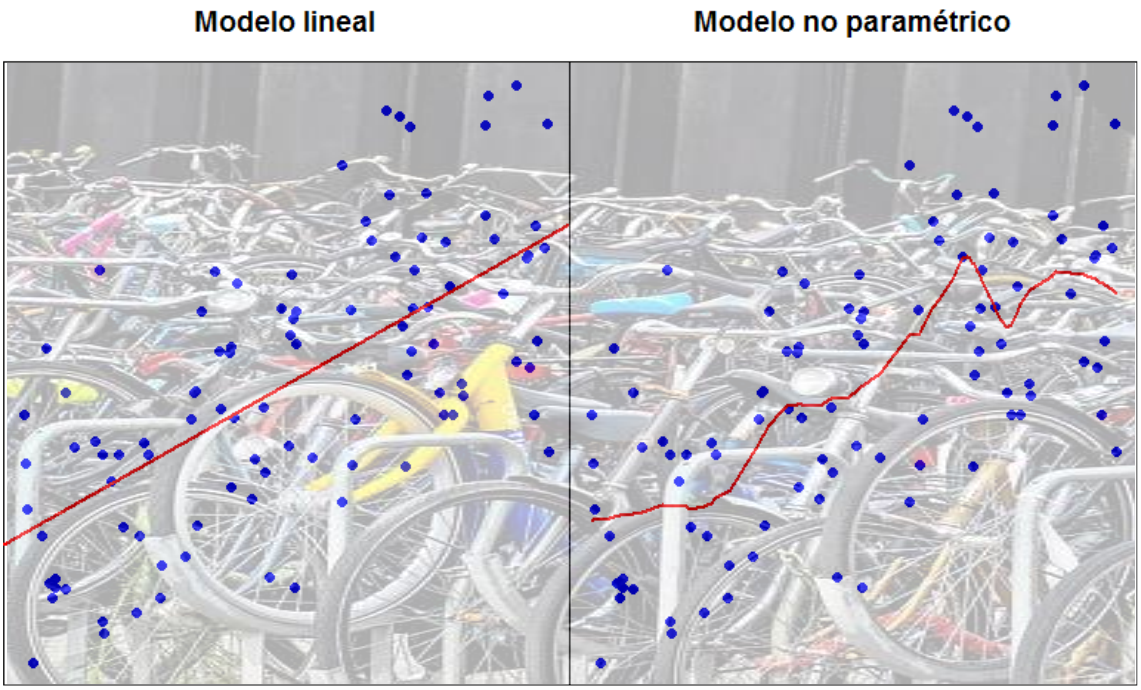
MASTER BIG DATA

ESTADÍSTICA: PRÁCTICA 1 (MODELO LINEAL)



Autor: Jaume Feliubadaló Rubio

Tutor: Jordi Cortés Martínez

Data: 02/12/2018

La Salle, Universidad Ramon Llull

ÍNDICE

ÍNDICE DE FIGURAS

1. INTRODUCCIÓN

1.1. Resumen

En este trabajo ponemos en práctica el concepto de modelo lineal. Un modelo simple, pero resultando ser una técnica de predicción con muy buena aproximación. Se parte de un conjunto de datos, los cuales están clasificados según variables independientes entre ellas. Hay dos tipos de variables: las predictivas y la respuesta. El proceso a seguir es construir un modelo a partir de datos de entrenamiento (son un tanto por ciento elevado de los datos totales) y predecir la respuesta con ese modelo en base a los datos test (tanto por ciento menor).
Cabe explicar que la variable respuesta a determinar para este modelo, va a ser el número de bicicletas alquiladas en una franja horaria (ver más adelante intervalos).
No sólo se tiene que construir el modelo, sino que hay que cumplir con ciertas premisas; Linealidad, Homoscedasticidad, Normalidad e Independencia.

