Entrega 3

Modelos de predicción Numéricos y Binarios

Alejandro Alarcón

23/12/2021

Contents

1 Introducción

1.1 Variables explicativas numéricas

Primero de todo, vamos a empezar aplicando una corrección para los valores de nuestras variables numéricas eliminando los valores 0 para poder eviar algunos errores que podrían salir a posteriori.

```
vars_con
ll<-which(df$age==0);ll
df$age[ll]<-0.5

ll<-which(df$tax==0);ll
df$tax[ll]<-0.5

ll<-which(df$mpg==0);ll
df$mpg[ll]<-0.5

ll<-which(df$mileage==0);ll
df$mileage[ll]<-0.5
```

2 Modelo de regresión lineal

2.1 Variables numéricas

Planteamos nuestro primer modelo basado en las variables numéricas. Con este modelo, pretendemos plantear una regresión lineal que tenga como target la variable price, y que use como variables explicativas mileage, tax, mpg y age.

```
m1<-lm(price~mileage+tax+mpg+age,data=df)
summary(m1)
```

```
Call:
lm(formula = price ~ mileage + tax + mpg + age, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-34504  -4799   -255   3259  69279

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.476e+04  1.627e+03  33.652  < 2e-16 ***
mileage     -4.516e-02  8.015e-03  -5.635 1.85e-08 ***
tax         -2.301e+01  9.385e+00  -2.452   0.0142 *
mpg         -4.138e+02  1.086e+01 -38.098  < 2e-16 ***
age         -1.883e+03  8.658e+01 -21.747  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7386 on 4995 degrees of freedom
Multiple R-squared:  0.5107,    Adjusted R-squared:  0.5103
F-statistic:  1303 on 4 and 4995 DF,  p-value: < 2.2e-16
```

En el summary que acabamos de mostrar, podemos apreciar varias cosas:

En primer lugar, podemos ver los errores residuales que se generan en el modelo.

También podemos ver los coeficientes que se plantean para las difernetes variables del modelo así como el término independiente (Intercept) 5.476e+04. Podemos ver como a priori, los coeficientes para todas las variables son significativamente distintos a 0.

Con el valor R-squared, podemos apreciar también que el modelo este modelo explica un 51% de la variabilidad de la variable price.

Por último, podemos ver también como el F-statistic nos indica que el p-valor de que todos los coeficientes sean iguales a 0 se puede rechazar.

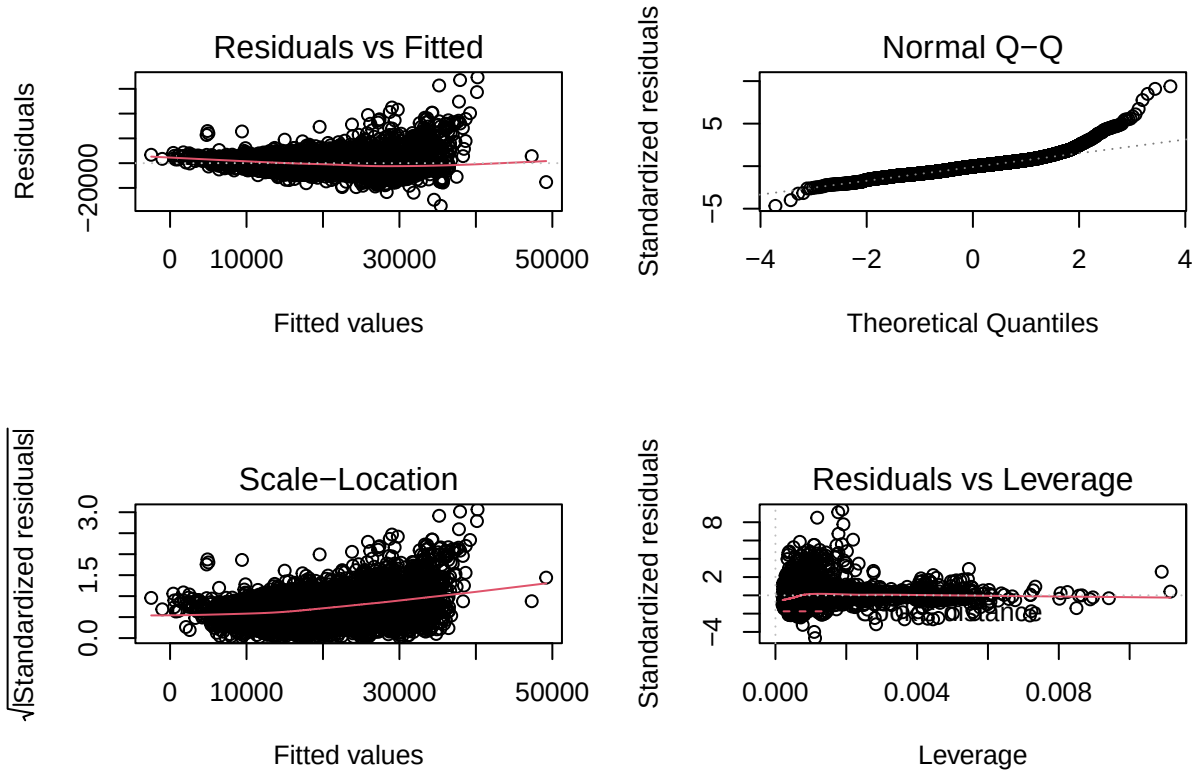
Si analizamos los valores de vif (variable inflation factor), podemos ver que los más altos corresponden a las variables de mileage y age, cercanos a 3. Estos valores nos indican la co-linealidad de las variables, y empezarían a ser preocupantes cuando se acercan al 5, de modo que de momento son correctos.

```
vif(m1)
```

```
 mileage      tax      mpg      age
2.774734 1.232502 1.348349 2.760318
```

En los siguientes gráficos, podemos apreciar algunas de las características del modelo.

```
par(mfrow=c(2,2));
plot(m1,id.n=0)
```



En el gráfico de Residuals vs. Fitted podemos apreciar como no existe homocedasticidad en el modelo, ya que los residuos estan distribuidos de manera heterogénea.

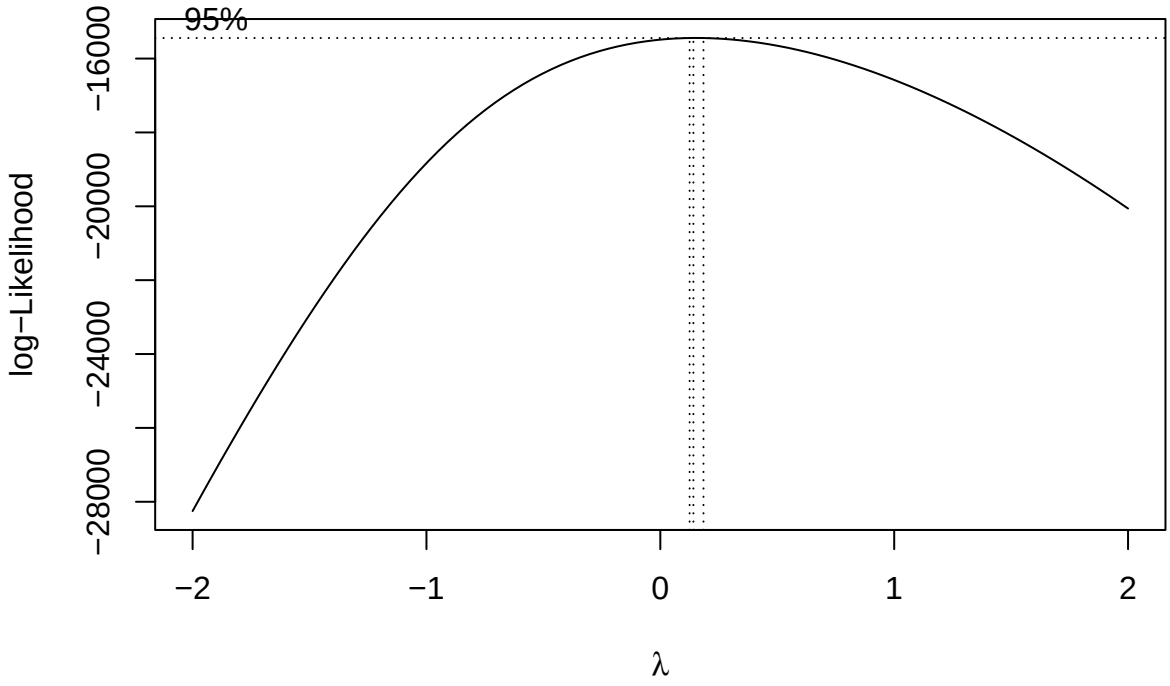
En el gráfico Normal Q-Q podemos ver como el modelo no presenta normalidad, ya que los residuos estandarizados se alejan de la recta que marca la normalidad.

En el gráfico Scale-Location, podemos volver a apreciar la heterosticidad del modelo, analizando la raiz de los residios estandarizados.

En el último gráfico, el de Residuals vs Leverage, podemos apreciar como a priori no parece que existan observaciones influyentes, ya que ninguno de los puntos que se muestran tiene una distancia de Cook que indique sobre-influencia.

A continuación, vamos a determinar si alguna de las variables que hemos introducido en el modelo anterior requiere algún tipo de transformación para lograr un mejor ajuste del modelo.

```
par(mfrow=c(1,1))
boxcox(price~mileage+tax+mpg+age,data=df)
```



En el gráfico podemos ver como el intervalo de lambda que aparece es cercano al 0, hecho que indicaría la necesidad de transformar nuestra variable target price a una escala logarítmica.

Vamos a proceder a plantear un segundo modelo que incluya esta transformación.

```
m2<-lm(log(price)~mileage+tax+mpg+age,data=df)
summary(m2)
```

```
Call:
lm(formula = log(price) ~ mileage + tax + mpg + age, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6717 -0.1762  0.0232  0.1961  1.2362

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.096e+01  6.903e-02 158.851   <2e-16 ***
mileage     -3.355e-06  3.400e-07  -9.868   <2e-16 ***
tax          7.973e-04  3.981e-04   2.003   0.0452 *
mpg         -1.376e-02  4.607e-04 -29.857   <2e-16 ***
age         -1.116e-01  3.673e-03 -30.376   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.3133 on 4995 degrees of freedom
Multiple R-squared:  0.5789,    Adjusted R-squared:  0.5786
F-statistic:  1717 on 4 and 4995 DF,  p-value: < 2.2e-16
```

Para este modelo, podemos apreciar como todos los coeficientes siguen siendo relevantes según el p-valor que se muestra en la última columna. Cabe destacar que para el caso de la variable tax, el p-valor ha subido y es cercano a 0.05.

Podemos ver también como el valor R-squared ha aumentado a 0.58, indicando que este nuevo modelo acumula más explicabilidad de nuestro target.

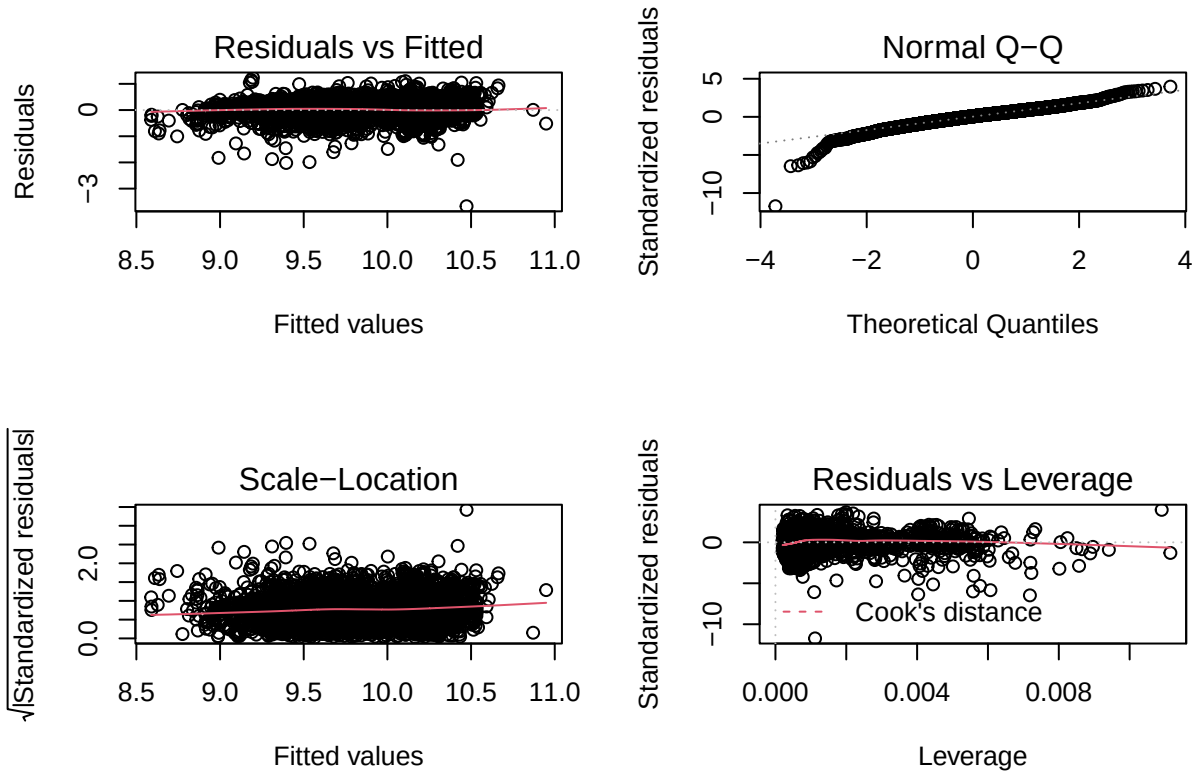
También podemos ver como los valores de vif se mantienen similares a los que aparecian en el modelo anterior, indicando que no parece existir co-linealidad entre las variables.

```
vif(m2)
```

```
 mileage      tax      mpg      age
2.774734 1.232502 1.348349 2.760318
```

Observemos algunos gráficos para analizar este nuevo modelo que hemos planeado.

```
par(mfrow=c(2,2));
plot(m2,id.n=0);
```




```
par(mfrow=c(1,1))
```

Podemos ver como según el primer gráfico, el modelo ha adquirido homocedasticidad, siendo la distribución de los residuos más homogénea.

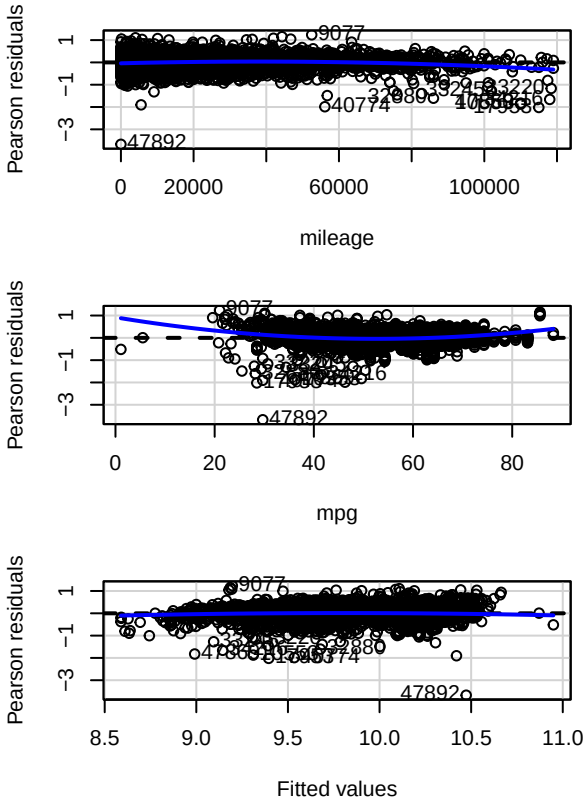
En lo que se refiere a la normalidad, podemos ver como este modelo presenta más normalidad que el anterior, al menos en lo que se refiere a los quantiles superiores. Sin embargo, para los cuantiles inferiores, aún aparece cierta distancia con la recta que describe la normalidad.

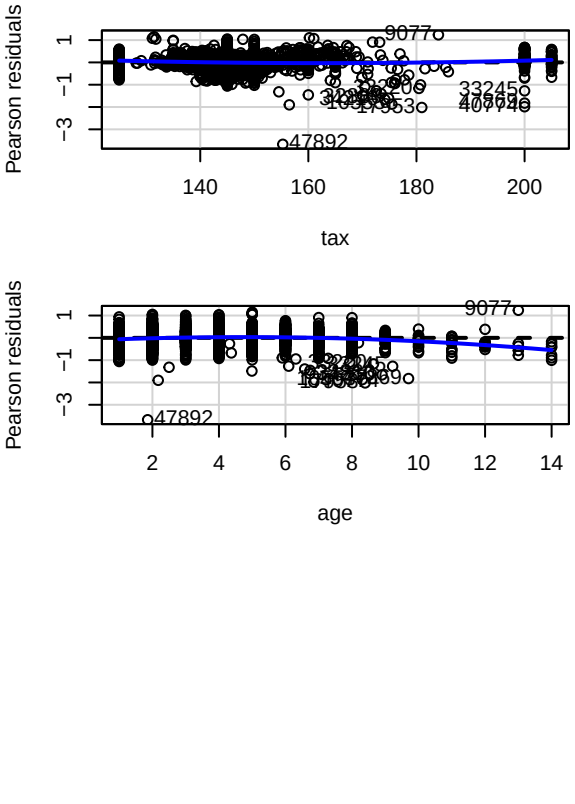
Por último, para el caso de la sobre-influencia en el modelo, podemos ver un resultado similar al del anterior modelo.

Vamos a proceder a realizar un análisis más en profuncidad del modelo.

En los siguientes gráficos podemos ver como se distribuyen los residuos.

```
residualPlots(m2,id=list(method=cooks.distance(m2),n=10))
```





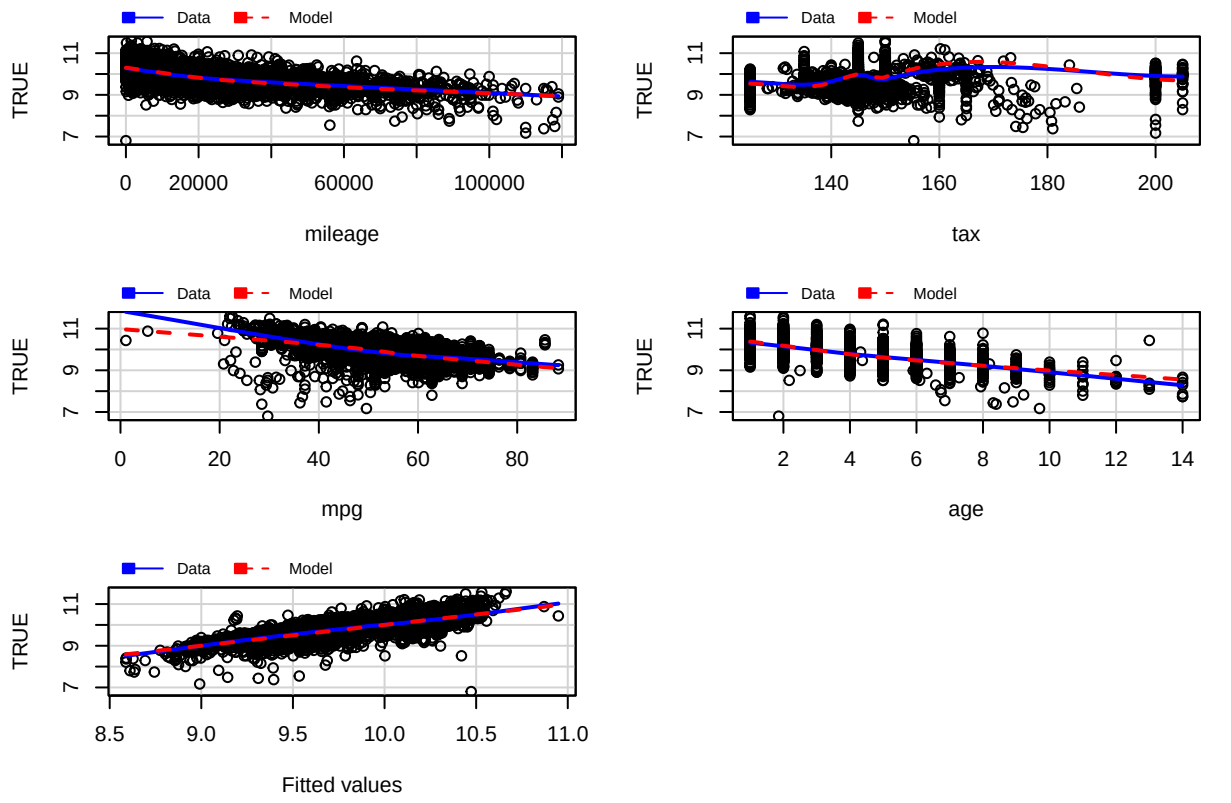

```
           Test stat Pr(>|Test stat|)
mileage      -8.9294        < 2.2e-16 ***
tax           6.7284        1.909e-11 ***
mpg          13.1778        < 2.2e-16 ***
age         -10.5882        < 2.2e-16 ***
Tukey test   -2.9525         0.003152 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Algunos casos a destacar son el de mileage, donde se puede apreciar una cierta acumulación en los valores bajos a la vez que aparecen puntos que se alejan de la nube y de la curva para valores más altos.

También podemos destacar que en el caso de age, que es una variable que originalmente generamos a partir de la variable year y se puede apreciar su distribución en columnas, aparece una nube de puntos cerca del centro, probablemente debido a la imputación a partir de PCA que se realizó en la primera entrega. Podemos ver algo parecido pero no tan evidente para la variable tax.

Con la siguiente función procederemos a ver el ajuste del modelo con los datos.

```
marginalModelPlots(m2)
```



Podemos ver como el ajuste para el caso de la variable mileage o age parecen casi perfectos, mientras que para las varibales mpg y tax existe una cierta desviación entre las curvas roja y azul.

En el caso de los Fitted values, las rectas se ajustan a la perfección.

Aplicando la función de boxTidwell podemos determinar las transformaciones que deberíamos aplicar a nuestro modelo para mejorarlo.

```
#boxTidwell(log(price)~mileage+tax+age+mpg,data=df[!df$mout=="YesMOut",], verbose=TRUE)
```

Sin embargo, cuando ejecutamos la función pasando como entrada el modelo que habíamos planteado, nos encontramos que la función falla. Aplicando el modo verbose, podemos ver como el motivo del fallo es la tendencia a +infinito del exponenete de la variable mpg.

Como no sé como proceder, vamos a realizar algunos tests:

En primer lugar, vamos a crear un nuevo modelo sin la variable mpg y lo vamos a evaluar con la función boxTidwell para determinar las transformaciones a aplicar a nuestras variables explicativas (aunque hemos tenido que aumentar el número máximo de iteraciones).

