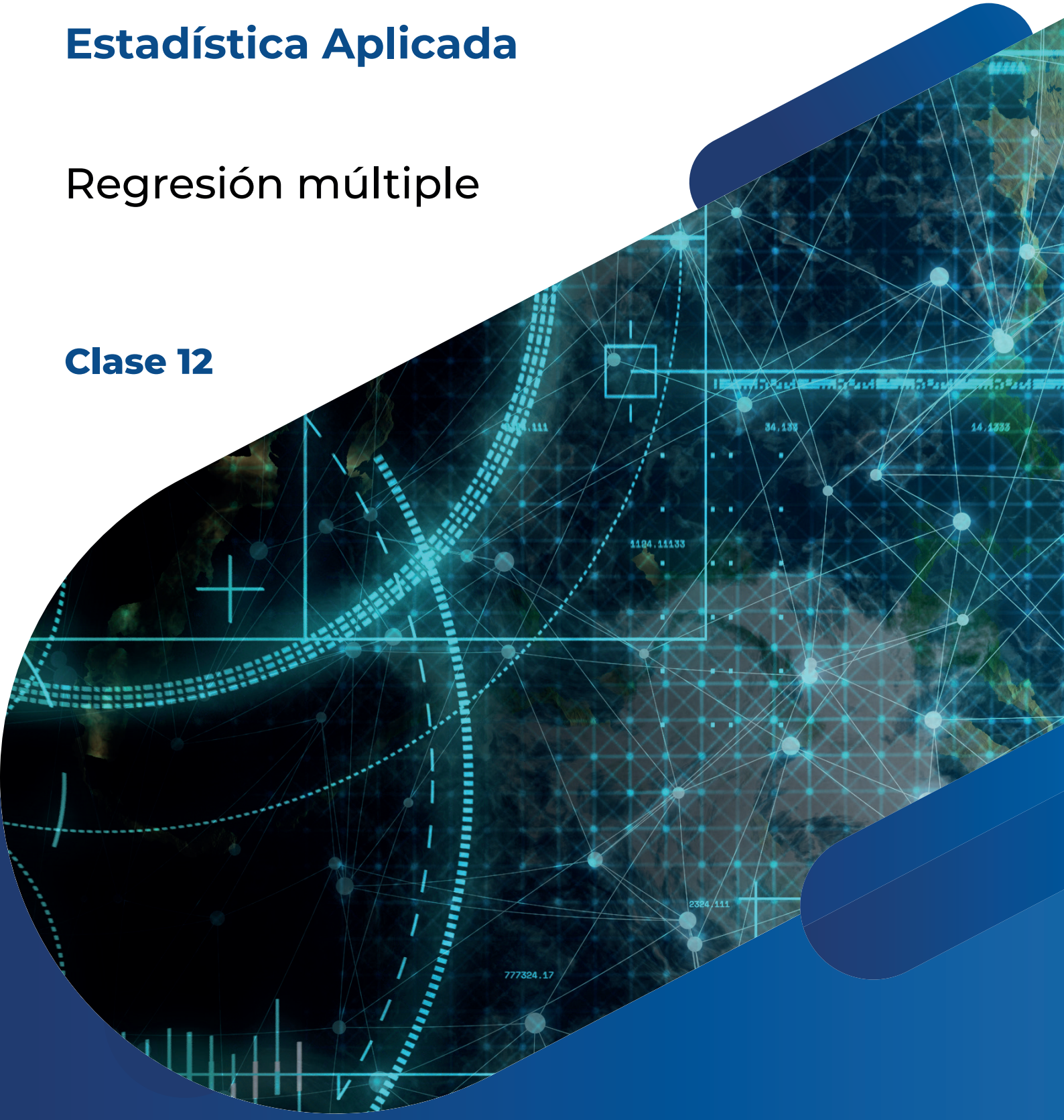
Estadística Aplicada

Regresión múltiple

Clase 12



Ingeniería en ciberseguridad

La excelencia no se improvisa

Clase 12:

Resultado o resultados de aprendizaje que será abordado con el contenido de la clase.

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 anterior, revisamos el concepto de la regresión lineal simple. Sin embargo, la mayoría de los negocios, procesos y eventos que se modelan contienen una gran cantidad de variables que afectan a la variable de respuesta, lo que hace que un modelo con una sola variable dé como resultado un error estándar considerable.

En estos modelos, además de identificar qué variables son importantes, vamos a considerar una serie de aspectos clave como la multicolinealidad y el sobreajuste, que, de no controlarse, podrían llevarnos a un error de predicción exagerado.