1.2 Objetivo

El principal objetivo es familiarizarse (ergo entender) con el concepto explicado ya en clase de regresión lineal múltiple. Este tipo de regresión es una técnica para analizar la relación entre dos o más variables a través de ecuaciones. Consta de variables independientes y de la variable respuesta.

$$Y_i = \left(B_0 + B_1 X_{1i} + B_2 X_{2i} + \cdots + B_p X_{pi}\right) + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma) \quad (k = 1 \ldots p\ ;\ i = 1 \ldots n)$$

$\boldsymbol{Y_i}$: Es el valor de la variable respuesta en el caso i-ésimo
$\boldsymbol{X_{ki}}$: Es el valor de la variable predictora k en el caso i-ésimo
$\boldsymbol{B_0}$: Término independiente
$\boldsymbol{B_k}$: Es el coeficiente (pendiente) de la variable X_k
$\boldsymbol{\epsilon_i}$: Error
$\boldsymbol{\sigma}$: Es la desviación típica de los errores (desviación residual)

1.3 Variables y premisas

Cabe destacar que alguna de las variables no es numérica, sino que son categóricas. Dichas variables en vez de estar representadas por un valor numérico cuantificativo, lo están por categorías. En nuestro caso un ejemplo muy claro es la variable '*weather'.* Son 4 los niveles que definen esta variable. Si introducimos variables categóricas en el modelo, R considera que un nivel es el de referencia y le atribuye el valor 0. El resto de niveles se comparando con dicho nivel de referencia.

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         -182.387     17.253 -10.571  < 2e-16 ***
year2012              89.166      2.570  34.696  < 2e-16 ***
season2               52.364      4.732  11.066  < 2e-16 ***
season3               51.105      5.571   9.174  < 2e-16 ***
season4               65.752      3.994  16.462  < 2e-16 ***
weather2             -17.143      3.042  -5.635 1.82e-08 ***
weather3             -78.424      4.934 -15.894  < 2e-16 ***
temp_Log              81.433      4.535  17.957  < 2e-16 ***
humidity_Log         -21.579      3.143  -6.866 7.13e-12 ***
hour_intervals[7,9)  256.136      5.057  50.653  < 2e-16 ***
hour_intervals[9,17) 186.525      3.496  53.350  < 2e-16 ***
hour_intervals[17,20) 352.853     4.499  78.422  < 2e-16 ***
hour_intervals[20,24) 111.155     3.962  28.053  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figura 1 Valor de referencia variable categórica

En la figura 1 se aprecia que se toma como referencia ‘*season1*’ y se da el valor del resto de niveles comparando como anteriormente he mencionado.

No solo tenemos que construir el modelo, sino que este tiene que cumplir unos requisitos. Dichos requisitos son cuatro premisas que a continuación explicaré y otra de vital importancia; la colinealidad. Se entiende por colinealidad como la correlación que tienen las variables predictoras entre ellas. Correlaciones altas indican que entre variables existe una gran cantidad de información compartida, eso quiere decir que nos explicaran de forma muy similar lo mismo de la variable respuesta. Para dotar dicha colinealidad existe el VIF, es un indicador de la colinealidad de la variable j con el resto de variables.

Las cuatro premisas son:

- Linealidad: que la recta se ajuste bien a los datos.
- Homoscedasticidad: que haya variabilidad constante
- Normalidad: en los residuos, que los errores sean normales.
- Independencia: la muestra debe ser aleatoria simple y que el resultado de una observación no condiciona el resto.

Estas premisas se deben verificar mediante el análisis de los residuos.

2. CONSTRUCCIÓN DEL MODELO

2.1 Descripción variables

“El objetivo de esta práctica es predecir la demanda en una serie de franjas horarias concretas, usando el conjunto de datos histórico como base para construir un modelo lineal”.

