PUCE Virtual
EDUCACIÓN FLEXIBLE

Estadística 1 para administración

Regresión múltiple

Clase 12

La excelencia no se improvisa

INTRODUCCIÓN DE LA CLASE

En la clase anterior revisamos el concepto de la regresión lineal simple, sin embargo, la mayoría de los negocios, procesos, eventos etc. a modelar contienen una gran cantidad de variables que afectan a la variable de respuesta, por lo que, hacer un modelo con una sola variable resultaría en un error estándar considerable.

En estos modelos, aparte de descubrir que variables son importantes, vamos a ver una serie de consideraciones que debemos tener en cuenta, entre las cuales se halla la multicolinealidad y el sobreajuste, que, de no controlarse, nos podrían llevar a un error de predicción exagerado.

Clase 12: Regresión múltiple

En general, cuando tratamos de modelar un proceso vamos a encontrar que existen múltiples variables que afectan su resultado. Un modelo de regresión lineal múltiple contendrá múltiples variables predictoras. Sin embargo, no todas las variables afectarán en igual medida al modelo.

Un modelo múltiple es básicamente el resultado de varios modelos de regresión simple combinados, por lo que, la fórmula general es la misma para cada una de las variables. En un modelo de regresión múltiple Y es una función de dos o más variables predictoras

Figura 1
Modelos de regresión como función de múltiples (p) variables
Adaptación de la fórmula según Webster capítulo 11

$$Y = f(X_1, X_2, X_3 \ldots . X_p)$$

Cada una de estas variables en realidad van a ser pequeños modelos simples que van a aportar al modelo múltiple a través de sus propios coeficientes, el error estándar y las observaciones. Matemáticamente podemos definir estas relaciones a través de la siguiente fórmula.

Notamos que las variables predictoras (X) se halla en mayúscula, indicando que éstos 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 3a). Los coeficientes (ahora llamados β) están representados por un vector de longitud p (figura 3b). El error de cada regresión es también un vector de longitud n, esto es un error por cada observación (figura 3c), la variable dependiente Y (en mayúscula) es también es una matriz que representa la contribución de cada variable (figura 3d).

Figura 2

Fórmula general de regresión múltiple

Creación del autor Alfonso Prado

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

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 3

Componentes de regresión múltiple

Creación del autor Alfonso Prado

(a)

$$\underset{n\times p}{\mathbf{X}} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}$$

(b)

$$\underset{p\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$$

(c)

$$\underset{n\times 1}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

(d)

$$\underset{n\times 1}{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

Funciones Programáticas de regresión múltiple

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

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

Por ejemplo, en el dataset de housing el valor de una propiedad esta dado por el área de la propiedad (GSqFt), el vecindario (Boro), el año de construcción (YearBuilt) entre otras variables, en cuyo caso el modelo sería creado asi: lm(ValuePerSqFt ~YearBuilt + GSqFt + Boro , data=housing)

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 la diabetes, se ha considerado como variable predictora el ejercicio que la persona realiza, como mecanismo para reducir la cantidad de glucosa en la sangre, pero 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), entonces el riesgo estará dado por la interacción de estas dos variables. El asterisco (*) se utiliza para indicar todos los efectos e interacciones principales 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 predict(), la siguiente figura muestra la sintaxis de esta función.

