Facultad de Ciencias.
Licenciatura en Actuaría.
Seminario de Estadística : Aprendizaje estadístico automatizado
Semestre 2020-1 Grupo 9269

Yanely Luna Gutiérrez

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Tarea 1
Regresión Lineal

Octubre 2020

1.

1.1. Ejercicio 9

del capítulo 3 Linear Regression, de James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning. With applications in R, Springer, ISL*

En este ejercicio se trabajará con la base de datos `auto`. Esta base de datos tiene información de 392 vehículos sobre 9 variables:

- `mpg`: Millas recorridas por galón.
- `cyliders`: Número de cilindros (enteros entre 3 y 8).
- `displacement`: Volumen útil de todos los cilindros en pulgadas cúbicas.
- `horsepower`: Potencia del motor (en caballos de fuerza).
- `weight`: Peso del vehículo en libras.
- `acceleration`: Tiempo en segundos que tarda en acelerar de 0 a 60mph.
- `year`: Año del modelo.
- `origin`: Origen del vehículo (1 Americano, 2 Europeo, 3 Japonés).
- `name`: Nombre del vehículo

Las variables discretas `cyliders` y `origin` fueron convertidas a factores (categóricas) y las demás son tratadas como numéricas.

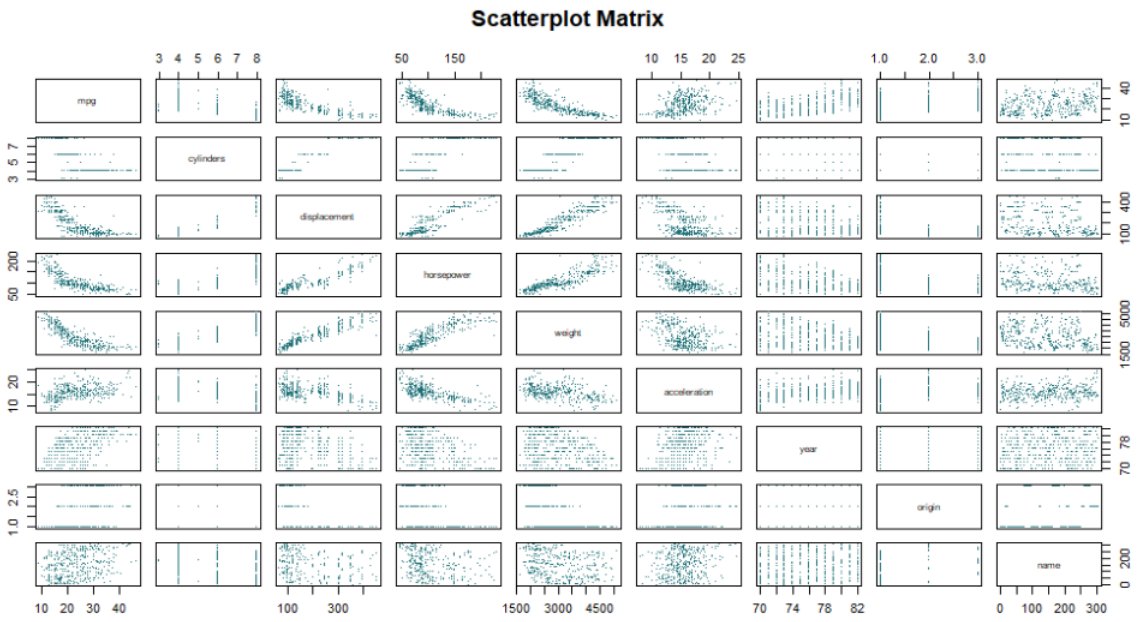

Figura 1: **Gráfica de dispersión por pares.** *Algunas variables parecen tener una relación lineal positiva (como* `displacement` *y* `horsepower`*), mientras que en otras se observa una relación no lineal negativa (como* `mpg` *y* `weight`*) y algunas parecen no tener una relación tan clara (en* `year` *y* `horsepower`*)*

En la gráfica de dispersión por pares podemos observar la relación gráfica entre cada par de variables de la base. Las relaciones con la variable `name` no nos dicen algo claro puesto que esta variable es solo el nombre del vehículo y no es muy relevante en el análisis que estamos haciendo.

Para notar cuantitativamente las relaciones entre cada variable obtenemos su matriz de correlaciones.