Dentro de los datos proporcionados hay las diferentes variables:

id: identificador de la franja horaria (no guarda relación con el orden temporal)
year: año (2011 o 2012)
hour: hora del día (0 a 23)
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 categorías (1 a 4) que van de mejor a peor tiempo
temp: temperatura en grados Celsius
atemp: sensación de temperatura en grados Celsius
***humidity*:** humedad relativa

***windspeed*:** velocidad del viento (km/h)
count (sólo en el conjunto de entrenamiento): número total de alquileres en esa franja

2.2 Modelo_entrenamiento

Definidas las variables se procede a seguir los pasos indicados en el script guía proporcionado por el tutor.

Ver Script R con indicaciones en cada etapa del proceso

Con tal de sintetizar, en este documento solo daré la información que creo necesaria, tal como la interpretación de resultados o las decisiones tomadas bajo criterio propio.

En primer lugar, procedo a explicar el porqué de los intervalos.

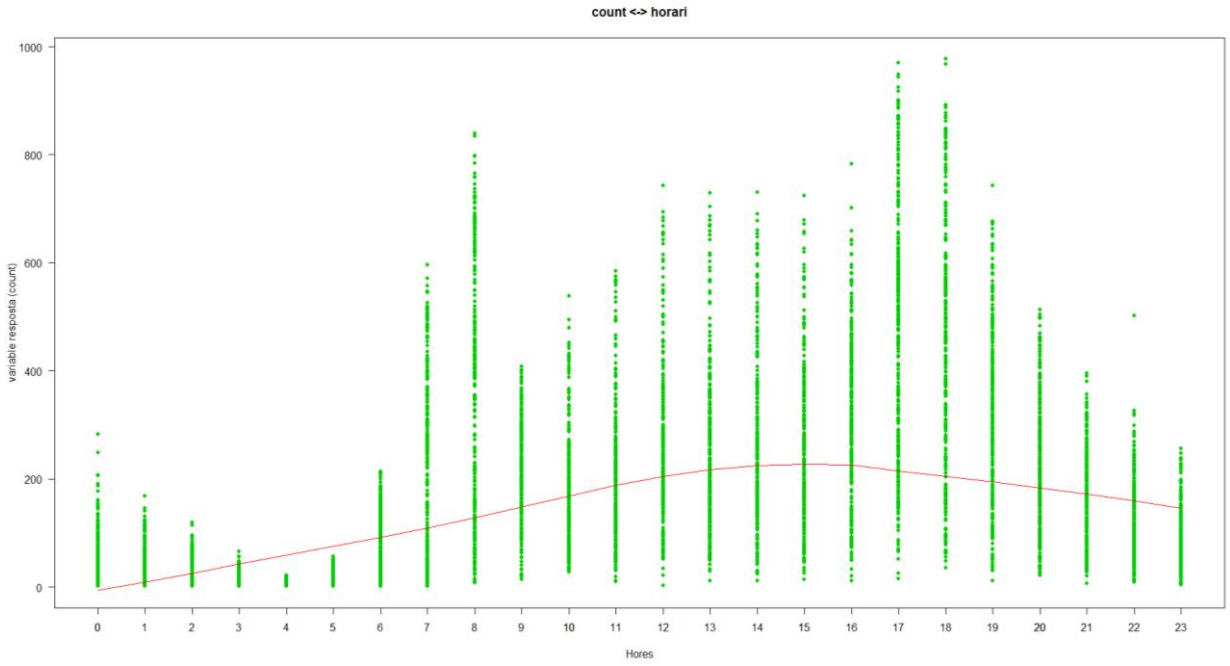


Figura 2 Count vs horari

Realizado el plot de la variable respuesta en función de las horas, deduzco los intervalos después de hacer el siguiente razonamiento: de 0h a 7h hay un comportamiento de disminución de demanda debido a que son altas horas de la noche. De 7h a 9h se produce un pico, ya que es cuando empieza la jornada laboral de la gran mayoría de trabajadores. El siguiente intervalo, de 9h a 17h viene determinado a que son horas donde supongo que la demanda de bicicletas va inducida por: simple desplazamiento, horario de comida y turismo. De 17h a 20h se produce otro pico debido al fin de la jornada laboral y de 20h a 24h se produce otra vez una disminución de la demanda de bicicletas.

