MBD –Estadística –Práctica 1 (ML)

La Salle, Universidad Ramon Llull, Carrer de Sant Joan de la Salle 42, 08022 Barcelona

Resumen: En este trabajo, vamos a estudiar una técnica llamada regresión lineal múltiple, donde a partir de unas variables independientes que se llaman predictivas se determina la variable respuesta del modelo. El modelo constará de dos partes, una primera parte donde se construye el modelo a partir del 70% de los datos en el cual se encuentra la variable respuesta y una segunda parte que contiene el 30% de los datos restantes donde se va a predecir la variable respuesta a partir del modelo construido de la primera parte. La variable respuesta del modelo a determinar va a ser el número de bicicletas alquiladas en una franja horaria determinada. Para que el modelo construido sea valido, deberá cumplir unas ciertas premisas, tales como: Linealidad, Homoscedasticidad, Normalidad e Independencia. Además, el modelo se regirá por el principio de parsimonia.

I. INTRODUCCIÓN:

El principal objetivo de este trabajo es entender el concepto de regresión lineal múltiple. La regresión lineal múltiple es una técnica que establece que a partir de unas variables independientes que comúnmente se llaman predictivas se puede analizar la influencia que tienen estas sobre la variable dependiente. Además, las variables predictivas pueden emplearse para determinar el valor de la variable respuesta. La formulación de la ecuación de este de los modelos lineales múltiples es el siguiente:

$$Y_i = \left(B_0 + B_1X_{1i} + B_2X_{2i} + \cdots + B_pX_{pi}\right) + \epsilon_i \sim N(0,\sigma) \quad (k = 1 \dots p);\ \ (i = 1 \dots n)$$

- $\boldsymbol{Y_i}$: Es el valor i-ésimo de la variable respuesta.
- $\boldsymbol{X_{ki}}$: Es el valor de la variable predictora k en el caso i-ésimo.
- $\boldsymbol{B_0}$: Término independiente o intersección en el eje de las ordenadas cuando los valores de la predictiva son 0.
- $\boldsymbol{B_k}$: Es el coeficiente (pendiente) de la variable X_k que representa el incremento de la variable respuesta dependiente respecto un incremento en una unidad de la variable predictora.
- $\boldsymbol{\epsilon_i}$: Error. Representa el residuo de la diferencia entre el valor observado y estimado.
- $\boldsymbol{\sigma}$: Es la desviación típica de los errores (da un valor residual).

Debemos tener en consideración que las variables predictivas no solamente pueden tener un carácter numérico sino que también pueden tener un carácter categórico, en este aspecto debemos entender el tratamiento que se hace sobre estas variables categóricas. Al introducir una variable predictiva con carácter categórico, la cual puede estar dividida en diferentes niveles, uno de estos niveles se considera el de referencia y se le atribuye el valor de 0, por consiguiente, el resto de niveles se comparan respecto a ese nivel de referencia. Entonces, los niveles restantes adquieren el valor de 0 o 1 en función si esta presente o no. Por ejemplo, si tratamos con una variable numérica llamada temperatura (temp) y otra variable categórica llamada season que contiene cuatro niveles (invierno, primavera, verano, otoño), la ecuación resultante es:

$$Y_i = (B_0 + B_1temp + B_2primavera + B_3verano + B_4otoño)$$

Si el valor de la variable categórica es verano, se codifica con un 1 y el resto con un 0. Además, los coeficientes que acompañan a los niveles representan el incremento de más respecto al nivel de referencia. El resultado final sería:

$$Y_i = (B_0 + B_1temp + B_3verano)$$

Para la construcción de un modelo de regresión múltiple se debe tener en cuenta que el modelo debe cumplir con una serie de requisitos. Es requisito indispensable que no exista colinealidad entre variables predictivas, que las variables sean independientes ya que no se puede distinguir el efecto que produce esa variable sobre la variable respuesta, por tanto se debe identificar la existencia de correlación entre variables predictivas y ver cual es el impacto de cada una de ellas sobre la variable respuesta, nos quedaremos con la que tenga más influencia sobre la variable respuesta. Una manera de visualizar la colinealidad es con el coeficiente VIF que más adelante entraremos en detalle. Otros requisitos que debe cumplir el modelo son los siguientes:

- **Linealidad:** El concepto de linealidad nos dice que las predictivas independientes que sean numéricas, deben tener una relación lineal con la respuesta. Se mide la linealidad representando los residuos en función de cada una de las predictivas.
- **Homoscedasticidad:** La dispersión de los residuos debe ser constante y no presentar formas cónicas. Varianza constante.
- **Normalidad:** Los residuos deben ser ajustados o se deben ajustar a la recta de Normalidad (media 0).
- **Independencia:** No debe existir ningún patrón de los residuos con el tiempo (creciente, decreciente, etc).

