

# MBD - Estadística - Práctica 1a (ML)

Arturo Menchaca y Víctor Juez

Noviembre 22, 2020

## Análisis de variables

El conjunto de datos consta de las siguientes variables:

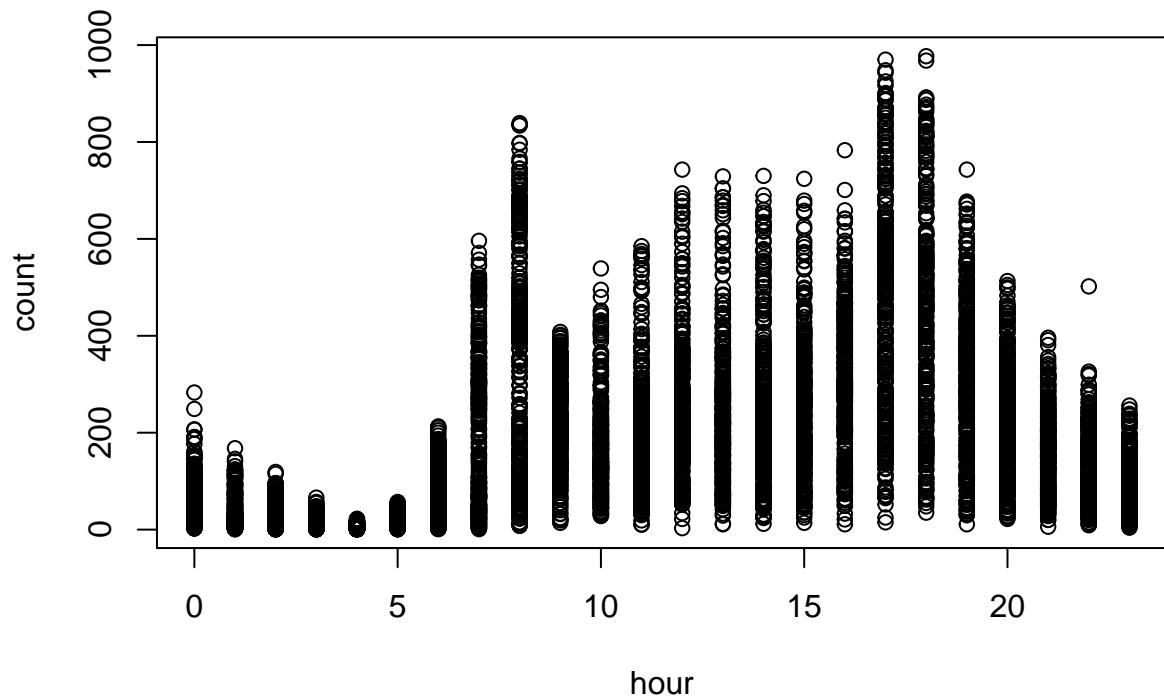
- id: identificador de la franja horaria (no guarda relación con el orden temporal)
- year: año (2011 o 2012)
- hour: hora del día (0 a 23)
- season: 1 = invierno, 2 = primavera, 3 = verano, 4 = otoño
- holiday: si el día era festivo
- workingday: si el día era laborable (ni festivo ni fin de semana)
- 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                   3: 586
##                  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
```