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin
mpg	1.00	-0.78	-0.81	-0.78	-0.83	0.42	0.58	0.57
cylinders	-0.78	1.00	0.95	0.84	0.90	-0.50	-0.35	-0.57
displacement	-0.81	0.95	1.00	0.90	0.93	-0.54	-0.37	-0.61
horsepower	-0.78	0.84	0.90	1.00	0.86	-0.69	-0.42	-0.46
weight	-0.83	0.90	0.93	0.86	1.00	-0.42	-0.31	-0.59
acceleration	0.42	-0.50	-0.54	-0.69	-0.42	1.00	0.29	0.21
year	0.58	-0.35	-0.37	-0.42	-0.31	0.29	1.00	0.18
origin	0.57	-0.57	-0.61	-0.46	-0.59	0.21	0.18	1.00

Figura 2: **Matriz de correlaciones.** *La correlación más alta en valor absoluto de la variable respuesta (*`mpg`*) es con la variable* `weight`*, la cual a su vez está altamente correlacionada con la variable* `displacement`*, lo que puede ser un indicio de problemas de multicolinealidad.*

En la matriz de correlaciones podemos confirmar lo que notamos en la gráfica anterior respecto a la relación lineal positiva entre `displacement` y `horsepower` pues su correlación es de 0.89, bastante cercana a 1.

Una vez exploradas nuestras variables, podemos ajustar un primer modelo de regresión para explicar la variable `mpg` en términos del resto de las variables de la base. Este primer modelo estima, además del intercepto, 4 parámetros

realicionados a las variables predictoras numéricas y 6 parámetros asociados a las dos variables categóricas (`cylinders` y `origin`). Para éstas últimas, los niveles de referencia son 3 y 1 respectivamente.

```
Call:
lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
    acceleration + year + origin, data = Auto)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6797 -1.9373 -0.0678  1.6711 12.7756

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.208e+01  4.541e+00  -4.862 1.70e-06 ***
cylinders4    6.722e+00  1.654e+00   4.064 5.85e-05 ***
cylinders5    7.078e+00  2.516e+00   2.813  0.00516 **
cylinders6    3.351e+00  1.824e+00   1.837  0.06701 .
cylinders8    5.099e+00  2.109e+00   2.418  0.01607 *
displacement  1.870e-02  7.222e-03   2.590  0.00997 **
horsepower   -3.490e-02  1.323e-02  -2.639  0.00866 **
weight       -5.780e-03  6.315e-04  -9.154  < 2e-16 ***
acceleration  2.598e-02  9.304e-02   0.279  0.78021
year          7.370e-01  4.892e-02  15.064  < 2e-16 ***
origin2       1.764e+00  5.513e-01   3.200  0.00149 **
origin3       2.617e+00  5.272e-01   4.964 1.04e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.098 on 380 degrees of freedom
Multiple R-squared:  0.8469, Adjusted R-squared:  0.8425
F-statistic: 191.1 on 11 and 380 DF,  p-value: < 2.2e-16
```

Figura 3: **Resumen del modelo 1.** *La variable `acceleration` es la que menos contribuye a explicar `mpg`.*

En el resumen del modelo podemos notar que sí hay una relación entre las variables predictoras y la variable respuesta pues la prueba de hipótesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_7 = 0 \text{ vs } H_A : \beta_i \neq 0 \text{ para alguna } i \in \{1, 2, ..., 7\}$$

tiene un p-value de $2{,}2 \cdot 10^{-16}$, lo cual nos dice que el modelo sí es significativo, es decir, al menos uno de las betas es estadísticamente distinta de cero.

Sin embargo, algunos predictores son más significativos que otros, por ejemplo, las variables `weight`, `year` y `origin3` son estadísticamente significativas a un nivel muy cercano a 0 mientras que `displacement`, `horsepower`, `cylinders5` y `origin2` lo son a un nivel de 0.001 y `cylinders6` lo es a un nivel de 0.1 pero no al nivel 0.5 . Por último `acceleration` no lo es ni siquiera a un nivel de 0.1 Para el caso de los parámetros asociados a los niveles de las variables categóricas, el hecho de que no sean significativos nos dice que en promedio no existe direfencia entre dicho nivel y el nivel de referencia de la correspondiente variable.

Podemos decir que la variable que menos contribuye a explicar las millas por galón del vehículo es la aceleración. Por otro lado, el año del modelo es una variable altamente significativa y al tener un coeficiente ajustado de 0.75, nos dice que en dos autos con las mismas características excepto en una diferencia de un año en el modelo, el vehículo más reciente será en promedio 0.75 millas por galón más rendidor que el auto que es un año más antiguo.