Figura 4

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, …)
```

La función lm retorna un objeto del tipo lm, mismo que contiene una cantidad de elementos como se muestra en la figura 5. Este objeto será pasado como argumento a la función predict para la predicción.

En la función predict, debemos indicar cuales son los valores de las variables predictoras para las cuales se debe realizar la predicción. Aquí se nos presenta 2 casos: el primero es realizar una predicción de las mismas observaciones y el segundo se trataría de predecir nuevos valores o valores a futuro.

En el primer caso, usaremos las mismas observaciones como valores de entrada para la predicción. Uno se 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 en el argumento newdata cuyas columnas contengan los valores a predecir.

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

Figura 5

Propiedades de objeto modelo

Creación del autor Alfonso Prado

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 a 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.

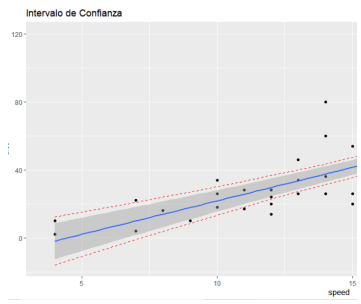
En el segundo caso, de acuerdo a Szretter M.(2017) , “..queremos predecir un resultado individual surgido a partir de la distribución de Y. Por supuesto, la gran mayoría de los resultados individuales se desvían de la respuesta media, y esto debe ser tenido en cuenta por el procedimiento para la predicción 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 5a) o “prediction” (figura 5b) y se lo puede visualizar mediante las rectas punteadas en rojo.

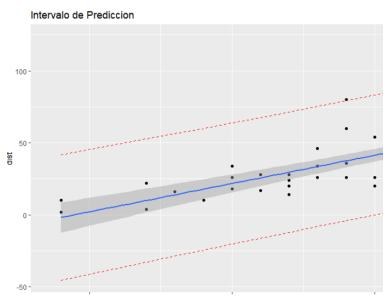
¿Cuál deberíamos usar? En general, estamos interesados en predicciones individuales específicas, por lo que un intervalo de predicción sería más apropiado.
De acuerdo a Bruce (2017) “El uso de un intervalo de confianza cuando debería usar un intervalo de predicción subestimará en gran medida la incertidumbre en un valor predicho dado “

Figura 6
Intervalos de Confianza y Predicción
Creación del autor Alfonso Prado

(a)



(b)



Diferencias de intervalos

Describe los usos de intervalo de predicción y confianza

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.