Se entrará en más detalle en algunos conceptos cuando discutamos el modelo. Sin embargo, existe un concepto también a destacar, el concepto de parsimonia. Este concepto nos dice que el mejor modelo que podamos construir es aquel que es capaz de explicar con la mayor precisión posible la variabilidad observada en la variable respuesta con el menor número de variables predictivas.

II. CONSTRUCCIÓN DEL MODELO:

- **Descripción**

Como sabemos, existe una red de estaciones de bicis distribuidas por toda la ciudad. Además, esta red dispone de un sistema de recogida y retorno de bicicletas. El modelo a construir se basa en predecir el número o la demanda de bicicletas en función de la franja horaria que nos encontremos. Para ello, contaremos con una serie de variables predictivas que siendo analizadas podremos ver como de influyentes son en la variable respuesta. El primer paso es construir nuestro modelo con el 70% de los datos (datos de entrenamiento), para posteriormente hacer una predicción con el 30% restante (datos test), ya que no contaremos con la variable respuesta.

Es de gran interés analizar estos datos y poder predecir cual será la demanda de bicicletas dentro de una franja horaria, ya que representará a posteriori una mayor calidad del servicio. Las variables que dispondremos vienes especificadas a continuación:

id: Identificador de la franja horaria
year: Año (2011 o 2012)
season: 1 = invierno; 2 = primavera; 3 = verano; 4 = otoño
holiday: Si el día era festivo
workingday: Si el día era laborable (ni festivo ni fin de semana)
weather: Cuatro niveles (1 a 4), van de mejor a peor
temp: Temperatura (ºC)
atemp: Sensación de temperatura (ºC)
humidity: Humedad relativa
windspeed: Velocidad del viento (km/h)
count: Número total de alquileres en esa franja (variable respuesta, solo está presente en los datos de entrenamiento)

- **Modelo entrenamiento**

Para empezar leemos los datos y vemos que hay presentes todas las variables, tanto las variables predictivas como la variable respuesta. El primer paso que hacemos es hacer una descriptiva de los datos (dimensión, cabecera, summary). Como ya sabíamos existen variables categóricas, por consiguiente las pasaremos a factores con la función factor() sin transformar la variable predictiva hour que de momento la dejaremos tal y como esta. Una vez realizado esta conversión hacemos una representación de la variable count en función de la variable predictiva hour para ver su comportamiento con un gráfico boxplot y plot. Dependiendo del comportamiento agruparemos las horas en intervalos, el resultado es el siguiente:

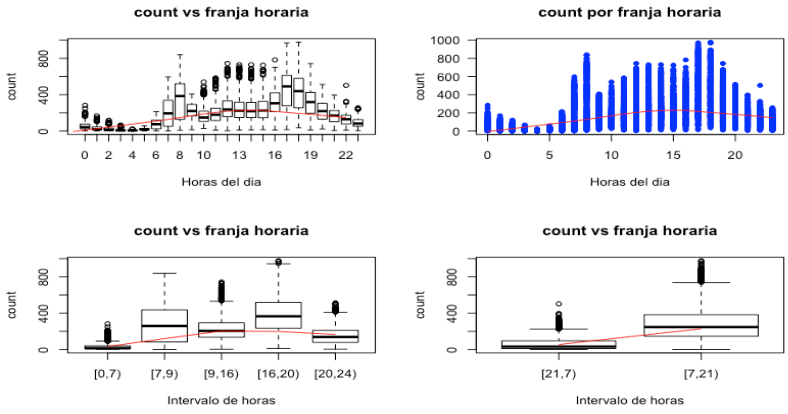

Figura 1. Boxplot y plot de count vs horas con horas del día y con diferentes intervalos. Con una línea de suavizado.

Podemos observar que no es adecuado agrupar solamente las horas en dos intervalos ya que hay horas del día que es muy significativa la demanda de bicis, por ese motivo hemos escogido los cinco intervalos que se representan en la figura 1. Aun así, haremos un análisis conjunto de las dos clases de intervalos para ver cual es mejor.