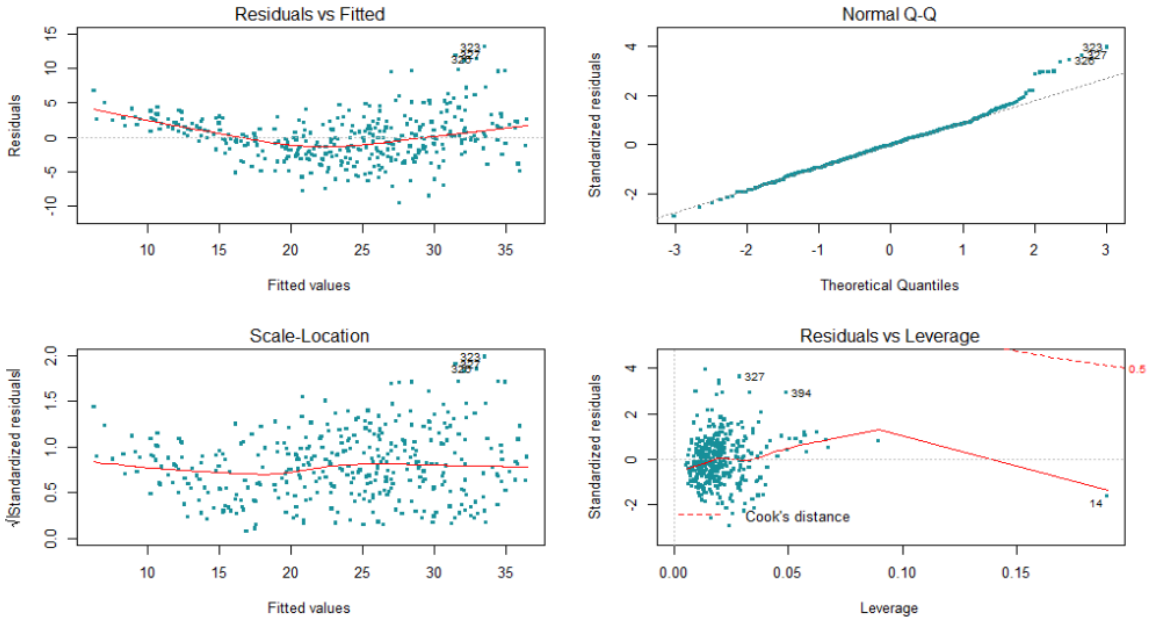

Figura 4: **Gráficas de los residuales del modelo 1.**

En las gráficas de diagnóstico de los residuales podemos notar que al comparar los residuales con los valores ajustados no se observa una dispersión aleatoria como se desea, pareciera que los residuales presentan una varianza que aumenta conforme incrementan los valores ajustados, lo cual incumple el supuesto de varianza constante. Además, en la comparación de cuantiles con la noraml vemos que los valores en la cola derecha difieren de los cuantiles teóricos de una forma significativa, sobre todo para las observaciones 323, 326 y 327, mientras que la distancia de Cook nos muestra que la observación 14 se aleja del conjunto de las demás observaciones, lo que nos sugiere probar con un modelo que no incluya dicha observación y ver si mejora el ajuste.

Como siguiente paso podemos construir un modelo que tome en cuenta las interacciones entre las variables predictoras. En la Tabla 1 se muestran los resultados de los modelos que tienen todas las variables predictoras del modelo 1 y también consideran la interacción (o interacciones) que se indican.

Modelo	Interacción	R^2 ajustada	p-value Anova
2	`year:origin`	0.8508	$1{,}284 \cdot 10^{-05}$
3	`weight:acceleration`	0.8511	$2{,}248 \cdot 10^{-6}$
4	`cylinders:displacement`	0.8625	$2{,}683 \cdot 10^{-11}$
5	`cylinders:displacement + weight:acceleration`	0.8628	$4{,}838 \cdot 10^{-11}$

Cuadro 1: **Comparación de los modelos con interacciones.** *La R^2 crece poco más de 1 % con respecto al modelo original, sin embargo el p-value nos dice que los parámetros que se añaden sí son estadísticamente significativos.*

En la columna **p-value Anova** se encuentra el p-value de la tabla anova, la cual compara el modelo 1 (el que toma en cuenta todas las variables) con el modelo correspondiente (el que además de todas las variables de la base añade la interacción que se indica). Cuando el p-value es menor al nivel de significancia, en este caso 0.05, nos dice que se rechaza la hipótesis de que el parámetro o parámetros extra que tiene un modelo con respecto al que se compara sean cero.

