APLICACION DE REGRESION LINEAL EN LA PREDICION DE PRECIOS DE UN HOTEL

Gonzalez David,Pasachova Garzon Jonathan, Tores Catalina
jonathana.pasachovag@ecci.edu.co,davids.gonzalezm@ecci.edu.co,catalina.torresa@ecci.edu.co

***Resumen*—Mediante el presente documento se hizo el analisis de la base de datos recolectada en un Hotel por las reservaciones de los cuartos, queriendo hacer prediciones de los precios de las habitaciones con conocimientos previos en regresion, para esto se construyo un modelo de regresion simple de igual manera haciendo todo el proceso de validacion de supuestos (Normalidad, Homocedasticidad, independencia) para asi tener mas certeza de que el modelo realizara predicciones correctas con un valor del significancia del 95 %.**

***Palabras clave*—**

Abstract This document attempts to determine the price of a hotel room based on the number of adults that will occupy it. The sample was estimated from 5000 reservations that was extracted from a Kaggle database. With the intention of explaining the relationship between variables, we assume that one of them will be dependent on the other and thus predict the behavior of these variables. a mathematical model used to establish the type of relationship between the dependency variable (response variable) between an independent variable.

Key words supuestos, huber normalidad,independencia,homocelasticidad.

I. Introducción

La regresión lineal es una técnica estadística que se utiliza para modelar la relación entre una variable dependiente y una o varias variables independientes. En otras palabras, la regresión lineal nos permite predecir el valor de una variable a partir de otra variable o varias variables.El objetivo de este proyecto es utilizar la regresión lineal simple para analizar un conjunto de datos y predecir el valor de una variable dependiente (precio promedio de la habitacion) en función de una o varias variables independientes. Para ello, se aplicará la teoría de la regresión lineal simple y se utilizarán diversas herramientas de programación en el leguaje R para implementar un modelo de regresión lineal,Diseñaremos tres modelos distintos en los cuales aplicamos herramientas como boxcox y regresión robusta para llegar a satisfacer los supuestos.

II. Concepto, metodología o técnica

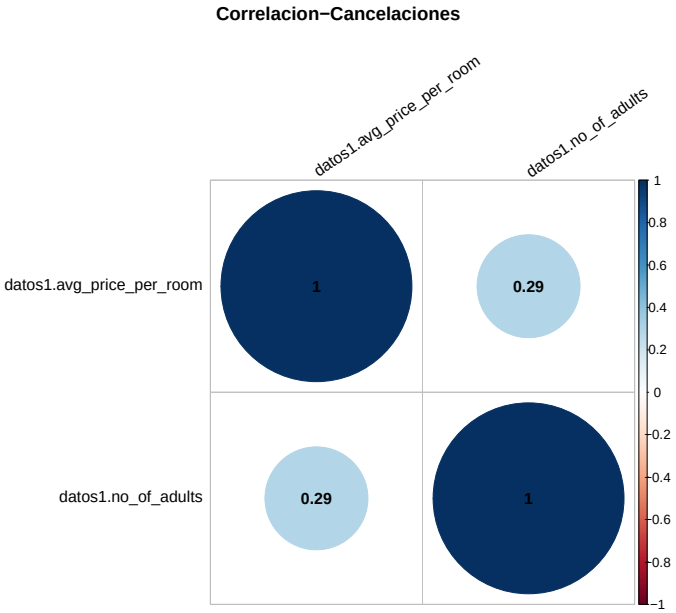
En cuanto al método utilizado en la regresión simple, utilizaremos el lenguaje de programación R para este fin. lo primero que haremos es tomar una muestra de 5000 unidades de nuestra base de datos y ajustar nuestro data frame.

```
### base de datos.
datos<- read.csv("Hotel Reservations.csv")
muestra<-sample(dim(datos)[1],5000)
datos1=datos[c(muestra),]

class(datos$booking_status)

## [1] "character"

datos$booking_status<-as.factor(datos$booking_stat
datos1<-datos1 %>% mutate(estado=case_when(booking_
```

II-A. Correlación

Por medio de un corplot podemos ver gráficamente la correlación entre las variables que queremos usar en nuestro modelo en este caso usaremos las variables 'avg price per room' que hace referencia al precio promedio pagado por habitación y 'no of adults' que se refiere al numero de adultos por reserva.

```
##                                datos1.avg_price_per_
## datos1.avg_price_per_room                      1.000
## datos1.no_of_adults                            0.290
```