Una vez hecho la descriptiva de la variable hour, vamos a realizar una descriptiva general con la función pair() para ver la correlación de las variables predictivas que solo sean numéricas, esto nos servirá para darnos cuenta si existen variables predictivas muy correlacionadas entre ellas. Podemos observar que existen dos variables muy correlacionadas, la variable temp y la variable atemp. Como las dos contienen gran parte de información de la otra vamos a hacer dos plots, count vs temp y count vs atemp de esta manera vamos a observar que variable predice peor la respuesta, entonces, podremos descartar la variable con el valor de R^2 mas bajo. Hacemos una regresión lineal con la función lm() y el resultado es el siguiente:

$$count = B_0 + B_1 temp + \epsilon \quad \rightarrow \quad B_0 = 4.00;\ B_1 = 9.247;\ \epsilon = 167.2;\quad R^2 = 15.78$$

$$count = B_0 + B_1 atemp + \epsilon \quad \rightarrow \quad B_0 = -7.30;\ B_1 = 8.386;\ \epsilon = 167.6;\quad R^2 = 15.37$$

Como sabemos que R^2 mide la capacidad predictiva de un modelo y que varia entre 0 (no predice nada) y 1 (predice el valor de la respuesta), podemos observar que la variable atemp predice peor la respuesta con un $R^2 = 15.37$, entonces vamos a prescindir de ella eliminándola.

Una vez eliminada la variable atemp, vamos a realizar una descriptiva bivariante de la variable respuesta count en función de las variables predictivas numéricas y una descriptiva bivariante de la variable respuesta count en función de las variables categóricas. Los plots y los boxplots de las descriptivas son los siguientes:

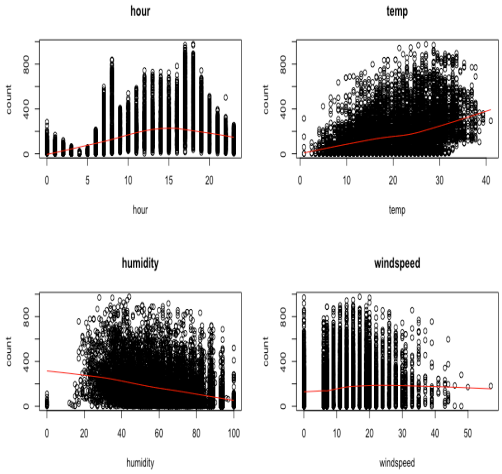

Figura 2. Plots de count vs variables predictivas numéricas.

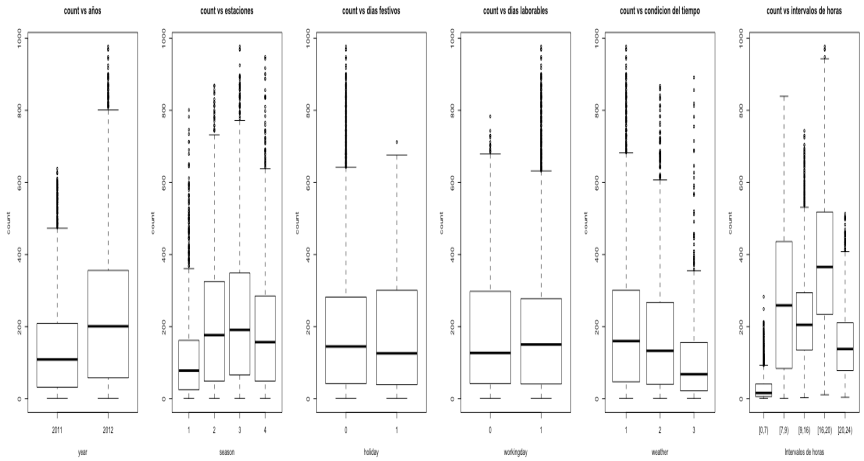

Figura 3. Boxplots de count vs variables predictivas categóricas.

Podemos observar en la figura 2, el plot de count vs hour no presenta una relación lineal ya que observamos una curvatura en el suavizado, mientras que en los otros tres plots, count vs temps, count vs humidity y count vs windspeed se observa que hay presente una relación lineal, aun así, se tendría que hacer un pequeño ajuste para que el suavizado fuera más una recta ya que como se observa presentan una pequeña curvatura.

