

MBD - Estadística - Práctica 1a (ML)

Arturo Menchaca y Víctor Juez

Noviembre 22, 2020

Contents

| | |
|---|-----------|
| 1 Conjunto de datos | 3 |
| 2 Análisis de las variables | 3 |
| 2.1 Categorización de la variable hora | 3 |
| 2.2 Descriptiva de las variables numericas | 5 |
| 2.3 Descriptiva de las variables categoricas | 6 |
| 3 Generación del modelo | 6 |
| 3.1 Modelo 1 - Utilizando todas las variables | 6 |
| 3.2 Modelo 2 - Selección automática de variables | 6 |
| 3.2.1 Análisis de colinealidad de las variables | 7 |
| 3.2.2 Validación de las premisas | 8 |
| 3.3 Modelo 3 - Transformación de la variable resuesta | 8 |
| 4 Modelo final - Modelo 3 | 9 |
| 4.1 Validación del modelo | 10 |
| 4.2 Efecto de las características sobre la variable respuesta | 11 |
| 5 Modelos descartados | 11 |
| 5.1 Modelo 5: Transformación polinómica | 11 |
| 6 Anexo | 13 |
| 6.1 Modelo 6: Eliminar observaciones influyentes | 13 |

1 Conjunto de datos

El conjunto de datos consta de las siguientes 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

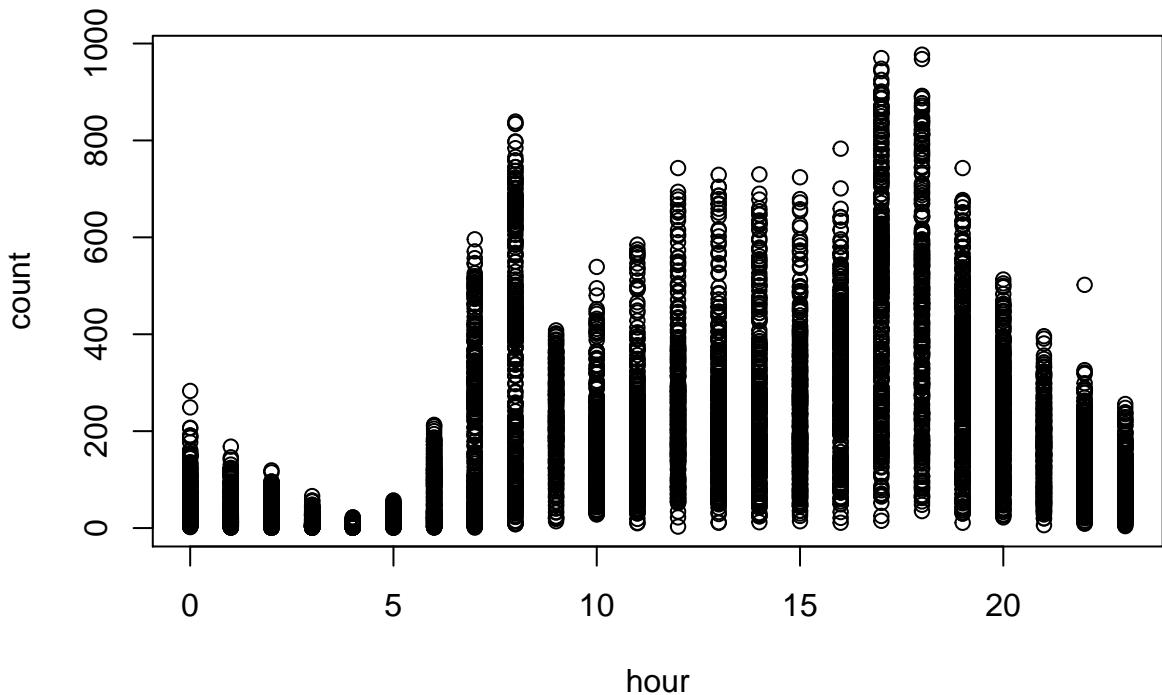
A continuación mostramos la descriptiva de los datos:

```
##      year          hour        season    holiday  workingday weather
## 2011:3879   Min. : 0.00  1:1901  0:7466  0:2481     1:5122
## 2012:3810   1st Qu.: 6.00  2:1920  1: 223   1:5208     2:1981
##                   Median :12.00  3:1943
##                   Mean   :11.57  4:1925
##                   3rd Qu.:18.00
##                   Max.   :23.00
##      temp         atemp       humidity      windspeed
##  Min.   : 0.82   Min.   : 0.76   Min.   : 0.00   Min.   : 0.000
##  1st Qu.:13.94   1st Qu.:16.66   1st Qu.: 46.00   1st Qu.: 7.002
##  Median :20.50   Median :24.24   Median : 62.00   Median :12.998
##  Mean   :20.27   Mean   :23.70   Mean   : 61.77   Mean   :12.802
##  3rd Qu.:26.24   3rd Qu.:31.06   3rd Qu.: 77.00   3rd Qu.:16.998
##  Max.   :41.00   Max.   :45.45   Max.   :100.00  Max.   :56.997
##      count
##  Min.   : 1.0
##  1st Qu.: 41.0
##  Median :145.0
##  Mean   :191.4
##  3rd Qu.:283.0
##  Max.   :977.0
```