```
m3<-lm(log(price)~mileage+tax+age, data=df)
boxTidwell(log(price)~mileage+tax+age, data=df[!df$mout=="YesMOut",], max.iter=100)
```

```
        MLE of lambda Score Statistic (z)  Pr(>|z|)
mileage       0.51623              4.3253 1.523e-05 ***
tax           4.35882              6.9709 3.149e-12 ***
age           1.13578             -2.1428   0.03213 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

iterations =  33
```

Como podemos ver, este nuevo modelo que excluye la variable mpg pierde explicabilidad, pero nos permite ejecutar la función boxTidwell para determinar las transformaciones que deberíamos aplicar a nuestras variables.

Vamos a aplicar las transformaciones que aparecen con la función boxTidwell para crear un nuevo modelo.

```
m4<-lm(log(price)~sqrt(mileage)+poly(tax,4)+age, data=df)
summary(m4)
```

```
Call:
lm(formula = log(price) ~ sqrt(mileage) + poly(tax, 4) + age,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6437 -0.1893  0.0094  0.1871  1.6992

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.5949900  0.0118047 897.521  < 2e-16 ***
sqrt(mileage) -0.0015588  0.0001148 -13.576  < 2e-16 ***
poly(tax, 4)1  4.9247954  0.3440468  14.314  < 2e-16 ***
poly(tax, 4)2  2.5889209  0.3494831   7.408 1.50e-13 ***
poly(tax, 4)3 -3.8225677  0.3698610 -10.335  < 2e-16 ***
poly(tax, 4)4  1.4936334  0.3362566   4.442 9.11e-06 ***
age           -0.1409327  0.0043431 -32.449  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3354 on 4993 degrees of freedom
Multiple R-squared:  0.5176,    Adjusted R-squared:  0.517
F-statistic: 892.8 on 6 and 4993 DF,  p-value: < 2.2e-16
```

```
#boxTidwell(log(price)~sqrt(mileage)+poly(tax,4)+age, data=df[!df$mout=="YesMOut",], verbose=TRUE)
```

Volviendo a aplicar la función de boxTidwell, podemos ver como, en este caso vuelve a fallar debido a que algunos de los coeficientes de los monómios que se han generado con la función poly(tax,4) son negativos. Sin embargo, el modelo que excluye la variable mpg, aún aplicando todas las transformaciones recomendadas, tiene un valor de R-squared inferior al original.

Si ejecutamos el test de Clarke que nos permite analizar modelos no anidados, (lo he encontrado por internet) podemos ver que, aparentemente, el modelo original, que incluye la variable mpg pero ninguna transformación a parte de la logaritmica para el target, es mejor que el modelo que excluye esta variable pero incluye las transformaciones.

```
library(games)
clarke(m2, m4)
```

```
Clarke test for non-nested models

Model 1 log-likelihood: -1289
Model 2 log-likelihood: -1629
Observations: 5000
Test statistic: 3205 (64%)

Model 1 is preferred (p < 2e-16)
```

Vamos a probar incluyendo la variable mpg en el modelo que ya habíamos planteado aplicando las transformaciones que aparecían con la función boxTidwell.

```
m5<-update(m4,~.+mpg)
summary(m5)
```

```
Call:
lm(formula = log(price) ~ sqrt(mileage) + poly(tax, 4) + age +
    mpg, data = df)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.6780 -0.1769  0.0206  0.1969  1.3587

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.1517010  0.0234285 475.988  < 2e-16 ***
sqrt(mileage) -0.0007257  0.0001117  -6.499  8.9e-11 ***
poly(tax, 4)1  0.4878986  0.3612542   1.351    0.177
poly(tax, 4)2  2.1950054  0.3269186   6.714  2.1e-11 ***
poly(tax, 4)3  0.5668394  0.3821265   1.483    0.138
poly(tax, 4)4  0.3166057  0.3172552   0.998    0.318
age           -0.1204402  0.0041294 -29.167  < 2e-16 ***
mpg           -0.0140643  0.0005222 -26.935  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3134 on 4992 degrees of freedom
Multiple R-squared:  0.5788,    Adjusted R-squared:  0.5782
F-statistic: 979.9 on 7 and 4992 DF,  p-value: < 2.2e-16
```

En este caso, se puede ver como la incorporación de la variable mpg sí que aumenta la explicabilidad del modelo, generando un modelo más completo.

Además, aplicando la función anova, podemos ver como se rechaza la hipótesis de equivalencia, de modo que el nuevo modelo es mejor.

```
anova(m4,m5)
```

```
Analysis of Variance Table

Model 1: log(price) ~ sqrt(mileage) + poly(tax, 4) + age
Model 2: log(price) ~ sqrt(mileage) + poly(tax, 4) + age + mpg
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   4993 561.63
2   4992 490.36  1    71.264 725.48 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sin embargo, si lo comparamos con el original:

```
anova(m2,m5)
```

```
Analysis of Variance Table

Model 1: log(price) ~ mileage + tax + mpg + age
Model 2: log(price) ~ sqrt(mileage) + poly(tax, 4) + age + mpg
  Res.Df    RSS Df Sum of Sq F Pr(>F)
1   4995 490.21
2   4992 490.36  3  -0.15517
```

```
clarke(m2,m5)
```

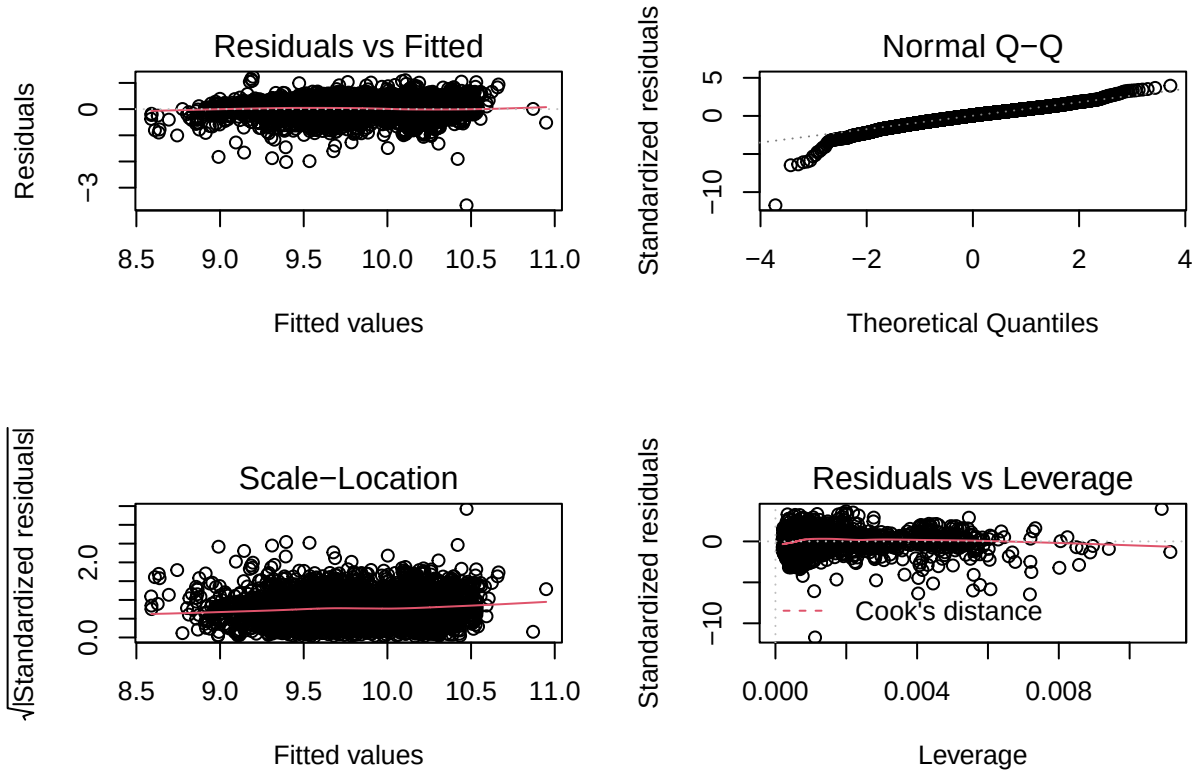
```
Clarke test for non-nested models

Model 1 log-likelihood: -1289
Model 2 log-likelihood: -1290
Observations: 5000
Test statistic: 2824 (56%)

Model 1 is preferred (p < 2e-16)
```

Podemos ver como estos dos tests determinan que ambos modelos son equivalentes, de modo que es preferible el más sencillo.

De modo que seguiremos avanzando con el segundo modelo que hemos planteado, que incluye la variable mpg pero no incluye ninguna transfromación a parte de la del target price.

Vamos a proceder a observar algunos gráficos para analizar este modelo:

```
par(mfrow=c(2,2));
plot(m2,id.n=0);
```




```
par(mfrow=c(1,1))
```

Podemos ver como según el primer gráfico, el modelo ha adquirido homocedasticidad, siendo la distribución de los residuos más homogénea.

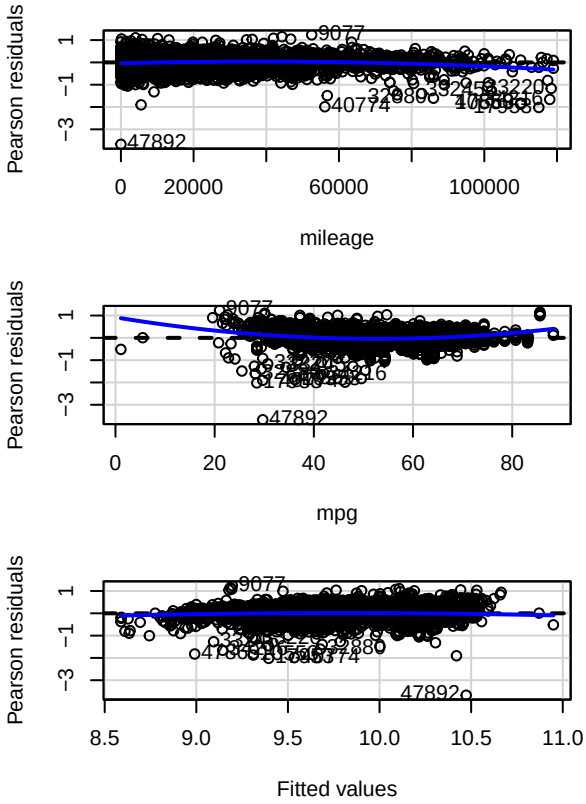
En lo que se refiere a la normalidad, podemos ver como para este modelo tiene más normalidad que el anterior, al menos en lo que se refiere a los cuantiles superiores. Sin embargo, para los cuantiles inferiores, aún aparece cierta distancia con la recta que describe la normalidad.

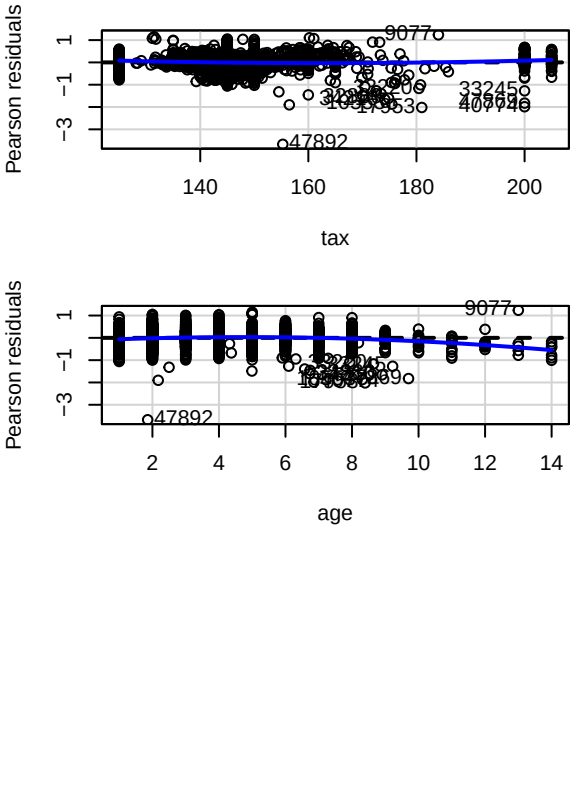
Por último, para el caso de la sobre-influencia en el modelo, podemos ver un resultado similar al del anterior modelo.

Vamos a proceder a realizar un análisis más en profuncidad del modelo.

En los siguientes gráficos podemos ver como se distribuyen los residuos.

```
residualPlots(m2,id=list(method=cooks.distance(m2),n=10))
```





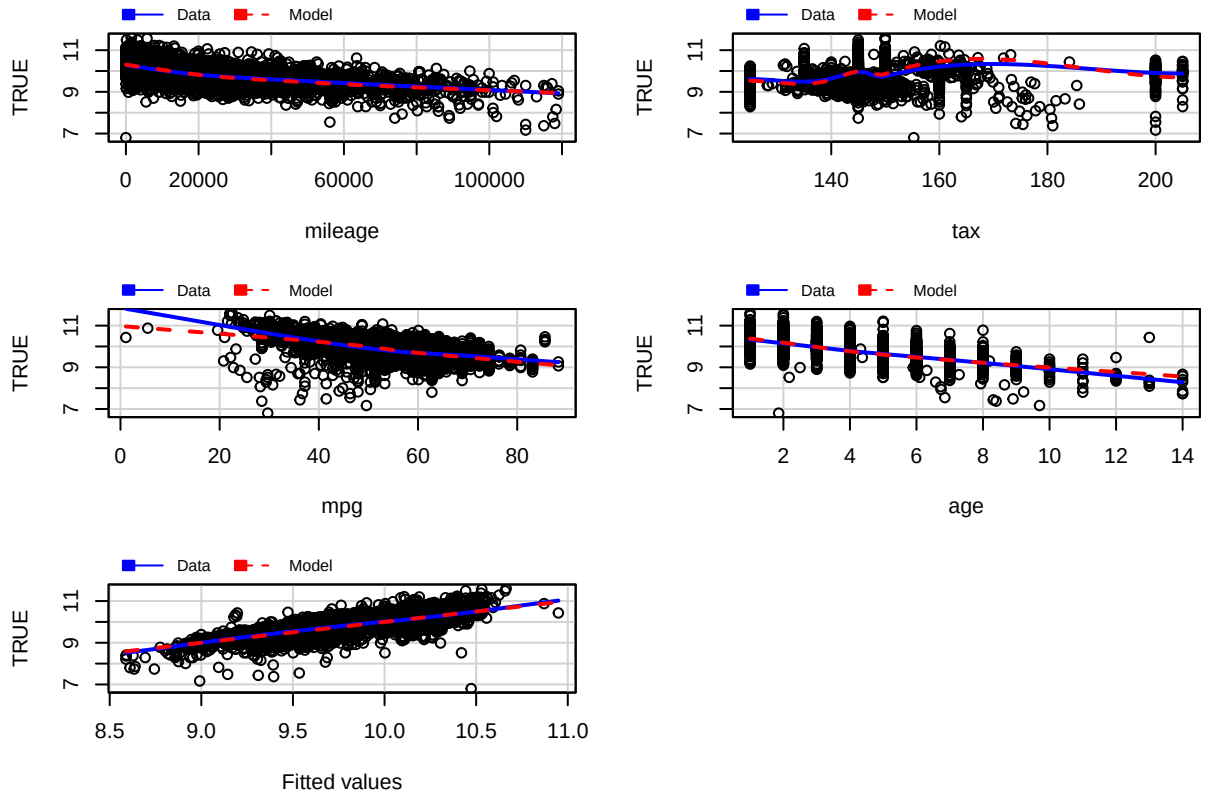

```
           Test stat Pr(>|Test stat|)
mileage      -8.9294        < 2.2e-16 ***
tax           6.7284        1.909e-11 ***
mpg          13.1778        < 2.2e-16 ***
age         -10.5882        < 2.2e-16 ***
Tukey test   -2.9525         0.003152 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Algunos casos a destacar son el de mileage, donde se puede apreciar una cierta acumulación en los valores bajos, a la vez que aparecen puntos que se alejan de la nube y de la curva para valores más altos.

También podemos destacar que en el caso de age, que es una variable que originalmente generamos a partir de la variable year tiene una distribución en columnas. Aparece una nube de puntos cerca del centro, probablemente debido a la imputación a partir de PCA que se realizó en la primera entrega. Podemos ver algo parecido pero no tan evidente para la variable tax.

Con la siguiente función procederemos a ver el ajuste del modelo con los datos.

```
marginalModelPlots(m2)
```



Podemos ver como el ajuste para el caso de la variable mileage o age parecen casi perfectos, mientras que para las varibales mpg y tax existe una cierta desviación entre las curvas roja y azul.

En el caso de los Fitted values, las rectas se ajustan casi a la perfección.

Vamos a proceder a volver a generar el modelo excluyendo los multivariant outliers.

```
m6<-update(m2, data=df[!df$mout=="YesMOut",])
summary(m6)
```

```
Call:
lm(formula = log(price) ~ mileage + tax + mpg + age, data = df[!df$mout ==
    "YesMOut", ])

Residuals:
     Min       1Q   Median       3Q      Max
-2.07052 -0.17091  0.02088  0.19062  1.12919

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.081e+01  6.594e-02 163.875  < 2e-16 ***
mileage     -2.090e-06  4.043e-07  -5.170 2.44e-07 ***
tax          1.574e-03  3.807e-04   4.136 3.60e-05 ***
mpg         -1.353e-02  4.547e-04 -29.761  < 2e-16 ***
age         -1.108e-01  3.986e-03 -27.794  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.2875 on 4754 degrees of freedom
Multiple R-squared:  0.564, Adjusted R-squared:  0.5636
F-statistic:  1537 on 4 and 4754 DF,  p-value: < 2.2e-16
```

En el caso del R-squared, podemos ver como este ha bajado, aunque infimamente, indicando que el modelo anterior aglutinaba más variabilidad de nuestro target.

Vamos a proceder a estudiar la validez de nuestro tercer modelo.

```
Anova(m6)
```

```
Anova Table (Type II tests)

Response: log(price)
          Sum Sq   Df F value    Pr(>F)
mileage     2.21    1  26.725 2.442e-07 ***
tax         1.41    1  17.106 3.597e-05 ***
mpg        73.19    1 885.725 < 2.2e-16 ***
age        63.84    1 772.510 < 2.2e-16 ***
Residuals 392.86 4754
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En este test podemos ver que todas las variables que aparecen son significaticas, de modo que ninguna de estas variables es prescindible.

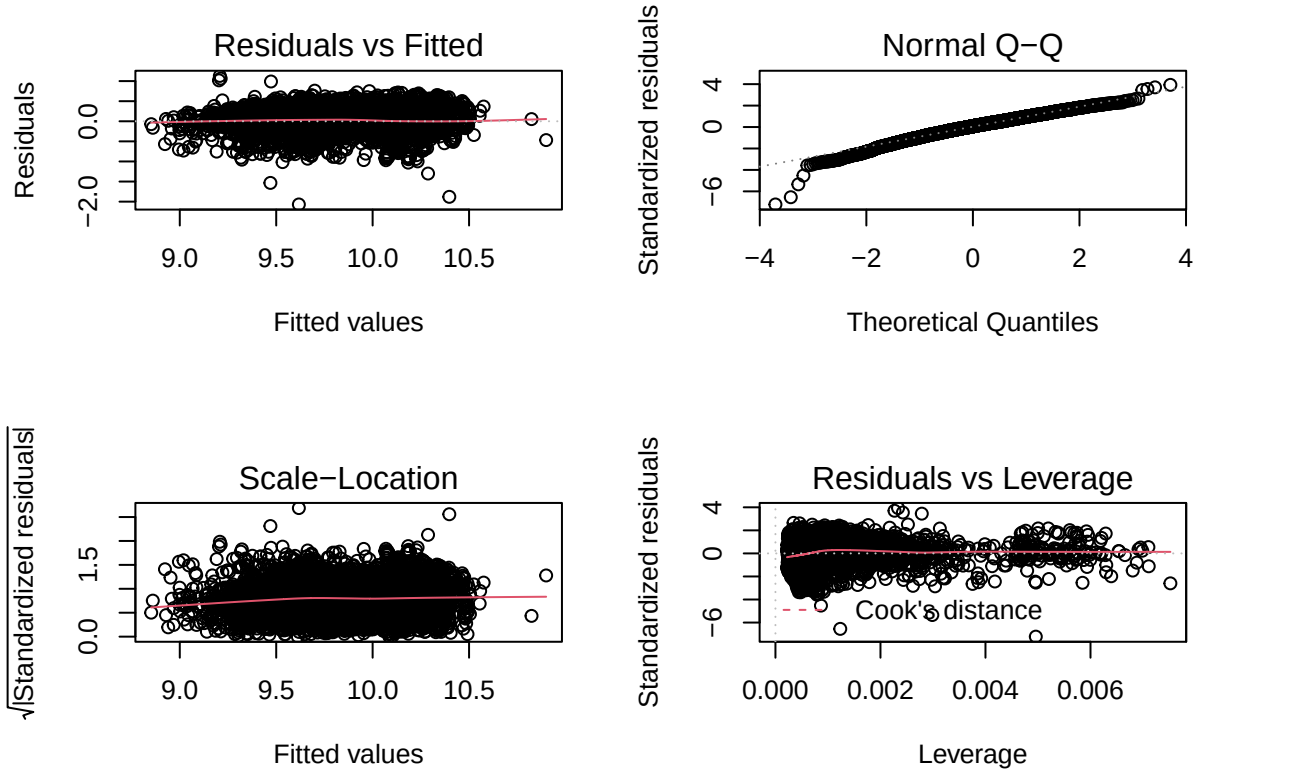
Con la función de VIF podemos ver como, los vifs asociados a mileage y age son mayores que 3, pero de momento no debería importarnos mucho.

```
vif(m6)
```

```
 mileage      tax      mpg      age
3.114409 1.163041 1.386670 3.137709
```

Vamos a analizar los gráficos del modelo:

```
par(mfrow=c(2,2))
plot(m6,id.n=0)
```



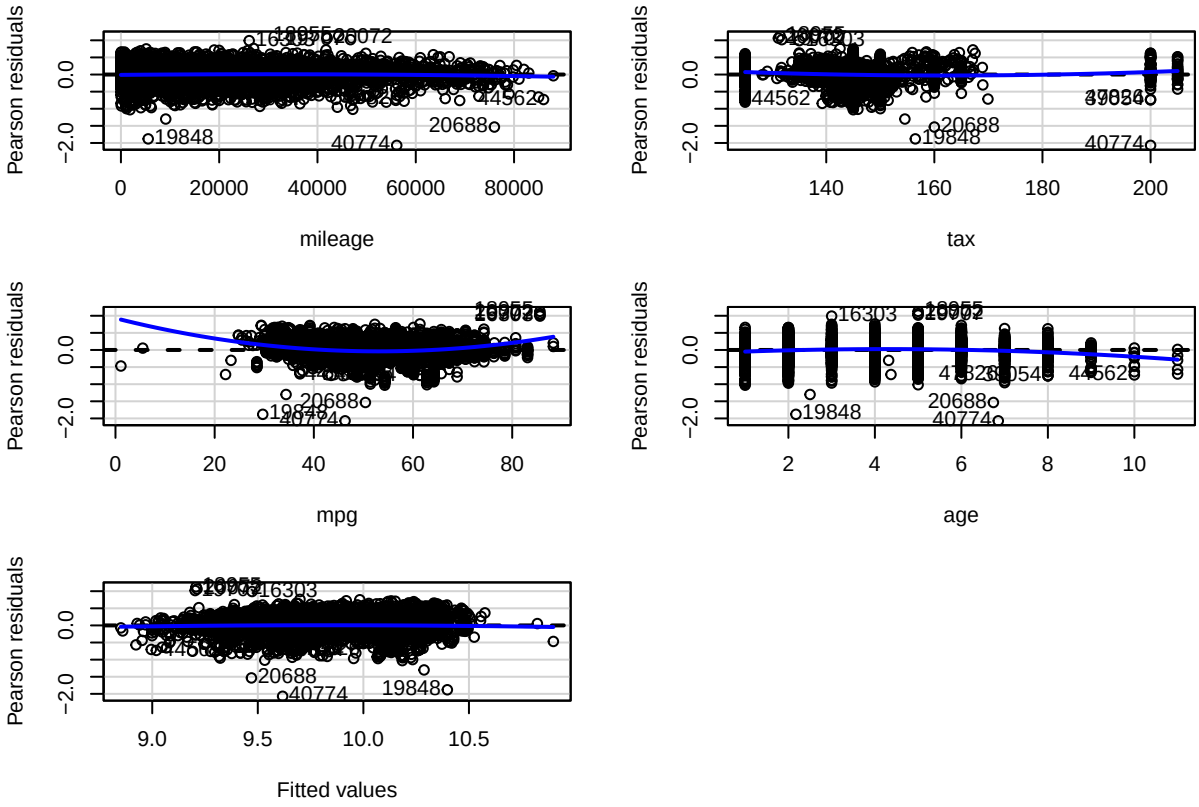
En el de Residuals vs Fitted y Scale-Location podemos ver como el modelo presenta bastante homocedasticidad.