Estos intervalos se deben a que hay horas, como he explicado que son significativas en cuanto a demanda de bicicletas se refiere.

Con la función pair() se ve gráficamente la correlación entre variables.

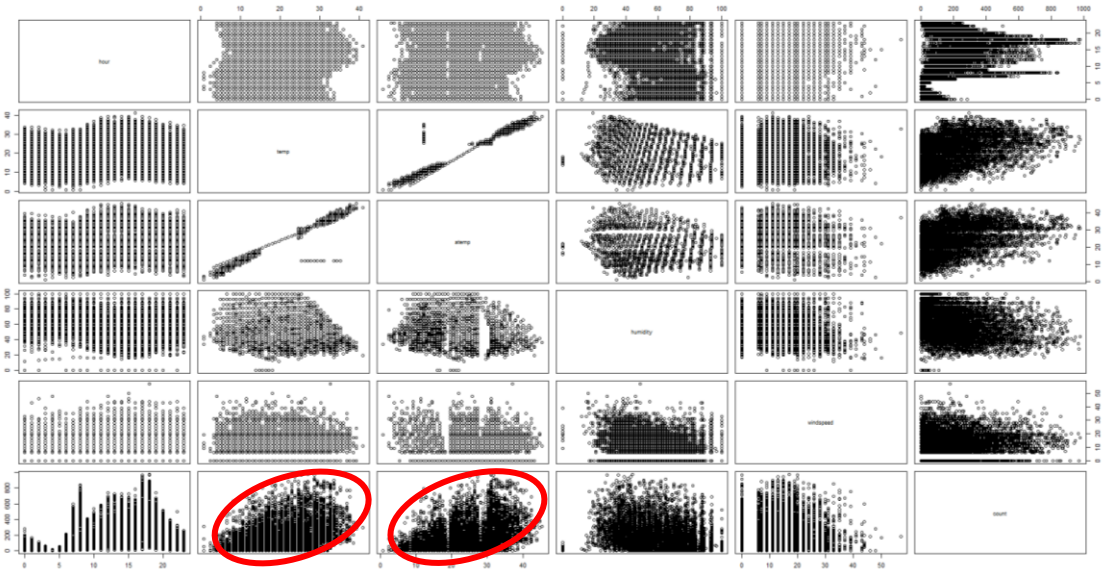


Figura 3 Pair()

Se aprecia que hay correlación entre las variables '*temp*' y '*atemp*'. Se tiene que descartar una de ellas, en el Script se explica el proceso y él porque de la decisión.

Eliminada la variable '*atemp*' se procede a realizar la descriptiva bivariante de la variable respuesta en función de las variables que predicen, tanto para las numéricas como para las categóricas. Dicha descriptiva se hace con plots y boxplots.