En la figura 3, el boxplot de holiday no influye si el día es festivo o no, lo mismo pasa en el bloxplot de workingday ya que no influye si es día laborable o no, en cambio en el bloxplot de year hubo un incremento en el alquiler de bicis en 2012 respecto a 2011, lo mismo pasa en el bloxplot de season y weather que influye en la respuesta (alquiler de bicis) dependiendo en que estación del año estamos o que climatología nos encontramos. Además, como hemos visto anteriormente, las franjas horarias también son influyentes.

A continuación, una vez visto todas las descriptivas anteriores, vamos a realizar una regresión lineal con todas las variables que tengamos. Debemos tener en cuenta que en el modelo nos consta la variable hour, la variable hour_intervals (5 intervalos) y variable hour2 (2 intervalos), por consiguiente, nos vamos a crear dos modelos: Uno tendrá todas las variables más la variable hour_intervals y menos las variables predictivas hour y hour2, el otro modelo tendrá todas las variables más la variable hour2 y menos las variables predictivas hour y hour_intervals. Este método lo hacemos para ver cual de los dos intervalos en las franjas horarias nos predice mejor la respuesta. Se puede ver una diferencia significativa en los dos modelos, en el modelo con hour_intervals (5 intervalos), la mayoría de variables tienen unos p-valores muy pequeños, del orden de 10^{-16}, lo que significa que esa variable es muy significativa. Además, la mayoría de variables también tienen t-valores grandes, lo que significa que el error asociado a esa variable no modifica mucho el valor de la variable. En cambio, el modelo con hour2 (2 intervalos) los p-valores y t-valores también son pequeños y grandes respectivamente pero fluctúan mas, incluso los coeficientes varían un poco siendo un poco más pequeños. Por otra parte, hay diferencias en los errores y los valores de R^2 son muy significativos:

$$R^2 = 0.6334; \quad \epsilon = 110.4 \quad \longrightarrow \quad Modelo\ con\ hour\_intervals \quad (5\ intervalos)$$
$$R^2 = 0.5432 \quad \epsilon = 123.2 \quad \longrightarrow \quad Modelo\ con\ hour2 \quad (2\ intervalos)$$

Podemos observar que nos predice mejor la variable hour_intervals con una $R^2 = 0.6334$ y un $\epsilon = 123.2$, por tanto nos quedamos con la variable que llamamos mod.var. Un a vez que tenemos un intervalo definido, procedemos a hacer una selección automática de variables con la función step().

La teoría de la información dice que métodos como el AIC (Akaike Information Criteria) y BIC (Bayesian Information Criteria), hacen una selección automática de variables que más se adecua al modelo.

$$AIC = 2k - 2\log(L)$$
$$BIC = \log(n) - 2\log(L)$$

Con la función step() hacemos uso de estos métodos. Hemos utilizado la función step() con los dos métodos y el método BIC nos indica que variables deben ser eliminadas del modelo con un AIC más pequeño. En este caso las variables workingday, windspeed y holiday. Entonces, el modelo sin estas variables queda:

$$count \sim year + season + weather + temp + humidity + hour\_intervals$$

De hecho si nos fijamos en el summary vemos que todas las variables son muy significativas con un p-valor muy pequeño a excepción de weather2, además el t-valor también es grande. Cabe mencionar que existen tres formas de selección de variables (Forward, Backward, Forward-Backward). Nosotros hemos utilizado la selección Backward, la cual tenemos en un principio todas las variables y el sistema va quitando las que minimizan el AIC o BIC. Una vez hecha la selección de variables, es momento de ver la colinealidad entre variables predictoras con el indicador VIF, el cual sigue el siguiente criterio:

$VIF = 1$: No existe colinealidad, total ausencia.
$1 < \text{VIF} < 5$: La regresión se puede ver afectada por cierta colinealidad.
$5 < \text{VIF} < 10$: Variable debe ser suprimida ya que un 80% de su información está contenida en las otras.

Observamos el VIF de todas las variables y vemos que todas las variables están contenidas en un rango VIF < 5, por consiguiente no vamos a eliminar ninguna variable, aunque podemos observar que la variable temp y season están en el rango 1 < VIF < 5 pero más cerca del 5.