En el de Normal Q-Q, similarmente al m2, podemos apreciar como en los cuantiles superiores aparece normalidad, mietras que para los inferiores, la normalidad de nuestro modelo se aleja de la recta.

Por último, en el gráfico de Residuals vs Leverage, podemos ver como para este modelo siguen sin aparecer puntos con una distancia de Cook relevante, de modo que no parece existir sobre-influencia de ninguna observación.

En los siguientes gráficos podemos apreciar como las rectas se ajustan bastante, excepto en el caso de la variable mpg, a la que, dados los problemas que aparecen con la tendencia a infinito de su exponente en las iteraciones de la función boxTidwell, no podemos determinar la transformación que se le debería aplicar para mejorar el modelo.

```
par(mfrow=c(2,3))
residualPlots(m6,id=list(method=cooks.distance(m6),n=10))
```




```
           Test stat Pr(>|Test stat|)
mileage      -2.0729          0.03824 *
tax           6.5007        8.815e-11 ***
mpg          12.7140        < 2.2e-16 ***
age          -7.0797        1.655e-12 ***
Tukey test   -1.3277          0.18427
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Podemos apreciar que hay tres individuos que aparecen constantemente fuera de las nubes de puntos, los 20688, 19848 y 40774.

Vamos a proceder a eliminar los elementos que aparecían en los plots anteriores constantemente alejados de las nubes de puntos así como los multivariant outliers.

```
df2 <- df[!df$mout=="YesMOut",]
df2 <- df2[row.names(df2)!="19848",]
df2 <- df2[row.names(df2)!="40774",]
df2 <- df2[row.names(df2)!="20688",]
```

Procederemos a replantear el modelo excluyendo las observaciones que hemos comentado previamente.

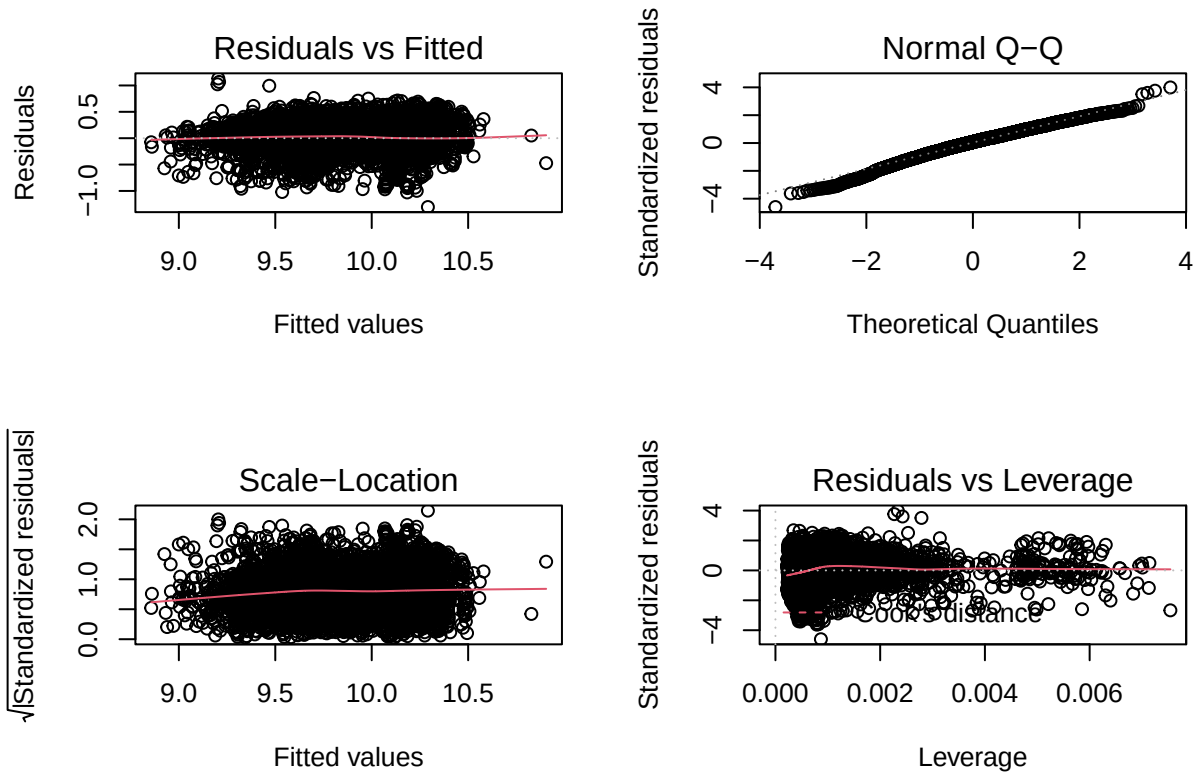
```r
m7<-update(m6,data=df2)
summary(m7)
```

```
Call:
lm(formula = log(price) ~ mileage + tax + mpg + age, data = df2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.30483 -0.17185  0.02041  0.18948  1.13222

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.079e+01  6.518e-02 165.469  < 2e-16 ***
mileage     -1.967e-06  3.995e-07  -4.925 8.74e-07 ***
tax          1.758e-03  3.765e-04   4.669 3.11e-06 ***
mpg         -1.365e-02  4.491e-04 -30.399  < 2e-16 ***
age         -1.111e-01  3.936e-03 -28.232  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2838 on 4751 degrees of freedom
Multiple R-squared:  0.5702,    Adjusted R-squared:  0.5698
F-statistic:  1575 on 4 and 4751 DF,  p-value: < 2.2e-16
```

Se puede apreciar como aumentan la explicabilidad pero no hay grandes cambios en la relevancia de los coeficientes.

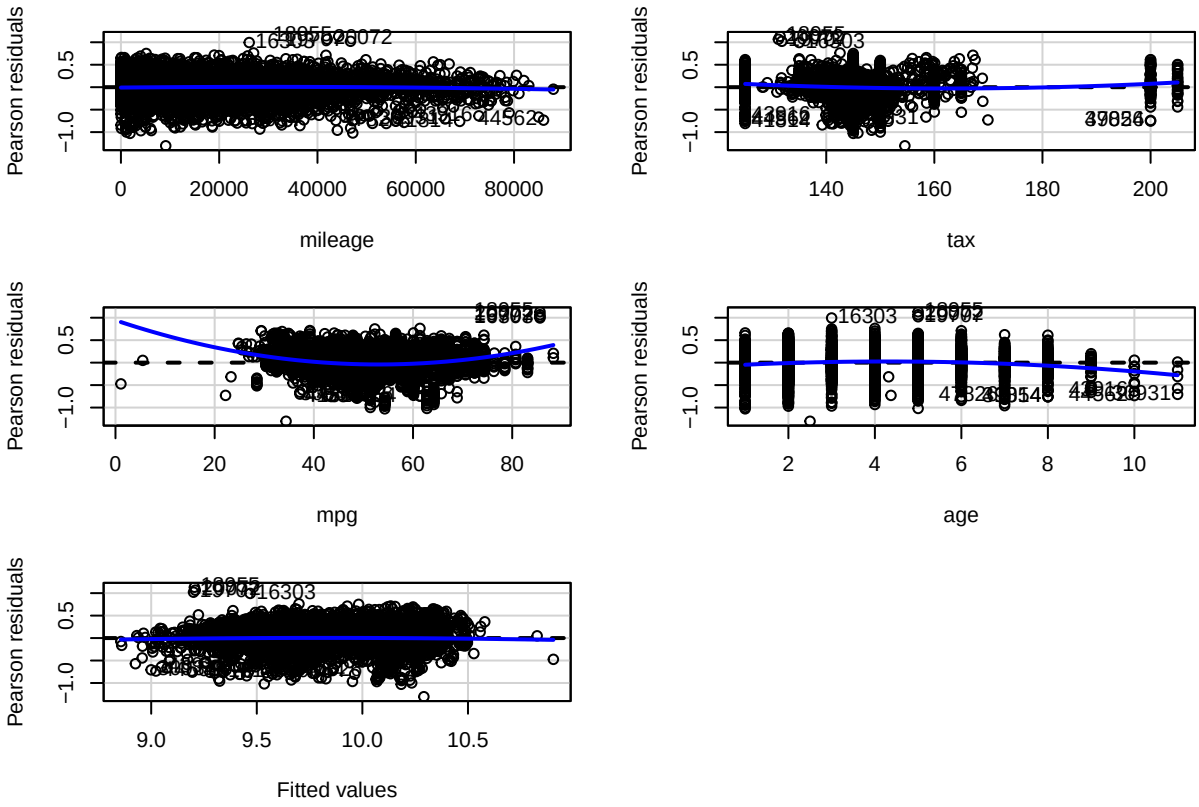
En los siguientes plots podemos ver como este pequeño cambio en el dataset hemos eliminado los indiviuos que constantemente se desviaban de las nubes de puntos. Este hecho se puede ver especialmente en el gráfico Normal Q-Q, donde hay un mejor ajuste a la recta para los cuantiles inferiores.

```r
par(mfrow=c(2,2))
plot(m7,id.n=0)
```



Si nos fijamos en los gráficos de los residuos, podemos ver como siguen existiendo algunos desajustes en las rectas, sobretodo para las variables mpg y age.

```
par(mfrow=c(2,3))
residualPlots(m7,id=list(method=cooks.distance(m7),n=10))
```




```
           Test stat Pr(>|Test stat|)
mileage      -1.6995          0.08928 .
tax           6.6737        2.780e-11 ***
mpg          13.0932        < 2.2e-16 ***
age          -7.1290        1.163e-12 ***
Tukey test   -1.1407          0.25400
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Si realizamos el test de Breusch-Pagan contra la heteroscedasticidad de nuestro modelo, obtenemos un p-valor de 0.0005, de modo que podemos rechazar la H_0 y confirmar que nuestro modelo es homocedástico.

```
library(lmtest)
```

```
Loading required package: zoo
```

```
Attaching package: ’zoo’
```

```
The following objects are masked from ’package:base’:

    as.Date, as.Date.numeric
```

```
bptest(m7)
```

```

    studentized Breusch-Pagan test

data:  m7
BP = 20.13, df = 4, p-value = 0.0004708
```

Vamos a proceder a mostrar los boxplots de los valores R-student, Hat y distancias de Cook de las observaciones del modelo.

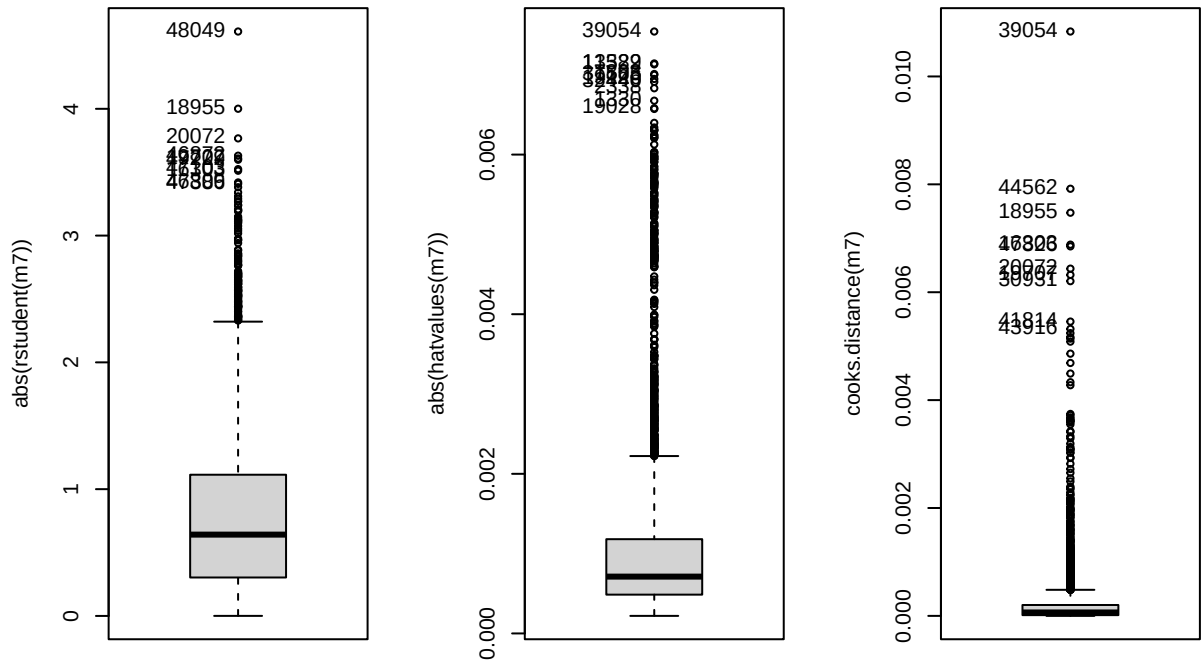
```
par(mfrow=c(1,3))
Boxplot(abs(rstudent(m7)),id=list(labels=row.names(df2)))
```

```
 [1] "48049" "18955" "20072" "46872" "19707" "47229" "47103" "16303" "46809"
[10] "47386"
```

```
Boxplot(abs(hatvalues(m7)),id=list(labels=row.names(df2)))
```

```
 [1] "39054" "11382" "13529" "7803"  "31585" "19126" "32440" "2338"  "1330"
[10] "19028"
```

```
Boxplot(cooks.distance(m7),id=list(labels=row.names(df2)))
```




```
 [1] "39054" "44562" "18955" "16303" "47826" "20072" "19707" "30931" "41814"
[10] "43916"
```

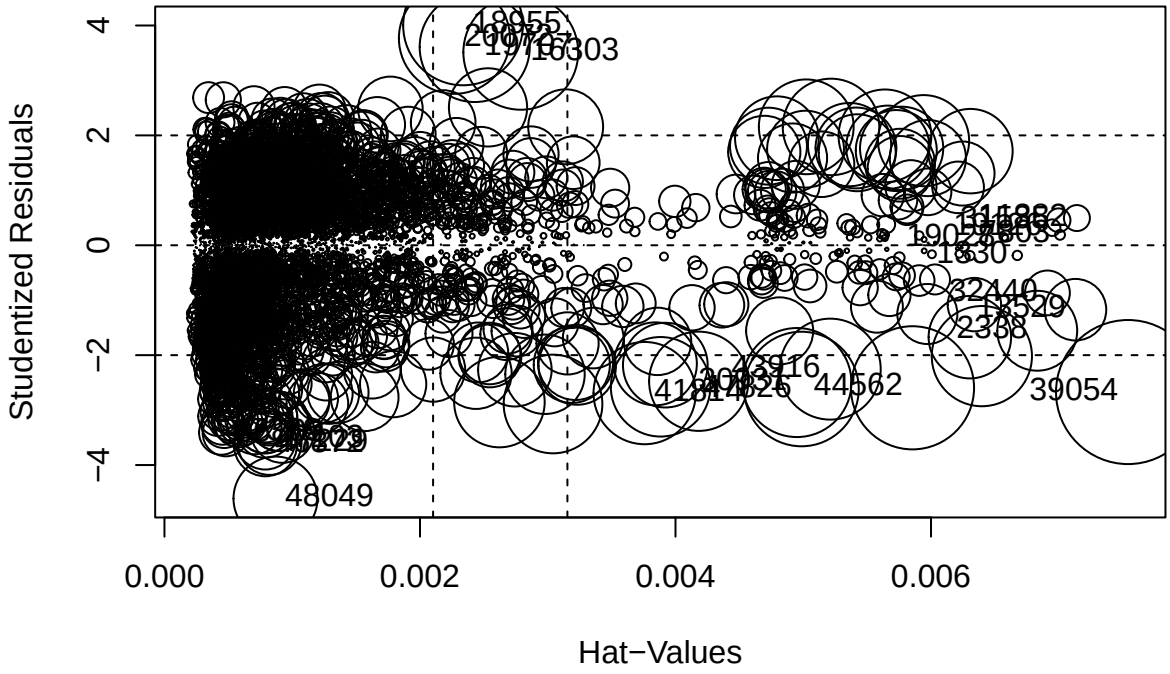
Con estos gráficos podemos detectar los valores para los cuales se rompe la cadena de puntos y podemos categorizar como outliers.

```
stu_out <- which(abs(rstudent(m7))>3.7);
cook_out <- which(abs(cooks.distance(m7))>0.0065);
hat_out <- which(abs(hatvalues(m7))>0.007);

outs<-unique(c(stu_out,cook_out,hat_out));outs
```

```
 [1] 1813 1910 4598 1549 3729 4267 4584  771 1069 1273
```

Si analizamos el gráfico de influencias, podemos ver como no existe una distribución aglomerada, tal vez en los individuos con valores de Hat muy bajos, pero en general existe basante dispersión.

```
par(mfrow=c(1,1));
outs2 <- influencePlot(m7, id=list(n=10));
```




```
outs2 <- labels(outs2)[[1]];
outs2 <- as.numeric(outs2);
outs3 <- unique(c(outs,outs2));outs3
```

```
 [1]  1813  1910  4598  1549  3729  4267  4584   771  1069  1273  1330  2338
[13]  7803 11382 13529 16303 18955 19028 19126 19707 20072 30931 31585 32440
[25] 39054 41814 43916 44562 46809 46872 47103 47229 47386 47826 48049
```

Vamos a terminar este análisis generando un modelo excluyendo los individuos que hemos detectado como outliers.

```
m8 <- update(m7, data=df2[-outs3,])
summary(m8)
```

```
Call:
lm(formula = log(price) ~ mileage + tax + mpg + age, data = df2[-outs3,
    ])

Residuals:
     Min       1Q   Median       3Q      Max
-1.02483 -0.17196  0.02111  0.19091  1.03263

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.078e+01  6.522e-02 165.298  < 2e-16 ***
mileage     -2.074e-06  3.979e-07  -5.211 1.96e-07 ***
tax          1.865e-03  3.774e-04   4.942 8.00e-07 ***
mpg         -1.395e-02  4.473e-04 -31.192  < 2e-16 ***
age         -1.091e-01  3.926e-03 -27.800  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
Residual standard error: 0.2815 on 4739 degrees of freedom
Multiple R-squared:  0.5747,    Adjusted R-squared:  0.5743
F-statistic:  1601 on 4 and 4739 DF,  p-value: < 2.2e-16
```

Vemos que la explicabilidad del modelo final es del 57,47%.

2.2 Factores

Vamos a empezar añadiendo un solo factor al modelo. En este caso empezaremos con los factores que determinamos que eran más influyentes en el análisis MCA de la entrega anterior. Los tres factores más relevantes eran f.price, transmission y fuelType. No añadiremos f.price ya que es un factor generado a partir de nuestra variable target.

De momento empezamos añadiendo fuelType.

```
m10<- update(m8, ~.+fuelType)
summary(m10)
```

```
Call:
lm(formula = log(price) ~ mileage + tax + mpg + age + fuelType,
    data = df2[-outs3, ])

Residuals:
     Min       1Q   Median       3Q      Max
-1.09807 -0.12596  0.00999  0.14957  1.09163

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.145e+01  5.334e-02 214.689  < 2e-16 ***
mileage                -4.522e-06  3.186e-07 -14.192  < 2e-16 ***
tax                     9.943e-04  2.997e-04   3.318 0.000915 ***
mpg                    -2.174e-02  3.841e-04 -56.613  < 2e-16 ***
age                    -8.596e-02  3.140e-03 -27.374  < 2e-16 ***
fuelTypef.Fuel-Petrol -3.886e-01  7.377e-03 -52.677  < 2e-16 ***
fuelTypef.Fuel-Hybrid  4.979e-02  3.129e-02   1.591 0.111586
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.223 on 4737 degrees of freedom
Multiple R-squared:  0.7333,    Adjusted R-squared:  0.733
F-statistic:  2171 on 6 and 4737 DF,  p-value: < 2.2e-16
```

Como podemos ver ya a simple vista, el R-squared del modelo ha aumentado significativamente, indicando mayor explicabilidad.