12.) Regresión múltiple

En general, al modelar un proceso, encontramos que existen múltiples variables que influyen en su resultado. Un modelo de regresión lineal múltiple contiene múltiples variables predictoras. Es importante notar que no todas las variables afectarán el modelo en igual medida.

Un modelo múltiple es básicamente el resultado de varios modelos de regresión simple combinados. Por lo tanto, la fórmula general es la misma para cada una de las variables. Matemáticamente, los coeficientes, el error estándar y las observaciones se pueden representar mediante vectores y matrices, como se muestra en la siguiente figura.

Figura 1

Fórmula general de regresión múltiple

Creación del autor Alfonso Prado

$$Y = X\beta + \epsilon$$

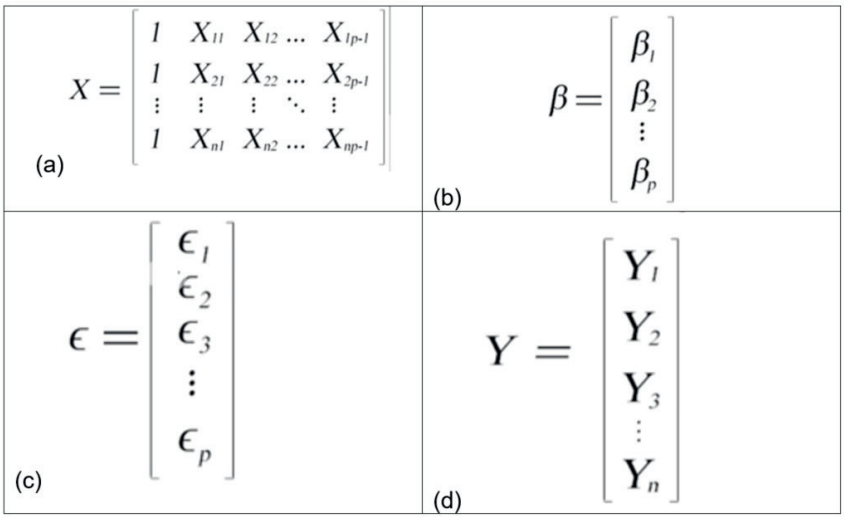
Notamos que las variable predictoras (X) se halla en mayúscula, indicando que estos representan una matriz, más específicamente una matriz de dimensión n*p, donde n es el número de observaciones y p la cantidad de predictores del modelo (figura 2a). Los coeficientes (ahora llamados β) están representados por un vector de longitud p (figura 2b). El error de cada regresión es también un vector de longitud p, es decir, un error por cada predictor (figura 2c). La variable dependiente Y (en mayúscula) es también una matriz que representa la contribución de cada variable (figura 2d).

Por último, también tenemos un vector ε una matriz de dimensión p X 1, que representa el error como explicamos en la clase anterior.

Figura 2

Componentes de regresión múltiple

Creación del autor Alfonso Prado



Funciones Programáticas de regresión múltiple

La función lm mencionada anteriormente es la misma utilizada para modelos de regresión múltiple; pero en este caso, incorporaremos más variables predictoras mediante los siguientes operadores:

El símbolo "+" es una solicitud implícita de calcular un(os) coeficiente(s) adicionales para otras variable(s) predictora(s).

Por ejemplo: lm(valxSqFt ~TUnits + GSqFt + Boro , data=housing) indicaría que el modelo contiene 3 variables predictoras (TUnits, GSqFt , y Boro)

Por otro lado, hay ocasiones donde nos interesa ver el efecto combinado de dos variables. Por ejemplo, en un modelo para predecir el riesgo de sufrir diabetes, se ha considerado como variable predictora el ejercicio que la persona realiza como mecanismo para reducir la cantidad de glucosa en la sangre. Sin embargo, el ejercicio por sí solo no es suficiente para predecir correctamente el riesgo, por lo que incorporamos una segunda variable, el índice de masa corporal (IMC o BMI en inglés). En este caso, el riesgo está dado por la interacción de estas dos variables. El asterisco (*) se utiliza para indicar todos los efectos principales e interacciones entre las variables a las que se une. Entonces, por ejemplo, el término "ejercicio * IMC" se expandiría a los tres términos: ejercicio, IMC, y ejercicio: IMC.

El símbolo ":" indica que se desea solo el efecto combinado de las variables. Por ejemplo, ejercicio:IMC representa solo el efecto combinado de estas variables y no el efecto de cada variable por separado.