## Categorización de la variable hora

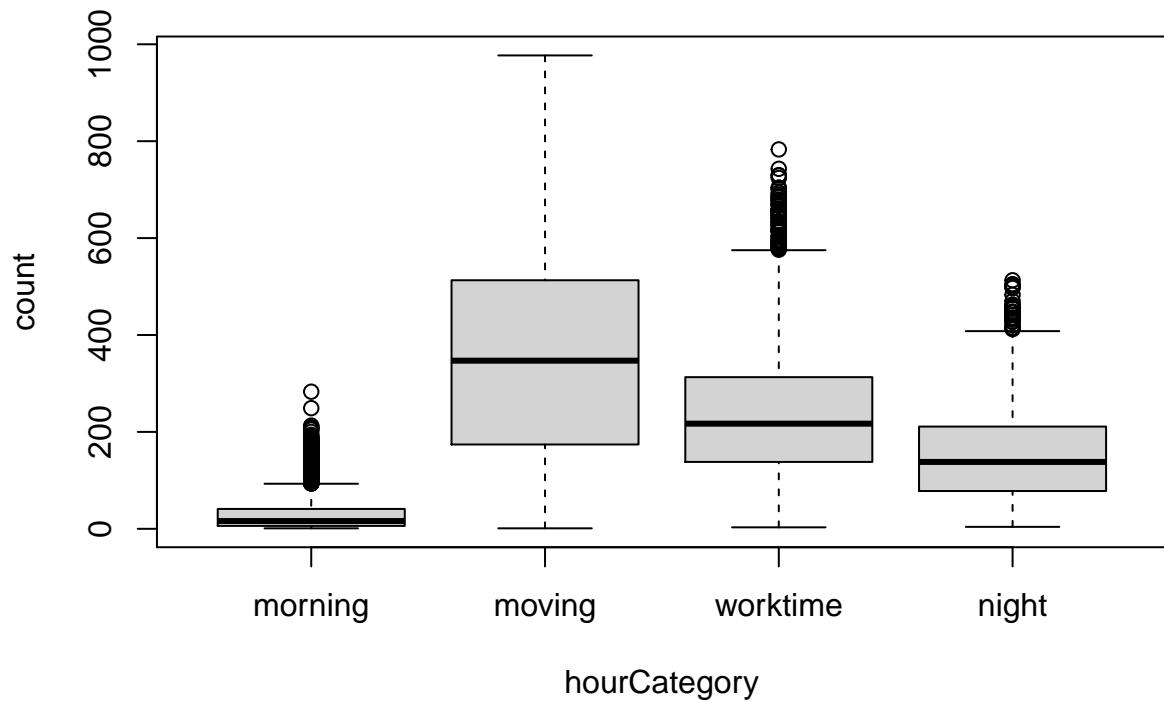
Descriptiva de la variable respuesta en función de la variable hora:



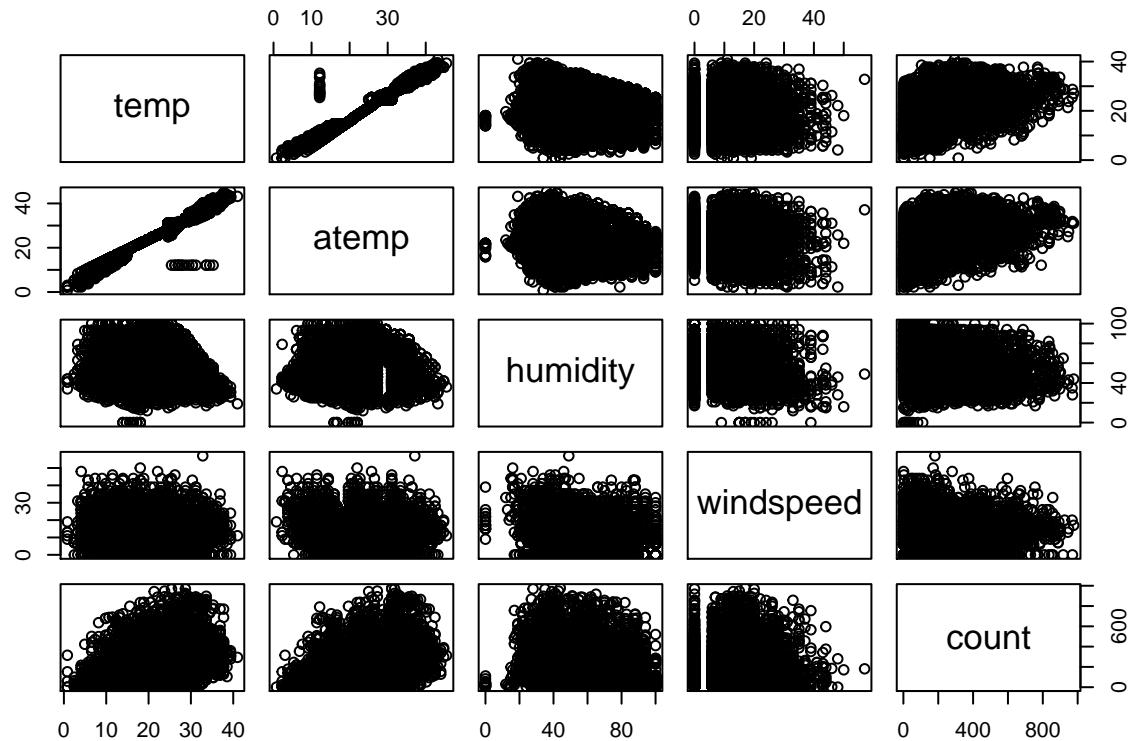
Decidimos agrupar la variable hora en los siguientes grupos:

- Morning: de 0:00h a 6:00h
- Moving: de 7:00h a 8:00h y de 17:00 a 19:00
- Worktime: de 9:00h a 16:00h
- Night: de 20:00h a 23:00h

A continuación la descriptiva de la variable hora categorizada:



## 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, así que vamos a analizar cual de las dos predice peor la respuesta para descartarla.

### Analisis de la variable Temp:

```
## 
## Call:
## lm(formula = count ~ temp, data = datos)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -293.73 -112.49  -32.32   79.01  741.84 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  4.0011    5.2937   0.756   0.45    
## temp         9.2474    0.2437  37.950  <2e-16 *** 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 167.2 on 7687 degrees of freedom
## Multiple R-squared:  0.1578, Adjusted R-squared:  0.1577 
## F-statistic: 1440 on 1 and 7687 DF, p-value: < 2.2e-16
```

### Analisis de la variable Atemp:

```
## 
## Call:
## lm(formula = count ~ atemp, data = datos)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -293.73 -112.49  -32.32   79.01  741.84 
```

```

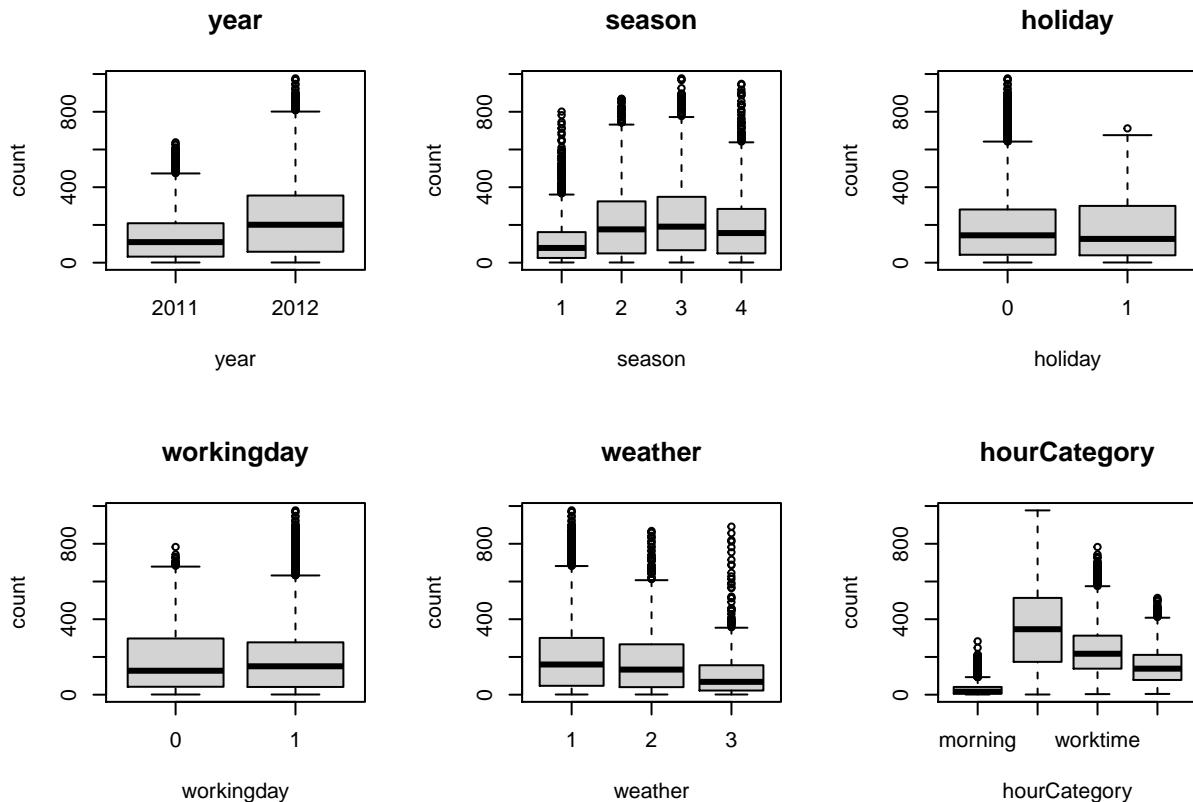
## -297.99 -113.88 -33.36 80.77 740.65
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.3044    5.6520 -1.292   0.196
## atemp        8.3862    0.2245 37.360 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 167.6 on 7687 degrees of freedom
## Multiple R-squared:  0.1537, Adjusted R-squared:  0.1536
## F-statistic: 1396 on 1 and 7687 DF, p-value: < 2.2e-16

```

Dados los resultados, eliminamos la variable Atemp ya que predice peor el resultado:

- Temp R-cuadrado: 0.1578
- ATemp R-cuadrado: 0.1537

## Variables categoricas



A simple vista holiday y workingday son las variables menos representativas ya que poseen boxcoxs muy similares para las distintas categorías.

# Generación del modelo

## Modelo 1: Todas las variables

Como modelo inicial partimos de todas las variables de las que disponemos, obteniendo el siguiente modelo:

```
##  
## Call:  
## lm(formula = count ~ ., data = datos)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -423.44  -68.88  -10.53   57.87  517.01  
##  
## Coefficients:  
##                               Estimate Std. Error t value Pr(>|t|)  
## (Intercept)           -75.70632    8.16201 -9.275 < 2e-16 ***  
## year2012              85.22421   2.57031 33.157 < 2e-16 ***  
## season2                37.25992   4.71127  7.909 2.96e-15 ***  
## season3                17.24770   6.02254  2.864  0.0042 **  
## season4                69.76330   3.91026 17.841 < 2e-16 ***  
## holiday1               -16.41548   7.81938 -2.099  0.0358 *  
## workingday1            2.28868   2.81270  0.814  0.4158  
## weather2               -6.45033   3.14801 -2.049  0.0405 *  
## weather3               -51.02053   5.30102 -9.625 < 2e-16 ***  
## temp                   6.92765   0.28540 24.274 < 2e-16 ***  
## humidity                -1.19059   0.08533 -13.952 < 2e-16 ***  
## windspeed               -0.24529   0.16817 -1.459  0.1447  
## hourCategorymoving     300.02671   3.83266 78.282 < 2e-16 ***  
## hourCategoryworktime   163.58095   3.67885 44.465 < 2e-16 ***  
## hourCategorynight      101.19095   3.98325 25.404 < 2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 111.3 on 7674 degrees of freedom  
## Multiple R-squared:  0.6276, Adjusted R-squared:  0.6269  
## F-statistic: 923.7 on 14 and 7674 DF,  p-value: < 2.2e-16
```

Vemos que las variables workingday y windspeed poseen un significancia muy baja para el modelo, aun asi no lo podemos asegurar.

## Modelo 2: Selección automática de variables

Para este modelo realizamos una selección automática de las variables usando la función `step`.