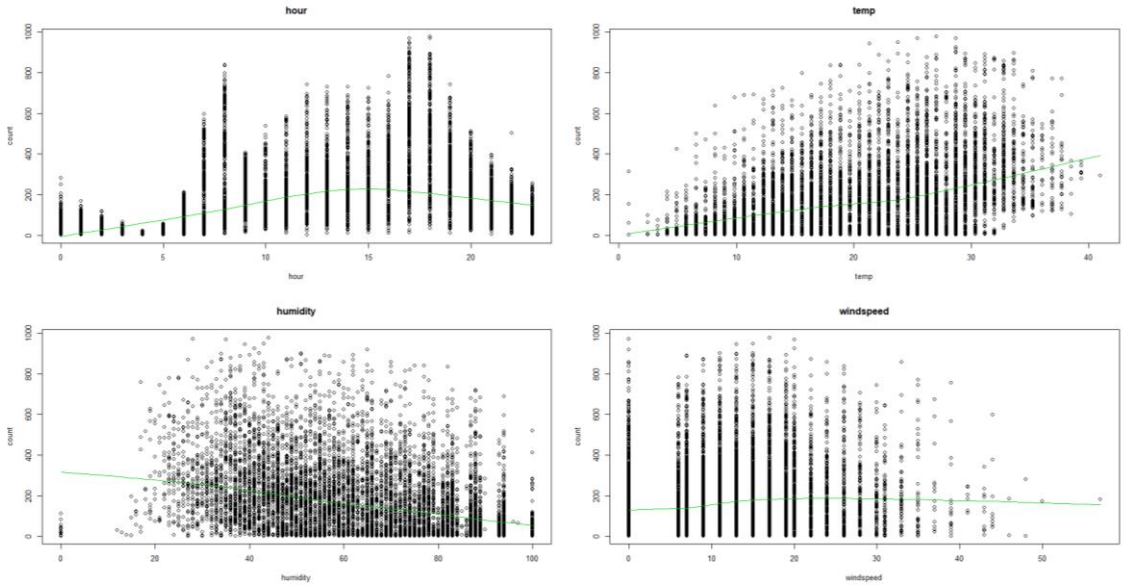


Figura 4 Plot count vs v.p. numéricas

Count vs hour no se aprecia una relación lineal, en los otros sí que se aprecia. Es en este apartado cuando me doy cuenta que habrá que ajustar la curvatura.

En cuanto al boxplot se refiere, si el día es holiday o workingday no influye. Se aprecia un aumento de demanda en el año 2012, este aumento también es visible para cierto periodo del año (season) y para según que meteorología tenemos.

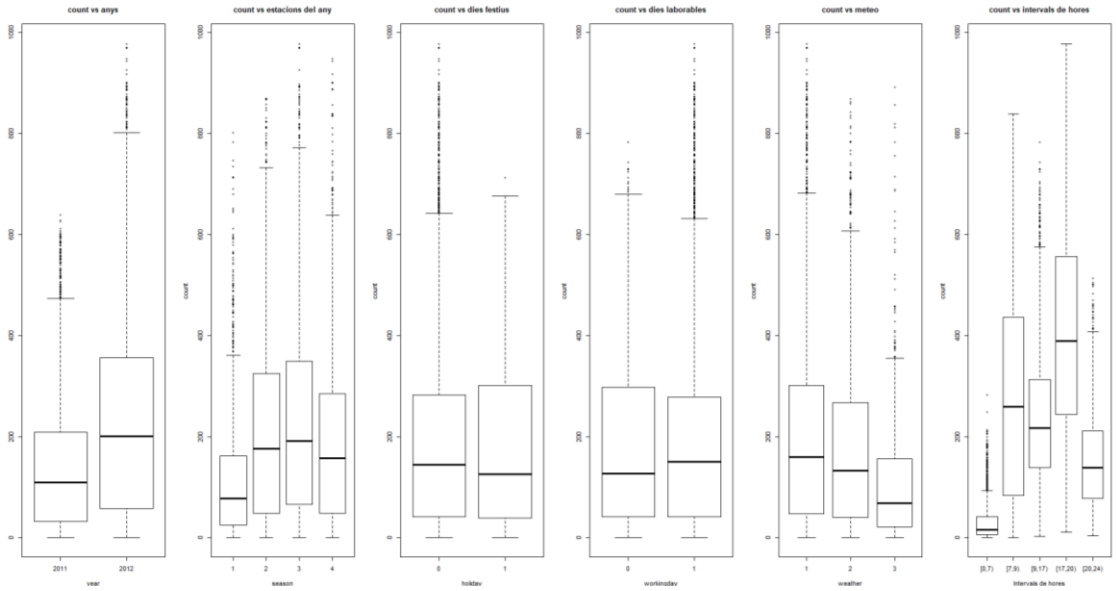


Figura 5 Boxplot count vs v.p. categóricas

Hecha la descriptiva, intuyo que el modelo dejará fuera las variables '*holiday*' y '*workingday*'.

Con la función lm() se construye el modelo. Procedo a explicar los resultados obtenidos tal como se indica en el script guía de la práctica.