Por lo tanto, cuando el p-value es menor a 0.05 nos conviene quedarnos con el modelo que tiene más variables, como sucede con las cuatro comparaciones hechas.

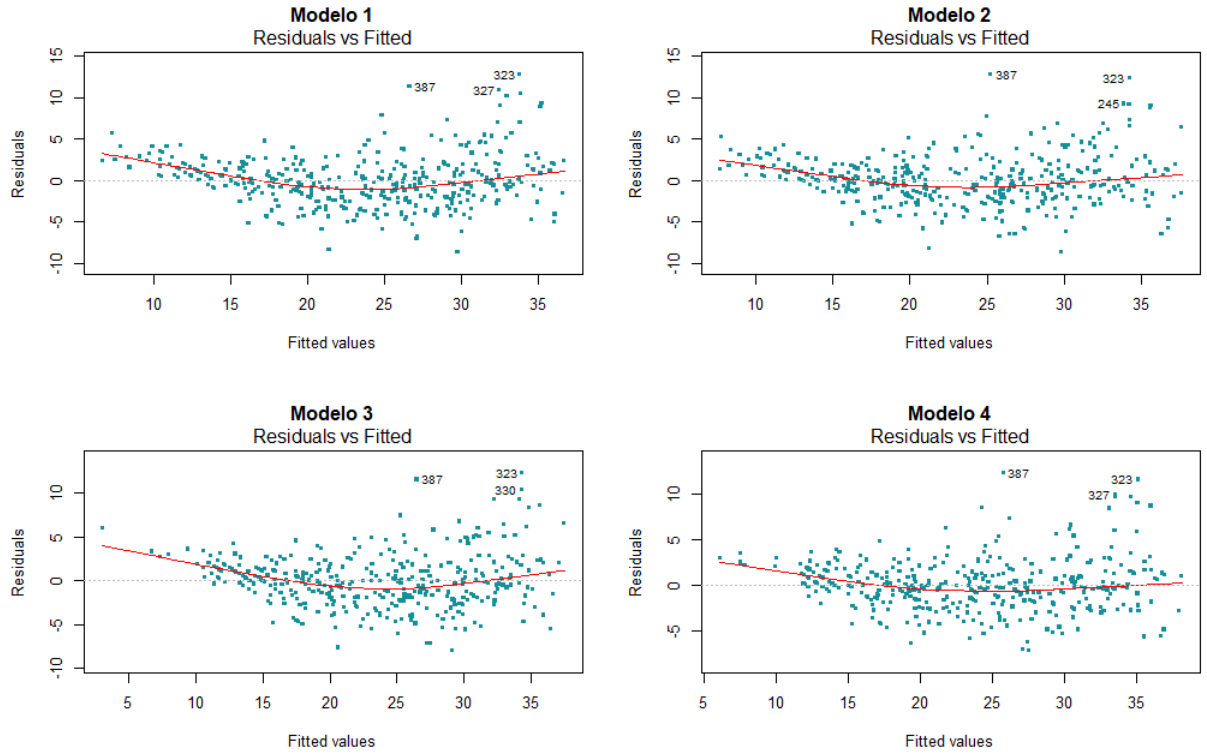

Figura 5: **Residuales de los modelos 1 al 4.** *El patrón de dispersión de los residuales es menos visible en los modelos 2 y 4, pero siguen siendo muy similares en general.*

Otra opción para mejorar el modelo es intentar con transformaciones de las variables predictoras. En los modelos 6, 7 y 8 se usaron las variables que ya se tenían en el modelo 3 pero se realizaron transformaciones a distintas variables originales y en la Tabla 2 se muestran los resultados.

Modelo	Transformación	R^2 ajustada
6	$\sqrt{\texttt{displacement}}$	0.8510
7	$\texttt{horsepower}^2$	0.8511
8	$log(\texttt{acceleration})$	0.8453

Cuadro 2: **Comparación de los modelos con transformaciones.***El valor de la R^2 ajustada crece muy poco respecto al modelo 3.*

Las gráficas de los residuales para los modelos 6, 7 y 8 casi no cambian con respecto a la gráfica del modelo 3 y dado que el aumento ganado en la R^2 es muy pequeño, considero que no es necesario complicar el modelo con transformaciones dado que no se gana mucho en el ajuste. El número de variables significativas en el modelo casi no cambia pues queda entre 9 y 11.

También se probó hacer estas transformaciones sobre los modelos 2 y 4 pero casi no aumentaba R^2 ajustada y en cambio pasabamos de 10 a solo 3 variables significativas en el modelo.