```
## Start:  AIC=72473.46  
## count ~ year + season + holiday + workingday + weather + temp +  
##        humidity + windspeed + hourCategory  
##  
##                               Df Sum of Sq    RSS    AIC  
## - workingday      1     8195  94994910 72472  
## <none>                      94986715 72473  
## - windspeed       1    26333  95013048 72474  
## - holiday         1    54551  95041266 72476  
## - weather         2   1152324  96139039 72562  
## - humidity        1   2409478  97396192 72664  
## - season          3    5004370  99991084 72862  
## - temp            1   7293058 102279772 73040  
## - year            1  13608065 108594779 73501
```

```

## - hourCategory 3 77155740 172142455 77039
##
## Step: AIC=72472.13
## count ~ year + season + holiday + weather + temp + humidity +
##      windspeed + hourCategory
##
##          Df Sum of Sq     RSS     AIC
## <none>             94994910 72472
## - windspeed      1    26243  95021153 72472
## - holiday        1    69968  95064877 72476
## - weather        2   1146064  96140974 72560
## - humidity       1   2414001  97408910 72663
## - season         3   5005857 100000767 72861
## - temp           1   7356385 102351295 73044
## - year           1  13603508 108598418 73499
## - hourCategory   3  77149370 172144280 77037

```

Como podemos observar, la variable workingday ha sido eliminada, mientras que windspeed se ha mantenido. Viendo el modelo resultante obtenemos:

```

##
## Call:
## lm(formula = count ~ year + season + holiday + weather + temp +
##      humidity + windspeed + hourCategory, data = datos)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -425.05 -68.49 -10.60  57.66 517.75
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -74.24749   7.96250 -9.325 < 2e-16 ***
## year2012                  85.20709   2.57017 33.152 < 2e-16 ***
## season2                   37.14065   4.70888  7.887 3.51e-15 ***
## season3                   17.02655   6.01627  2.830  0.00467 **
## season4                   69.70831   3.90959 17.830 < 2e-16 ***
## holiday1                 -18.00294   7.57192 -2.378  0.01745 *
## weather2                  -6.39429   3.14718 -2.032  0.04221 *
## weather3                 -50.82987   5.29572 -9.598 < 2e-16 ***
## temp                      6.94282   0.28478 24.379 < 2e-16 ***
## humidity                  -1.19158   0.08532 -13.966 < 2e-16 ***
## windspeed                 -0.24487   0.16817 -1.456  0.14540
## hourCategorymoving       299.98334   3.83220 78.280 < 2e-16 ***
## hourCategoryworktime    163.51235   3.67780 44.459 < 2e-16 ***
## hourCategorynight        101.17379   3.98310 25.401 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 111.3 on 7675 degrees of freedom
## Multiple R-squared:  0.6275, Adjusted R-squared:  0.6269
## F-statistic: 994.7 on 13 and 7675 DF,  p-value: < 2.2e-16

```