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -83.53082    8.07792 -10.341  < 2e-16 ***
year2012              86.25825    2.53888  33.975  < 2e-16 ***
season2               42.91494    4.66897   9.192  < 2e-16 ***
season3               26.52521    5.98271   4.434 9.39e-06 ***
season4               69.89589    3.86084  18.104  < 2e-16 ***
holiday1             -15.98280    7.72059  -2.070   0.0385 *
workingday1            2.69002    2.77729   0.969   0.3328
weather2              -7.81917    3.10973  -2.514   0.0119 *
weather3             -57.21355    5.25241 -10.893  < 2e-16 ***
temp                   6.20447    0.28642  21.662  < 2e-16 ***
humidity              -0.93641    0.08616 -10.868  < 2e-16 ***
windspeed             -0.34401    0.16619  -2.070   0.0385 *
hour_intervals[7,9)   254.31524    4.98324  51.034  < 2e-16 ***
hour_intervals[9,17)  171.31816    3.67356  46.635  < 2e-16 ***
hour_intervals[17,20) 337.50467    4.62458  72.981  < 2e-16 ***
hour_intervals[20,24) 104.77314    3.94109  26.585  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 109.8 on 7673 degrees of freedom
Multiple R-squared:  0.637,     Adjusted R-squared:  0.6363
F-statistic: 897.6 on 15 and 7673 DF,  p-value: < 2.2e-16
```

Figura 6 Características modelo lineal

Hay variables con p-valores próximos a 0.05 y la variable *workingday* mayor a 0.05, el resto son del orden de 10^{-16}. P-valores pequeños indican que la variable es significativa.

Para valores altos en t-valor, el error no modifica el valor de la variable de forma significativa.

Observo con el summary que el modelo hecho predice la variable con un $R^2 = 0.637$ y un $\epsilon = 109.8$.