2 Análisis de las variables

2.1 Categorización de la variable hora

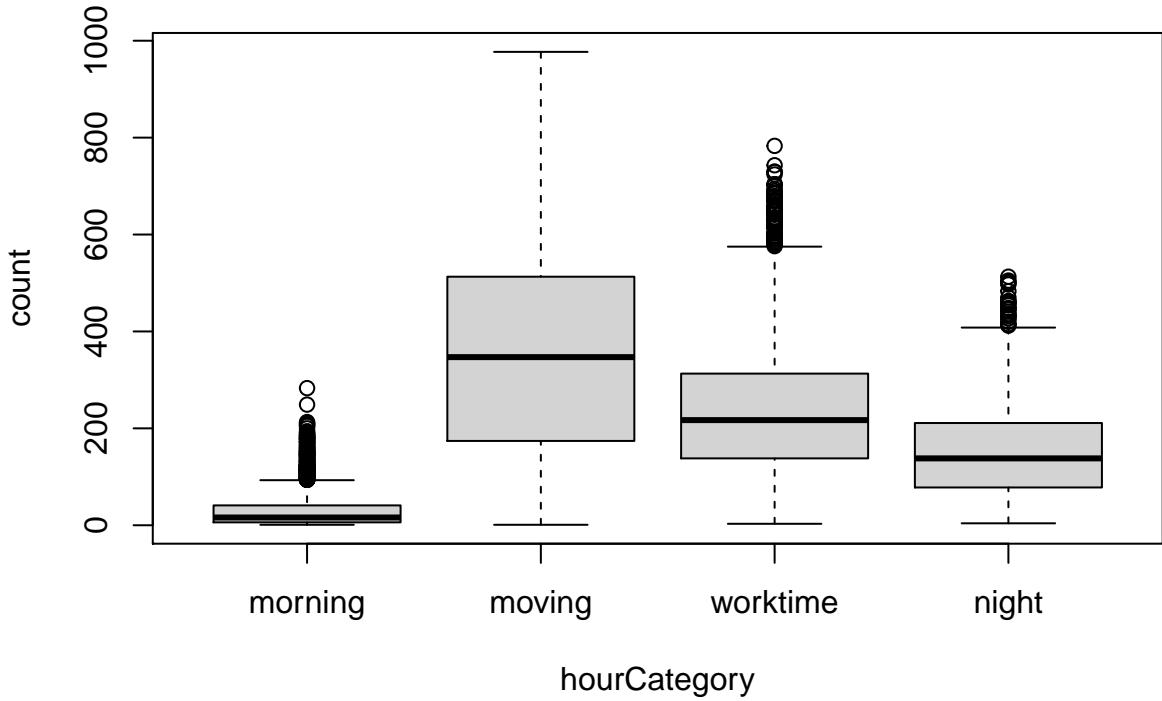
Descriptiva de la variable respuesta en función de la variable hora:



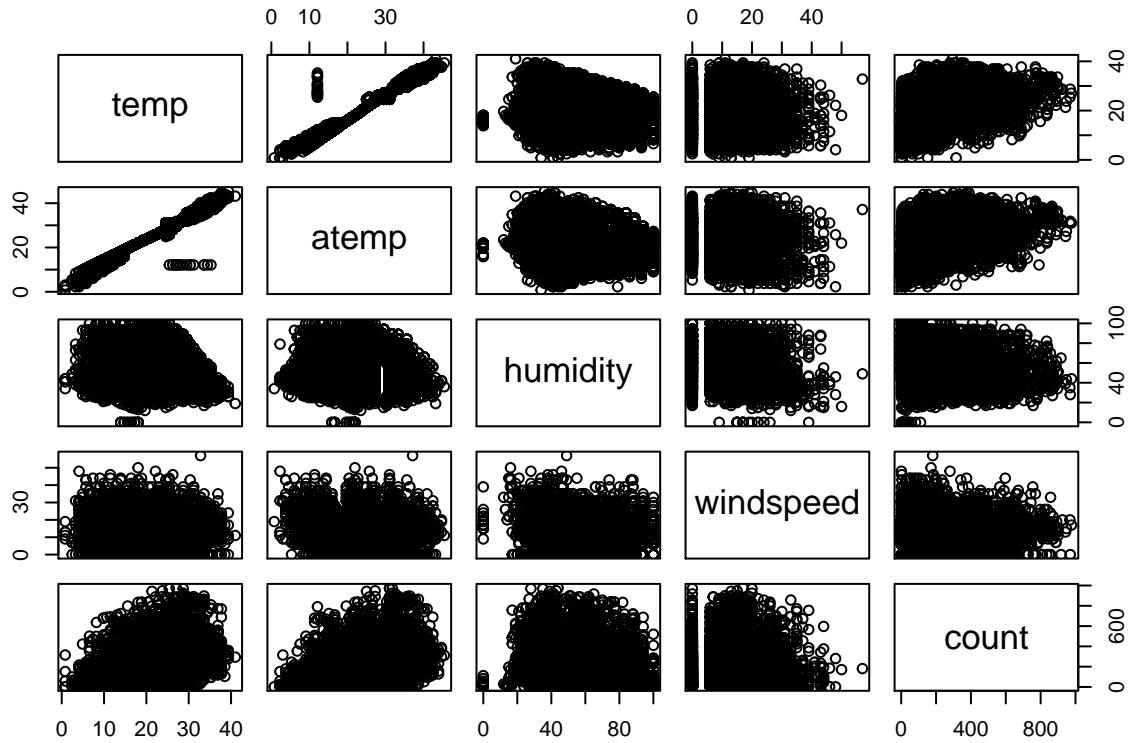
Decidimos agrupar la variable hora en los siguientes grupos:

- Morning: de 0:00h a 6:00h
- Moving: de 6:00h a 8:00h y de 16:00 a 19:00
- Worktime: de 8:00h a 16:00h
- Night: de 19:00h a 23:00h

A continuación la descriptiva de la variable hora categorizada:



2.2 Descriptiva de las variables numericas

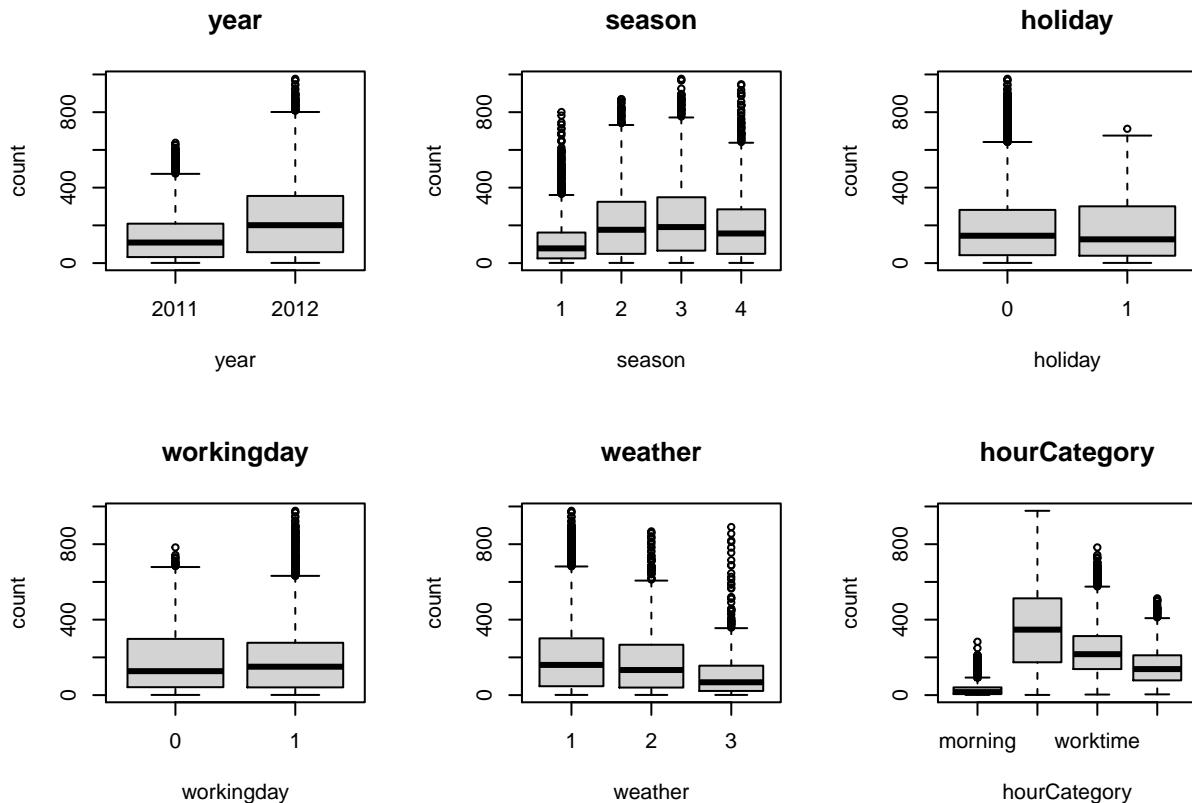


Como podemos observar en la descriptiva que mostramos a arriba, vemos que la variable temp y atemp estan muy relacionadas. Generamos un modelo utilizando cada una de las variables por separado para ver cual de las dos predice peor el resultado y eliminarla.

| | Modelo utilizando Atemp | Modelo utilizando Temp |
|-----------|-------------------------|------------------------|
| R-squared | 0.1578 | 0.1537 |

Atemp describe mejor el resultado (R-squared mayor), eliminamos la variable Temp.

2.3 Descriptiva de las variables categoricas



Vemos que a primera vista parece que algunas categorías van a influir más en la respuesta que otras. Las que tienen boxplots muy similares entre categorías menos representativas van a ser, como es el caso de Holiday y Working day.

3 Generación del modelo

3.1 Modelo 1 - Utilizando todas las variables

- Variables utilizadas: year, season, holiday, workingday, weather, temp, humidity, windspeed, hourCategory

| Propiedad | Valores |
|-------------------------|-----------|
| Residual standard error | 111.3 |
| Multiple R-squared | 0.6276 |
| p-value | < 2.2e-16 |

3.2 Modelo 2 - Selección automática de variables

Hemos utilizado el método matemático AIC (Akaike Information Criterion) para determinar qué conjunto de variables es el óptimo para explicar el modelo y cuales sería conveniente eliminar. Recordemos que cuanto menor es el AIC mejor.

| Variable a eliminar | AIC eliminando la variable |
|---------------------|----------------------------|
| workingday | 72472 |
| <ninguna> | 72473 |
| windspeed | 72474 |
| holiday | 72476 |
| weather | 72562 |
| humidity | 72664 |
| season | 72862 |
| temp | 73040 |
| year | 73501 |
| hourCategory | 77039 |

Eliminamos la variable `workingday` y generamos otro modelo. A continuación el resultado.

| Propiedad | Valores |
|-------------------------|-----------|
| Residual standard error | 111.3 |
| Multiple R-squared | 0.6275 |
| p-value | < 2.2e-16 |

- Vemos un resultado prácticamente idéntico al del Modelo 1 pero utilizando una variable menos.