II-B. Planteamiento del modelo-Modelo 1

```
datos1<-na.omit(datos1)
modelo<-lm(datos1$avg_price_per_room~datos1$no_of_adults)
for(i in 1:length(datos1$avg_price_per_room)){
if(datos1$avg_price_per_room[i]==0){
  datos1$avg_price_per_room[i]=0.05
}
}
```

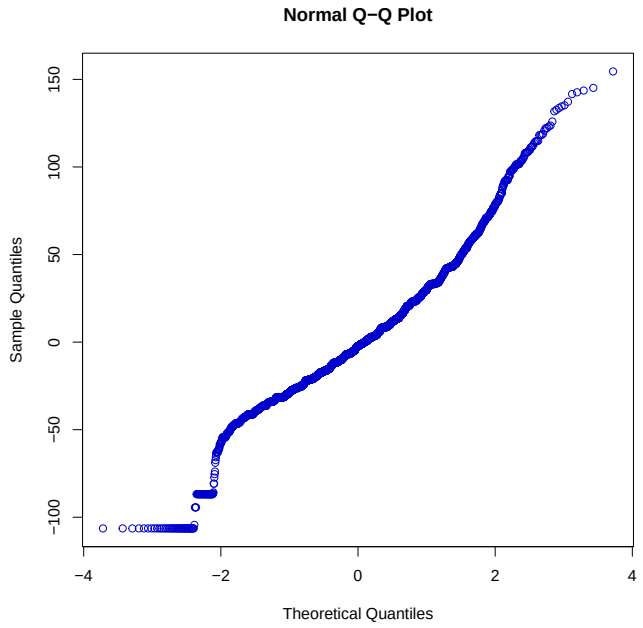
III. Supuestos

III-A. Normalidad

Primero provamos normalidad en nuestro modelo para esto usasaremos la prueba de shapiro-wilk que nos permite medir normalidad con alta presicion en conjuntos pequeños de datos.

Normalidad:
$Hi = \sigma\mathcal{E}$i ˜N(0,σ^2)
$Ha = \sigma\mathcal{E}$i —˜N(0,σ^2)
Homocedasticidad
$Hi = \sigma_1^2 = \sigma_2^2$
$Ha = \sigma_1^2 \neq \sigma_2^2$
Hay igualdad de varianzas
Independencia:
$\mathcal{E}$i $independiente\mathcal{E}$i
los residuos deben ser independientes entre sí.

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo$residuals
## W = 0.96501, p-value < 2.2e-16
```

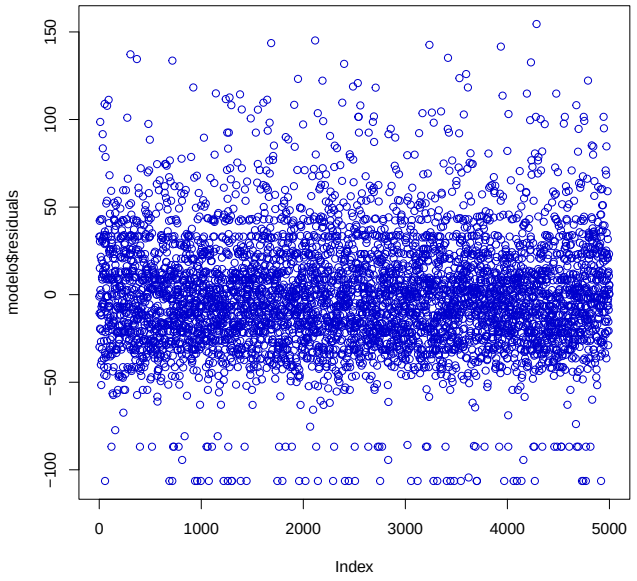


como podemos ver el p valor es muy pequeño es decir $p \leq \alpha$ por lo tanto se rechaza H_0 y por tanto se rechaza que los datos sigan una distribucion normal , este supuesto establece que los errores o residuos de la regresión deben seguir una distribución normal con media cero y varianza constante para que los resultados obtenidos sean válidos y confiables. En otras palabras, se espera que los errores estén distribuidos simétricamente alrededor de cero y que la mayoría de ellos se concentren cerca de cero, con menos errores a medida que se alejan de cero.

III-B. Homocedasticidad

Para el supuesto de homocedasticidad utilizaremos la prueba propuesta por Breusch y Pagan (Breusch-Pagan.test)

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo
## BP = 0.23769, df = 1, p-value = 0.6259
```