PREDICCIÓN:

Una vez establecido el modelo (lm), podemos obtener predicciones para los valores observados y nuevos valores, para lo cual usaremos la función prodict(). La siguiente figura muestra la sintaxis de esta función.

Figura 3

La función de predicción

Fuente: www.rdocumentation.org

```
# S3 method for lm
predict(object, newdata, se.fit = FALSE, scale = NULL, df = Inf,
        interval = c("none", "confidence", "prediction"),
        level = 0.95, type = c("response", "terms"),
        terms = NULL, na.action = na.pass,
        pred.var = res.var/weights, weights = 1, …)
```

Esta función retorna un objeto del tipo lm, que contiene una cantidad de elementos como se muestra en la figura 4. Este objeto será pasado como argumento a la función predict para la predicción.

Además, debemos indicar cuáles son los valores de las variables predictoras para los cuales se debe realizar la predicción. Hay dos casos: el primero es realizar una predicción de las mismas observaciones y el segundo consiste en predecir nuevos valores o valores a futuro.

En el primer caso, usaremos las mismas observaciones como valores de entrada para la predicción. Uno podría preguntarse: ¿A qué propósito sirve esto? La respuesta es simple: es una forma rápida de evaluar la bondad del modelo.

En el segundo caso, debernos ingresar un dataframe (newdata) cuyas columnas contengan los valores a predecir.

Figura 4

Características de un modelo

Creación del autor Alfonso Prado

```
> names(padre_hijoLM)
 [1] "coefficients"  "residuals"
 [3] "effects"       "rank"
 [5] "fitted.values" "assign"
 [7] "qr"            "df.residual"
 [9] "xlevels"       "call"
[11] "terms"         "model"
```

Adicionalmente, como hemos visto, las predicciones nunca serán exactas, existe una cierta incertidumbre dentro del cual estimamos que estará la predicción correcta. Otra vez, aquí tenemos 2 opciones.

El primer caso se trata de predecir la respuesta para un valor medio. De acuerdo con Szretter M.(2017) "Nos interesa estimar la respuesta media o esperada cuando (X1, . . . , Xp−1) toma el valor dado (Xh1, . . . , Xh,p−1) . Notamos a esta respuesta media por E (Y_h) o bien E (Y_h | (Xh1, . . . , Xh,p−1)) "

Por ejemplo, (simplificando con un modelo de regresión simple) podemos tener algunas observaciones para el mismo valor de la variable predictora, con distintos resultados de la variable de respuesta. Esto nos dará un intervalo de confianza dentro del cual asumimos que se incluye el valor de la predicción.

En el segundo caso, de acuerdo con Szretter M.(2017) , " queremos predecir un resultado individual surgido a partir de la distribución de Y. Por supuesto, la gran mayoría de los resultados individuales se desvían de la respuesta media, y esto debe ser tenido en cuenta por el procedimiento para la predicción de la Y_h(nueva)." Por lo tanto, un intervalo de predicción será generalmente mucho más amplio que un intervalo de confianza para el mismo valor.

El argumento de "Interval" permite seleccionar el caso, mediante el valor de "confidence" (figura 4a) o Prediction" (figura 4b) y se lo puede visualizar mediante las rectas punteadas en rojo.

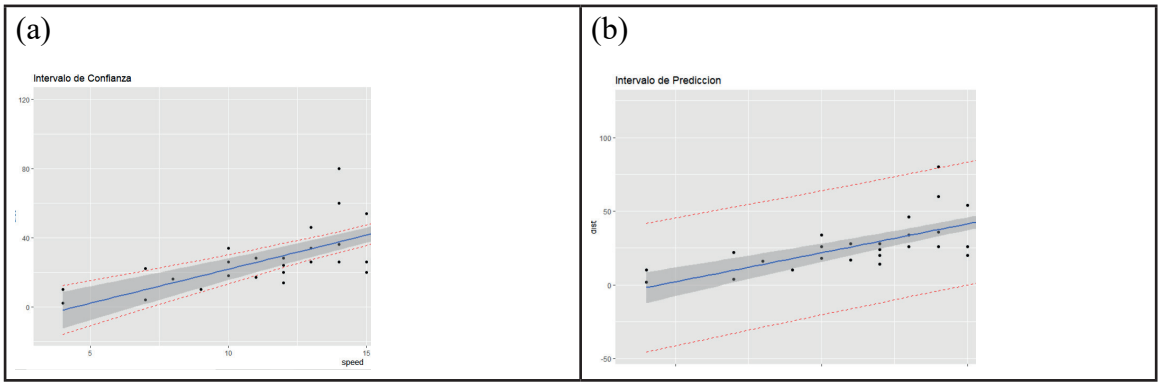
¿Cuál deberíamos usar? En general, estamos interesados en predicciones individuales específicas, por lo que un intervalo de predicción sería más apropiado.