Si analizamos los vif, vemos que no ha habido cambios significativos respecto al modelo anterior. Los vifs se mantinenen en valores inferiores a 5.

```
vif(m10)
```

```
             GVIF Df GVIF^(1/(2*Df))
mileage  3.188292  1        1.785579
tax      1.163380  1        1.078601
mpg      1.628981  1        1.276315
age      3.205863  1        1.790492
fuelType 1.258266  2        1.059115
```

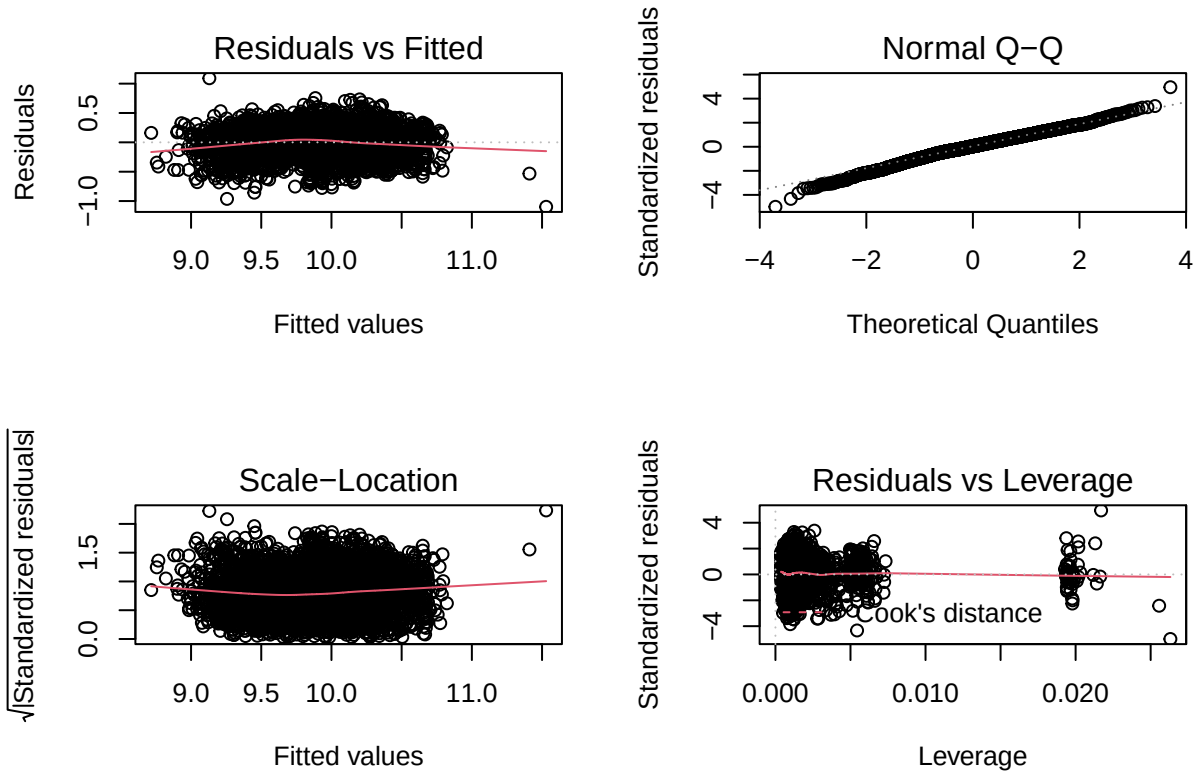
Con el test anova podemos ver como todas las variables mantienen su significatividad.

```
Anova(m10)
```

```
Anova Table (Type II tests)

Response: log(price)
             Sum Sq   Df  F value    Pr(>F)
mileage      10.015    1  201.425 < 2.2e-16 ***
tax           0.547    1   11.006 0.0009149 ***
mpg         159.350    1 3205.010 < 2.2e-16 ***
age          37.256    1  749.319 < 2.2e-16 ***
fuelType    140.111    2 1409.026 < 2.2e-16 ***
Residuals   235.520 4737
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Si analizmos los plots del modelo, podemos ver como el modelo parece haber perdido homocedasticidad y normalidad en los extremos, y la distribución de las distancias de Cook es algo peculiar, generando diversas acumulaciones de puntos.

```
par(mfrow=c(2,2))
plot(m10,id.n=0)
```



Vamos a proceder a incorporar también el factor transmission.

```
m11 <- update(m10, ~.+transmission)
summary(m11)
```

```
Call:
lm(formula = log(price) ~ mileage + tax + mpg + age + fuelType +
    transmission, data = df2[-outs3, ])

Residuals:
     Min       1Q   Median       3Q      Max
-0.86376 -0.12096  0.00155  0.13213  0.96270
```

```
Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                   1.110e+01  4.845e-02 229.162   <2e-16 ***
mileage                      -4.341e-06  2.832e-07 -15.331   <2e-16 ***
tax                           5.112e-04  2.666e-04   1.917   0.0552 .
mpg                          -1.784e-02  3.600e-04 -49.555   <2e-16 ***
age                          -8.072e-02  2.795e-03 -28.877   <2e-16 ***
fuelTypef.Fuel-Petrol        -3.019e-01  7.028e-03 -42.954   <2e-16 ***
fuelTypef.Fuel-Hybrid        -8.895e-04  2.784e-02  -0.032   0.9745
transmissionf.Trans-SemiAuto  2.476e-01  7.283e-03  33.998   <2e-16 ***
transmissionf.Trans-Automatic 2.243e-01  8.181e-03  27.415   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1981 on 4735 degrees of freedom
Multiple R-squared:  0.7896,    Adjusted R-squared:  0.7892
F-statistic:  2221 on 8 and 4735 DF,  p-value: < 2.2e-16
```

Con el aumento del R-squared, podemos apreciar un nuevo aumento en la explicabilidad del modelo.

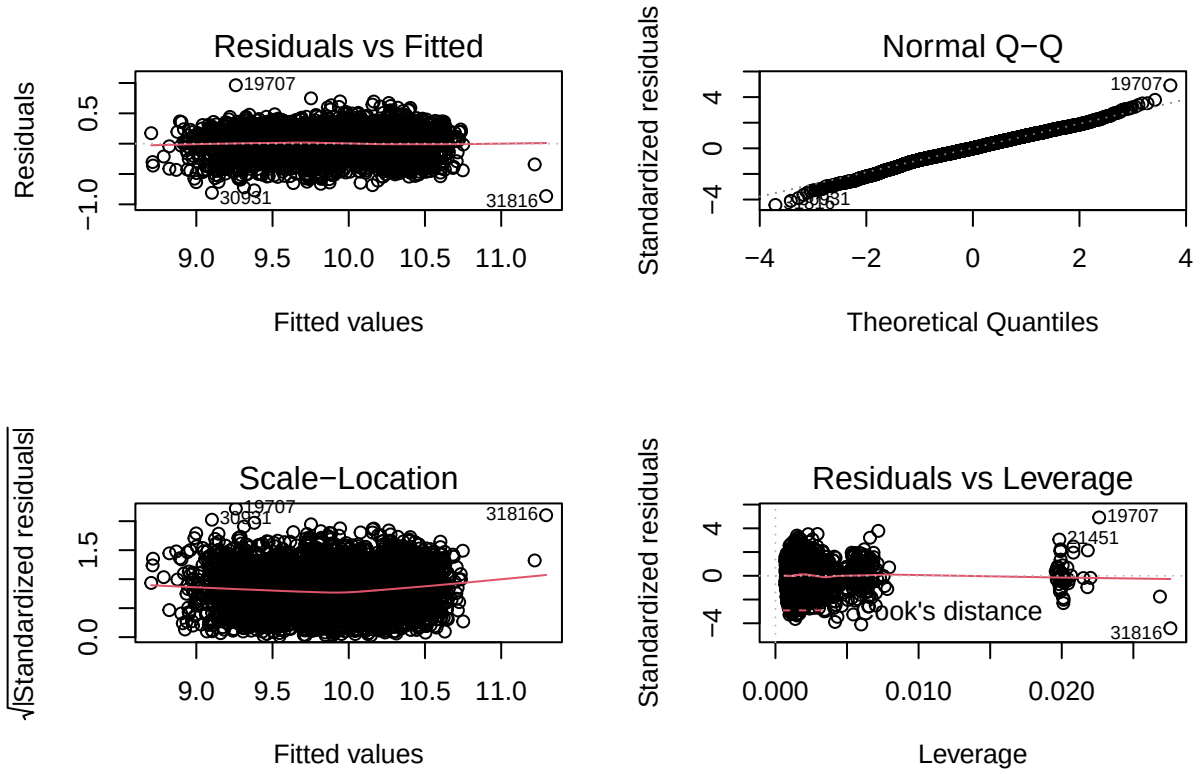
El test de anova nos indica que el nuevo modelo es preferible al anteior.

```
anova(m10,m11)
```

```
Analysis of Variance Table

Model 1: log(price) ~ mileage + tax + mpg + age + fuelType
Model 2: log(price) ~ mileage + tax + mpg + age + fuelType + transmission
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   4737 235.52
2   4735 185.84  2    49.683 632.94 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Si analizamos los plots de este modelo, podemos ver como no aparecen cambios significativos. Tal vez podemos apreciar algo más de normalidad, pero la homocedasticidad y las distancias de Cook no parecen haber cambiado mucho.

```
par(mfrow=c(2,2))
plot(m11)
```



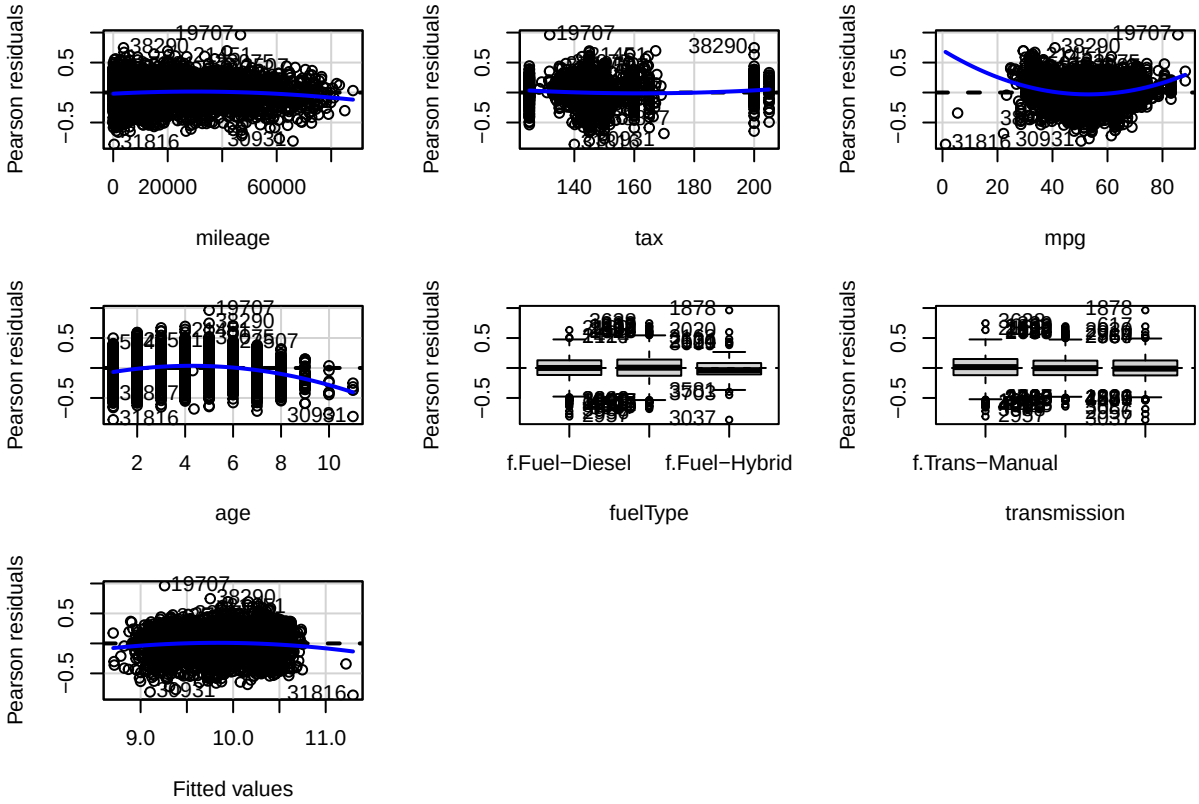
En lo que se refiere a los vifs, no hay cambios significativos, indicando que no aparece co-linealidad en las variables.

```
vif(m11)
```

```
                GVIF Df GVIF^(1/(2*Df))
mileage     3.190266  1        1.786131
tax         1.166405  1        1.080002
mpg         1.812822  1        1.346411
age         3.218176  1        1.793928
fuelType    1.457174  2        1.098697
transmission 1.276301  2        1.062890
```

Si analizamos los plots de los residuos, para las variables numéricas no apreciamos grandes cambios,

```
par(mfrow=c(1,1))
residualPlots(m11,id=list(method=cooks.distance(m11),n=10))
```




```
             Test stat Pr(>|Test stat|)
mileage        -5.6207        2.011e-08 ***
tax             4.4690        8.041e-06 ***
mpg            13.9839        < 2.2e-16 ***
age           -14.9839        < 2.2e-16 ***
fuelType
transmission
Tukey test     -4.3469        1.380e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.3 Interacciones

2.3.1 Interacciones entre factores

Vamos a proceder a incorporar la interacció entre los factores transmission y fuelType.

```
m13 <- update(m11, ~.+transmission*fuelType)
summary(m13)
```

```
Call:
lm(formula = log(price) ~ mileage + tax + mpg + age + fuelType +
    transmission + fuelType:transmission, data = df2[-outs3,
    ])

Residuals:
    Min      1Q  Median      3Q     Max
-0.8810 -0.1188  0.0013  0.1344  0.9343

Coefficients: (1 not defined because of singularities)
                                                    Estimate Std. Error
(Intercept)                                        1.111e+01  4.820e-02
mileage                                           -4.458e-06  2.821e-07
tax                                                5.831e-04  2.653e-04
```

```
mpg                                                 -1.765e-02  3.590e-04
age                                                 -8.041e-02  2.780e-03
fuelTypef.Fuel-Petrol                               -3.582e-01  1.023e-02
fuelTypef.Fuel-Hybrid                                3.858e-02  4.082e-02
transmissionf.Trans-SemiAuto                         2.051e-01  9.606e-03
transmissionf.Trans-Automatic                        1.799e-01  1.030e-02
fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto   8.813e-02  1.360e-02
fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto  -4.905e-02  5.556e-02
fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic  1.003e-01  1.582e-02
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic         NA         NA
                                                    t value Pr(>|t|)
(Intercept)                                         230.587  < 2e-16 ***
mileage                                             -15.806  < 2e-16 ***
tax                                                   2.198    0.028 *
mpg                                                 -49.156  < 2e-16 ***
age                                                 -28.925  < 2e-16 ***
fuelTypef.Fuel-Petrol                               -35.015  < 2e-16 ***
fuelTypef.Fuel-Hybrid                                 0.945    0.345
transmissionf.Trans-SemiAuto                         21.352  < 2e-16 ***
transmissionf.Trans-Automatic                        17.458  < 2e-16 ***
fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto    6.483 9.92e-11 ***
fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto   -0.883    0.377
fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic   6.343 2.47e-10 ***
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.197 on 4732 degrees of freedom
Multiple R-squared:  0.7921,    Adjusted R-squared:  0.7916
F-statistic:  1639 on 11 and 4732 DF,  p-value: < 2.2e-16
```

Podemos ver como el R-squared apenas aumenta y se mantiene la explicabilidad de las variables.

Si realizamos el test de anova, podemos ver como, a pensar de que a partir del summary los modelos parecen similares, no lo son, y en realidad el nuevo modelo es mejor que el anterior.

```
anova(m11,m13)
```

```
Analysis of Variance Table

Model 1: log(price) ~ mileage + tax + mpg + age + fuelType + transmission
Model 2: log(price) ~ mileage + tax + mpg + age + fuelType + transmission +
    fuelType:transmission
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   4735 185.84
2   4732 183.59  3    2.2501 19.332 1.866e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.3.2 Interacciones Factor-Numéricas

Vamos a probar a añadir las interacciones de la variable numérica age y el factor transmission. Podemos ver como los dos modelos no son equivalentes, y según el test de anova, el nuevo modelo es más completo.

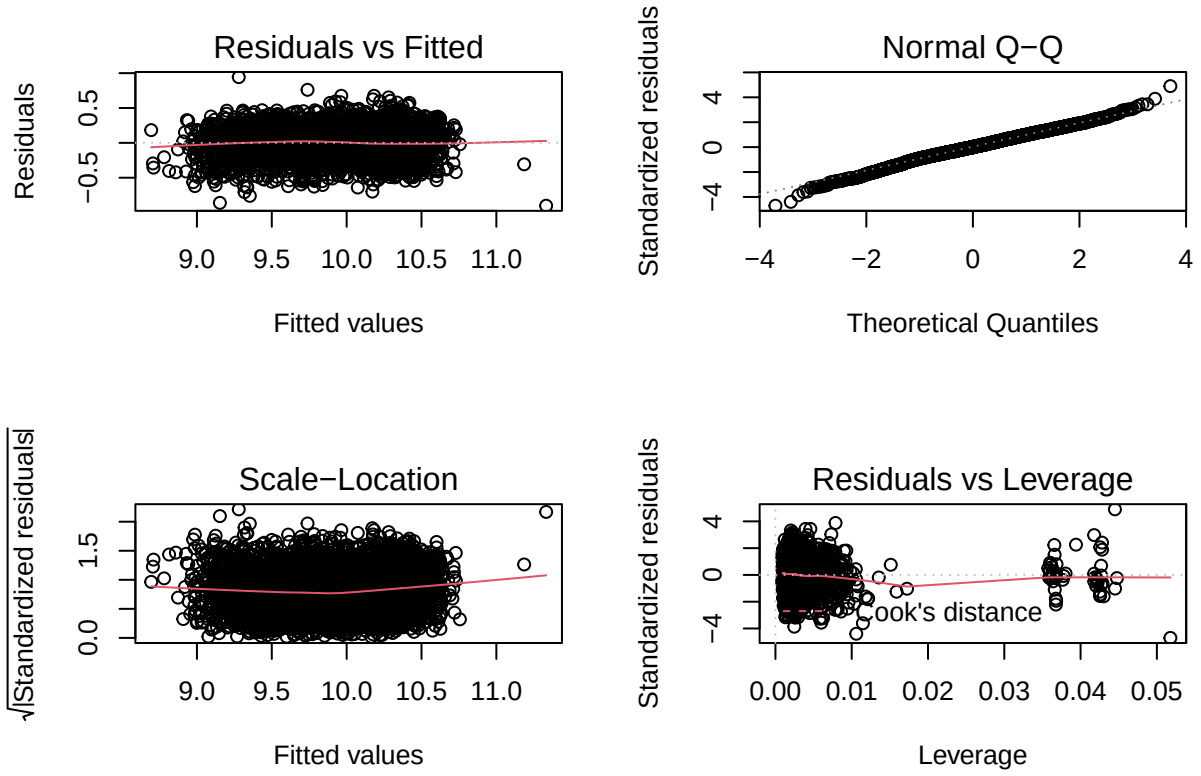
```
m14 <- update(m13, ~.+age*transmission)
anova(m13,m14)
```

```
Analysis of Variance Table

Model 1: log(price) ~ mileage + tax + mpg + age + fuelType + transmission +
    fuelType:transmission
Model 2: log(price) ~ mileage + tax + mpg + age + fuelType + transmission +
```

```
    fuelType:transmission + age:transmission
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1   4732 183.59
2   4730 183.27  2   0.31699 4.0906 0.01679 *
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Si analizamos los plots del modelo, podemos ver como existe homocedasticidad y aparece bastante normalidad.

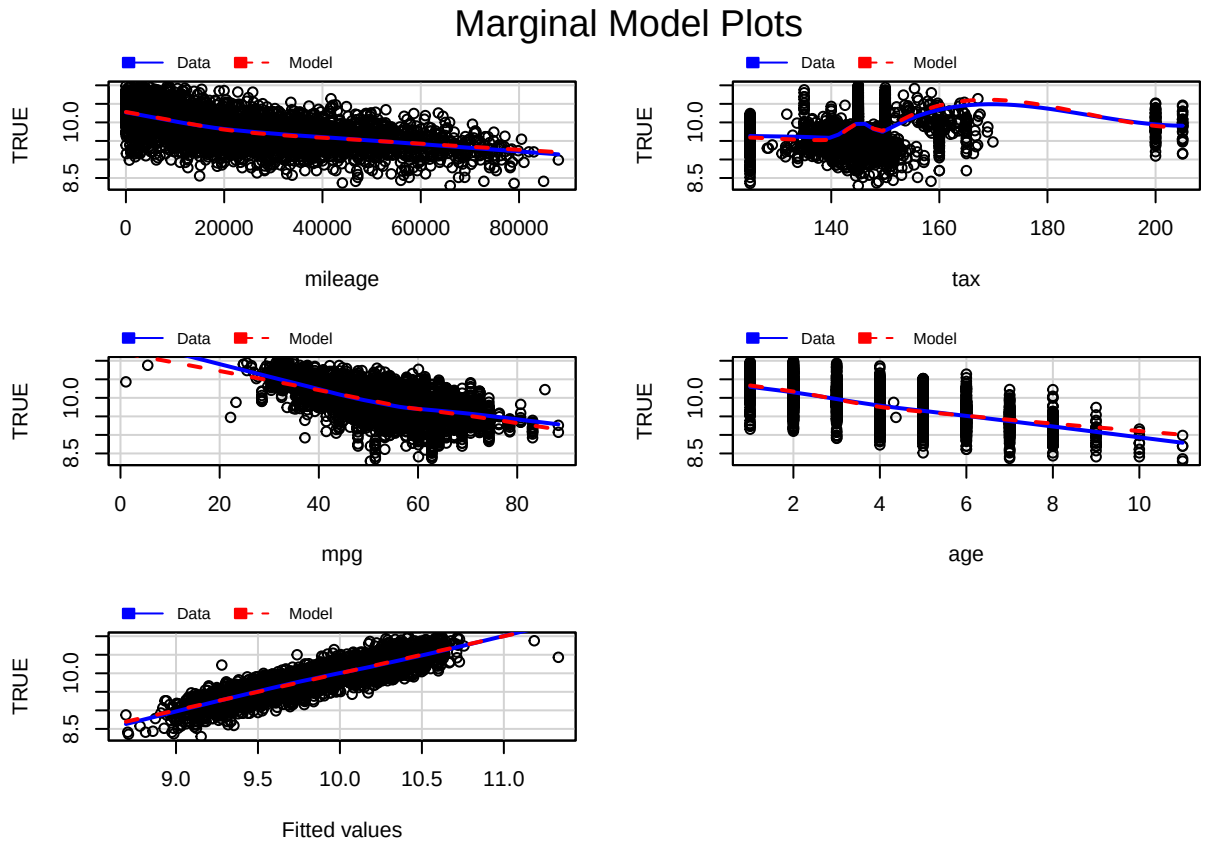
```
par(mfrow=c(2,2))
plot(m14, id.n=0)
```



Si analizamos los plots de los residuos, podemos ver como el ajuste de las regresiones entre el modelo y los datos es muy preciso.

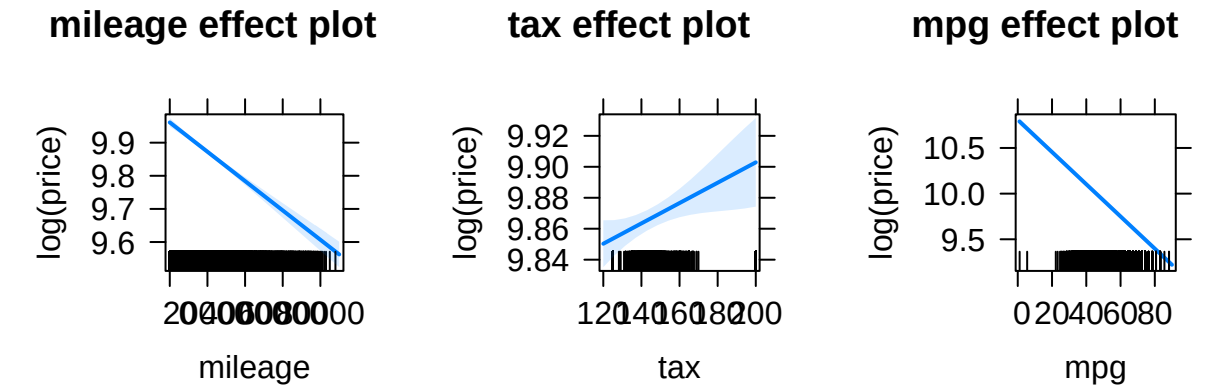
```
marginalModelPlots(m14)
```

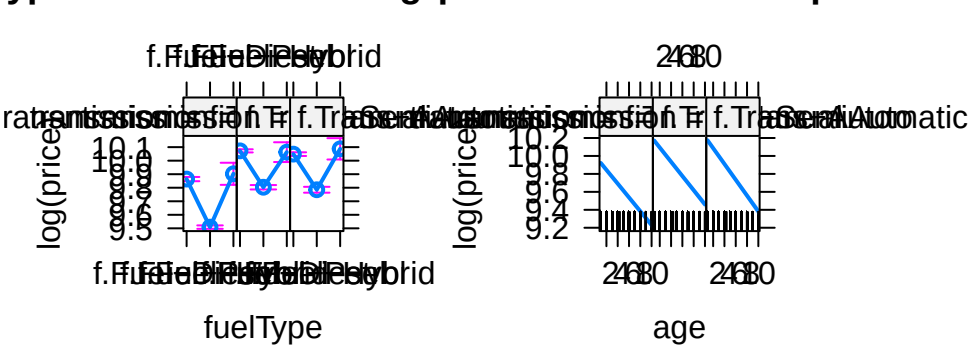
```
Warning in mmps(...): Interactions and/or factors skipped
```



Podemos ver claramente fuerte relación entre la variable mpg y nuestro target numérico.