2.3 Premisas modelo + colinealidad VIF

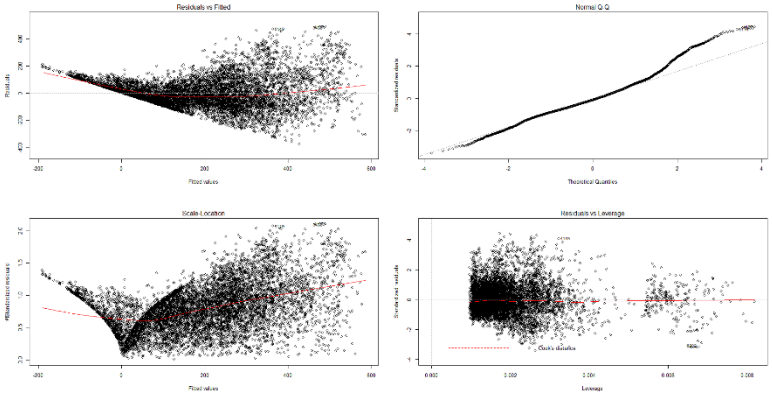


Figura 7 Plot para validación premisas

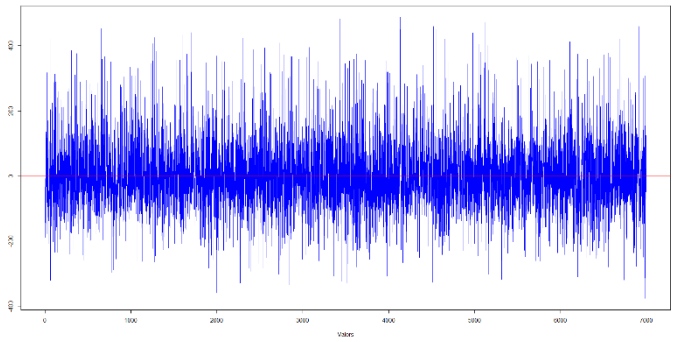


Figura 8 Plot validación independencia

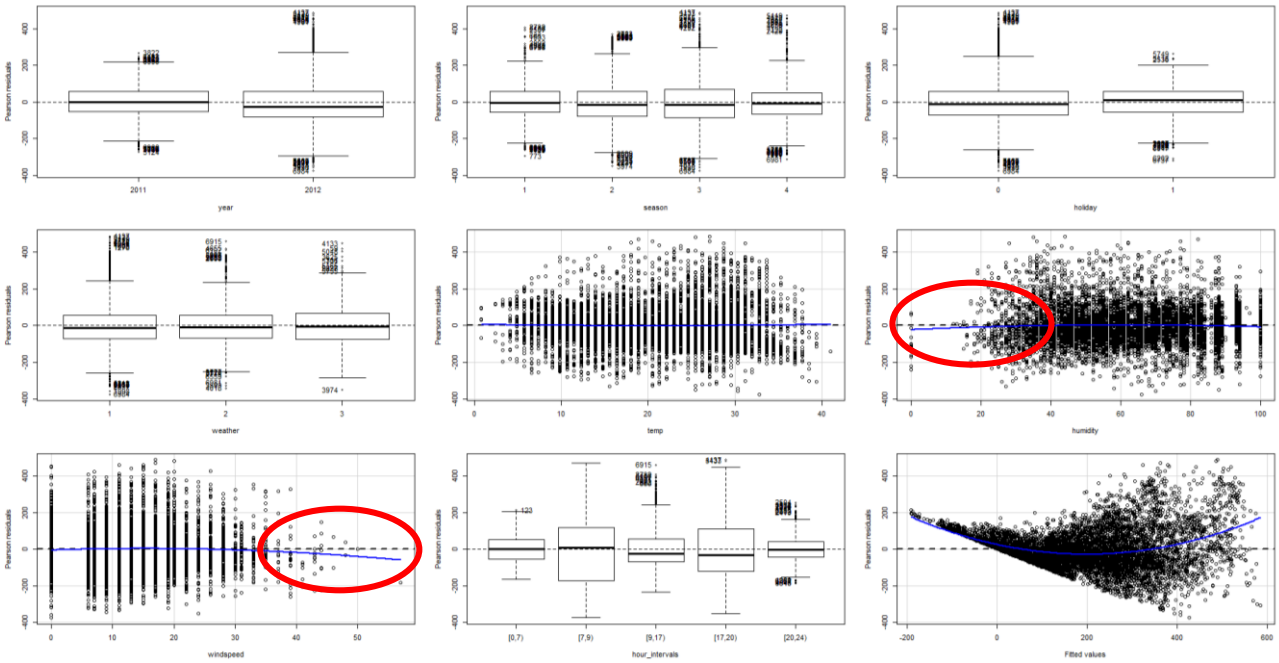


Figura 9 Plots residuos de cada variable

	GVIF	Df	GVIF^(1/(2*Df))
year	1.026682	1	1.013253
season	3.226926	3	1.215621
holiday	1.003031	1	1.001514
weather	1.301286	2	1.068054
temp	3.185649	1	1.784839
humidity	1.761795	1	1.327326
windspeed	1.177177	1	1.084978
hour_intervals	1.456753	4	1.048150

Figura 10 VIF

Figura 7, se aprecia que no se cumplen alguna de las premisas. Hay una distribución de los puntos en forma de cono (o embudo) en el primer plot, esto indica que no se cumple la linealidad ni homoscedasticidad. Tampoco se cumple la normalidad, los puntos/valores no se ajustan a la recta. Si se cumple la premisa de independencia, no hay patrón alguno. Para la independencia, he hecho el plot con 500 valores y posteriormente con 7000.

Ya que no se cumplen ciertas premisas, se decide hacer una transformación del tipo Box-Cox.

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

Para ello deberemos calcularnos la lamba. Cabe destacar que la transformación se lleva a cabo para la variable respuesta. Hecha la transformación se hace el plot de residuos para la validación de las premisas anteriormente mencionadas.

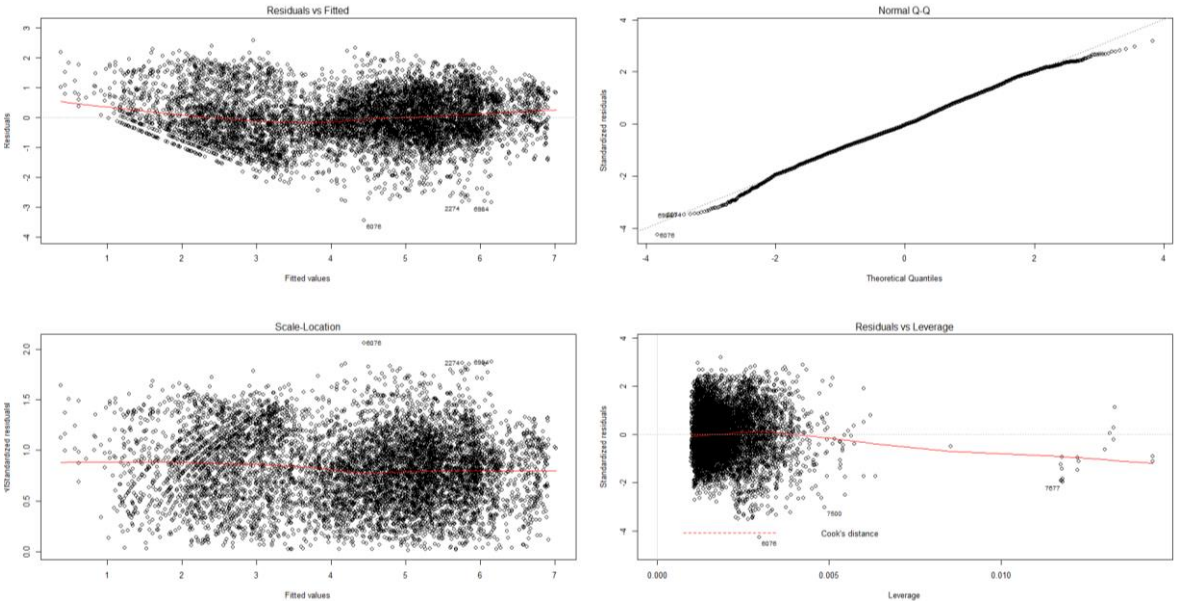


Figura 11 Validación premisas modelo transformado con Box-cox

Obtenemos los siguientes resultados, los cuales muestro en forma de formula.

