Universidad Andrés Bello®

Actividad Evaluativa #2

Grupo 5

Universidad Andrés Bello
Facultad de Ingeniería
Ingeniería Civil Informática
Ingeniería Civil Industrial

Martín Fernández 1[1], Tomás Moya 2, Wesly Ocampo 3[1], Alan Tovar 4[1]

Abstract

Describir brevemente en qué consiste este primer análisis de los datos, e incluir el objetivo del estudio. (máximo 150 palabras)

*Corresponding author
Email addresses: nombre@uandresbello.edu (Martín Fernández 1), nombre@uandresbello.edu (Tomás Moya 2), nombre@uandresbello.edu (Wesly Ocampo 3), nombre@uandresbello.edu (Alan Tovar 4)

Introducción

En el presente informe se presentan los resultados de un análisis de dos modelos de regresión lineal múltiple con el objetivo de reducir las variables no significativas y trabajar con una base de datos más pequeña.

La reducción de dimensionalidad es la forma de convertir un conjunto de datos de dimensiones elevadas en un conjunto de datos de dimensiones menores, asegurando que la información que proporciona es similar en ambos casos. Esta técnica nos permite obtener un modelo predictivo más ajustado mientras se resuelven los problemas de regresión y clasificación que presentan los algoritmos.

Por otro lado, la regresión logística resulta útil para los casos en los que se desea predecir la presencia o ausencia de una característica o resultado según los valores de un conjunto de predictores. Es similar a un modelo de regresión lineal pero está adaptado para modelos en los que la variable dependiente es dicotómica, es decir que puede tomar solo dos valores. Los coeficientes de regresión logística pueden utilizarse para estimar la razón de probabilidad de cada variable independiente del modelo. La regresión logística se puede aplicar a un rango más amplio de situaciones de investigación que el análisis discriminante.

Desarrollo

Análisis de Componentes

Para el análisis de componentes principales no tomamos en cuenta la variable cantidad de bicicletas rentadas ya que corresponde a la variable dependiente y eliminando las variables categóricas dejando únicamente las variables numéricos, obteniendo un total de 10 variables, observando el gráfico de proporción de varianzas a la izquierda podemos ver que tiene porcentajes bastante disparejos presentando una diferencia de porcentajes de alrededor del 27% entre el primer y último componente, según el gráfico de varianza acumulada con 4 componentes se puede lograr el 70% de la variabilidad total de los datos.

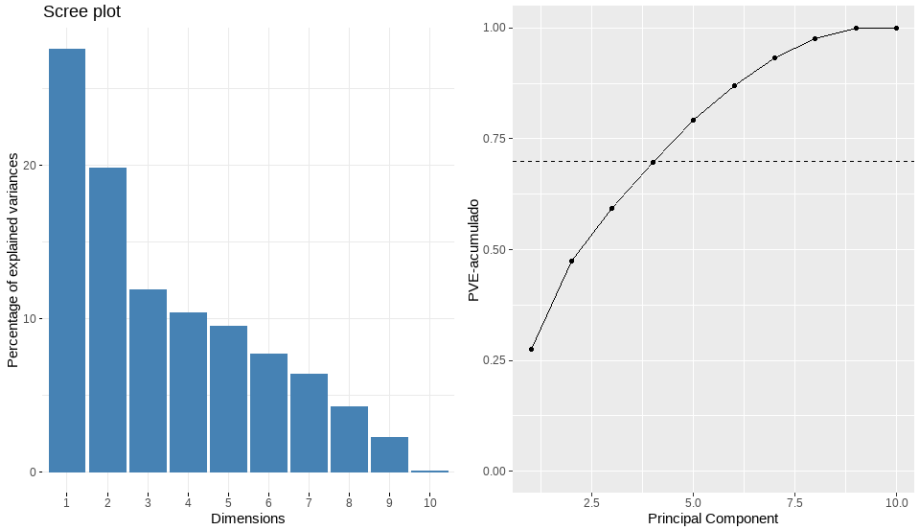

Figure 1: Análisis de Componentes.