De acuerdo con Bruce (2017) "El uso de un intervalo de confianza cuando debería usar un intervalo de predicción subestimará en gran medida la incertidumbre en un valor predicho dado "

Figura 5

Intervalos de Confianza y Predicción

Creación del autor Alfonso Prado



Diferencias de intervalos

Describe los usos de intervalo de predicción y confianza

Enlace

12.1) El error estándar de la estimación

Dentro de la salida de la regresión encontramos el término error estándar residual, es una medida de la variabilidad de los residuos de un modelo lineal.

Figura 6

Error estándar de los residuos

Creación del autor Alfonso Prado

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.3080154 1.5789429 5.262 8.54e-07
gestage 0.4487328 0.0672460 6.673 1.56e-09
birthwt 0.0047123 0.0006312 7.466 3.60e-11
---
Residual standard error: 1.274 on 97 degrees of freedom
Multiple R-squared: 0.752, Adjusted R-squared: 0.7469
F-statistic: 147.1 on 2 and 97 DF, p-value: < 2.2e-16
```

Este estadístico sirve para comparar que tan bueno es este modelo, pero no sabemos contra que debemos comparar. Para cada modelo estimaremos la suma de cuadrados de los residuos, esto estaría dado por la siguiente fórmula

Figura 7

Suma de cuadrados de los residuos

Creación del autor Alfonso Prado

$$\Delta\text{modelo} = \Sigma(\text{observados} - \text{modelo})^2 = \text{SSRes}$$

Entonces, usamos las sumas de cuadrados para calcular el ajuste del

modelo más básico (que contiene un solo parámetro que ajuste a todas las observaciones). Es decir, elegimos el valor de μ tal que minimice la suma de cuadrados con respecto a la media, y esta es calcu-

lada sin tener en cuenta para nada los valores de las covariables (X1, . . . , Xp−1). En otras palabras, el modelo más básico solo toma en cuenta el intercept y no las variables, y la resta de los dos valores indicará que tan bueno es el modelo con respecto a la media, como se muestra en la siguiente figura.

Figura 8

Error Estándar residual

Creación del autor Alfonso Prado

$$\Delta_{\text{modelo básico}} - \Delta_{\text{regresión lineal}} = \text{SSTo} - \text{SSRes}$$

Donde:

SSTo = suma de cuadrados totales

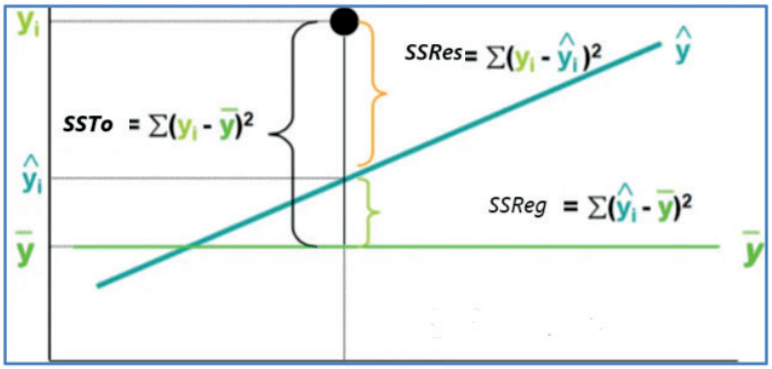
SSRes= suma de cuadrados de la predicción a la media

El siguiente gráfico clarifica estos conceptos

Figura 9

Relación entre SSTo, SSRes, SSReg

Creación del autor Alfonso Prado



Básicamente, si el modelo lineal es razonablemente bueno ajustará a los datos significativamente mejor que el modelo básico.

12.2) El problema de la multicolinealidad

Además del tema relacionado con los valores atípicos, existen 2 problemas que pueden afectar el rendimiento del modelo.

Sobreajuste

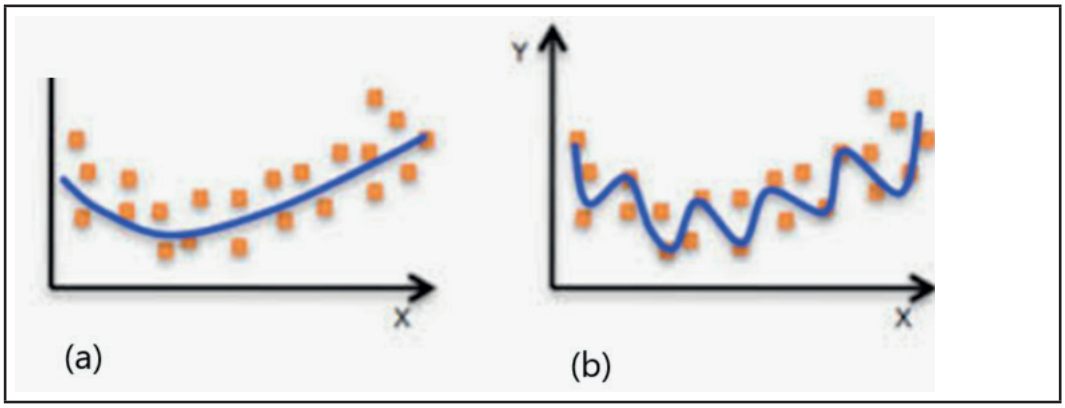
El sobreajuste es un error de modelado que ocurre cuando un modelo se ajusta demasiado a las observaciones. Generalmente, esto sucede al crear un modelo demasiado complejo para explicar los datos observados, lo cual conlleva a un mal desempeño al predecir valores no observados o futuros.

Como consecuencia, existirá sobreajuste entre el modelo y la muestra cuando lo que se busca es un modelo que, si bien ha sido entrenado con una muestra específica, sea capaz de realizar buenas predicciones sobre valores no contenidos en ella. La siguiente figura explica este concepto:

Figura 10

Comparación de modelos normal (a) y sobre ajustado (b)

Fuente AWS documentation



Algunos factores que contribuyen al sobre ajuste son:

- ✓ La cantidad de variables incorporadas al modelo es alta
- ✓ La muestra es demasiado pequeña y no contiene posibles valores futuros
- ✓ El dato de la muestra contiene información aleatoria que sobre la cual se entrena el modelo

De existir sobre ajuste, podemos utilizar el mecanismo de validación cruzada herramienta que per-

mite limitar la data de entrenamiento. Existen numerosas técnicas de validación cruzada, pero la más efectiva es la llamada k-fold. Este mecanismo lo veremos en detalle en la clase 16.

Multicolinealidad

De acuerdo con Webster (2000) “Este problema surge cuando unas de las variables independientes están relacionadas están relacionadas entre sí. Esta situación contraviene una de las condiciones de la regresión múltiple que es que las variables predictoras son independientes entre sí. Esto produce la incapacidad de separar los efectos individuales de cada variable independiente sobre Y. Ante la presencia de la multicolinealidad es imposible desenmarañar los efectos de cada X.”

Multicolinealildad

Describe en detalle este fenómeno

Enlace

Detección de la Multicolinealidad con la función VIF

VIF es un diagnóstico simple, conocido como factor de inflación de la varianza.

Hay varios paquetes que disponen de esta función; el más común es HH, el cual calcula el VIF para cada predictor y lo compara con el resto.

El R^2 de un predictor se compara con los predictores restantes, y, por lo tanto, se determina si la presencia del predictor i en el modelo es redundante.

Los valores calculados de VIF superiores a 5 se consideran sospechosos, y valores por encima de 10 son evidencia clara de multicolinealidad.

Comparación de modelos

En esta sección daremos algunas de las mejores prácticas para el desarrollo de modelos.

Es importante entender que el desarrollo de un modelo de predicción es un proceso iterativo, en el cual el analista prueba distintas combinaciones de variables predictoras, valida las métricas de bondad del modelo, realiza ajustes y vuelve a iniciar el ciclo.

La selección correcta de variables basados en la correlación, incluir variables con baja correlación solo aumentará el error. Por otro lado, es importante validar la linealidad de la relación, para lo cual la visualización mediante diagramas de caja y dispersión son las herramientas principales

Evaluar la métrica de R2 ajustado que es el principal indicador de la calidad del modelo, cuyo propósito es obtener el mejor modelo con la menor cantidad de variables, esto es fácil de deducir en base a

su fórmula de cálculo, que incluye el valor K, que representa la cantidad de variables del modelo.

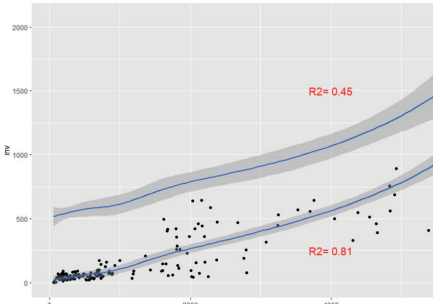
Más aún, al aumentar la cantidad de variables, es posible que el modelo sea afectado por la multicolinealidad. Las pruebas de VIF y análisis de correlación ayudan al analista a corregir este problema, encontrando el subconjunto óptimo de variables.

La siguiente figura muestra las observaciones de 2 modelos de regresión múltiple. Note como el modelo que obtiene R^2=0.81 se ajusta mejor a las observaciones que aquella que obtiene un R^2=0.45

Figura 12

Ajuste de 2 modelos

Creación del autor Alfonso Prado



Por último, es importante validar si el modelo resultante puede contener un sobre ajuste, esto lo haremos mediante mecanismos de validación cruzada, los cuales ayudará a crear un modelo más equilibrado en cuanto a la predicción de observaciones vs. valores nuevos.

Veamos un ejemplo completo