Aqui podemos ver que este modelo presenta el mismo R-cuadrado y error estándar, pero al trabajar con menos variables nos quedamos con este modelo.

## Análisis de colinealidad (vif)

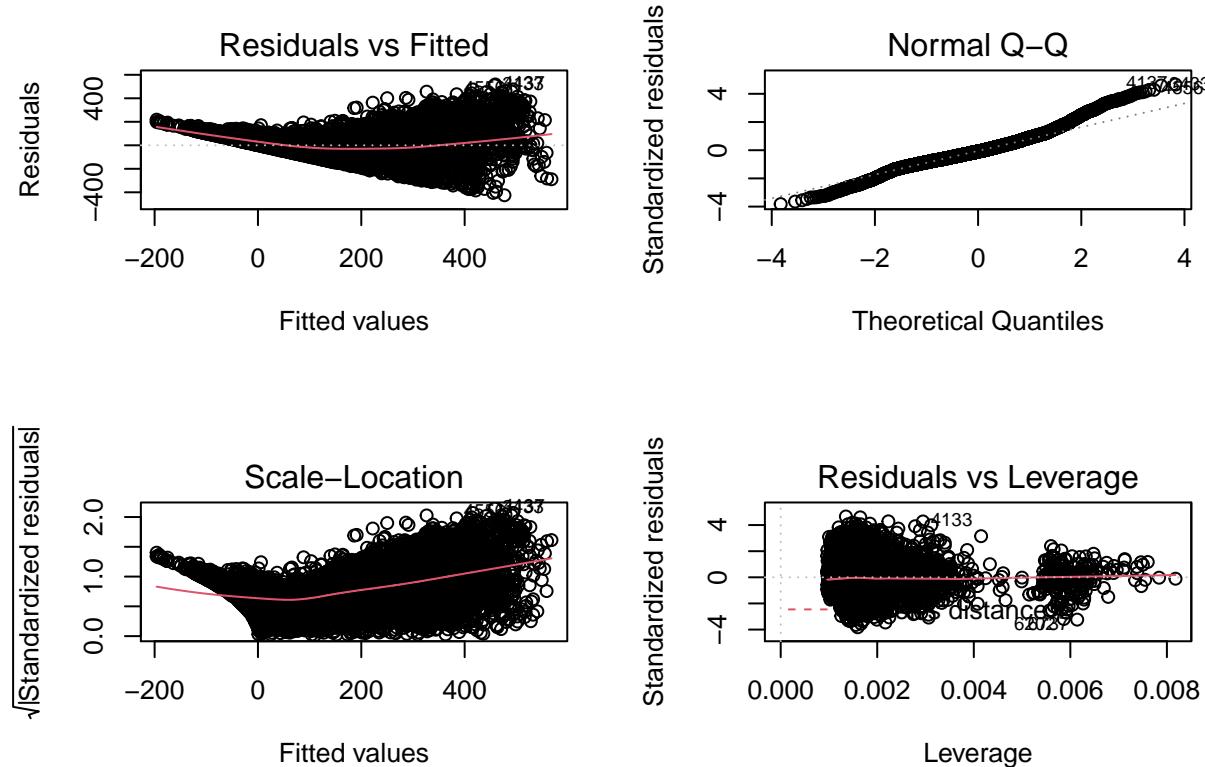
Antes de seguir mejorando el modelo, revisamos si existen algunas variables muy dependientes del resto, para esto usamos la función `vif` de la librería `car`:

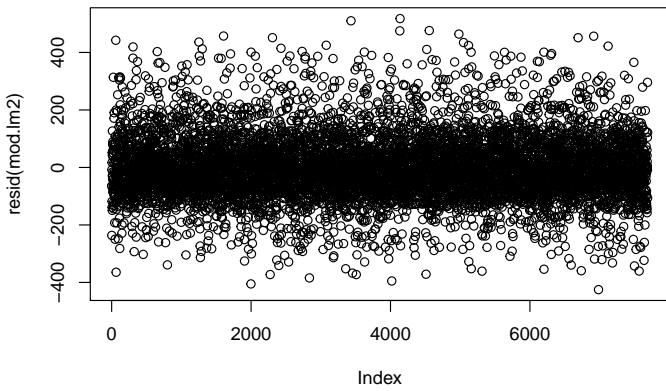
```
## Loading required package: carData
##          GVIF Df GVIF^(1/(2*Df))
## year      1.025830  1     1.012833
## season    3.169114  3     1.211964
## holiday   1.003029  1     1.001513
## weather   1.292185  2     1.066181
## temp      3.083793  1     1.756073
## humidity  1.684752  1     1.297980
## windspeed 1.175091  1     1.084016
## hourCategory 1.319474  3     1.047290
```

Como podemos observar, no existe ninguna variable lo suficientemente dependiente para eliminarla, `season` y `temp` son las que poseen un `vif` más alto pero no alcanzan el umbral de 5 como para considerar eliminarlas.

## Validación de las premisas

Para validar el modelo, analizamos las premisas de linealidad, homoscedasticidad, normalidad de residuos e independencia.





Dado los gráficos anteriores poder llegar a las siguientes conclusiones acerca de las premisas:

- Linealidad: Se observa una curvatura de los puntos donde no se mantiene la uniformidad a lo largo de los valores predichos, por lo que asumimos que no cumple esta premisa.
- Homocedasticidad: Claramente se ve una forma de cono donde no se mantiene la distribución de los puntos a lo largo de los valores predichos, por lo que tampoco se cumple esta premisa.
- Normalidad: Los puntos se distribuyen bastante uniforme a lo largo de la recta de normalidad, pero hay cierto grado de desviación en los extremos que se podría mejorar.
- Independencia: Se cumple, vemos la misma varianza de los residuos a lo largo del orden en que aparecen, sin ningún patrón que indique dependencia de las muestras.

## Transformación de la variable respuesta

Para tratar de mejorar estos valores de las premisas probamos aplicando distintas transformaciones a la respuesta para medir si ofrecen mejores resultados:

### Modelo 3: Transformación BoxCox

Aplicamos la transformación de BoxCox a la variable respuesta, obteniendo el siguiente modelo:

```
## 
## Call:
## lm(formula = countBC ~ year + season + holiday + weather + temp +
##     humidity + windspeed + hourCategory, data = datos)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.7752 -0.3683  0.0015  0.4019  1.8997 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.5161674  0.0423634 35.790 < 2e-16 ***
## year2012    0.4364839  0.0136742 31.920 < 2e-16 ***
## season2     0.2660031  0.0250530 10.618 < 2e-16 ***
## season3     0.1770850  0.0320088  5.532 3.26e-08 ***
## season4     0.4738698  0.0208005 22.782 < 2e-16 ***
## holiday1    -0.1075865  0.0402854 -2.671  0.00759 ** 
## weather2    -0.0291989  0.0167442 -1.744  0.08123 .  
## weather3    -0.3931704  0.0281752 -13.954 < 2e-16 ***
## temp        0.0380311  0.0015152 25.100 < 2e-16 ***
## humidity    -0.0056539  0.0004540 -12.455 < 2e-16 ***
## windspeed   -0.0022970  0.0008947 -2.567  0.01027 *  
## hourCategorymoving 2.1240396  0.0203887 104.177 < 2e-16 ***
## hourCategoryworktime 1.6613631  0.0195673  84.905 < 2e-16 ***
```

```

## hourCategorynight      1.3035639  0.0211916  61.513 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5919 on 7675 degrees of freedom
## Multiple R-squared:  0.7418, Adjusted R-squared:  0.7413
## F-statistic:  1696 on 13 and 7675 DF,  p-value: < 2.2e-16

```

#### Modelo 4: Transformación logarítmica

Aplicamos una transformación logarítmica a la variable respuesta, obteniendo el siguiente modelo:

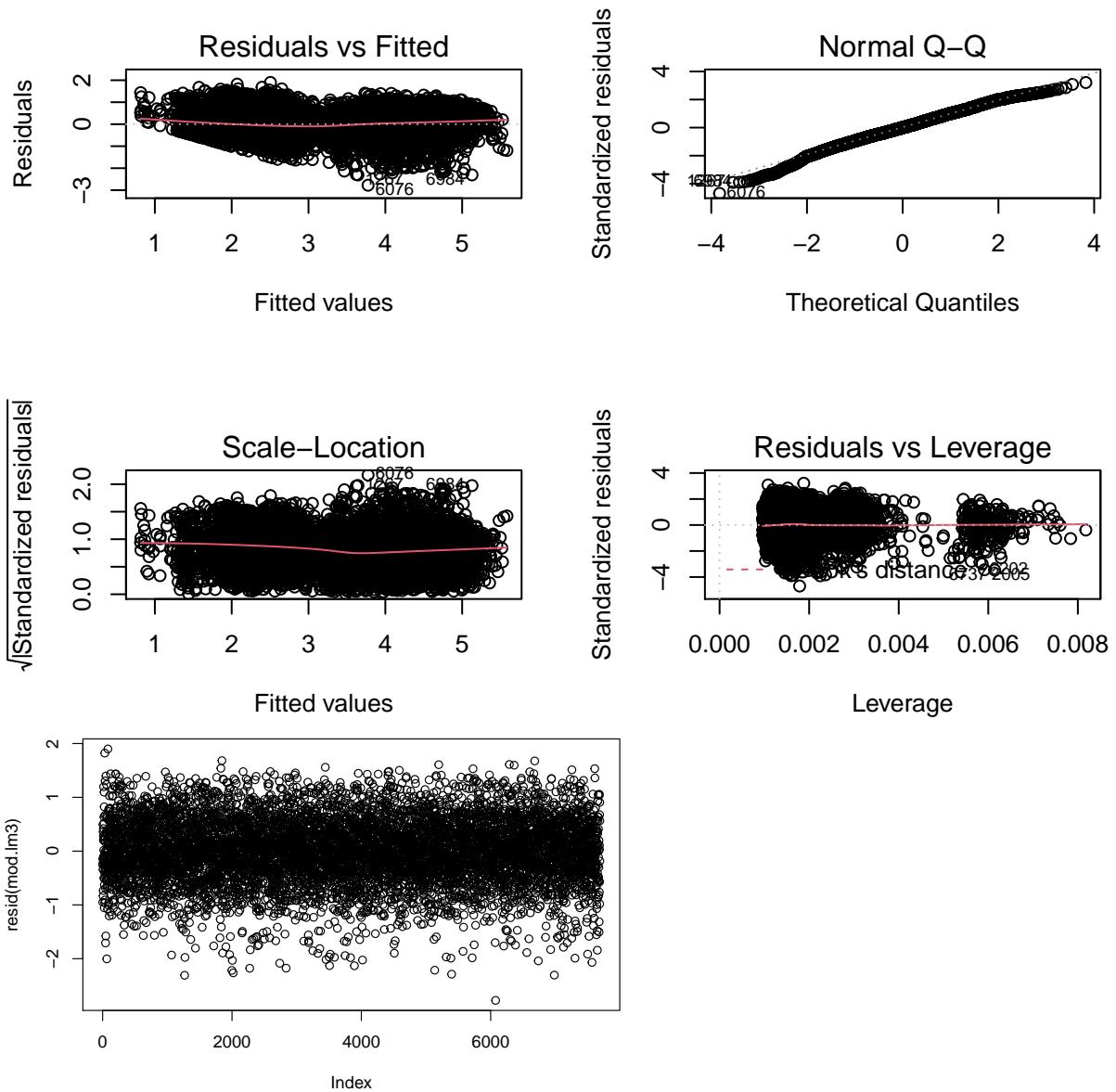
```