```
library(effects);
plot(allEffects(m14))
```





Además, si nos fijamos en la interaccions de age y transmission, podemos ver como para todos los tipos de transmisión se respeta la relación inversa con el precio: cuantos más años tiene el coche, más barato es, y viceversa. Por el contrario, en el caso de fuelType, podemos ver como los coches de gasolina son más baratos que los diesel o híbridos.

Si hacemos una rápida búsqueda en internet (https://www.motor.mapfre.es/consejos-practicos/consejos-para-ahorrar/diesel-o-gasolina/) podemos ver como esto este hecho cuadra con la realidad.

El problema aparece cuando intentamos aplicar la función vif para comprobar la no co-linealidad de las variables del modelo, ya que salta un error que nos indica que si que hay variables co-lineales. Sin embargo, si realizamos un análisis de las varianzas, todas las variables parecen significativas. Las que menos significación adquieren son tax y la interacción entre age y transmission.

```
#vif(m14)
Anova(m14);
```

```
Note: model has aliased coefficients
      sums of squares computed by model comparison

Anova Table (Type II tests)

Response: log(price)
                       Sum Sq   Df   F value    Pr(>F)
mileage                 9.550    1  246.4666 < 2.2e-16 ***
tax                     0.233    1    6.0218   0.01417 *
mpg                    93.525    1 2413.7764 < 2.2e-16 ***
age                    32.459    1  837.7290 < 2.2e-16 ***
fuelType               72.422    2  934.5719 < 2.2e-16 ***
transmission           49.683    2  641.1261 < 2.2e-16 ***
fuelType:transmission   1.945    3   16.7327 8.232e-11 ***
age:transmission        0.317    2    4.0906   0.01679 *
Residuals             183.270 4730
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vamos a proceder a aplicar la función step para tratar de eliminar esta co-linealidad.

```
m15 <- step( m14, k=log(nrow(df2)))
```

```
Start:  AIC=-15316.89
log(price) ~ mileage + tax + mpg + age + fuelType + transmission +
    fuelType:transmission + age:transmission

                       Df Sum of Sq    RSS    AIC
- age:transmission      2     0.317 183.59 -15326
- tax                   1     0.233 183.50 -15319
<none>                              183.27 -15317
- fuelType:transmission 3     1.945 185.22 -15292
- mileage               1     9.550 192.82 -15084
- mpg                   1    93.525 276.80 -13369

Step:  AIC=-15325.63
log(price) ~ mileage + tax + mpg + age + fuelType + transmission +
    fuelType:transmission

                       Df Sum of Sq    RSS    AIC
- tax                   1     0.187 183.77 -15329
<none>                              183.59 -15326
- fuelType:transmission 3     2.250 185.84 -15293
- mileage               1     9.693 193.28 -15090
- age                   1    32.459 216.05 -14562
- mpg                   1    93.745 277.33 -13377

Step:  AIC=-15329.26
log(price) ~ mileage + mpg + age + fuelType + transmission +
    fuelType:transmission

                       Df Sum of Sq    RSS    AIC
```

```
<none>                              183.77 -15329
- fuelType:transmission  3     2.207 185.98 -15298
- mileage                1     9.533 193.31 -15098
- age                    1    32.282 216.06 -14570
- mpg                    1   107.017 290.79 -13161
```

```
#vif(m15)
```

Si realizamos el test de anova, podemos ver como el nuevo modelo no es equivalente al anterior, y de hecho, sigue apareciendo co-linealidad en las variables, de modo que el m14 es preferible.

```
anova(m15,m14)
```

```
Analysis of Variance Table

Model 1: log(price) ~ mileage + mpg + age + fuelType + transmission +
    fuelType:transmission
Model 2: log(price) ~ mileage + tax + mpg + age + fuelType + transmission +
    fuelType:transmission + age:transmission
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1   4733 183.77
2   4730 183.27  3   0.50443 4.3396 0.004631 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Si aplicamos la función alias, podemos ver que esta colinealidad parece generada por la aparición de las variables transmission y fuelType. Supongo que es debido a que los coches híbridos suelen tener transmisión automática, de modo que el conjunto Hybrid:SemiAuto queda vacío, mientras que hybrid y hybrid&Automatic son el mismo conjunto.

```
alias(m14)
```

```
Model :
log(price) ~ mileage + tax + mpg + age + fuelType + transmission +
    fuelType:transmission + age:transmission

Complete :
                                                    (Intercept) mileage tax mpg
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic  0           0       0   0
                                                    age fuelTypef.Fuel-Petrol
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic  0   0
                                                    fuelTypef.Fuel-Hybrid
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic  1
                                                    transmissionf.Trans-SemiAuto
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic  0
                                                    transmissionf.Trans-Automatic
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic  0
                                                    fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic  0
                                                    fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic -1
                                                    fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic  0
                                                    age:transmissionf.Trans-SemiAuto
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic  0
                                                    age:transmissionf.Trans-Automatic
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic  0
```

```
summary(m14)
```

```
Call:
```

```
lm(formula = log(price) ~ mileage + tax + mpg + age + fuelType +
    transmission + fuelType:transmission + age:transmission,
    data = df2[-outs3, ])

Residuals:
    Min      1Q  Median      3Q     Max
-0.9014 -0.1188 -0.0002  0.1338  0.9426

Coefficients: (1 not defined because of singularities)
                                                    Estimate Std. Error
(Intercept)                                         1.108e+01  5.046e-02
mileage                                            -4.430e-06  2.822e-07
tax                                                 6.569e-04  2.677e-04
mpg                                                -1.763e-02  3.589e-04
age                                                -7.685e-02  3.432e-03
fuelTypef.Fuel-Petrol                              -3.550e-01  1.035e-02
fuelTypef.Fuel-Hybrid                               3.912e-02  4.080e-02
transmissionf.Trans-SemiAuto                        2.173e-01  1.809e-02
transmissionf.Trans-Automatic                       2.260e-01  1.978e-02
fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto  8.600e-02  1.385e-02
fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto -4.860e-02  5.554e-02
fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic 9.166e-02  1.612e-02
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic         NA         NA
age:transmissionf.Trans-SemiAuto                   -2.325e-03  3.816e-03
age:transmissionf.Trans-Automatic                  -1.127e-02  4.069e-03
                                                   t value Pr(>|t|)
(Intercept)                                        219.665  < 2e-16 ***
mileage                                            -15.699  < 2e-16 ***
tax                                                  2.454  0.01417 *
mpg                                                -49.130  < 2e-16 ***
age                                                -22.396  < 2e-16 ***
fuelTypef.Fuel-Petrol                              -34.288  < 2e-16 ***
fuelTypef.Fuel-Hybrid                                0.959  0.33771
transmissionf.Trans-SemiAuto                        12.011  < 2e-16 ***
transmissionf.Trans-Automatic                       11.426  < 2e-16 ***
fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto   6.208 5.81e-10 ***
fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto  -0.875  0.38156
fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic  5.687 1.37e-08 ***
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic     NA       NA
age:transmissionf.Trans-SemiAuto                    -0.609  0.54243
age:transmissionf.Trans-Automatic                   -2.769  0.00564 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1968 on 4730 degrees of freedom
Multiple R-squared:  0.7925,    Adjusted R-squared:  0.7919
F-statistic:  1389 on 13 and 4730 DF,  p-value: < 2.2e-16
```

Finalmente, si echamos un último vistazo al modelo, podemos ver que hemos conseguido aglomerar una explicabilidad del 80%.

3 Modelo de regresión Binaria

Para plantear el modelo de regresión binaria, vamos a proceder primero a separar nuestro dataframe en dos subconjuntos de entrenamiento y de validación.

```
llwork <- sample(1:nrow(df2),round(0.80*nrow(df2),0))

df_train <- df2[llwork,]
df_test <-df2[-llwork,]
```

3.1 Variables numéricas

Vamos a empezar el proceso de generación del modelo de regresión binaria incorporando las variables numéricas y planteando un modelo incial.

```
m20<-glm(Audi~mileage+tax+mpg+age,family="binomial",data=df_train)
summary(m20)
```

```
Call:
glm(formula = Audi ~ mileage + tax + mpg + age, family = "binomial",
    data = df_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4107  -0.7182  -0.6163  -0.4782   2.1847

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.429e-01  5.811e-01   0.934   0.3502
mileage      7.221e-06  3.866e-06   1.867   0.0618 .
tax         -2.663e-04  3.288e-03  -0.081   0.9355
mpg         -4.310e-02  4.523e-03  -9.529   <2e-16 ***
age          7.272e-02  3.806e-02   1.911   0.0560 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3919.0  on 3804  degrees of freedom
Residual deviance: 3807.9  on 3800  degrees of freedom
AIC: 3817.9

Number of Fisher Scoring iterations: 4
```

Como podemos ver, la variable más significativa de nuestro modelo incial parece ser mpg, mientras que mileage o tax parecen no ser significativas. También podemos ver como el valor AIC es 3810.1, nuestro objetivo será reducirlo.

Si miramos los valores vif de nuestro modelo, podemos ver como no reflejan co-linealidad entre las variables.

```
vif(m20)
```

```
 mileage      tax      mpg      age
3.264049 1.180556 1.394644 3.323606
```

Vamos a plantear un nuevo modelo que incluya solo las variables mpg y age, que son las que aparecían como significativas en el modelo anterior. Si realizamos el test anova de los dos modelos, podemos ver como los modelos parecen equivalentes, de modo que será preferible trabajar con el más sencillo.

```
m21 <- glm(Audi~ mpg + age,family="binomial",data=df_train)
anova(m21,m20,test = "LR");
```

```
Analysis of Deviance Table

Model 1: Audi ~ mpg + age
Model 2: Audi ~ mileage + tax + mpg + age
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      3802     3811.4
2      3800     3807.9  2   3.4707   0.1763
```

Probaremos a añadir los polinomios de segundo grado de las variables del modelo que acabamos de generar. El test de anova nos vuelve a indicar que los modelos son equivalentes, de modo que continuaremos trabajando con el más simple.

```r
m22 <- glm(Audi~ poly(mpg,2) + poly(age,2), family="binomial", data=df_train)
anova(m21,m22, test = "LR")
```

```
Analysis of Deviance Table

Model 1: Audi ~ mpg + age
Model 2: Audi ~ poly(mpg, 2) + poly(age, 2)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      3802     3811.4
2      3800     3800.2  2   11.237 0.003631 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finalmente, este es el modelo con el que continuaremos trabajando, ya que es el más sencillo y equivalente a otros modelos más complejos que hemos planteado.

```r
summary(m21)
```

```
Call:
glm(formula = Audi ~ mpg + age, family = "binomial", data = df_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3890  -0.7168  -0.6215  -0.4838   2.1732

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.402981   0.200162   2.013   0.0441 *
mpg         -0.042125   0.004195 -10.043  < 2e-16 ***
age          0.128142   0.022828   5.613 1.98e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3919.0  on 3804  degrees of freedom
Residual deviance: 3811.4  on 3802  degrees of freedom
AIC: 3817.4

Number of Fisher Scoring iterations: 4
```

Podemos ver como el AIC de este modelo es 3807.8.

3.2 Factores

Vamos a proceder a añadir los factores que resultaron más significativos en el análisis MCA de la segunda entrega.

Como el factor engineSize tiene muchos niveles, vamos a proceder a transformala para que adquiera solo 3 niveles y poder simplificar el modelo.

```r
df_train$engineSize_f <- as.integer(df_train$engineSize)
par(mfrow=c(1,1))
hist(df_train$engineSize_f)
```


```
quantile(df_train$engineSize_f, c(0.3333333,0.6666666,1))
```

```
33.33333% 66.66666%      100%
        6         9        24
```

```
df_train$engineSize_f <- factor(cut(df_train$engineSize_f, breaks = c(0,8,9,20)))
df_test$engineSize_f <- as.integer(df_test$engineSize)
df_test$engineSize_f <- factor(cut(df_test$engineSize_f, breaks = c(0,8,9,20)))
table(df_train$engineSize_f)
```

```
 (0,8]  (8,9] (9,20]
  1429   1644    727
```

Añadimos los factores fuelType y transmission al modelo. Podemos ver como añadiendo estos dos factores hemos conseguido una reducción del AIC. Además, con la función vif podemos ver que no hay aparente co-linealidad entre las variables explicativas del modelo.

```
m24 <- update(m21, ~.+fuelType+transmission)
summary(m24)
```

```
Call:
glm(formula = Audi ~ mpg + age + fuelType + transmission, family = "binomial",
    data = df_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3678  -0.7304  -0.6087  -0.4338   2.4437

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                 1.0401445  0.2948654   3.528 0.000419 ***
mpg                        -0.0479344  0.0049230  -9.737  < 2e-16 ***
age                         0.1187835  0.0230820   5.146 2.66e-07 ***
```

```
fuelTypef.Fuel-Petrol          0.0009893  0.0942448   0.010 0.991625
fuelTypef.Fuel-Hybrid         -1.1313091  0.6236571  -1.814 0.069679 .
transmissionf.Trans-SemiAuto  -0.4825999  0.1020516  -4.729 2.26e-06 ***
transmissionf.Trans-Automatic -0.4445760  0.1152824  -3.856 0.000115 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3919.0  on 3804  degrees of freedom
Residual deviance: 3777.1  on 3798  degrees of freedom
AIC: 3791.1

Number of Fisher Scoring iterations: 5
```

```
vif(m24)
```

```
                 GVIF Df GVIF^(1/(2*Df))
mpg          1.627325  1        1.275666
age          1.219371  1        1.104251
fuelType     1.351054  2        1.078123
transmission 1.295185  2        1.066800
```

Si analizamos la varianza del modelo, podemos ver como todas las variables de nuestro modelo son sifnificativas.

```
Anova(m24, test="LR")
```

```
Analysis of Deviance Table (Type II tests)

Response: Audi
             LR Chisq Df Pr(>Chisq)
mpg            97.507  1  < 2.2e-16 ***
age            26.092  1  3.255e-07 ***
fuelType        4.426  2     0.1094
transmission   24.860  2  3.997e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Con el test anova podemos ver como los modelos no son equivalentes, de modo que nos quedaremos con el que incluye los factores, ya que presenta un AIC inferior.

```
anova(m21,m24, test="LR")
```

```
Analysis of Deviance Table

Model 1: Audi ~ mpg + age
Model 2: Audi ~ mpg + age + fuelType + transmission
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      3802     3811.4
2      3798     3777.1  4   34.309 6.439e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ahora añadiremos el factor derivado de engineSize que hemos generado antes. Podemos ver que este nuevo modelo vuelve a presentar un AIC inferior al del anterior. Además, si usamos la función vif podemos ver como no existe co-linealidad entre las variables.

```
m25 <- update(m24, ~.+engineSize_f)
summary(m25)
```

```
Call:
glm(formula = Audi ~ mpg + age + fuelType + transmission + engineSize_f,
    family = "binomial", data = df_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5299  -0.7335  -0.5881  -0.3415   2.5736

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   2.181371   0.361888   6.028 1.66e-09 ***
mpg                          -0.067118   0.005809 -11.554  < 2e-16 ***
age                           0.176770   0.024168   7.314 2.59e-13 ***
fuelTypef.Fuel-Petrol        -0.379701   0.119036  -3.190  0.00142 **
fuelTypef.Fuel-Hybrid        -1.672927   0.665290  -2.515  0.01192 *
transmissionf.Trans-SemiAuto -0.284352   0.107201  -2.653  0.00799 **
transmissionf.Trans-Automatic -0.223225   0.120091  -1.859  0.06306 .
engineSize_f(8,9]            -0.204113   0.117238  -1.741  0.08168 .
engineSize_f(9,20]           -1.383476   0.178119  -7.767 8.03e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3914.0  on 3799  degrees of freedom
Residual deviance: 3682.3  on 3791  degrees of freedom
  (5 observations deleted due to missingness)
AIC: 3700.3

Number of Fisher Scoring iterations: 5
```

```
vif(m25)
```

```
                GVIF Df GVIF^(1/(2*Df))
mpg          2.105830  1        1.451148
age          1.344941  1        1.159716
fuelType     2.110583  2        1.205315
transmission 1.412056  2        1.090092
engineSize_f 2.207332  2        1.218897
```

Si observamos el análisis de la varianza, vemos que todos los componentes de nuestro modelo son representativos. Por otro lado, no podemos usar las funciones anova ni AIC para comparar ambos modelos, ya que al incluir el factor EngineSize, se han generado algunos missings en los datos que imposibilitan la comparación.

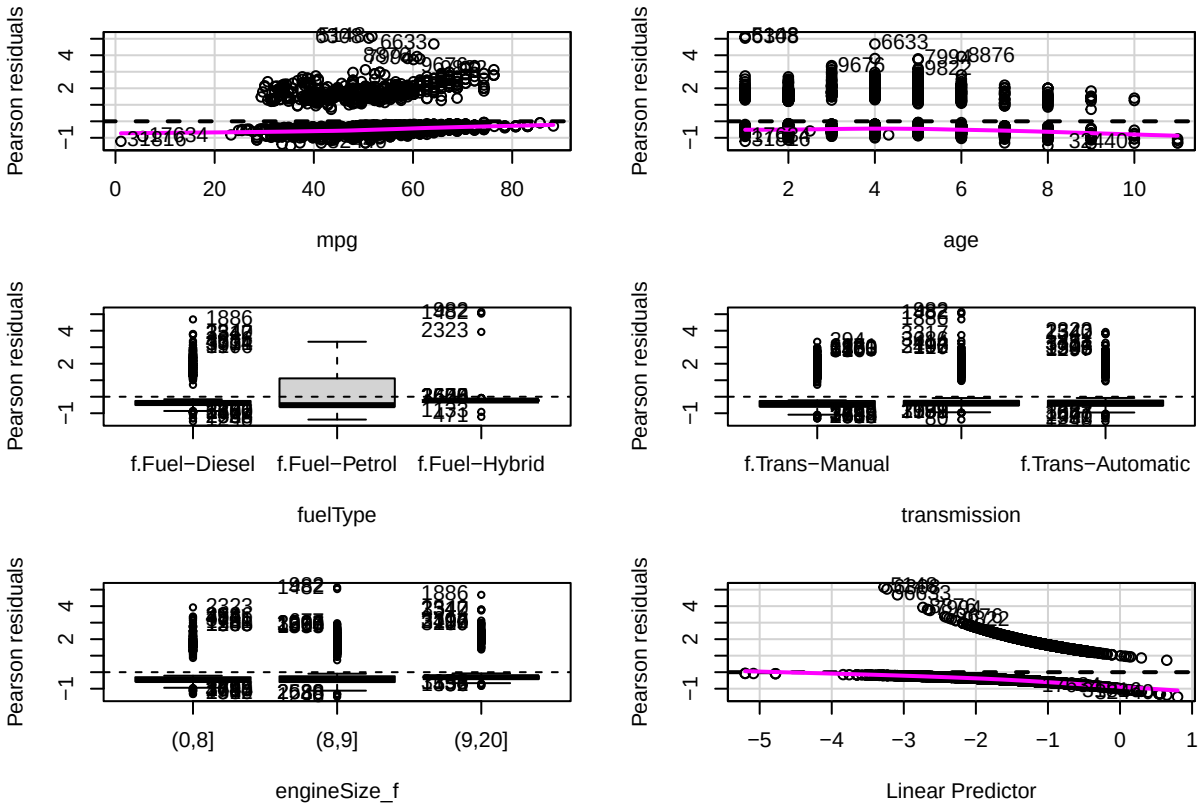
```
Anova(m25, test="LR")
```

```
Analysis of Deviance Table (Type II tests)

Response: Audi
             LR Chisq Df Pr(>Chisq)
mpg           141.100  1  < 2.2e-16 ***
age            53.013  1  3.313e-13 ***
fuelType       17.060  2  0.0001974 ***
transmission    7.267  2  0.0264195 *
engineSize_f   87.902  2  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#anova(m24,m25, test="LR")    #No se pueden hacer porque han aparecido missings en el dataset cuando he
#AIC(m24,m25)
```

Sin embargo, seguiremos avanzando con el nuevo modelo que hemos generado ya que tiene un AIC inferior.