```
housing <- read.table("housing.csv" , sep=",", header=TRUE, stringsAsFactors=FALSE)
names(housing)
```

```
 [1] "sector"       "class"        "TUnits"       "Year"         "GSqFt"
 [6] "Est_GIncome"  "GIncomexSqFt" "Est_Expense"  "ExpSqFt"      "NetIncome"
[11] "FullValue"    "valxSqFt"     "Boro"
```

```
#La variable dependiente es valxSqFt
#Las siguientes variables parecen ser de interés 3,5:12
#Obteniendo correlación
cor_housing  <- cor(housing[,c(3,5:12)])
#Creamos unas funciones para visualizar la correlación
get_lower_tri<-function(x){
  x[upper.tri(x)] <- NA
  return(x)
}
get_upper_tri <- function(x){
  x[lower.tri(x)]<- NA
  return(x)
}
reorder_cor_mat <- function(x){
  dd <- as.dist((1-x)/2)
  hc <- hclust(dd)
  x <-x[hc$order, hc$order]
}
#Visualizando
```

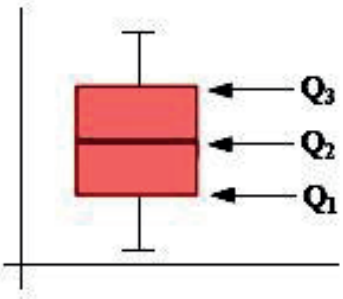
Referencias citadas en la Clase 12.

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

Szretter, M. (2017), Apunte de Regresión Lineal, Universidad de Buenos Aires

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 1.

COEFICIENTE DE DETERMINACIÓN AJUSTADO **r^2**	El R cuadrado ajustado es una versión modificada del R cuadrado que tiene en cuenta los predictores que no son significativos en un modelo de regresión. En otras palabras, el R cuadrado ajustado muestra si la adición de predictores adicionales mejora o no un modelo de regresión. El coeficiente de determinación mide la proximidad del ajuste de la ecuación de regresión de la muestra a los valores observados de la variable dependiente.
DIAGRAMA DE CAJAS	Conocido también como **BOXPLOT.** Es un importante gráfico del análisis exploratorio de datos. Al igual que el histograma, permite tener una idea visual de la distribución de los datos. Permite determinar si hay simetría, ver el grado de variabilidad existente y detectar los "outliers" (datos muy diferentes al conjunto de información), es decir la existencia de posibles datos discordantes. Además, el Boxplot es bien útil para comparar grupos. Es un diagrama que muestra la distancia en que se encuentran los datos y cómo están distribuidos equitativamente. 

SUMA DE CUADRADOS DE RESIDUOS	La suma de cuadrados residuales , también conocida como la suma de los residuos al cuadrado o la suma de la estimación al cuadrado de los errores , es la suma de los cuadrados de los residuos (desviaciones predichas de los valores empíricos reales de los datos) . Es una medida de la discrepancia entre los datos y un modelo de estimación. Un pequeño SSRes indica un ajuste ajustado del modelo a los datos. Se utiliza como criterio de optimización en la selección de parámetros y la selección de modelos. En el contexto de ANOVA este estadístico se llama SST Dado por la fórmula siguiente: $SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$

Profundización Clase 4.

Estadistica-clase12

PUCE Virtual

La excelencia no se improvisa

síguenos