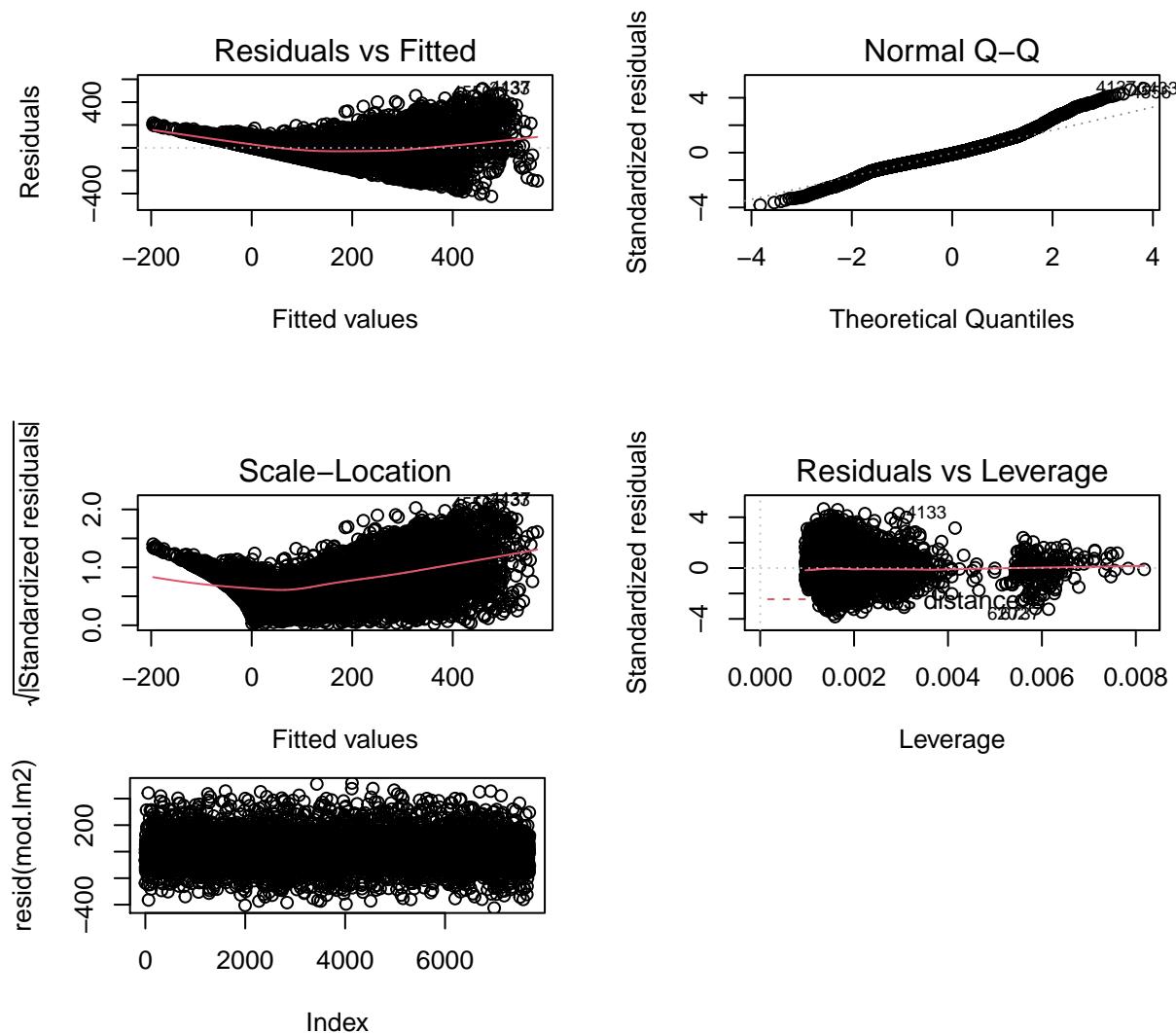
3.2.1 Análisis de colinealidad de las variables

Utilizamos el indicador de VIF para analizar la colinealidad de las variables restantes por si tuviéramos que eliminar alguna más. Consideramos un valor de VIF < 5 como bueno.

| Variable | VIF |
|--------------|----------|
| year | 1.025830 |
| season | 3.169114 |
| holiday | 1.003029 |
| weather | 1.292185 |
| temp | 3.083793 |
| humidity | 1.684752 |
| windspeed | 1.175091 |
| hourCategory | 1.319474 |

Podemos observar que el indicador VIF de todas las variables se mantiene por debajo del 5, lo que nos indica que hay poca colinealidad entre las variables y que no tendríamos que eliminar ninguna.

3.2.2 Validación de las premisas



- **Homocedasticidad:** Vemos una forma de cono clara, no se cumple.
- **Linealidad:** Hay curvatura, tampoco se cumple.
- **Normalidad de los residuos:** Desviación de la normal tanto en valores pequeños y grandes, no se cumple.
- **Independencia:** Se cumple, vemos la misma varianza de los residuos a lo largo del orden en que aparecen en el conjunto de datos.

3.3 Modelo 3 - Transformación de la variable resuesta

Hemos generado dos nuevos modelos utilizando la transformación logarítmica y la de BoxCox de la variable respuesta. A continuación los resultados.

| | Transformación logarítmica | Transformación BoxCox |
|-------------------------|----------------------------|-----------------------|
| R-squared | 0.7184 | 0.7418 |
| Residual standard error | 0.7899 | 0.5919 |

- En general vemos una mejora sustancial utilizando una transformación en la variable respuesta.