Una vez llegado a este punto, es hora de hacer una validación con un análisis de las premisas que debe cumplir nuestro modelo. Para ello vamos a realizar un análisis de los residuos haciendo una representación gráfica:

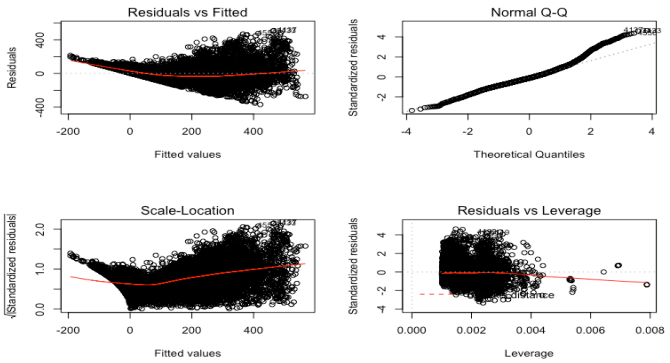

Figura 4. Plots de residuos. Validación de las tres premisas.

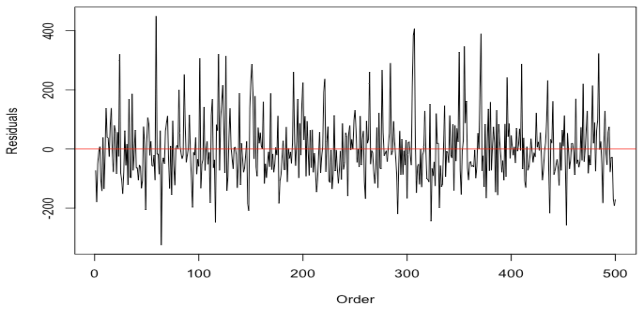

Figura 5. Plot de residuos vs order (500 puntos). Independencia de los residuos con el tiempo.

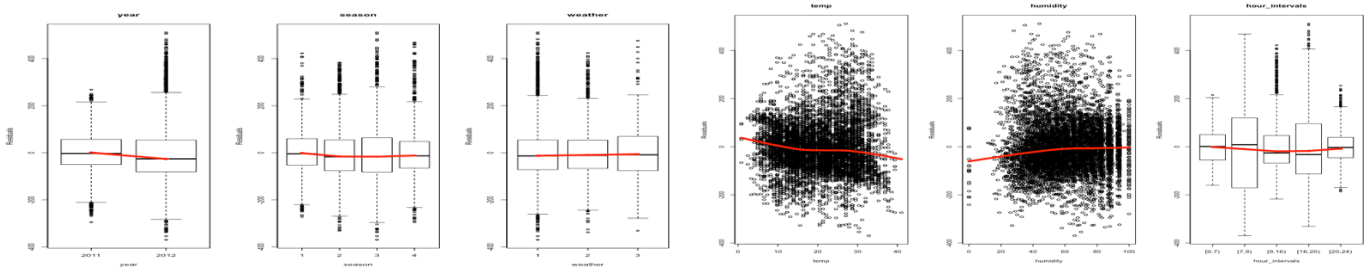

Figura 6. Plots de los residuos vs cada una de las variables predictivas.

Podemos observar que no se cumplen algunas premisas. En la figura 4, el plot de Residuals vs Fitted values se observa una curvatura en la línea, además presenta una distribución de puntos en forma de embudo, por tanto no cumple las premisas de linealidad ni Homoscedasticidad. Por otra parte tampoco cumple la premisa de Normalidad porque los puntos no se ajustan a la recta de Normalidad. Sin embargo, en la figura 5, observamos que cumple la premisa de independencia de los residuos ya que son uniformes en el tiempo. Podemos observar también en la figura 6 que no se cumple la linealidad de los residuos con las variables predictivas. Entonces, deberemos probar algunas transformaciones. El hecho que nos se cumplan las premisas nos indica que se requiere de alguna transformación. La transformación que proponemos, es hacer una transformación de la variable respuesta. Para ello, haremos la transformación de Box-Cox:

$$Y' = \begin{cases} Y^{\lambda} - 1) & si\ \lambda \neq 0 \\ \log(Y) & si\ \lambda = 0 \end{cases}$$

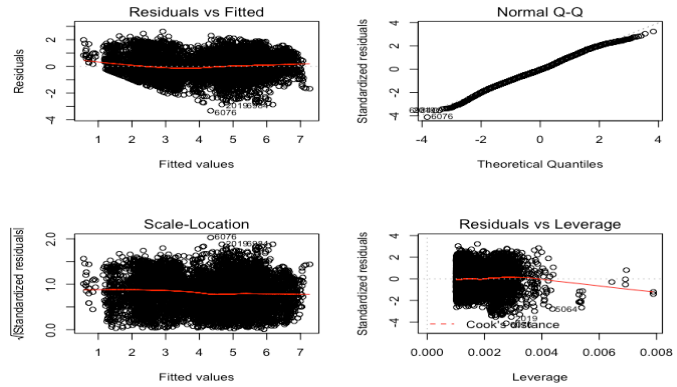

Figura 7. Plots de residuos. Validación de las tres premisas. Transformación Box-Cox.