```
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
```

Figura 7

Error estándar de los residuos

Creación del autor Alfonso Prado

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 8

Suma de cuadrados de los residuos

Creación del autor Alfonso Prado

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

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 ésta es calculada 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 9

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

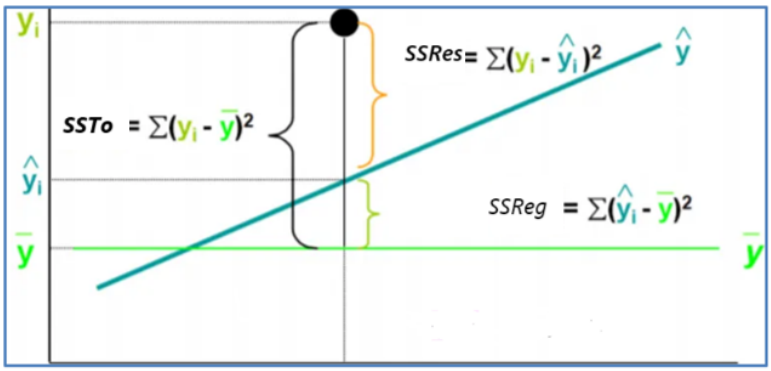
SSRes= suma de cuadrados de los residuos

El siguiente gráfico clarifica estos conceptos

Figura 10

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.

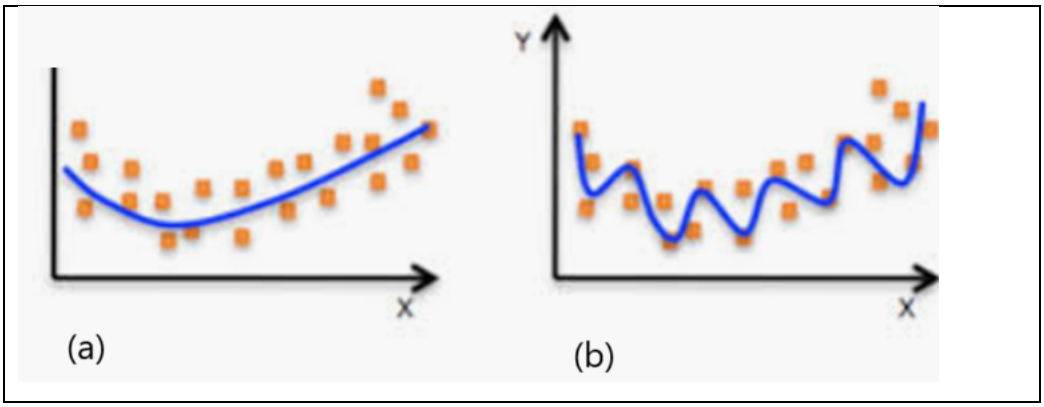
El problema de la multicolinealidad

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

Sobre ajuste

El sobre ajuste es un error de modelado que ocurre cuando un modelo se ajusta demasiado a las observaciones. El sobreajuste del modelo generalmente toma la forma de hacer un modelo demasiado complejo para explicar los datos observados, pero tiene una mala predicción sobre valores no observados o futuros.
Como consecuencia existirá un sobre ajuste entre el modelo y la muestra, cuando lo que se busca es un modelo que, si bien ha sido entrenado con una muestra específica, nos de una buena predicción sobre valores no contenidos en la muestra. La siguiente figura explica el concepto.

Figura 11
Comparación de modelos normal (a) y sobre ajustado (b)
Fuente https://docs.aws.amazon.com/



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 sirve para limitar la data de entrenamiento, si se limita la data de entrenamiento mal puede el modelo ajustarse a datos que no figuran en subset de entrenamiento.

Multicolinealidad

De acuerdo a 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."

Detección de la Multicolinealidad con la función VIF

Vif es un diagnóstico simple conocido como factor de inflación de varianza.
Hay varios paquetes que disponen de la función, el más común es HH, el cual calcula el Vif para cada predictor y lo compara con el resto.
El R^2 de cada predictor se compara con los predictores restantes y por lo tanto 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 sobre 10 son evidencia 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 las distintas combinaciones de variables predictoras, valida las métricas de bondad del modelo, realiza correcciones 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 de regresión múltiple, 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.

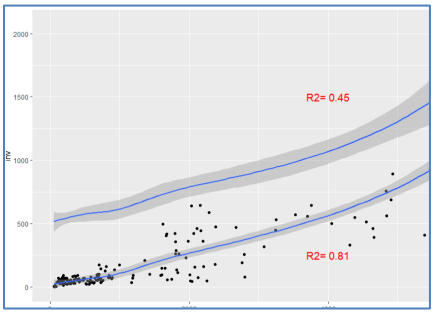
Figura 12
Formula de R^2 ajustado
Creación del autor Alfonso Prado

$$\bar{R}2 = 1 - \frac{SCE(n-k-1)}{SCT/(n-1)}$$

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

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 13
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.

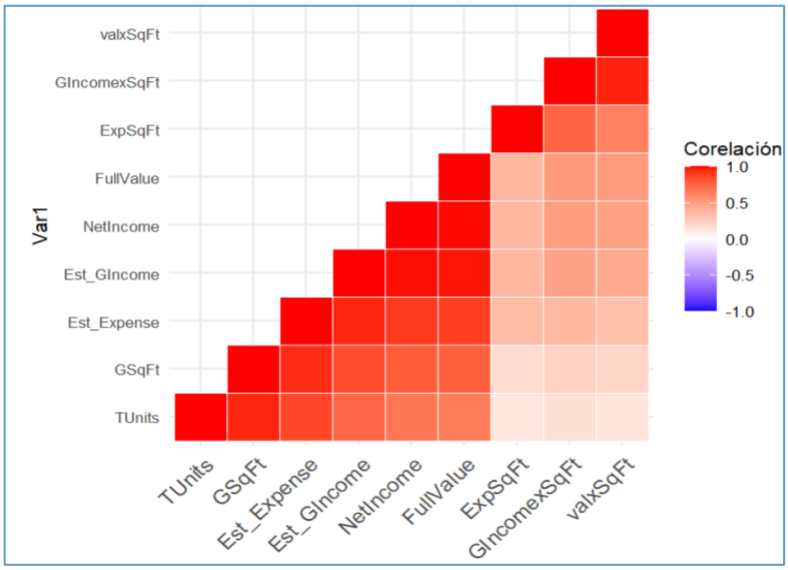
Veamos un ejemplo completo