$Y^{\lambda} = countBC \quad siendo \quad \lambda = 0.3030303$

Con un $R^2 = 0.748 \quad y \quad \epsilon = 0.8091$

La formulación del modelo queda tal que:

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

También he realizado una transformación del tipo log. pero la descarto ya que tiene un R^2 menor al de la transformación del tipo Box-Cox.

Como anteriormente he observado que las variables '*temp*' y '*humidity*' presentaban curvatura, decido hacer una transformación polinómica sobre ellas. El resultado es el siguiente:

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

Con $R^2 = 0.748$ y $\epsilon = 0.8091$

Se cumplen las premisas, tenemos una linealidad correcta junto a una normalidad ajustada. La homoscedasticidad es buena, no hay presencia de forma de cono.

En el apartado 18 del script guía se realiza una mejora del modelo, ya que dicha mejora no es significativa no procedo a su debida explicación.

3. CONLUSIÓN

Hemos obtenido unos resultados en mi opinión muy correctos. Se aprecia que el modelo predice con una exactitud alta $R^2 = 0.748$. Destacar el cumplimiento de todas las premisas y en especial el principio de parsimonia, el cual se hace hincapié al principio del script guía. El modelo es simple, la complejidad reside en la interpretación de los conceptos técnicos asociados a la creación de un modelo lineal.

Se observa claramente que gracias a las transformaciones realizadas el modelo ha mejorado de forma significativa.

Se ha realizado un profundo análisis del temario correspondiente al modelo lineal, lo que me ha ayudado a entender mejor los conceptos aprendidos con anterioridad en clase.