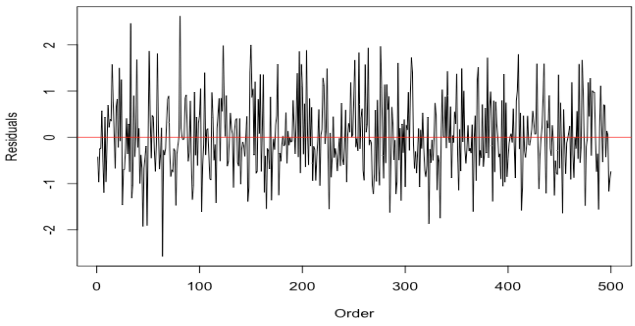

Figura 8. Independencia de los residuos con el tiempo (500 puntos). Transformación Box-Cox.

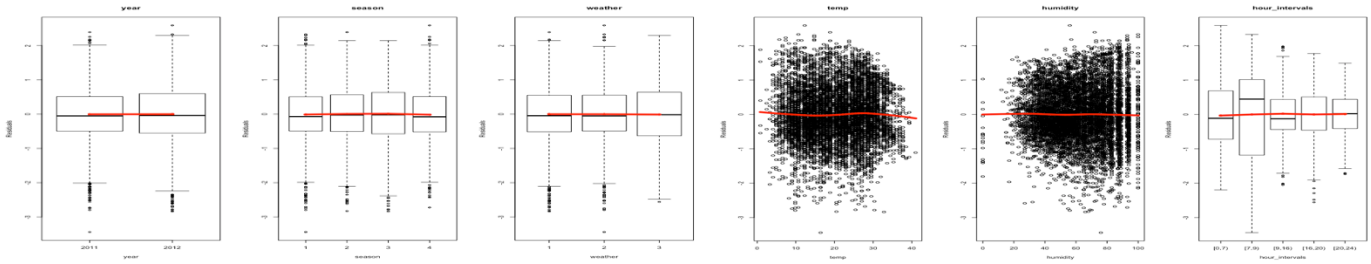

Figura 9. Plots de los residuos vs cada una de las variables predictivas con transformación countBC.

Hemos realizado las dos transformaciones de Y'. Una variable la hemos llamado countBC con $\lambda \neq 0$ y la otra countLog con el logaritmo de la variable count. Obtenemos unos mejores resultados cuando aplicamos la transformación countBC (figura5, figura 6). Además, observamos que nos predice mejor la variable respuesta. Podemos observar que hemos ajustado bastante las tres premisas, presenta mejor linealidad, hemos ajustado la línea a la recta de Normalidad y los residuos presentan una distribución homogénea, por consiguiente también se ha ajustado la Homoscedasticidad. Los residuos se mantienen constantes con el tiempo, por consiguiente también cumple la premisa de independencia. Además, hemos mejorado la linealidad de los residuos con las variables predictivas. De momento la ecuación resultante es la siguiente:

$$\begin{cases} Y^{\lambda} = countBC & \lambda = 0.3030 \quad \longrightarrow \quad R^2 = 0.7461 \quad \epsilon = 0.8121 \\ \log(Y) = countLog & \quad \longrightarrow \quad R^2 = 0.7218 \quad \epsilon = 0.7851 \end{cases}$$

$$countBC \sim year + season + weather + temp + humidity + hour\_intervals$$

Aunque los resultados son bastante buenos para los dos modelos, realizaremos unas transformaciones a las variables predictivas temp y humidity para intentar ajustar un poco mas los resultados y ver como reaccionan los dos modelos. Vamos a probar a introducir polinomios de grado superior (2,3,4..) y otras transformaciones tipo 1/Y, 1/X, raíces, transformaciones logarítmicas a las variables predictoras, etc. Después de todas estas pruebas, los resultados que se ajustan mas son transformaciones polinómicas de grado dos a las variables temp y humidity, entonces el resultado es el siguiente:

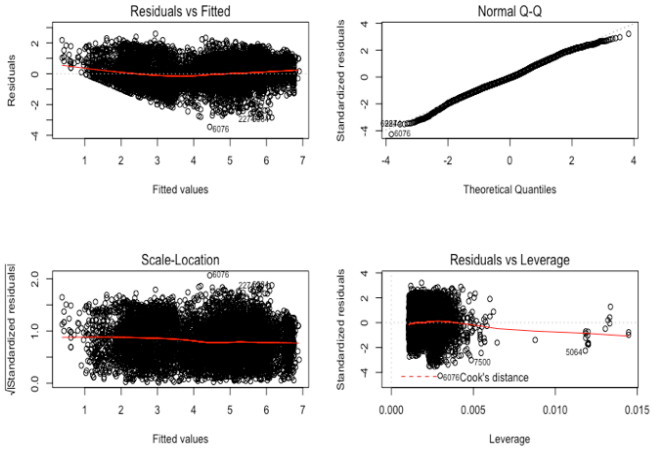

Figura 10. Plots de residuos. Validación de las tres premisas. Transformación poly de grado 2.

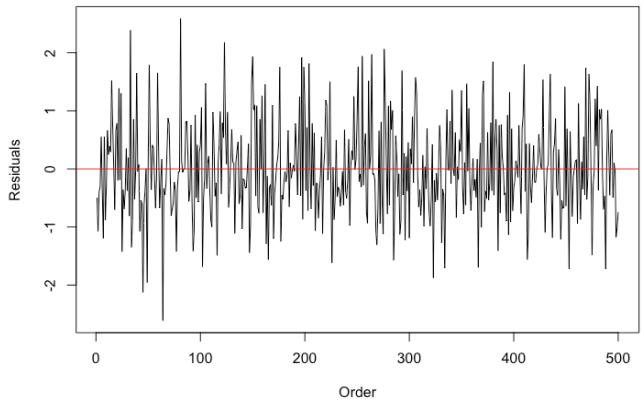

Figura 11. Independencia de los residuos con el tiempo (500 puntos). Transformación poy de grado 2.

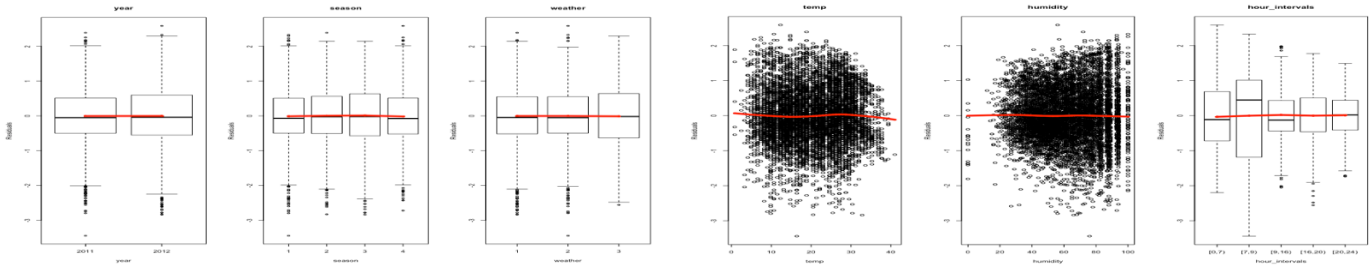

Figura 12. Plots de los residuos vs cada una de las variables predictivas con transformación countBC.

Podemos observar que se mantienen las tres premisas, presenta buena linealidad, buen ajuste a la recta de Normalidad y buena homoscedasticidad. Con los polinomios de grado dos, los residuos a cada una de las variables predictivas ha mejorado ligeramente la pequeña curvatura que presentaban, es decir, son más lineales. La premisa de independencia también se cumple. La conclusión es que de momento nos quedamos con este modelo, entonces, el modelo que en el script se llama mod.var3 es el siguiente:

$$countBC \sim year + season + weather + poly(temp, 2) + poly(humidity, 2) + hour\_intervals$$

$$R^2 = 0.7489 \qquad \epsilon = 0.8077$$

Una manera de ajustar el modelo es observar que puntos son influyentes y en función de estos podemos decidir eliminarlos para ver su impacto o influencia en la variable respuesta. Un método para hacer esto es mirando la distancia de Cook. Entonces, realizamos un análisis y nos da el siguiente el resultado:

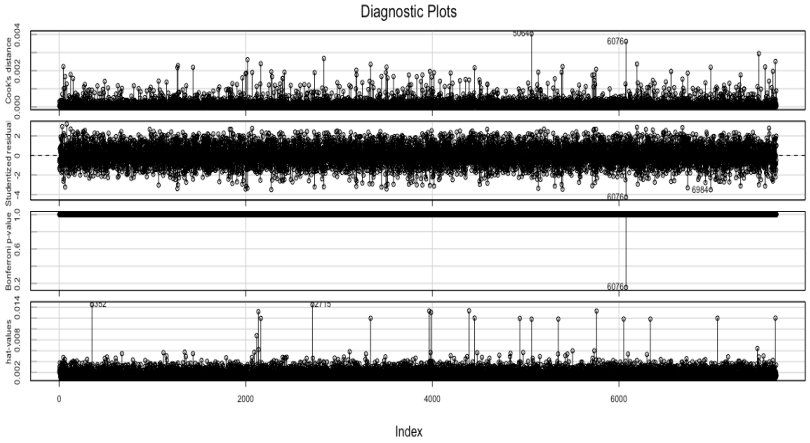

Figura 13. Distancia de Cook de las observaciones.

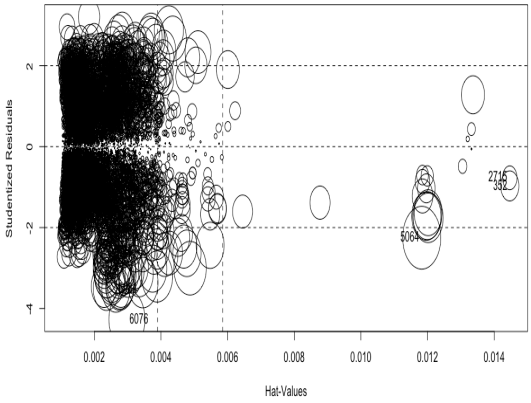

Figura 14. Observaciones influyentes.

Las observaciones 6076 y 5064 en principio tienen mucha influencia a posteriori. La 352 y 2715 tienen mucha influencia a priori. Hemos eliminado las observaciones 6076 y 5064 y posteriormente hemos hacho una regresión lineal del modelo para ver su comportamiento. El resultado no tiene un impacto muy grande, la conclusión es que nos quedamos con el modelo antiguo.

$$\begin{cases} Modelo\ antiguo\ (mod.var3)\ con\ las\ observaciones & \longrightarrow\ [R^2 = 0.7489] \\ Modelo\ \text{sin}\ las\ \ observaciones\ influyentes & \longrightarrow\ R^2 = 0.7495 \end{cases}$$

Una vez realizado la construcción del modelo entrenamiento, toca realizar la predicción de la variable respuesta count con los datos test que representan el 30% de los datos. Para hacer la predicción, antes de todo nos creamos los mismos intervalos en las franjas horarias que en el modelo entrenamiento (5 intervalos). Luego, eliminamos las variables que no estén presente en el modelo de entrenamiento, posteriormente pasaremos a factor las variables que sean categóricas. Una vez preparados los datos, hacemos la predicción de la variable respuesta count con la función predict() que incluye nuestro modelo final y los datos test. Elevaremos los datos obtenidos a la $Y^{1/\lambda}$ y redondearemos el resultado. Los datos obtenidos los guardaremos en un fichero con el id y la predicción.

III. CONCLUSIÓN

Los resultados obtenidos son bastante buenos. El modelo predice bastante bien con una $R^2 = 0.7489$, además hemos conseguido que se cumplan las premisas con buen acierto. Por otra parte, siempre nos hemos regido por el principio de parsimonia a la hora de escoger el menor número de variables predictivas e intervalos en las franjas horarias. A medida que íbamos mejorando el modelo los p-valores iban disminuyendo, tenían un valor más significativo, igual a los t-valores que iban aumentando. En referente a los coeficientes, han ido variando, de hecho han ido disminuyendo su valor. Todos ellos son buenos indicadores para ver como varían los coeficientes de las variables.

Cabe destacar que el hecho de no pasar a factor la variable hour de buen inicio es un buen acierto. Si hubiéramos pasado a factor la variable hour y no hubiéramos hecho los intervalos hubiéramos obtenido una $R^2 > 0.80$, pero me he dado cuenta que si lo hubiéramos hecho, nuestro modelo estaría sobrestimado ya que nuestro modelo dependería exactamente de las horas del modelo entrenamiento y podría ser que el número de bicis alquiladas en esas horas en los datos test fuera ligeramente diferente, entonces la predicción podría no estar bien. Nuestro acierto o validez de nuestro modelo, dependerá ahora del error cuadrático medio logarítmico entre la predicción y el valor real.