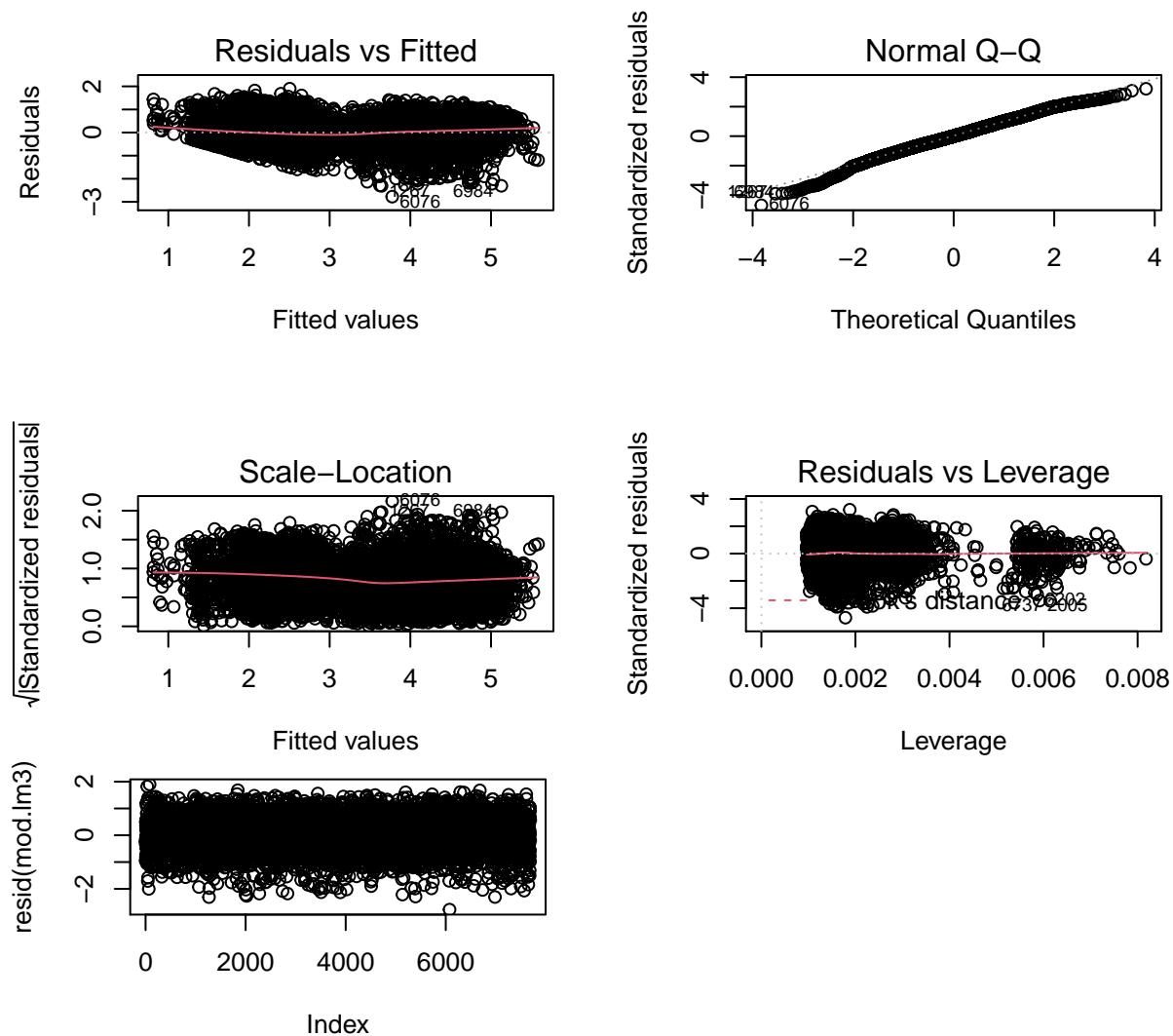
- De las dos transformaciones, nos quedamos con la de BoxCox ya que nos da un mejor resultado.

4 Modelo final - Modelo 3

- Variables utilizadas: `year`, `season`, `holiday`, `weather`, `temp`, `humidity`, `windspeed`, `hourCategory`
- Transformaciones:
 - BoxCox en la variable respuesta
 - Categorización de la variable hora
- Resultado:

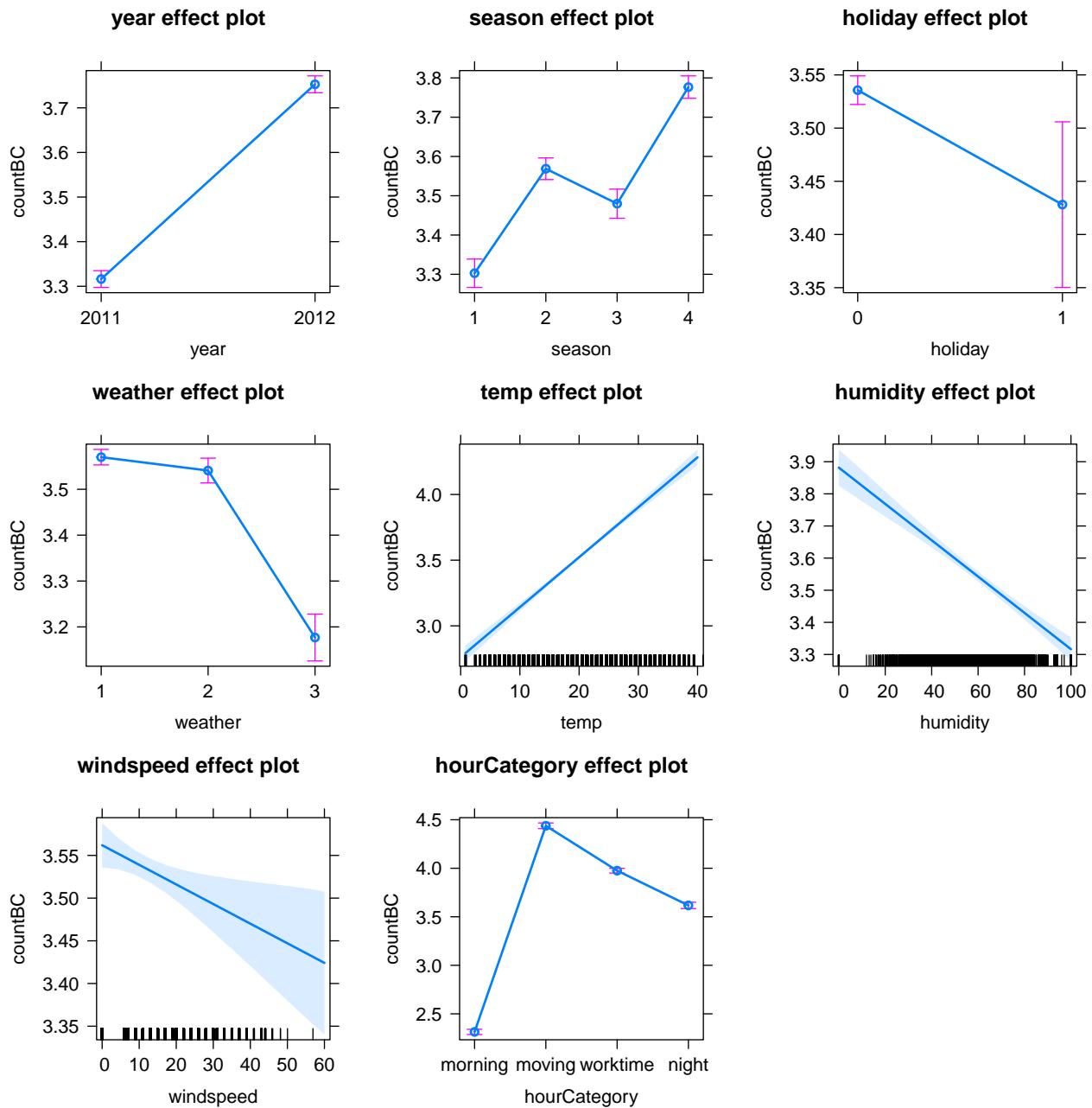
| Propiedades | Valores |
|-------------------------|-----------|
| Residual standard error | 0.5919 |
| Multiple R-squared | 0.7418 |
| p-value | < 2.2e-16 |

4.1 Validación del modelo



- **Homocedasticidad:** Se cumple, sigue habiendo más dispersión en el centro que en los extremos, pero ya no tenemos la forma de cono que teníamos en el Modelo 2.
- **Linealidad:** Se cumple, hay curvatura pero muy leve.
- **Normalidad de los residuos:** Sigue desviándose en valores altos pero podríamos considerar que ahora se cumple, ha mejorado respecto al Modelo 2.
- **Independencia:** Se cumple, vemos la misma varianza de los residuos a lo largo del orden en que aparacene en el conjunto de datos.

4.2 Efecto de las características sobre la variable respuesta



5 Modelos descartados

5.1 Modelo 5: Transformación polinómica

Hemos realizado una transformación polinómica a las variables numéricas y generado un nuevo modelo con ellas. Vemos el resultado a continuación.