```
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
cor_mat <-reorder_cor_mat(cor_housing)
#Obtenenos el triángulo superior
upper_tri <- get_upper_tri(cor_mat)
# Transponemos esta matriz
melted_cor_mat <- melt(get_upper_tri(cor_mat), na.rm = TRUE)
# Creando mapa de calor
ggplot(melted_cor_mat, aes(Var2, Var1, fill = value))+
```

```
geom_tile(color = "white")+
scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                     midpoint = 0, limit = c(-1,1), space = "Lab",
                     name="CorelaciÃ³n Ordenada") +
theme_minimal()+ # minimal theme
theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                 size = 12, hjust = 1))+
coord_fixed()
```




```
#Que observamos en este gráfico.
#Vemos que las variables que mayor correlación tienen con la respuest son
#GIncomexSqFt, ExpSqFt, FullValue, NetIncome.
#Y las variables de menor importancia serían TUnits y  GSqFt
#Para la selección deberemos tambien validar que las variables predictoras no esten correlacionadas entre si.


#Creamos el modelo
house1 <- lm(valxSqFt ~TUnits+ GSqFt +Boro  , data=housing)
#Boro es una variable categórica que ha sido codificada (Veremos esto es siguientes módulos)
summary(house1)
```

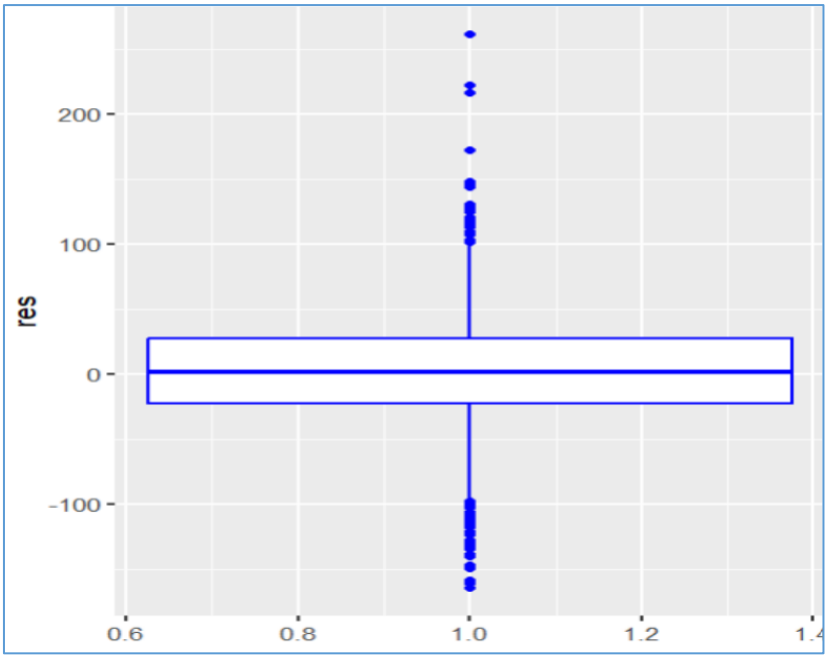
```
Residuals:
     Min       1Q   Median       3Q      Max
-164.418  -22.692    1.416   26.972  261.122