En el gráfico de contribución de cada variable podemos observar que en las 2 primeras dimensiones tenemos un total del 47,5% de la variabilidad total de los datos, dentro de la dimensión 1 los componentes que contribuyen significativamente de forma positiva son radiación solar con gran contribución, mientras que en menor medida Hora, velocidad del viento, visibilidad y temperatura. De forma negativa en la dimensión 1 serian humedad con gran contribución y en menor contribución serian lluvia y nevada. En la dimensión 2 los componentes que contribuyen significativamente de forma positiva temperatura de punto de rocío y temperatura con gran contribución y en menor contribución, humedad y lluvia. De forma negativa en la dimensión 2 serian nevada y visibilidad con poca contribución

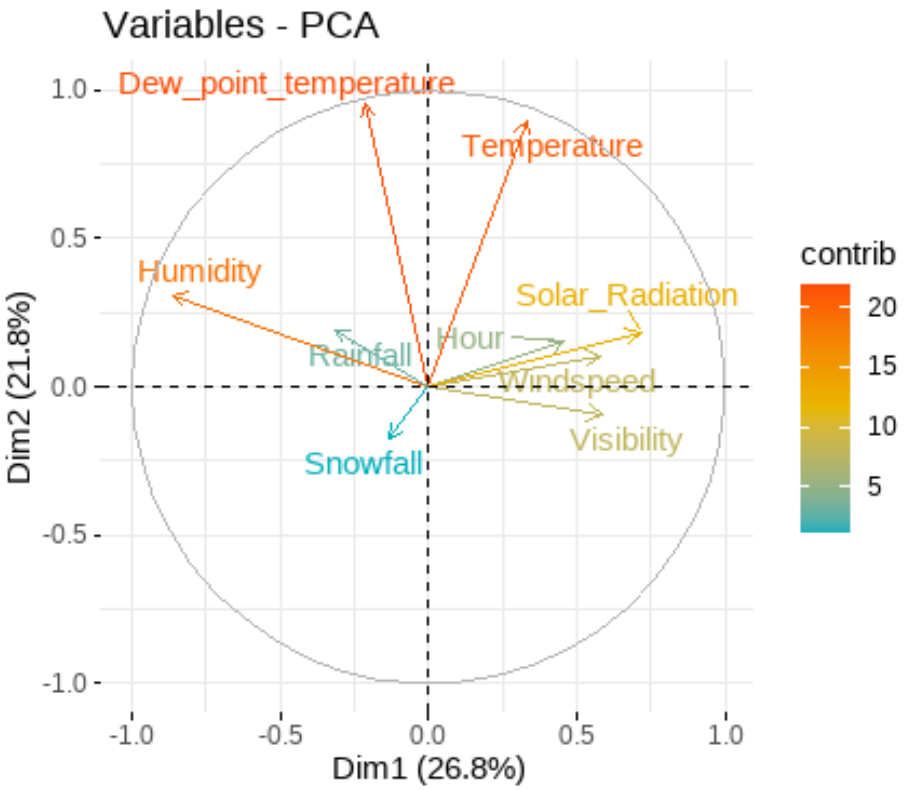

Figure 2: Análisis de Componentes.

Table 1: Modelo 1.

	Variables	Estimación	Std.Error	t.value	P-valor
(Intercept)	(Intercept)	862.41	143.67	6.00	0.00
Hour	Hour	39.27	1.22	32.27	0.00
Temperature	Temperature	16.48	5.64	2.92	0.00
Humidity	Humidity	-12.22	1.57	-7.76	0.00
Windspeed	Windspeed	23.55	8.85	2.66	0.01
Visibility	Visibility	0.00	0.02	0.16	0.87
Dew_point_temperature	Dew_point_temperature	6.07	5.80	1.05	0.30
Solar_Radiation	Solar_Radiation	-108.85	11.44	-9.52	0.00
Rainfall	Rainfall	-55.82	6.08	-9.19	0.00
Snowfall	Snowfall	-2.48	24.79	-0.10	0.92

Regresión Lineal Múltiple.