como vemos el p valor es menor que la significacncia por tanto rechazamos la hipotesis nula $H_0 : \sigma_a^2 = \sigma_b^2 = \sigma_c^2$ y concluimos que no se cumple el supuesto.este supuesto establece que la varianza del error aleatorio de la variable dependiente es constante para todos los valores de la variable independiente. Esto significa que la dispersión de los residuos es igual en toda la gama de los valores de la variable independiente. En otras palabras, la dispersión de los errores alrededor de la línea de regresión es constante y no cambia a medida que se mueve a lo largo del rango de los valores de la variable independiente. Si este supuesto no se cumple, puede haber heterocedasticidad en los datos, lo que puede conducir a conclusiones erróneas al hacer inferencias y estimaciones con el modelo de regresión lineal.

III-C. Independencia

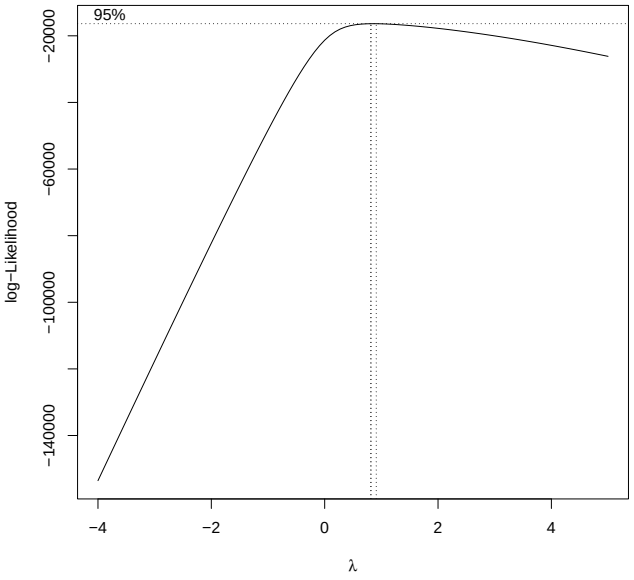
Para este supuesto utilizaremos el test de independencia de Durbin Watson

```
##
##  Durbin-Watson test
##
## data:  modelo
## DW = 1.971, p-value = 0.1525
## alternative hypothesis: true autocorrelat
```

Como vemos existe una correlación positiva entre las variables in embargo al no tener satisfecho el supuesto de normalidad no podemos tener independencia.