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.329e+01  5.330e+00   8.122 6.97e-16 ***
TUnits           -1.881e-01  2.210e-02  -8.511  < 2e-16 ***
GSqFt             2.103e-04  2.087e-05  10.079  < 2e-16 ***
BoroBrooklyn      3.456e+01  5.535e+00   6.244 4.95e-10 ***
BoroManhattan     1.310e+02  5.385e+00  24.327  < 2e-16 ***
BoroQueens        3.299e+01  5.663e+00   5.827 6.35e-09 ***
BoroStaten Island -3.630e+00  9.993e+00  -0.363    0.716
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43.35 on 2619 degrees of freedom
Multiple R-squared:  0.6009,    Adjusted R-squared:    0.6
F-statistic: 657.2 on 6 and 2619 DF,  p-value: < 2.2e-16
```

#analizando residuos

resdf <- data.frame(res=house1$residuals)

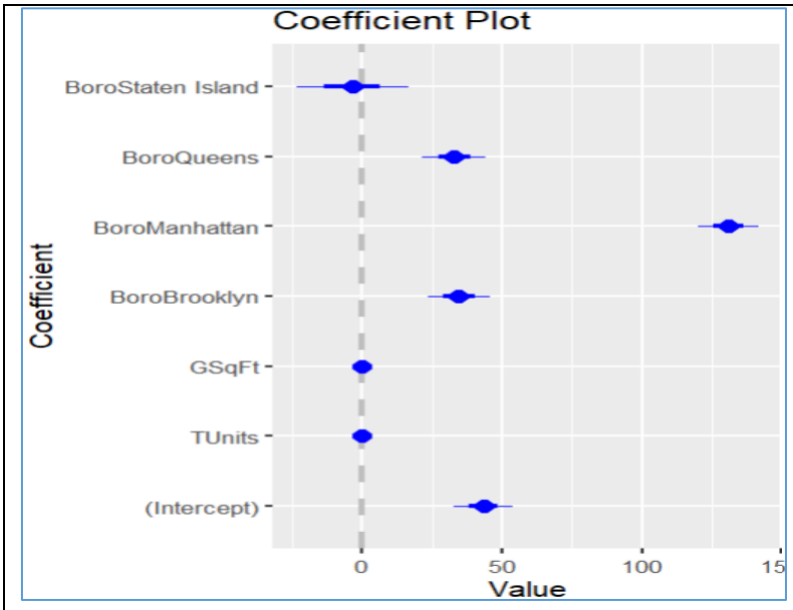
ggplot(data=resdf ,aes (y=res, x=1))+

geom_boxplot(color="blue")



#Analizando coeficientes

coefplot(house1)



#Analizando Intervalos de confianza

confint(house1)

	2.5 %	97.5 %
(Intercept)	3.283971e+01	5.374202e+01
TUnits	-2.314538e-01	-1.447764e-01
GSqFt	1.694025e-04	2.512432e-04
BoroBrooklyn	2.370859e+01	4.541571e+01
BoroManhattan	1.204338e+02	1.415509e+02
BoroQueens	2.189012e+01	4.409723e+01
BoroStaten Island	-2.322513e+01	1.596463e+01

#Analizando multicolinealidad

house1 <- lm(valxSqFt ~TUnits+ GSqFt +Boro , data=housing)

length(coef(house1))-1

vars <- c("TUnits" , "GSqFt" , "BoroBrooklyn" , "BoroManhattan" ,
"BoroQueens" , "BoroStaten Island")

house1vif <- HH::vif(house1)

house1r2 <- 1/(1- summary(house1)$r.squared)

sprintf("%s %f debe ser menor que %f para no multicolinealidad", vars , house1vif, house1r2)

#Obtenemos

[1] "TUnits 12.46 debe ser menor que 2.51 para no multicolinealidad"

[2] "GSqFt 12.62 debe ser menor que 2.51 para no multicolinealidad"

#Concluimos que TUnits y GsqFt tienen multicolinealidad

Referencias citadas en la Clase 12

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