Modelo 1.

Para el primer modelo podemos ver como las variables Visibility, Dew_point_temperature y Snowfall no son significativas ya que su p-valor es mayor a 0.05 o 5%.

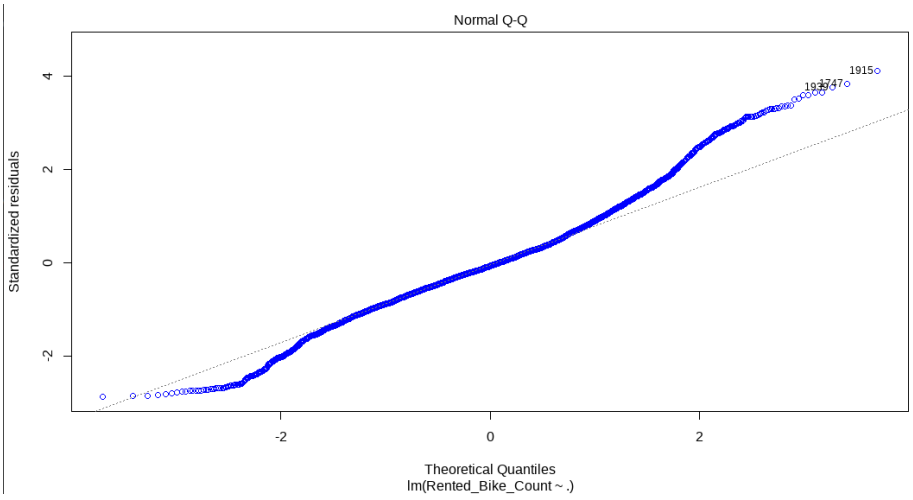

Figure 3: Q-Q plot Modelo 1.

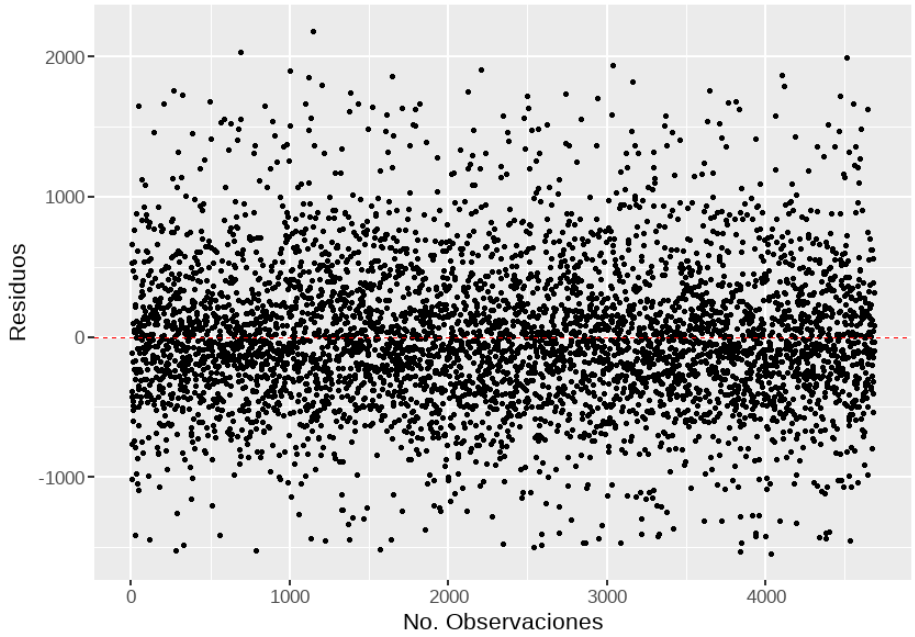

Figure 4: Residuos Modelo 1.

```
##
##  Shapiro-Wilk normality test
##
## data:  dat.train$resid
## W = 0.97858, p-value < 2.2e-16
```

Además no se cumplen los supuestos de normalidad.