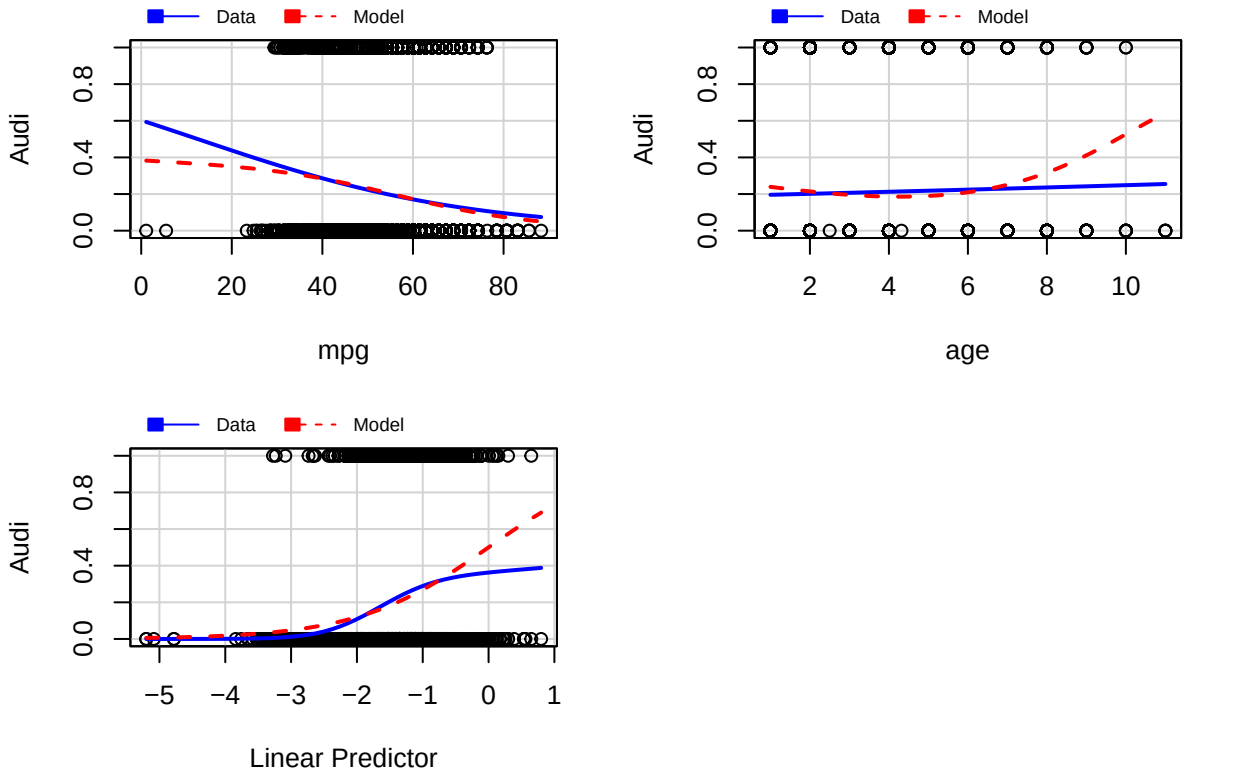
```
residualPlots(m25,id=list(method=cooks.distance(m25),n=10))
```



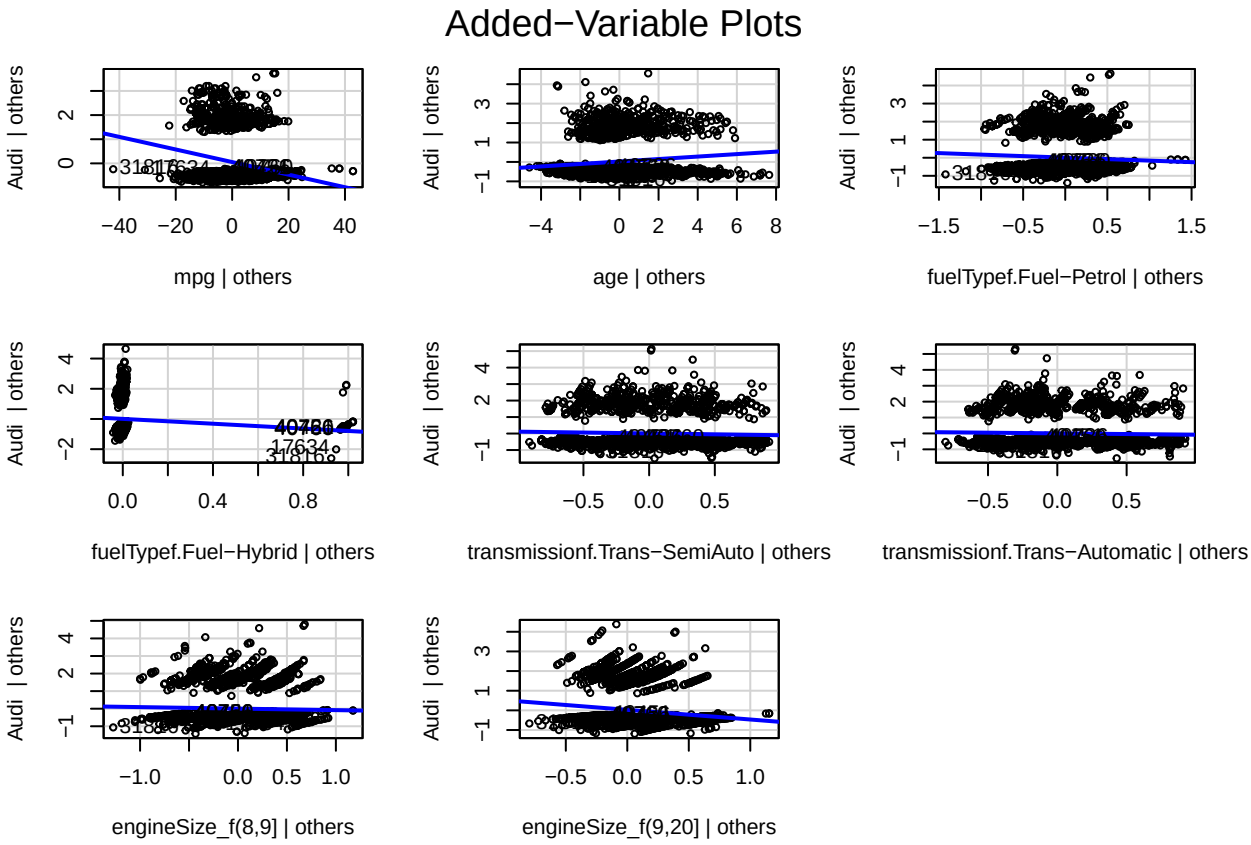

```
             Test stat Pr(>|Test stat|)
mpg             2.3281           0.1271
age            36.1804        1.799e-09 ***
fuelType
transmission
engineSize_f
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
marginalModelPlots(m25)
```

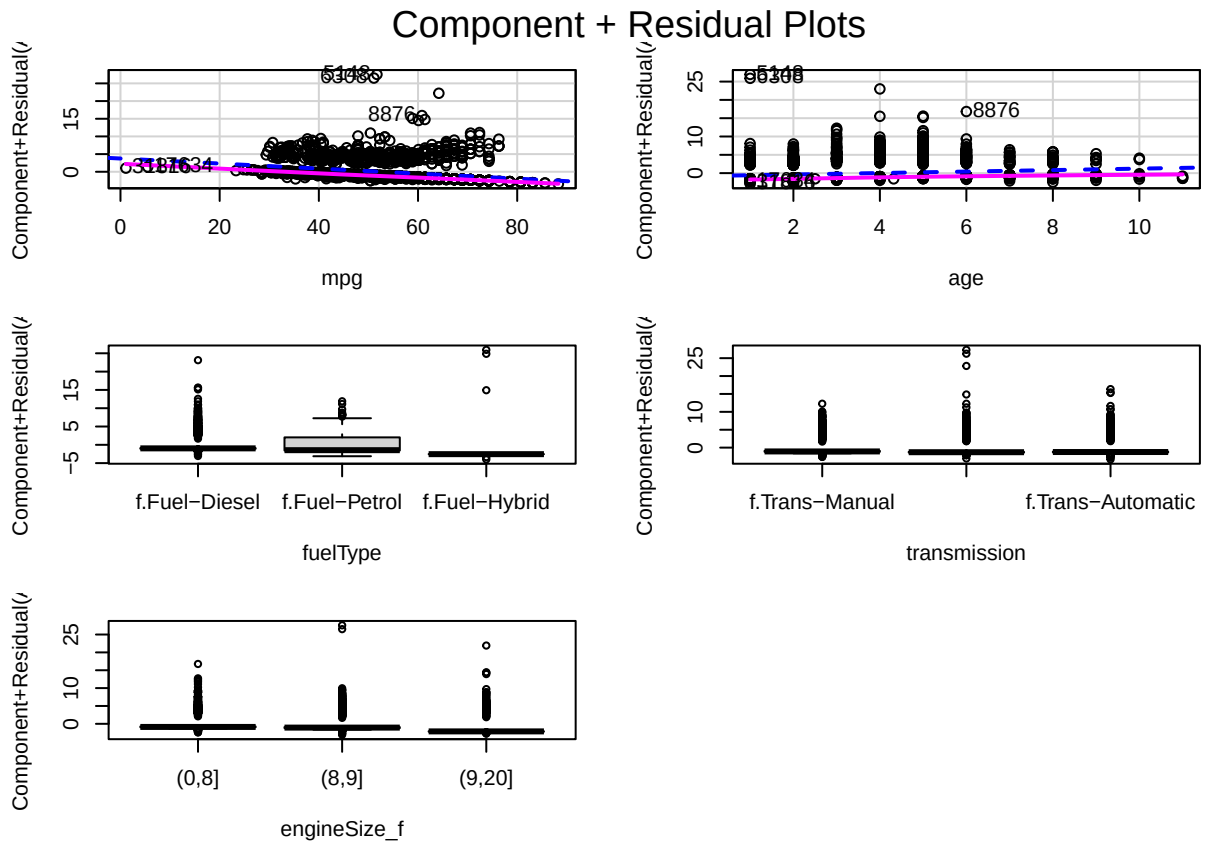
```
Warning in mmps(...): Interactions and/or factors skipped
```



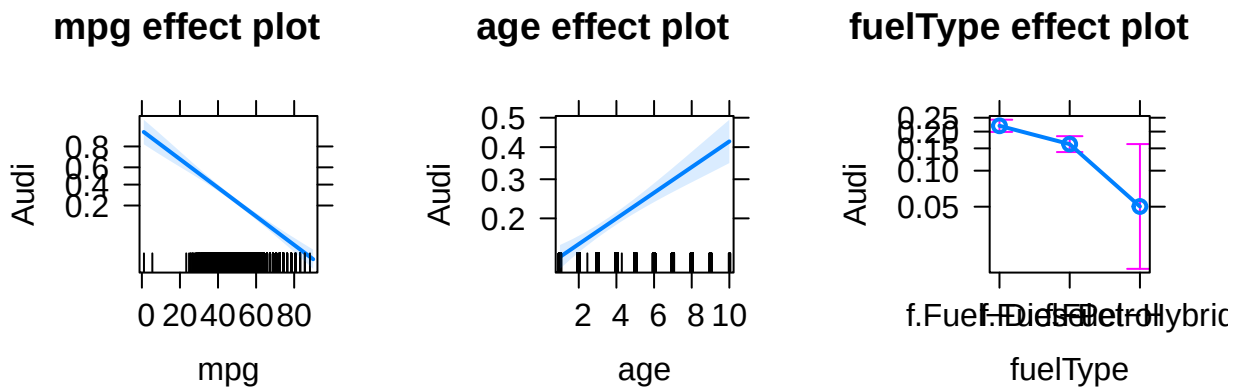

```
avPlots(m25,id=list(method=hatvalues(m25),n=5))
```

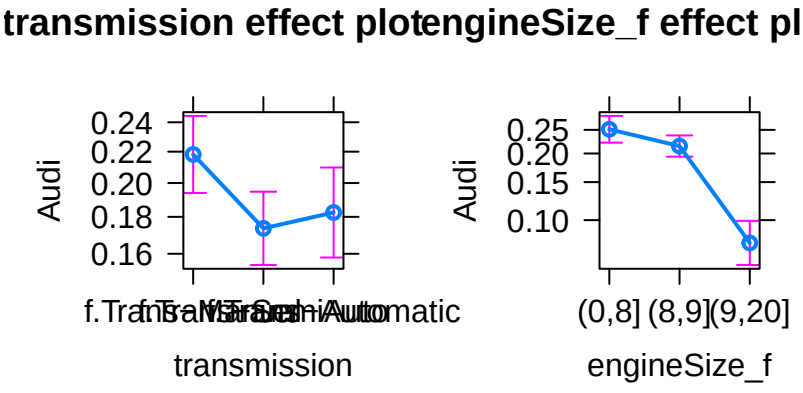


```
crPlots(m25,id=list(method=cooks.distance(m25),n=5))
```




```
# library(effects)
plot(allEffects(m25))
```





3.3 Interacciones

Procederemos a incorporar las interacciones entre los factores fuelType y transmission. Podemos ver como el AIC del modelo se reduce inmendiatamente. Sin embargo, cuando intentamos usar la función vif para estudiar la co-linealidad de las variables, nos genera un error ya que parece que sí que existe co-linealidad.

```
m26 <- update(m25, ~.*(fuelType+transmission)^2,data=df_train)
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(m26)
```

```
Call:
glm(formula = Audi ~ mpg + age + fuelType + transmission + engineSize_f +
    fuelType:transmission + mpg:fuelType + mpg:transmission +
    age:fuelType + age:transmission + fuelType:engineSize_f +
    transmission:engineSize_f + mpg:fuelType:transmission + age:fuelType:transmission +
    fuelType:transmission:engineSize_f, family = "binomial",
    data = df_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8341  -0.7507  -0.5940  -0.2219   2.7263

Coefficients: (5 not defined because of singularities)
                                                                  Estimate
(Intercept)                                                     1.144e+00
mpg                                                            -4.567e-02
age                                                             1.299e-01
fuelTypef.Fuel-Petrol                                          -1.291e+00
fuelTypef.Fuel-Hybrid                                          -1.911e+02
transmissionf.Trans-SemiAuto                                    1.754e+00
transmissionf.Trans-Automatic                                   2.060e+00
engineSize_f(8,9]                                               1.202e-01
engineSize_f(9,20]                                             -1.754e+01
fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto              6.404e-01
fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto              1.779e+02
fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic             -1.328e+00
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic                     NA
mpg:fuelTypef.Fuel-Petrol                                       1.814e-02
mpg:fuelTypef.Fuel-Hybrid                                       8.580e-01
mpg:transmissionf.Trans-SemiAuto                               -5.755e-02
mpg:transmissionf.Trans-Automatic                              -3.949e-02
age:fuelTypef.Fuel-Petrol                                      -3.033e-02
age:fuelTypef.Fuel-Hybrid                                       2.579e+01
age:transmissionf.Trans-SemiAuto                                1.667e-01
age:transmissionf.Trans-Automatic                              -4.303e-02
fuelTypef.Fuel-Petrol:engineSize_f(8,9]                         9.289e-02
fuelTypef.Fuel-Hybrid:engineSize_f(8,9]                        -2.499e+01
fuelTypef.Fuel-Petrol:engineSize_f(9,20]                       -4.439e-01
fuelTypef.Fuel-Hybrid:engineSize_f(9,20]                       -1.596e+01
transmissionf.Trans-SemiAuto:engineSize_f(8,9]                  6.911e-02
transmissionf.Trans-Automatic:engineSize_f(8,9]                -3.858e-01
transmissionf.Trans-SemiAuto:engineSize_f(9,20]                 1.639e+01
transmissionf.Trans-Automatic:engineSize_f(9,20]                1.607e+01
mpg:fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto          1.633e-03
mpg:fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto         -1.111e-01
mpg:fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic         2.851e-02
mpg:fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic                 NA
age:fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto          5.287e-02
age:fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto         -6.147e+01
age:fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic         3.866e-02
```

```
age:fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic                       NA
fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto:engineSize_f(8,9]   -1.122e+00
fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto:engineSize_f(8,9]    5.192e+01
fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic:engineSize_f(8,9]   6.135e-02
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic:engineSize_f(8,9]          NA
fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto:engineSize_f(9,20]  -1.635e-01
fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto:engineSize_f(9,20]   1.734e+02
fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic:engineSize_f(9,20]  1.498e+00
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic:engineSize_f(9,20]         NA
                                                                       Std. Error
(Intercept)                                                             7.715e-01
mpg                                                                     1.295e-02
age                                                                     5.425e-02
fuelTypef.Fuel-Petrol                                                   1.045e+00
fuelTypef.Fuel-Hybrid                                                   7.965e+03
transmissionf.Trans-SemiAuto                                            1.141e+00
transmissionf.Trans-Automatic                                           1.139e+00
engineSize_f(8,9]                                                       2.281e-01
engineSize_f(9,20]                                                      1.018e+03
fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto                      1.606e+00
fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto                      8.254e+03
fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic                     1.761e+00
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic                            NA
mpg:fuelTypef.Fuel-Petrol                                               1.919e-02
mpg:fuelTypef.Fuel-Hybrid                                               1.027e+02
mpg:transmissionf.Trans-SemiAuto                                        1.840e-02
mpg:transmissionf.Trans-Automatic                                       1.852e-02
age:fuelTypef.Fuel-Petrol                                               7.508e-02
age:fuelTypef.Fuel-Hybrid                                               9.806e+02
age:transmissionf.Trans-SemiAuto                                        7.979e-02
age:transmissionf.Trans-Automatic                                       8.193e-02
fuelTypef.Fuel-Petrol:engineSize_f(8,9]                                 4.060e-01
fuelTypef.Fuel-Hybrid:engineSize_f(8,9]                                 1.628e+03
fuelTypef.Fuel-Petrol:engineSize_f(9,20]                                2.846e+03
fuelTypef.Fuel-Hybrid:engineSize_f(9,20]                                4.921e+03
transmissionf.Trans-SemiAuto:engineSize_f(8,9]                          5.143e-01
transmissionf.Trans-Automatic:engineSize_f(8,9]                         5.137e-01
transmissionf.Trans-SemiAuto:engineSize_f(9,20]                         1.018e+03
transmissionf.Trans-Automatic:engineSize_f(9,20]                        1.018e+03
mpg:fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto                  3.057e-02
mpg:fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto                  1.064e+02
mpg:fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic                 3.438e-02
mpg:fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic                        NA
age:fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto                  1.188e-01
age:fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto                  1.371e+03
age:fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic                 1.278e-01
age:fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic                        NA
fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto:engineSize_f(8,9]    6.683e-01
fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto:engineSize_f(8,9]    2.578e+03
fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic:engineSize_f(8,9]   6.962e-01
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic:engineSize_f(8,9]          NA
fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto:engineSize_f(9,20]   2.846e+03
fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto:engineSize_f(9,20]   9.443e+03
fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic:engineSize_f(9,20]  2.846e+03
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic:engineSize_f(9,20]         NA
                                                                        z value
(Intercept)                                                               1.482
mpg                                                                      -3.527
age                                                                       2.395
fuelTypef.Fuel-Petrol                                                    -1.235
fuelTypef.Fuel-Hybrid                                                    -0.024
transmissionf.Trans-SemiAuto                                              1.537
transmissionf.Trans-Automatic                                             1.808
engineSize_f(8,9]                                                         0.527
```

engineSize_f(9,20]	-0.017
fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto	0.399
fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto	0.022
fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic	-0.754
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic	NA
mpg:fuelTypef.Fuel-Petrol	0.945
mpg:fuelTypef.Fuel-Hybrid	0.008
mpg:transmissionf.Trans-SemiAuto	-3.127
mpg:transmissionf.Trans-Automatic	-2.132
age:fuelTypef.Fuel-Petrol	-0.404
age:fuelTypef.Fuel-Hybrid	0.026
age:transmissionf.Trans-SemiAuto	2.089
age:transmissionf.Trans-Automatic	-0.525
fuelTypef.Fuel-Petrol:engineSize_f(8,9]	0.229
fuelTypef.Fuel-Hybrid:engineSize_f(8,9]	-0.015
fuelTypef.Fuel-Petrol:engineSize_f(9,20]	0.000
fuelTypef.Fuel-Hybrid:engineSize_f(9,20]	-0.003
transmissionf.Trans-SemiAuto:engineSize_f(8,9]	0.134
transmissionf.Trans-Automatic:engineSize_f(8,9]	-0.751
transmissionf.Trans-SemiAuto:engineSize_f(9,20]	0.016
transmissionf.Trans-Automatic:engineSize_f(9,20]	0.016
mpg:fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto	0.053
mpg:fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto	-0.001
mpg:fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic	0.829
mpg:fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic	NA
age:fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto	0.445
age:fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto	-0.045
age:fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic	0.303
age:fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic	NA
fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto:engineSize_f(8,9]	-1.679
fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto:engineSize_f(8,9]	0.020
fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic:engineSize_f(8,9]	0.088
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic:engineSize_f(8,9]	NA
fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto:engineSize_f(9,20]	0.000
fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto:engineSize_f(9,20]	0.018
fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic:engineSize_f(9,20]	0.001
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic:engineSize_f(9,20]	NA
	Pr(>|z|)
(Intercept)	0.13826
mpg	0.00042
age	0.01660
fuelTypef.Fuel-Petrol	0.21683
fuelTypef.Fuel-Hybrid	0.98086
transmissionf.Trans-SemiAuto	0.12431
transmissionf.Trans-Automatic	0.07065
engineSize_f(8,9]	0.59839
engineSize_f(9,20]	0.98625
fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto	0.69009
fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto	0.98280
fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic	0.45096
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic	NA
mpg:fuelTypef.Fuel-Petrol	0.34467
mpg:fuelTypef.Fuel-Hybrid	0.99334
mpg:transmissionf.Trans-SemiAuto	0.00177
mpg:transmissionf.Trans-Automatic	0.03300
age:fuelTypef.Fuel-Petrol	0.68624
age:fuelTypef.Fuel-Hybrid	0.97902
age:transmissionf.Trans-SemiAuto	0.03667
age:transmissionf.Trans-Automatic	0.59943
fuelTypef.Fuel-Petrol:engineSize_f(8,9]	0.81905
fuelTypef.Fuel-Hybrid:engineSize_f(8,9]	0.98776
fuelTypef.Fuel-Petrol:engineSize_f(9,20]	0.99988
fuelTypef.Fuel-Hybrid:engineSize_f(9,20]	0.99741
transmissionf.Trans-SemiAuto:engineSize_f(8,9]	0.89311

```
transmissionf.Trans-Automatic:engineSize_f(8,9]                          0.45259
transmissionf.Trans-SemiAuto:engineSize_f(9,20]                          0.98715
transmissionf.Trans-Automatic:engineSize_f(9,20]                         0.98741
mpg:fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto                   0.95741
mpg:fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto                   0.99917
mpg:fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic                  0.40687
mpg:fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic                       NA
age:fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto                   0.65625
age:fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto                   0.96423
age:fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic                  0.76225
age:fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic                       NA
fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto:engineSize_f(8,9]     0.09314
fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto:engineSize_f(8,9]     0.98393
fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic:engineSize_f(8,9]    0.92978
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic:engineSize_f(8,9]         NA
fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto:engineSize_f(9,20]    0.99995
fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto:engineSize_f(9,20]    0.98535
fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic:engineSize_f(9,20]   0.99958
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic:engineSize_f(9,20]        NA

(Intercept)
mpg                                                                      ***
age                                                                      *
fuelTypef.Fuel-Petrol
fuelTypef.Fuel-Hybrid
transmissionf.Trans-SemiAuto
transmissionf.Trans-Automatic                                            .
engineSize_f(8,9]
engineSize_f(9,20]
fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto
fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto
fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic
mpg:fuelTypef.Fuel-Petrol
mpg:fuelTypef.Fuel-Hybrid
mpg:transmissionf.Trans-SemiAuto                                         **
mpg:transmissionf.Trans-Automatic                                        *
age:fuelTypef.Fuel-Petrol
age:fuelTypef.Fuel-Hybrid
age:transmissionf.Trans-SemiAuto                                         *
age:transmissionf.Trans-Automatic
fuelTypef.Fuel-Petrol:engineSize_f(8,9]
fuelTypef.Fuel-Hybrid:engineSize_f(8,9]
fuelTypef.Fuel-Petrol:engineSize_f(9,20]
fuelTypef.Fuel-Hybrid:engineSize_f(9,20]
transmissionf.Trans-SemiAuto:engineSize_f(8,9]
transmissionf.Trans-Automatic:engineSize_f(8,9]
transmissionf.Trans-SemiAuto:engineSize_f(9,20]
transmissionf.Trans-Automatic:engineSize_f(9,20]
mpg:fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto
mpg:fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto
mpg:fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic
mpg:fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic
age:fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto
age:fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto
age:fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic
age:fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic
fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto:engineSize_f(8,9]     .
fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto:engineSize_f(8,9]
fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic:engineSize_f(8,9]
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic:engineSize_f(8,9]
fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto:engineSize_f(9,20]
fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto:engineSize_f(9,20]
fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic:engineSize_f(9,20]
```

```
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic:engineSize_f(9,20]
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3914.0  on 3799  degrees of freedom
Residual deviance: 3584.6  on 3760  degrees of freedom
  (5 observations deleted due to missingness)
AIC: 3664.6

Number of Fisher Scoring iterations: 17
```

```
#vif(m26)
```