##
## Call:
## lm(formula = countLog ~ year + season + holiday + weather + temp +
##     humidity + windspeed + hourCategory, data = datos)
##
## Residuals:
##    Min      1Q  Median      3Q     Max
## -4.8157 -0.3914  0.0504  0.4416  2.5127
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.9782698  0.0565327 34.993 < 2e-16 ***
## year2012                 0.4688089  0.0182478 25.691 < 2e-16 ***
## season2                  0.3457507  0.0334325 10.342 < 2e-16 ***
## season3                  0.2687093  0.0427148  6.291 3.33e-10 ***
## season4                  0.5866748  0.0277576 21.136 < 2e-16 ***
## holiday1                -0.1218181  0.0537597 -2.266  0.02348 *
## weather2                 -0.0249622  0.0223446 -1.117  0.26397
## weather3                 -0.4872327  0.0375989 -12.959 < 2e-16 ***
## temp                      0.0413001  0.0020219 20.426 < 2e-16 ***
## humidity                 -0.0056929  0.0006058 -9.398 < 2e-16 ***
## windspeed                 -0.0032861  0.0011940 -2.752  0.00593 **
## hourCategorymoving        2.6604263  0.0272081 97.781 < 2e-16 ***
## hourCategoryworktime     2.2826958  0.0261119 87.420 < 2e-16 ***
## hourCategorynight         1.9179892  0.0282795 67.823 < 2e-16 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7899 on 7675 degrees of freedom
## Multiple R-squared:  0.7184, Adjusted R-squared:  0.7179
## F-statistic:  1506 on 13 and 7675 DF,  p-value: < 2.2e-16

```

Como podemos ver, en ambas transformaciones se mejora en gran medida el modelo obtenido, teniendo un valor R-cuadrado superior a 0.7. En el caso de la transformación BoxCox con un medida superior de 0.7418, por lo que nos quedamos con esta transformación de BoxCox para nuestro modelo.

**Validación de las premisas** Validamos una vez mas el modelo, analizando las premisas:

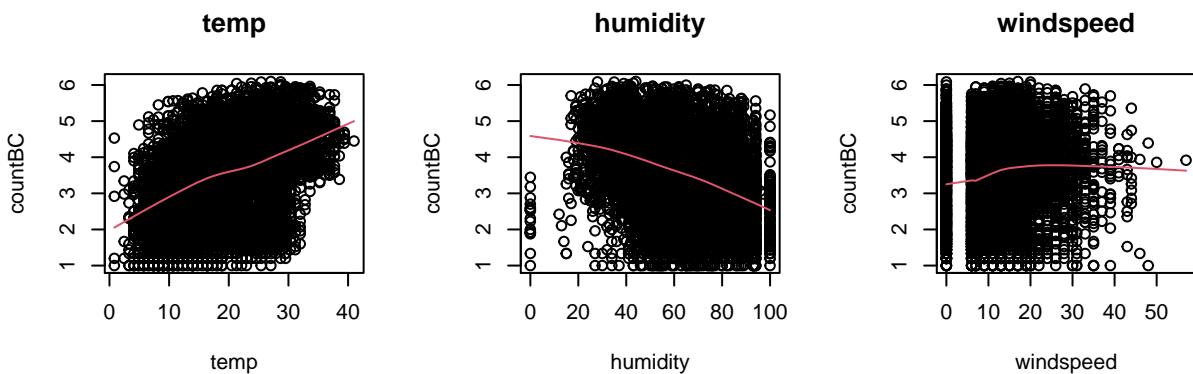


En este nuevo modelo podemos ver las mejoras sobre las premisas:

- Linealidad: Aunque aun mantiene una leve curvatura en los puntos, se observa una gran linealidad sobre la distribución a lo largo de los valores predichos, por lo que asumimos que se cumple.
- Homocedasticidad: Los puntos mantienen una mejor distribución, eliminándose en gran medida la forma de cono, por lo que igualmente asumimos que se cumple.
- Normalidad: Los puntos se distribuyen uniforme a todo lo largo, solo con una pequeña desviación pero muy poca.
- Independencia: Se cumple, al igual que en el modelo anterior.

## Modelo 5: Transformaciones polinómicas

Con el objetivo de mejorar las premisas, aplicamos transformaciones a las variables para ver si mejoran los resultados. Para esto analizamos la relación entre las variables y la respuesta:



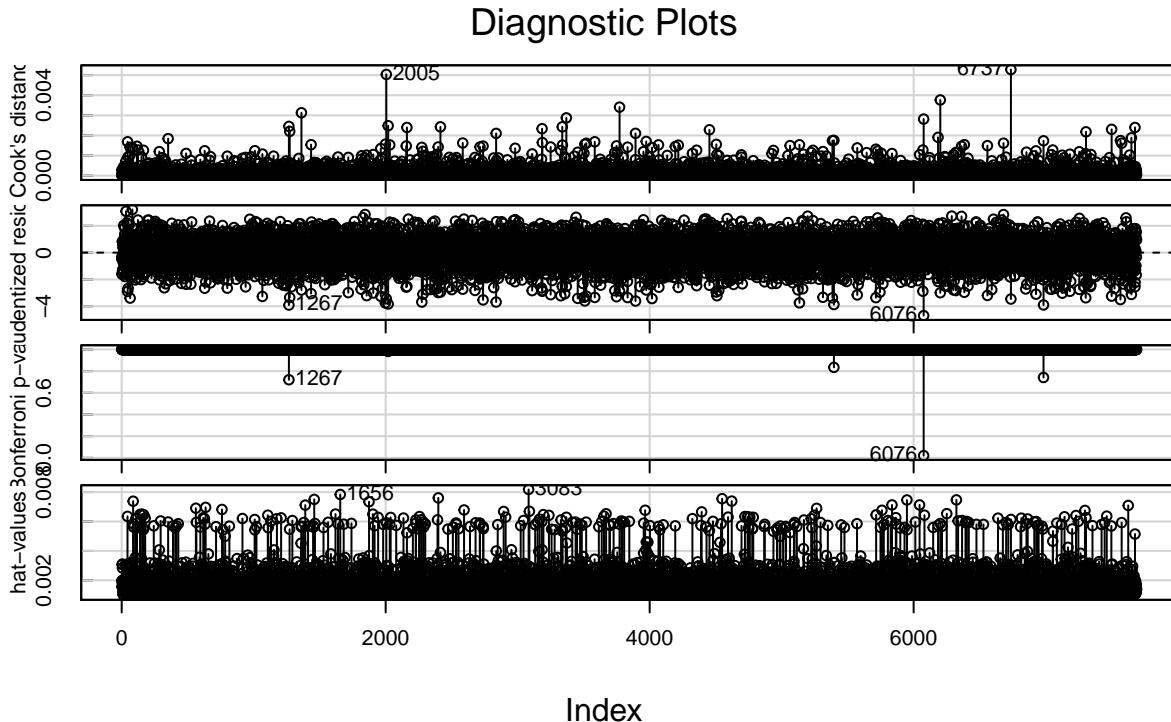
Visto que en las variables se puede observar cierta desviación, aplicamos una transformación cuadrática tratando de linearizar mas esta relación:

```
##  
## Call:  
## lm(formula = countBC ~ year + season + holiday + weather + poly(temp,  
##      2) + poly(humidity, 2) + poly(windspeed, 2) + hourCategory,  
##      data = datos)  
##  
## Residuals:  
##       Min     1Q   Median     3Q    Max  
## -2.8633 -0.3665 -0.0065  0.4038  1.8798  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1.93846  0.02478 78.239 < 2e-16 ***  
## year2012    0.42327  0.01366 30.981 < 2e-16 ***  
## season2     0.21887  0.02542  8.611 < 2e-16 ***  
## season3     0.17185  0.03229  5.323 1.05e-07 ***  
## season4     0.42424  0.02137 19.850 < 2e-16 ***  
## holiday1    -0.10128  0.04008 -2.527 0.01152 *  
## weather2    -0.02617  0.01669 -1.568 0.11687  
## weather3    -0.34638  0.02906 -11.921 < 2e-16 ***  
## poly(temp, 2)1 25.89651  1.04436 24.797 < 2e-16 ***  
## poly(temp, 2)2 -4.90709  0.68118 -7.204 6.41e-13 ***  
## poly(humidity, 2)1 -11.12672  0.78521 -14.170 < 2e-16 ***  
## poly(humidity, 2)2 -3.99788  0.62870 -6.359 2.15e-10 ***  
## poly(windspeed, 2)1 -2.04138  0.64086 -3.185 0.00145 **  
## poly(windspeed, 2)2 -1.36602  0.59656 -2.290 0.02206 *  
## hourCategorymoving 2.12385  0.02029 104.688 < 2e-16 ***  
## hourCategoryworktime 1.66115  0.01947  85.303 < 2e-16 ***  
## hourCategorynight   1.28844  0.02113  60.964 < 2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.5884 on 7672 degrees of freedom  
## Multiple R-squared:  0.7449, Adjusted R-squared:  0.7444  
## F-statistic: 1400 on 16 and 7672 DF, p-value: < 2.2e-16
```

Podemos ver que el modelo mejora con respecto al anterior, pero es despreciable teniendo en cuenta la complejidad que agrega transformar las variables, por lo que nos mantenemos con el modelo anterior.

## Observaciones influyentes

Otro punto a tener en cuenta, son las observaciones influyentes, estas las podemos obtener usando la función `influenceIndexPlot` y eliminarlas a la hora de generar el modelo:



### Modelo 6: Eliminar observaciones influyentes

Observamos que los puntos 2005 y 6737 presentan una gran influencia y los quitamos a ver si mejora la estimación:

```
##
## Call:
## lm(formula = countBC ~ year + season + holiday + weather + temp +
##     humidity + windspeed + hourCategory, data = datos2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.77644 -0.36729  0.00086  0.40071  1.90117
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.5159019  0.0423075 35.831 < 2e-16 ***
## year2012    0.4376392  0.0136567 32.046 < 2e-16 ***
## season2     0.2656383  0.0250174 10.618 < 2e-16 ***
## season3     0.1777568  0.0319638  5.561 2.77e-08 ***
## season4     0.4728267  0.0207719 22.763 < 2e-16 ***
## holiday1    -0.0891707  0.0404032 -2.207 0.02734 *  
## weather2    -0.0303112  0.0167221 -1.813 0.06993 .  
## weather3    -0.3940882  0.0281360 -14.007 < 2e-16 ***
## temp        0.0379711  0.0015131 25.096 < 2e-16 ***
## humidity    -0.0056410  0.0004534 -12.442 < 2e-16 ***
## windspeed   -0.0023051  0.0008936 -2.580 0.00991 ** 
## hourCategorymoving 2.1268040  0.0203679 104.419 < 2e-16 ***
## hourCategoryworktime 1.6618415  0.0195405  85.046 < 2e-16 ***
## hourCategorynight   1.3036725  0.0211615  61.606 < 2e-16 ***
```

```

## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5911 on 7673 degrees of freedom
## Multiple R-squared: 0.7425, Adjusted R-squared: 0.7421
## F-statistic: 1702 on 13 and 7673 DF, p-value: < 2.2e-16

```

Al igual que antes, este modelo presenta una mejora, pero muy pequeña en comparación a la complejidad que agrega al modelo, por lo que no la tendremos en cuenta.

## Modelo Final: Modelo 3 BoxCox

Como modelo final nos quedamos con el modelo número 3, el cual ofrece un valor alto de R-cuadrado y mantiene la simplicidad del modelo.

```

##
## Call:
## lm(formula = countBC ~ year + season + holiday + weather + temp +
##      humidity + windspeed + hourCategory, data = datos)
##
## Residuals:
##    Min      1Q  Median      3Q     Max
## -2.7752 -0.3683  0.0015  0.4019  1.8997
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             1.5161674  0.0423634 35.790 < 2e-16 ***
## year2012              0.4364839  0.0136742 31.920 < 2e-16 ***
## season2                0.2660031  0.0250530 10.618 < 2e-16 ***
## season3                0.1770850  0.0320088  5.532 3.26e-08 ***
## season4                0.4738698  0.0208005 22.782 < 2e-16 ***
## holiday1               -0.1075865  0.0402854 -2.671 0.00759 **
## weather2               -0.0291989  0.0167442 -1.744 0.08123 .
## weather3               -0.3931704  0.0281752 -13.954 < 2e-16 ***
## temp                   0.0380311  0.0015152 25.100 < 2e-16 ***
## humidity               -0.0056539  0.0004540 -12.455 < 2e-16 ***
## windspeed              -0.0022970  0.0008947 -2.567 0.01027 *
## hourCategorymoving    2.1240396  0.0203887 104.177 < 2e-16 ***
## hourCategoryworktime  1.6613631  0.0195673  84.905 < 2e-16 ***
## hourCategorynight     1.3035639  0.0211916  61.513 < 2e-16 ***
##
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5919 on 7675 degrees of freedom
## Multiple R-squared: 0.7418, Adjusted R-squared: 0.7413
## F-statistic: 1696 on 13 and 7675 DF, p-value: < 2.2e-16
##
## Registered S3 methods overwritten by 'lme4':
##   method                  from
##   cooks.distance.influence.merMod car
##   influence.merMod          car
##   dfbeta.influence.merMod   car
##   dfbetas.influence.merMod  car
##
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.

```