- Modelo de referencia: Modelo 3
- Variables utilizadas: `year`, `season`, `holiday`, `weather`, `temp`, `humidity`, `windspeed`, `hourCategory`

- Transformaciones:
 - BoxCox en la variable respuesta
 - Categorización de la variable hora
 - Transformación polinómica de las variables `temp`, `humidity` y `windspeed`
- Resultado:

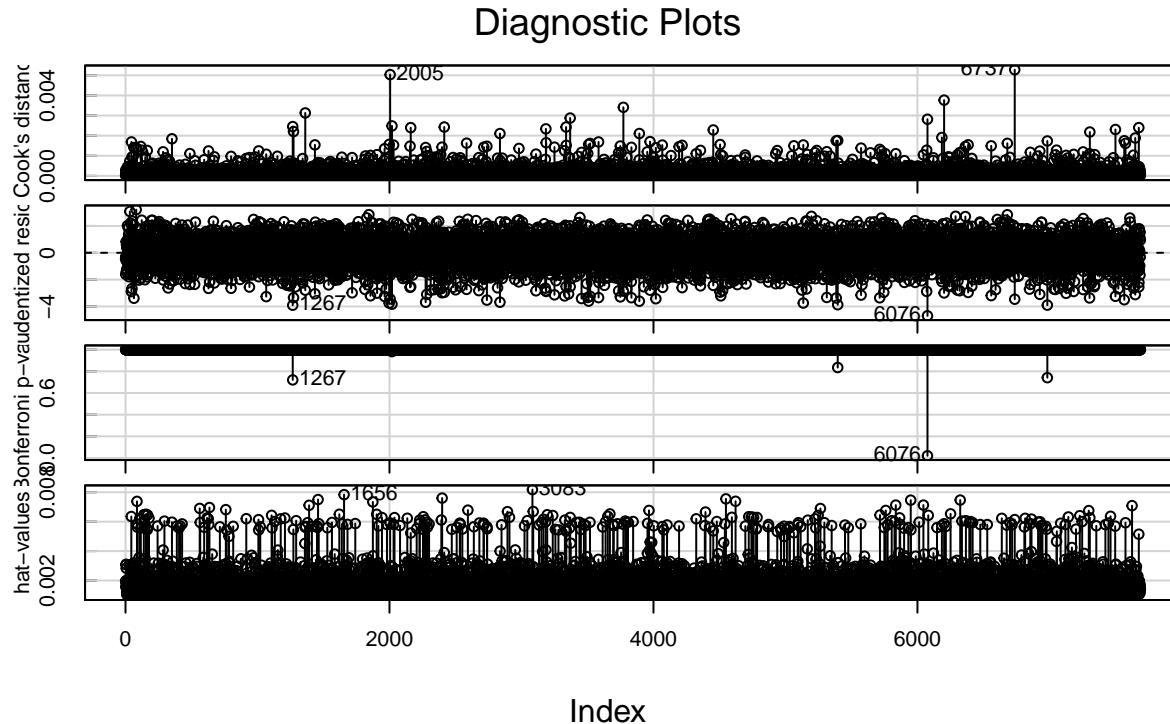
| Propiedades | Valores |
|-------------------------|-----------|
| Residual standard error | 0.5884 |
| Multiple R-squared | 0.7449 |
| p-value | < 2.2e-16 |

- Vemos que hay una diferencia insignificante respecto al Modelo 3, por lo que no vemos que sea necesario utilizar la transformación polinómica

6 Anexo

6.1 Modelo 6: Eliminar observaciones influyentes

Analizamos las observaciones influyentes del Modelo 3:



Vemos que las observaciones 2005 y 6737 son las que tienen una distancia de Cook mayor, por lo que son las más influyentes. A continuación sus distancias de Cook respectivas:

| Observación | Distancia de Cook |
|-------------|-------------------|
| 2005 | 0.0050 |
| 6737 | 0.0052 |

Eliminamos estas dos observaciones y generamos un nuevo modelo.

- Modelo de referencia: Modelo 3
- Variables utilizadas: `year`, `season`, `holiday`, `weather`, `temp`, `humidity`, `windspeed`, `hourCategory`
- Transformaciones:
 - Eliminación de las observaciones influyentes 2005 y 6737
- Resultado:

| Propiedades | Valores |
|-------------------------|-----------|
| Residual standard error | 0.5911 |
| Multiple R-squared | 0.7425 |
| p-value | < 2.2e-16 |

- De nuevo, la diferencia es insignificante respecto el Modelo 3, por lo que no interesa hacer esta transformación