Si realizamos un análisis de la varianza del modelo, vemos como hay algunas interacciones que no son significativas, como pueden ser mpg:fuelType:transmission, age:fuelType:transmission o fuelType:transmission.

```
Anova(m26, test="LR")
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: algorithm did not converge

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
Analysis of Deviance Table (Type II tests)

Response: Audi
                                  LR Chisq Df Pr(>Chisq)
mpg                                128.129  1  < 2.2e-16 ***
age                                 43.774  1  3.685e-11 ***
fuelType                            15.671  2  0.0003955 ***
transmission                         5.804  2  0.0549088 .
engineSize_f                        81.512  2  < 2.2e-16 ***
fuelType:transmission                1.626  3  0.6535800
mpg:fuelType                        14.939  2  0.0005701 ***
mpg:transmission                    14.910  2  0.0005786 ***
age:fuelType                        22.199  1  2.458e-06 ***
age:transmission                    14.086  2  0.0008734 ***
fuelType:engineSize_f                3.883  4  0.4221197
transmission:engineSize_f           14.961  4  0.0047823 **
mpg:fuelType:transmission            0.746  3  0.8624137
age:fuelType:transmission            0.220  3  0.9742490
fuelType:transmission:engineSize_f   4.128  6  0.6593613
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Usaremos la función step para plantear el mejor modelo posible a partir de una simplificación del que ya tenemos. Si intentamos volver a ejecutar la función vif, vuelve a generar un error que indica que sigue existiendo colinealidad entre las variables de nuestro modelo, pero en el summary podemos ver como el AIC de este modelo vuelve a ser inferior al del anterior.

```
m27 <- step( m26 )
```

```
Start:  AIC=3664.58
Audi ~ mpg + age + fuelType + transmission + engineSize_f + fuelType:transmission +
    mpg:fuelType + mpg:transmission + age:fuelType + age:transmission +
    fuelType:engineSize_f + transmission:engineSize_f + mpg:fuelType:transmission +
    age:fuelType:transmission + fuelType:transmission:engineSize_f

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

                                    Df Deviance    AIC
- fuelType:transmission:engineSize_f  6   3588.7 3656.7
- age:fuelType:transmission           3   3584.8 3658.8
- mpg:fuelType:transmission           3   3585.3 3659.3
<none>                                    3584.6 3664.6

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Step:  AIC=3656.71
Audi ~ mpg + age + fuelType + transmission + engineSize_f + fuelType:transmission +
    mpg:fuelType + mpg:transmission + age:fuelType + age:transmission +
    fuelType:engineSize_f + transmission:engineSize_f + mpg:fuelType:transmission +
    age:fuelType:transmission

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
                            Df Deviance    AIC
- mpg:fuelType:transmission  3   3589.8 3651.8
- fuelType:engineSize_f      4   3592.6 3652.6
<none>                           3588.7 3656.7
- transmission:engineSize_f  4   3603.7 3663.7
- age:fuelType:transmission  3   3610.3 3672.3

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Step:  AIC=3651.76
Audi ~ mpg + age + fuelType + transmission + engineSize_f + fuelType:transmission +
    mpg:fuelType + mpg:transmission + age:fuelType + age:transmission +
    fuelType:engineSize_f + transmission:engineSize_f + age:fuelType:transmission


Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

                            Df Deviance    AIC
- fuelType:engineSize_f      4   3592.9 3646.9
<none>                           3589.8 3651.8
- transmission:engineSize_f  4   3605.3 3659.3
- mpg:transmission           2   3604.2 3662.2
- age:fuelType:transmission  3   3611.7 3667.7
- mpg:fuelType               2   3611.5 3669.5

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Step:  AIC=3646.9
Audi ~ mpg + age + fuelType + transmission + engineSize_f + fuelType:transmission +
    mpg:fuelType + mpg:transmission + age:fuelType + age:transmission +
    transmission:engineSize_f + age:fuelType:transmission


Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

                            Df Deviance    AIC
<none>                           3592.9 3646.9
- transmission:engineSize_f  4   3610.6 3656.6
- mpg:transmission           2   3608.1 3658.1
- age:fuelType:transmission  3   3615.4 3663.4
- mpg:fuelType               2   3615.7 3665.7
```

```
#vif(m27)
summary(m27)
```

```

Call:
glm(formula = Audi ~ mpg + age + fuelType + transmission + engineSize_f +
    fuelType:transmission + mpg:fuelType + mpg:transmission +
    age:fuelType + age:transmission + transmission:engineSize_f +
    age:fuelType:transmission, family = "binomial", data = df_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
```

```
-1.9049  -0.7432  -0.5920  -0.2368   2.7126

Coefficients: (2 not defined because of singularities)
                                                      Estimate Std. Error
(Intercept)                                           1.257e+00  6.622e-01
mpg                                                  -4.811e-02  1.100e-02
age                                                   1.340e-01  5.218e-02
fuelTypef.Fuel-Petrol                                -1.511e+00  6.637e-01
fuelTypef.Fuel-Hybrid                                -1.123e+03  2.460e+04
transmissionf.Trans-SemiAuto                          2.380e+00  8.913e-01
transmissionf.Trans-Automatic                         1.565e+00  9.311e-01
engineSize_f(8,9]                                     1.447e-01  1.896e-01
engineSize_f(9,20]                                   -1.959e+01  2.575e+03
fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto   -7.500e-02  4.825e-01
fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto    9.839e+02  2.150e+04
fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic   7.474e-02  5.144e-01
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic           NA         NA
mpg:fuelTypef.Fuel-Petrol                             2.292e-02  1.152e-02
mpg:fuelTypef.Fuel-Hybrid                             4.621e+00  1.029e+02
mpg:transmissionf.Trans-SemiAuto                     -5.636e-02  1.459e-02
mpg:transmissionf.Trans-Automatic                    -3.139e-02  1.536e-02
age:fuelTypef.Fuel-Petrol                            -3.702e-02  6.956e-02
age:fuelTypef.Fuel-Hybrid                             1.430e+02  3.205e+03
age:transmissionf.Trans-SemiAuto                      1.653e-01  7.645e-02
age:transmissionf.Trans-Automatic                    -6.678e-02  7.898e-02
transmissionf.Trans-SemiAuto:engineSize_f(8,9]       -7.288e-01  2.795e-01
transmissionf.Trans-Automatic:engineSize_f(8,9]      -3.097e-01  3.184e-01
transmissionf.Trans-SemiAuto:engineSize_f(9,20]       1.781e+01  2.575e+03
transmissionf.Trans-Automatic:engineSize_f(9,20]      1.844e+01  2.575e+03
age:fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto  6.090e-02  1.046e-01
age:fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto -2.246e+02  4.770e+03
age:fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic  9.230e-02  1.151e-01
age:fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic         NA         NA
                                                     z value Pr(>|z|)
(Intercept)                                            1.899 0.057579 .
mpg                                                   -4.372 1.23e-05 ***
age                                                    2.567 0.010248 *
fuelTypef.Fuel-Petrol                                 -2.276 0.022845 *
fuelTypef.Fuel-Hybrid                                 -0.046 0.963595
transmissionf.Trans-SemiAuto                           2.670 0.007576 **
transmissionf.Trans-Automatic                          1.681 0.092837 .
engineSize_f(8,9]                                      0.763 0.445344
engineSize_f(9,20]                                    -0.008 0.993930
fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto    -0.155 0.876483
fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto     0.046 0.963504
fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic    0.145 0.884489
fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic        NA       NA
mpg:fuelTypef.Fuel-Petrol                              1.989 0.046689 *
mpg:fuelTypef.Fuel-Hybrid                              0.045 0.964171
mpg:transmissionf.Trans-SemiAuto                      -3.862 0.000112 ***
mpg:transmissionf.Trans-Automatic                     -2.043 0.041041 *
age:fuelTypef.Fuel-Petrol                             -0.532 0.594594
age:fuelTypef.Fuel-Hybrid                              0.045 0.964407
age:transmissionf.Trans-SemiAuto                       2.163 0.030573 *
age:transmissionf.Trans-Automatic                     -0.846 0.397828
transmissionf.Trans-SemiAuto:engineSize_f(8,9]        -2.607 0.009128 **
transmissionf.Trans-Automatic:engineSize_f(8,9]       -0.973 0.330607
transmissionf.Trans-SemiAuto:engineSize_f(9,20]        0.007 0.994482
transmissionf.Trans-Automatic:engineSize_f(9,20]       0.007 0.994286
age:fuelTypef.Fuel-Petrol:transmissionf.Trans-SemiAuto  0.582 0.560509
age:fuelTypef.Fuel-Hybrid:transmissionf.Trans-SemiAuto -0.047 0.962443
age:fuelTypef.Fuel-Petrol:transmissionf.Trans-Automatic  0.802 0.422626
age:fuelTypef.Fuel-Hybrid:transmissionf.Trans-Automatic     NA       NA
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3914.0  on 3799  degrees of freedom
Residual deviance: 3592.9  on 3773  degrees of freedom
  (5 observations deleted due to missingness)
AIC: 3646.9

Number of Fisher Scoring iterations: 19
```

Si realizamos un análisis de la varianza, podemos ver como algunas interacciones parecen no ser significativas.

```
Anova(m27, test="LR")
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Analysis of Deviance Table (Type II tests)

Response: Audi
                         LR Chisq Df Pr(>Chisq)
mpg                       128.823  1  < 2.2e-16 ***
age                        45.056  1  1.915e-11 ***
fuelType                   15.671  2  0.0003955 ***
transmission                6.762  2  0.0340175 *
engineSize_f               82.754  2  < 2.2e-16 ***
fuelType:transmission       1.429  3  0.6988030
mpg:fuelType               22.838  2  1.098e-05 ***
mpg:transmission           15.213  2  0.0004971 ***
age:fuelType                2.427  2  0.2971581
age:transmission           13.597  2  0.0011152 **
transmission:engineSize_f  17.751  4  0.0013805 **
age:fuelType:transmission  22.520  3  5.083e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vamos a proceder a generar un nuevo modelo que excluya las interacciones no rsignificativas del modelo anterior. Nos encontramos que cuando analizamos este modelo, vuelve a aparecer co-linealidad en las variables, de modo que realizaremos una step regression.

```
m28 <- update(m27, ~.-fuelType:transmission - age:fuelType)
#vif(m28) -> Salta error
```

En este caso, si que podemos ver que ya ha desaparecido la co-linealidad en las variables, y si realizamos el test de anova para los dos modelos, podemos ver que son equivalentes.

```
m29 <- step(m28)
```

```
Start:  AIC=3662.25
```

```
Audi ~ mpg + age + fuelType + transmission + engineSize_f + mpg:fuelType +
    mpg:transmission + age:transmission + transmission:engineSize_f +
    age:fuelType:transmission

                            Df Deviance    AIC
- age:fuelType:transmission  5   3619.5 3657.5
<none>                           3614.3 3662.3
- mpg:fuelType               2   3625.8 3669.8
- transmission:engineSize_f  4   3631.6 3671.6
- mpg:transmission           2   3630.1 3674.1

Step:  AIC=3657.45
Audi ~ mpg + age + fuelType + transmission + engineSize_f + mpg:fuelType +
    mpg:transmission + age:transmission + transmission:engineSize_f

                            Df Deviance    AIC
<none>                           3619.5 3657.5
- mpg:fuelType               2   3628.3 3662.3
- age:transmission           2   3633.1 3667.1
- transmission:engineSize_f  4   3642.9 3672.9
- mpg:transmission           2   3647.1 3681.1
```

```
vif(m29)
```

```
                                  GVIF Df GVIF^(1/(2*Df))
mpg                       4.805983e+00  1        2.192255
age                       3.086963e+00  1        1.756976
fuelType                  5.208927e+02  2        4.777350
transmission              1.267480e+03  2        5.966715
engineSize_f              9.821768e+06  2       55.981872
mpg:fuelType              5.038798e+02  2        4.737855
mpg:transmission          1.209959e+03  2        5.897835
age:transmission          3.605848e+01  2        2.450484
transmission:engineSize_f 6.286284e+07  4        9.436246
```

```
anova(m29,m28, test="LR")
```

```
Analysis of Deviance Table

Model 1: Audi ~ mpg + age + fuelType + transmission + engineSize_f + mpg:fuelType +
    mpg:transmission + age:transmission + transmission:engineSize_f
Model 2: Audi ~ mpg + age + fuelType + transmission + engineSize_f + mpg:fuelType +
    mpg:transmission + age:transmission + transmission:engineSize_f +
    age:fuelType:transmission
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      3781     3619.5
2      3776     3614.3  5   5.1984   0.3921
```

3.4 Diagnóstico

En primer lugar, vamos a ver el summary de nuestro modelo.

```
summary(m29)
```

```
Call:
glm(formula = Audi ~ mpg + age + fuelType + transmission + engineSize_f +
    mpg:fuelType + mpg:transmission + age:transmission + transmission:engineSize_f,
    family = "binomial", data = df_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
```

```
-1.8525  -0.7456  -0.5883  -0.2666   2.6955

Coefficients:
                                                  Estimate Std. Error z value
(Intercept)                                       0.907267   0.491049   1.848
mpg                                              -0.042037   0.008760  -4.798
age                                               0.109937   0.037054   2.967
fuelTypef.Fuel-Petrol                            -1.265374   0.467898  -2.704
fuelTypef.Fuel-Hybrid                            -5.878925   2.451683  -2.398
transmissionf.Trans-SemiAuto                      2.530705   0.585758   4.320
transmissionf.Trans-Automatic                     2.297330   0.632083   3.635
engineSize_f(8,9]                                 0.236566   0.165797   1.427
engineSize_f(9,20]                              -14.506289 211.634172  -0.069
mpg:fuelTypef.Fuel-Petrol                         0.018190   0.009359   1.944
mpg:fuelTypef.Fuel-Hybrid                         0.085857   0.043598   1.969
mpg:transmissionf.Trans-SemiAuto                 -0.058949   0.011703  -5.037
mpg:transmissionf.Trans-Automatic                -0.043161   0.012492  -3.455
age:transmissionf.Trans-SemiAuto                  0.189089   0.057271   3.302
age:transmissionf.Trans-Automatic                -0.011553   0.060682  -0.190
transmissionf.Trans-SemiAuto:engineSize_f(8,9]   -0.800706   0.216033  -3.706
transmissionf.Trans-Automatic:engineSize_f(8,9]  -0.552259   0.263474  -2.096
transmissionf.Trans-SemiAuto:engineSize_f(9,20]  12.749327 211.634271   0.060
transmissionf.Trans-Automatic:engineSize_f(9,20] 13.126489 211.634333   0.062
                                                 Pr(>|z|)
(Intercept)                                      0.064659 .
mpg                                              1.60e-06 ***
age                                              0.003008 **
fuelTypef.Fuel-Petrol                            0.006843 **
fuelTypef.Fuel-Hybrid                            0.016489 *
transmissionf.Trans-SemiAuto                     1.56e-05 ***
transmissionf.Trans-Automatic                    0.000278 ***
engineSize_f(8,9]                                0.153625
engineSize_f(9,20]                               0.945352
mpg:fuelTypef.Fuel-Petrol                        0.051937 .
mpg:fuelTypef.Fuel-Hybrid                        0.048920 *
mpg:transmissionf.Trans-SemiAuto                 4.73e-07 ***
mpg:transmissionf.Trans-Automatic                0.000550 ***
age:transmissionf.Trans-SemiAuto                 0.000961 ***
age:transmissionf.Trans-Automatic                0.849012
transmissionf.Trans-SemiAuto:engineSize_f(8,9]   0.000210 ***
transmissionf.Trans-Automatic:engineSize_f(8,9]  0.036076 *
transmissionf.Trans-SemiAuto:engineSize_f(9,20]  0.951963
transmissionf.Trans-Automatic:engineSize_f(9,20] 0.950543
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3914.0  on 3799  degrees of freedom
Residual deviance: 3619.5  on 3781  degrees of freedom
  (5 observations deleted due to missingness)
AIC: 3657.5

Number of Fisher Scoring iterations: 14
```

Como podemos ver, su valor de AIC asociado es de 3639.7. Aunque no es el mejor resultado que hemos conseguido, ya que el m28 tenia 3646.8, este no presenta co-linealidad.