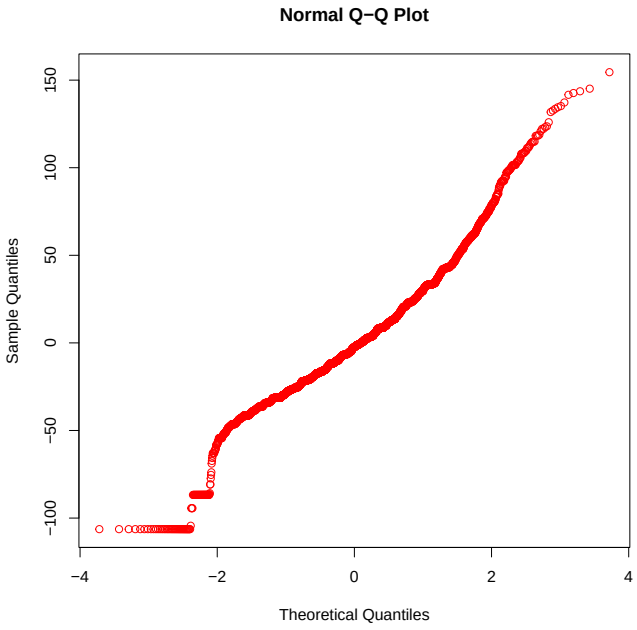
IV. Modelo 2

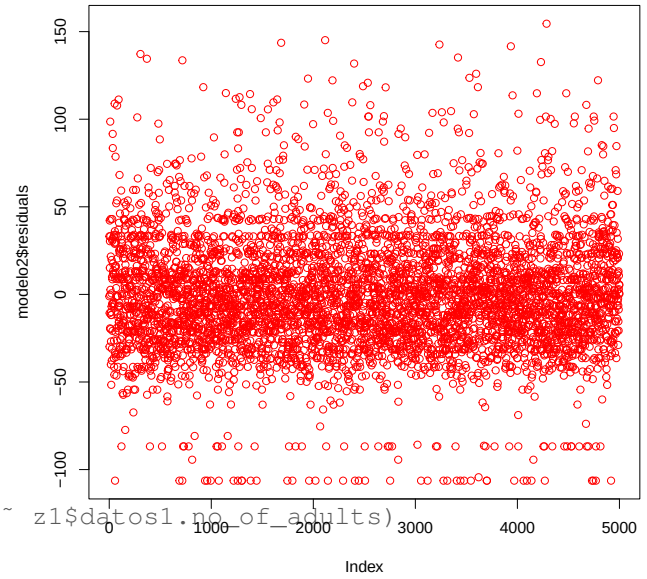
Para este modelo usaremos la técnica de Box-Cox, esta es una técnica estadística utilizada para transformar una variable de respuesta que no sigue una distribución normal en una variable que se aproxima a una distribución normal. Esta técnica fue propuesta por los estadísticos George Box y David Cox en 1964. La transformación Box-Cox implica la elección de un parámetro lambda () que determina el tipo y la magnitud de la transformación. El valor de lambda puede ser cualquier número real, pero se busca un valor que haga que los datos transformados se aproximen a una distribución normal.




```
##
## Call:
## lm(formula = z1$datos1.avg_price_per_room ~ z1$datos1.no_of_adults)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -106.327  -20.877   -2.377   17.638  154.523
##
```

```
## Coefficients:
##                          Estimate Std. Error t va
## (Intercept)               67.2585     1.7365   38
## z1$datos1.no_of_adults    19.5592     0.9106   21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.0
##
## Residual standard error: 33.17 on 4998 degrees
## Multiple R-squared:  0.08452,Adjusted R-squared
## F-statistic: 461.4 on 1 and 4998 DF,  p-value:
```





En este caso el Box-Cox no será suficiente para acercarnos a la distribución normal por lo tanto recurrimos al método de la regresión robusta.

V. Modelo 3

la pérdida de Huber es una función de pérdida particular (introducida por primera vez en 1964 por Peter Jost Huber, un matemático suizo) que se usa ampliamente para problemas de regresión robustos, situaciones en las que hay valores atípicos que pueden degradar el rendimiento y la precisión de los mínimos cuadrados.

VI. Resultados

Aqui realizamos un modelo robusto de regresion que explicacion anteriormente. como vemos que el t-value es mayor que 2 vemos que es sifnificativo para el modelo y el estimado positivo entonces concluimos que ante mayor numero de adultos el precio de la habitacion aumenta 19.000 al precio de cada habitacion.

```
##
## Call: rlm(formula = datos1$avg_price_per_room ~ datos1$no_of_adults)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -104.747  -19.797   -1.092   19.203  156.103
##
## Coefficients:
##                     Value   Std. Error t value
## (Intercept)         67.8962  1.5392    44.1103
## datos1$no_of_adults 18.4503  0.8071    22.8596
##
## Residual standard error: 29.17 on 4998 degrees of freedom
```

VII. Conclusiones

En conclusion podemos afirmar que el regresion simple tomando como ajercicio casos de la vida real es mas complicado construir modelos que puedan cumplir suspuestos y de igual manera que nos funcione para unas predicciones optimas, en estos casos podemos tomar como opcion alternativa modelos robustos para lograr el valor esperado.

Referencias

[1] Inicial1. Apellido1 and Inicial2. Apellido2, *Nombre de libro*, #edición ed. Ciudad, País: Editorial, año.
[2] Freddy Hernadez,Mauricio MazoLaTeX, 3rd ed.*analisis de regresion con R* #edición ed.
[3] Rpuds. https://rpubs.com/gpadmaperuma/695180#:~:text=Robust%20regression%20is%20an%20alternative,to%20see%20how%20them%20behave.. Recuperado el 30 de Enero de 2017.