Table 2: Modelo 2.

	Variables	Estimación	Std.Error	t.value	P-valor
(Intercept)	(Intercept)	733.12	41.70	17.58	0.00
Hour	Hour	39.18	1.21	32.38	0.00
Temperature	Temperature	22.36	1.03	21.70	0.00
Humidity	Humidity	-10.76	0.50	-21.54	0.00
Windspeed	Windspeed	23.14	8.84	2.62	0.01
Solar_Radiation	Solar_Radiation	-112.11	10.81	-10.37	0.00
Rainfall	Rainfall	-56.67	6.01	-9.43	0.00

Modelo 2.

Para este modelo 2 todas nuestras variables son significativas.

Table 3: Medidas de Comparación de los Modelos.

Medidas	Modelo_1	Modelo_2
R2.ajustado	41.12	41.15
COR	0.63	0.63
BIAS	-1.31	-1.49
RMSE	532.97	532.82

Comparación de Modelos.

R cuadrado ajustado nos indica el porcentaje de variabilidad total de la variable de respuesta, en este caso no hay una gran diferencia pero el modelo 2 explica en un 41,14% el índice de bicicletas rentadas, es decir un 0.02% más que el modelo 1. Por otro lado la medida cor nos indica la correlación entre lo predicho y lo observado, es decir entre más cercano a 1 sea es mejor, por lo que igual que en el punto anterior el dos es mejor.

El promedio del error (BIAS), nos indica la relación entre el error esperado y el obtenido, aquí al obtener valor negativo nos indica que obtuvimos un valor mayor al esperado. Y el error cuadrático medio RMSE indica que entre menor sea su valor es mejor nuestro modelo, y en este caso nuevamente el segundo modelo es mejor que el primero.

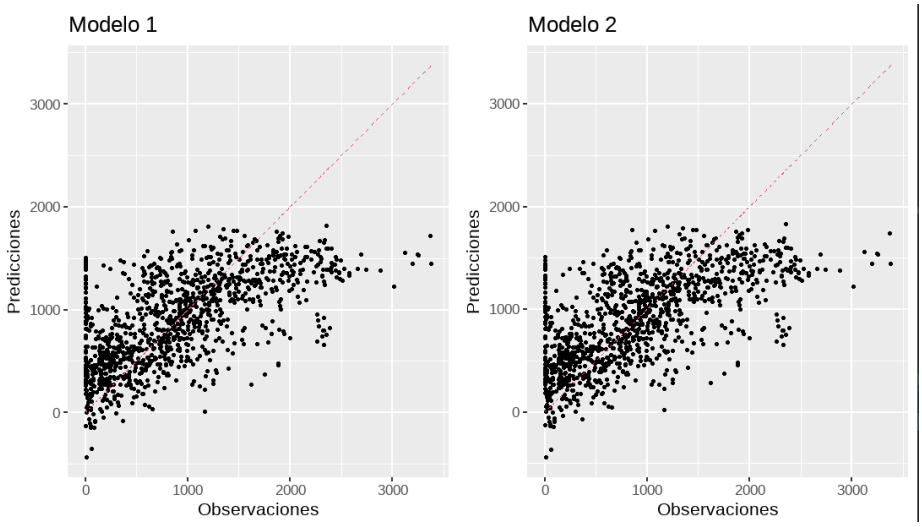

Figure 5: Comparación de los Modelos.

Conclusiones

Luego de este análisis podemos concluir que los modelos son similares y no son viables. Esto se debe a que la variable dependiente con la que trabajamos (Bicicletas rentadas) tiene una distribucion asimetrica positiva, por la regresion lineal multiple no arroja un buen modelo, para obtener un modelo viable tendríamos que convertir la variable para que tenga una distribución normal y luego realizar nuevamente la regresión.