Como hemos comentado anteriormente, con la función vif podemos ver como no aparece co-linealidad entre las variables del modelo. Si analizamos la varianza del modelo, veremos como todas las variables que aparecen son significativas.

```
vif(m29)
```

```
                                GVIF Df GVIF^(1/(2*Df))
mpg                     4.805983e+00  1        2.192255
age                     3.086963e+00  1        1.756976
fuelType                5.208927e+02  2        4.777350
transmission            1.267480e+03  2        5.966715
engineSize_f            9.821768e+06  2       55.981872
mpg:fuelType            5.038798e+02  2        4.737855
mpg:transmission        1.209959e+03  2        5.897835
age:transmission        3.605848e+01  2        2.450484
transmission:engineSize_f 6.286284e+07  4        9.436246
```

```
Anova(m29)
```

```
Analysis of Deviance Table (Type II tests)

Response: Audi
                         LR Chisq Df Pr(>Chisq)
mpg                       133.980  1  < 2.2e-16 ***
age                        44.994  1  1.977e-11 ***
fuelType                   15.671  2  0.0003955 ***
transmission                6.874  2  0.0321588 *
engineSize_f               98.982  2  < 2.2e-16 ***
mpg:fuelType                8.842  2  0.0120238 *
mpg:transmission           27.667  2  9.820e-07 ***
age:transmission           13.639  2  0.0010923 **
transmission:engineSize_f  23.411  4  0.0001048 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vamos a proceder a analizar algunos gráficos del modelo que hemos generado.

Si analizamos los boxplots de las los valores de Hat y las distancias de Cook, podemos ver como hay una serie de elementos que aparecen muy separados, algunos coinciden en los tres gráficos.

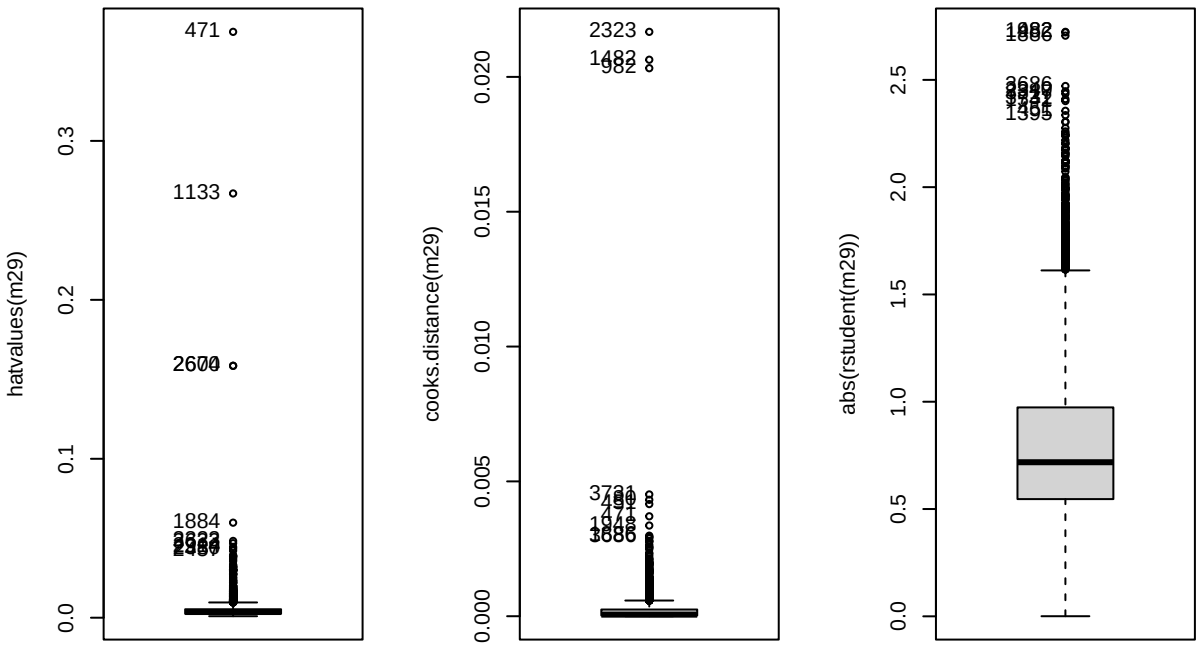
```
par(mfrow=c(1,3))
Boxplot(hatvalues(m29),id=c(labels=row.names(df_train)))
```

```
 [1]  471 1133 2600 2674 1884 2323 3632 2914 2380 2457
```

```
Boxplot(cooks.distance(m29),id=c(labels=row.names(df_train)))
```

```
 [1] 2323 1482  982 3731   80  451  471 1948 1886 3686
```

```
Boxplot(abs(rstudent(m29)),id=c(labels=row.names(df_train)))
```




```
 [1]  982 1482 1886 3686 2340 3317 3731 1542  451 1395
```

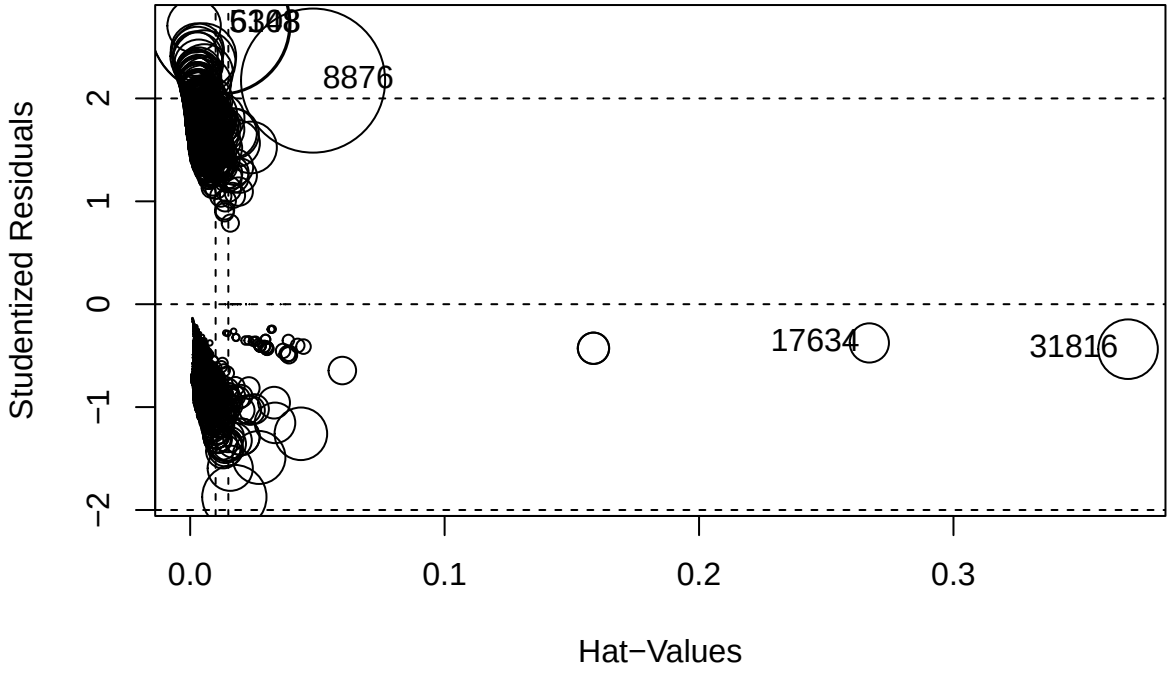
En primer lugar, vamos a observar los valores de los residuos de student. Al haber generado el gráfico con el valor absoluto, solo tendremos que filtrar la parte superior del boxplot. Podemos ver como la cadena de puntos se rompe aproximadamente cerca del 2.3, de manera que vamos a considerar los individuos con residuo de student fuera del intervalo [-2.3, 2.3] como outliers. Haremos lo mismo para las observaciones con distancias de Cook superiores a 0.0035 y Hat values superiores a 0.07.

```
stu_out <- which(abs(rstudent(m29))>2.3);
cook_out <- which(abs(cooks.distance(m29))>0.0035);
hat_out <- which(abs(hatvalues(m29))>0.07);

outs<-unique(c(stu_out,cook_out,hat_out));outs
```

```
 [1]  451  982 1060 1395 1482 1542 1555 1886 2340 3317 3686 3731   80  471 2323
[16] 1133 2600 2674
```

Si echamos un vistazo al plot que nos marca la influencia que tienen las distinas observaciones hacia el modelo, podemos ver como aparecen algunos puntos bastante alejados de las nubes que se crean, algunos con bastante influencia.

```
par(mfrow=c(1,1));
outs2 <- influencePlot(m29, id=c(labels=row.names(df_train)));
```




```
outs2 <- labels(outs2)[[1]];
outs2 <- as.numeric(outs2);
outs3 <- unique(c(outs,outs2));outs3
```

```
 [1]   451   982  1060  1395  1482  1542  1555  1886  2340  3317  3686  3731
[13]    80   471  2323  1133  2600  2674 31816  5148 17634  6308  8876
```

Vamos a proceder a crear un nuevo modelo que excluya las observaciones que hemos considerado como outliers.

```
m30 <- update(m29,data=df_train[-outs3,])
summary(m30)
```

```
Call:
glm(formula = Audi ~ mpg + age + fuelType + transmission + engineSize_f +
    mpg:fuelType + mpg:transmission + age:transmission + transmission:engineSize_f,
    family = "binomial", data = df_train[-outs3, ])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6093  -0.7462  -0.5866  -0.2627   2.7114

Coefficients:
                                              Estimate Std. Error z value
(Intercept)                                   0.956291   0.493736   1.937
mpg                                          -0.042994   0.008824  -4.872
age                                           0.111463   0.037352   2.984
fuelTypef.Fuel-Petrol                        -1.329694   0.469637  -2.831
fuelTypef.Fuel-Hybrid                        -5.945097   2.448001  -2.429
transmissionf.Trans-SemiAuto                  2.587878   0.589103   4.393
transmissionf.Trans-Automatic                 2.291013   0.636000   3.602
engineSize_f(8,9]                             0.229136   0.167109   1.371
engineSize_f(9,20]                          -14.511852 213.902011  -0.068
mpg:fuelTypef.Fuel-Petrol                     0.019282   0.009390   2.054
mpg:fuelTypef.Fuel-Hybrid                     0.086973   0.043529   1.998
```

```
mpg:transmissionf.Trans-SemiAuto                  -0.060569   0.011803  -5.132
mpg:transmissionf.Trans-Automatic                 -0.043208   0.012555  -3.441
age:transmissionf.Trans-SemiAuto                   0.198183   0.057778   3.430
age:transmissionf.Trans-Automatic                 -0.011940   0.060861  -0.196
transmissionf.Trans-SemiAuto:engineSize_f(8,9]    -0.790322   0.217072  -3.641
transmissionf.Trans-Automatic:engineSize_f(8,9]   -0.528720   0.265685  -1.990
transmissionf.Trans-SemiAuto:engineSize_f(9,20]   12.727706 213.902109   0.060
transmissionf.Trans-Automatic:engineSize_f(9,20]  13.137264 213.902172   0.061
                                                  Pr(>|z|)
(Intercept)                                       0.052764 .
mpg                                               1.10e-06 ***
age                                               0.002844 **
fuelTypef.Fuel-Petrol                             0.004636 **
fuelTypef.Fuel-Hybrid                             0.015159 *
transmissionf.Trans-SemiAuto                      1.12e-05 ***
transmissionf.Trans-Automatic                     0.000316 ***
engineSize_f(8,9]                                 0.170320
engineSize_f(9,20]                                0.945910
mpg:fuelTypef.Fuel-Petrol                         0.040020 *
mpg:fuelTypef.Fuel-Hybrid                         0.045714 *
mpg:transmissionf.Trans-SemiAuto                  2.87e-07 ***
mpg:transmissionf.Trans-Automatic                 0.000579 ***
age:transmissionf.Trans-SemiAuto                  0.000603 ***
age:transmissionf.Trans-Automatic                 0.844469
transmissionf.Trans-SemiAuto:engineSize_f(8,9]    0.000272 ***
transmissionf.Trans-Automatic:engineSize_f(8,9]   0.046588 *
transmissionf.Trans-SemiAuto:engineSize_f(9,20]   0.952552
transmissionf.Trans-Automatic:engineSize_f(9,20] 0.951027
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3892.3  on 3781  degrees of freedom
Residual deviance: 3595.0  on 3763  degrees of freedom
  (5 observations deleted due to missingness)
AIC: 3633

Number of Fisher Scoring iterations: 14
```

Podemos ver como el valor AIC del modelo ha disminuido.

Con la función vif vemos que no existen variables co-lineales en el modelo. Si analizamos la varianza, vemos que todas las variables son significativas.

```
vif(m30)
```

```
                                GVIF Df GVIF^(1/(2*Df))
mpg                     4.824980e+00  1        2.196584
age                     3.105035e+00  1        1.762111
fuelType                5.191643e+02  2        4.773382
transmission            1.280708e+03  2        5.982222
engineSize_f            1.008267e+07  2       56.349991
mpg:fuelType            5.022510e+02  2        4.734021
mpg:transmission        1.226350e+03  2        5.917709
age:transmission        3.654090e+01  2        2.458639
transmission:engineSize_f 6.532354e+07  4        9.481645
```

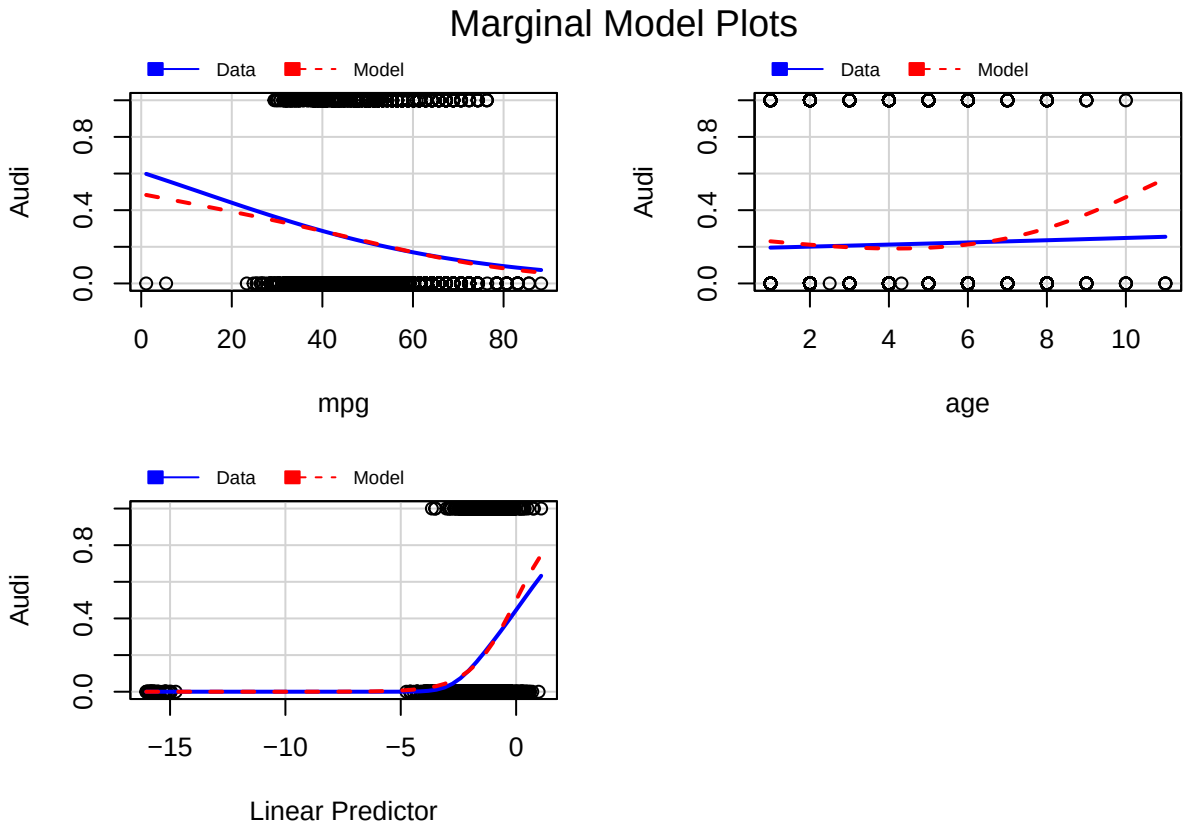
```
Anova(m30, test="LR")
```

```
Analysis of Deviance Table (Type II tests)

Response: Audi
```

```
                          LR Chisq Df Pr(>Chisq)
mpg                        137.001  1  < 2.2e-16 ***
age                         46.453  1  9.383e-12 ***
fuelType                    16.183  2  0.0003061 ***
transmission                 6.263  2  0.0436575 *
engineSize_f               100.689  2  < 2.2e-16 ***
mpg:fuelType                 9.389  2  0.0091445 **
mpg:transmission            28.487  2  6.517e-07 ***
age:transmission            14.740  2  0.0006298 ***
transmission:engineSize_f   22.586  4  0.0001532 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
marginalModelPlots(m30)
```

```
Warning in mmps(...): Interactions and/or factors skipped
```




```
m0<-glm(Audi ~ 1, family="binomial", data=df_train[-outs3,])
```

3.5 Bondad del ajuste y capacidad de predicción

Vamos a estudiar la bondad de nuestro modelo y su capacidad de predicción.

En primer lugar, vamos a empezar planteando la distribucion del modelo de forma asimptótica con el test de chi-cuadrado.

```
Anova(m30)
```

```
Analysis of Deviance Table (Type II tests)

Response: Audi
                          LR Chisq Df Pr(>Chisq)
mpg                        137.001  1  < 2.2e-16 ***
```

```
age                          46.453  1  9.383e-12 ***
fuelType                     16.183  2  0.0003061 ***
transmission                  6.263  2  0.0436575 *
engineSize_f                100.689  2  < 2.2e-16 ***
mpg:fuelType                  9.389  2  0.0091445 **
mpg:transmission             28.487  2  6.517e-07 ***
age:transmission             14.740  2  0.0006298 ***
transmission:engineSize_f    22.586  4  0.0001532 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
1-pchisq(m30$deviance, m30$df.residual)
```

```
[1] 0.9748901
```

Como podemos ver, el p-valor de nuestra hipótesis nula es de 0.94, de modo que podemos refutarla y podemos afirmar que, en efecto, el modelo no se ajusta bien a los datos.

Similarmente, si planteamos el estadístico de Pearson X2, nos encontramos que en este caso, aplicando un intervalo de confianza del 95%, deberíamos recharzar nuestra hipótesis nula y afirmar que el modelo NO se ajusta bien a los datos.

```r
X2m30<-sum((resid(m30,"pearson")^2))
1-pchisq( X2m30, m30$df.res)
```

```
[1] 0.9551595
```

Si aplicamos el test de Pseudo R2, que tiene un rol similar a la suma de los cuadrados de los residuos en una regresion clásica, podemos ver como existen claras discrepancias entre si podemos o no aceptar nuestra hipótesis nula.

```r
library(DescTools)
```

```
Attaching package: 'DescTools'

The following object is masked from 'package:games':

    Mode

The following object is masked from 'package:car':

    Recode
```

```r
PseudoR2(m30, which='all')
```

```
       McFadden     McFaddenAdj        CoxSnell      Nagelkerke   AldrichNelson
   7.636954e-02    6.660663e-02    7.558707e-02    1.176100e-01    7.286916e-02
VeallZimmermann           Efron McKelveyZavoina            Tjur             AIC
   1.436737e-01    6.559726e-02    4.754666e-01    6.928751e-02    3.633028e+03
           BIC          logLik         logLik0              G2
   3.751550e+03   -1.797514e+03   -1.946140e+03    2.972516e+02
```

Sin embargo, debemos recordar que estos test no funcionan con conjuntos de datos agrupados, como pueden ser los que aparecen en nuestra variable engineSize, o en los factores que hemos incluido en nuestro modelo.

Si planteamos el test de Hoslem, podemos ver como el p-valor de la hipótesis nula es de 0.011, y podemos aceptar nuestra hipótesis nula, afirmando que el modelo SÍ que se ajusta bien a los datos.

```
library(ResourceSelection)
```

```
ResourceSelection 0.3-5      2019-07-22
```

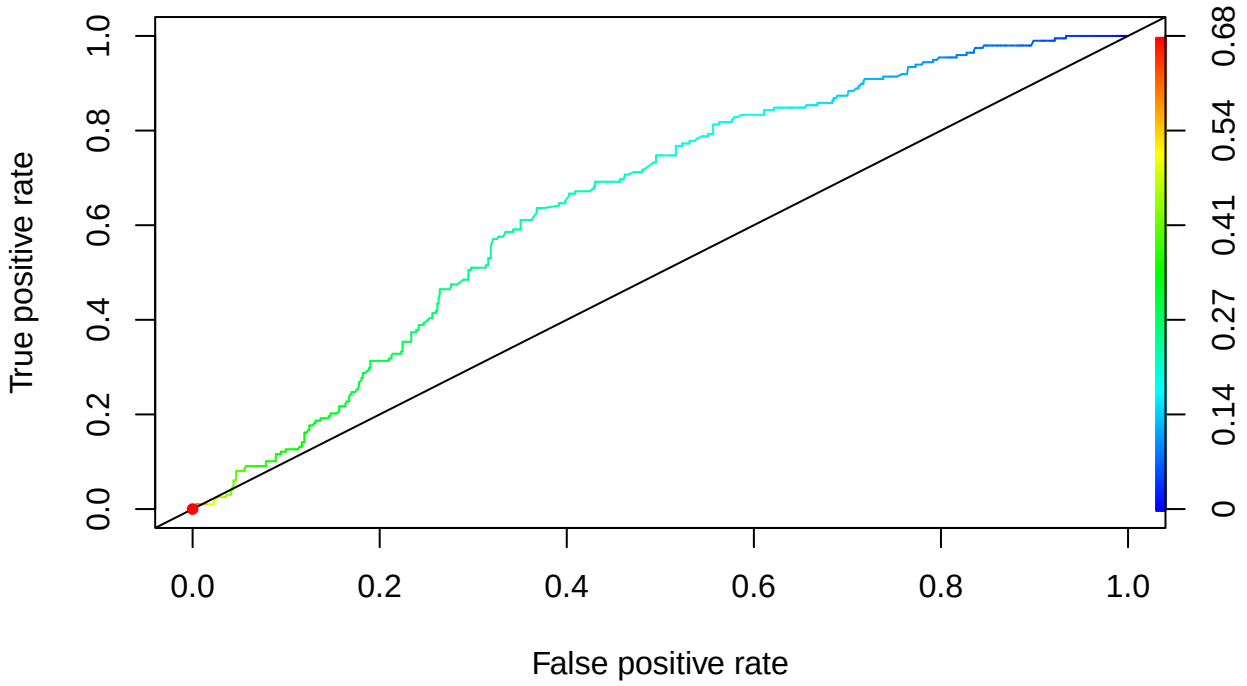
```
ll <- which( is.finite(df_test$engineSize_f) )
pred_test <- predict(m30, newdata=df_test[ll,], type="response")
ht <- hoslem.test(as.numeric(df_test$Audi[ll])-1,  pred_test)
ht
```

```
    Hosmer and Lemeshow goodness of fit (GOF) test

data:  as.numeric(df_test$Audi[ll]) - 1, pred_test
X-squared = 19.46, df = 8, p-value = 0.01258
```

A continuación vamos a general la curva de ROC que nos ayudará a ver de manera gráfica la bondad del ajuste del modelo.

De manera indicativa, un modelo excelente, se acercaría mucho al punto (0,1), mientras que un modelo que se acerca a la recta con y = x sería un modelo malo.

```
pred <- prediction(pred_test, df_test$Audi[ll])
perf <- performance(pred,measure="tpr",x.measure="fpr")

plot(perf,colorize=TRUE,type="l")
abline(a=0,b=1)

# Área bajo la curva
AUC       <- performance(pred,measure="auc")
AUCaltura <- AUC@y.values

# Punto de corte óptimo
cost.perf <- performance(pred, measure ="cost")
opt.cut   <- pred@cutoffs[[1]][which.min(cost.perf@y.values[[1]])]
#coordenadas del punto de corte óptimo
x<-perf@x.values[[1]][which.min(cost.perf@y.values[[1]])]
y<-perf@y.values[[1]][which.min(cost.perf@y.values[[1]])]
points(x,y, pch=20, col="red")
```



Como podemos ver, nuestro modelo se acerca más a la recta x=y que al punto (0,1), indicando que es bastante mejorable.

Vamos a analizar algunos valores característicos de esta curva.

```
# Área bajo la curva
AUC       <- performance(pred,measure="auc")
AUCaltura <- AUC@y.values

cat("AUC:", AUCaltura[[1]]);
```

```
AUC: 0.6500228
```

```
cat("Punto de corte óptimo:",opt.cut)
```

```
Punto de corte óptimo: Inf
```

Podemos ver que el área bajo la curva es de 0.687, indicando que es un modelo bastante malo. Además, el punto de córte óptimo se situa en el (0,0) (No se muy bien como interpretar este resultado, pero muy positivo no debe ser...)

3.6 Matriz de confusión

Vamos a generar la matriz de confusión del modelo que hemos planteado.

```
audi.est <- ifelse(pred_test<0.4,0,1)
tt<-table(audi.est,df_test$Audi[ll]);tt;
```

```
audi.est Audi No Audi Yes
       0     718       184
       1      35        14
```

Si aplicamos las definiciones: Sensibilidad: 12 / (12 + 184) = 0.06 Especificidad: 726 / (28 + 721) = 0.97

Con estos dos conceptos podemos concluir que el modelo responde que NO a casi todo. Vemos como de las 210 observaciones que SÍ que eran Audi, solo ha respondido correctamente al 5%. Por otro lado, vemos como ha acertado casi todos los NO Audi...

Finalmente, si estudiamos la tasa de acierto de nuestro modelo, podemos ver que con el conjunto de validación ha acertado el 77.68% de las veces.

```
100*sum(diag(tt))/sum(tt)
```

```
[1] 76.97161
```