Por lo tanto, elegí como el modelo final de regresión lineal múltiple para la base Auto al modelo 3.

```
Call:
lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
    acceleration + weight:acceleration + year + origin, data = Auto)

Residuals:
    Min      1Q  Median      3Q     Max
-7.9994 -1.9358 -0.0442  1.4969 12.2548

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -4.094e+01  5.907e+00  -6.930 1.82e-11 ***
cylinders4           6.331e+00  1.610e+00   3.933 9.98e-05 ***
cylinders5           7.286e+00  2.447e+00   2.978 0.003089 **
cylinders6           4.036e+00  1.779e+00   2.268 0.023881 *
cylinders8           5.226e+00  2.050e+00   2.549 0.011205 *
displacement         7.072e-03  7.427e-03   0.952 0.341613
horsepower          -4.653e-02  1.308e-02  -3.557 0.000423 ***
weight               1.634e-03  1.661e-03   0.984 0.325948
acceleration         1.146e+00  2.501e-01   4.582 6.25e-06 ***
year                 7.639e-01  4.789e-02  15.951  < 2e-16 ***
origin2              1.531e+00  5.381e-01   2.846 0.004671 **
origin3              2.174e+00  5.208e-01   4.174 3.72e-05 ***
weight:acceleration -4.206e-04  8.756e-05  -4.804 2.25e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.011 on 379 degrees of freedom
Multiple R-squared:  0.8557,    Adjusted R-squared:  0.8511
F-statistic: 187.3 on 12 and 379 DF,  p-value: < 2.2e-16
```

Figura 6: **Resumen del modelo 3.**

1.2. Ejercicio 15

del capítulo 3 Linear Regression, de James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning. With applications in R, Springer, ISL*

En este ejercicio se trabajará con la base de datos `Boston` de la biblioteca **MASS** la cual contiene información de 14 variables sobre 506 comunidades en el estado de Boston. Las variables son:

- `crim`: Tasa de crimen per cápita.
- `zn`: Proporción de zona residencial por cada 25,000 pies cuadrados.
- `indus`: Proporción de acres ocupados por negocios no minoristas.
- `chas`: 1 si la comunidad limita con el río, 0 si no.
- `nox`: Concentración de óxido de nitrógeno medido en partes por 10 millones.
- `rm`: Promedio de habitaciones por vivienda.
- `dis`: Promedio ponderado de distancias a cinco centros de trabajo en Boston.
- `rad`: Índice de accesibilidad a carreteras radiales.
- `tax`: Tasa de impuestos por cada $10,000USD.
- `ptratio`: Cociente de alumnos-maestros.

- `black` $= 1000(Bk - 0{,}63)^2$ donde Bk es la proporción de personas de raza negra por ciudad.
- `lstat`: Porcentaje de personas de clase baja.
- `medv`: Valor medio de las casas ocupadas por su propetario en miles de dólares.

En esta base de datos todas las variables están registradas como numéricas pero dado que `chas` y `rad` son en realidad categóricas (con 2 y 9 niveles respectivamente) fueron convertidas a factores, con 1 como nivel de referencia de ambas variables.

a) Ajustamos modelos de regresión lineal simple con `crim` como la variable respuesta y cada una de las demás variables como variable predictora, por lo que en total tendremos 13 modelos de regresión lineal simple.

Los resultados del ajuste de estos modelos se encuentra en el Cuadro 3

Variable predictora	R2	sigma_estimada	Intercepto	p_value_B0	beta_1	p_value_B1
zn	0.04	8.44	4.45	0.00	-0.07	0.00
indus	0.17	7.87	-2.06	0.00	0.51	0.00
chas	0.00	8.60	3.74	0.00	-1.89	0.21
nox	0.18	7.81	-13.72	0.00	31.25	0.00
rm	0.05	8.40	20.48	0.00	-2.68	0.00
age	0.12	8.06	-3.78	0.00	0.11	0.00
dis	0.14	7.97	9.50	0.00	-1.55	0.00
rad	0.40	6.71	0.04	0.98	0.05	0.98
tax	0.34	7.00	-8.53	0.00	0.03	0.00
ptratio	0.08	8.24	-17.65	0.00	1.15	0.00
black	0.15	7.95	16.55	0.00	-0.04	0.00
lstat	0.21	7.66	-3.33	0.00	0.55	0.00
medv	0.15	7.93	11.80	0.00	-0.36	0.00

Cuadro 3: **Resumen de los modelos ajustados.** *El valor de R^2 es bajo para la mayoría de los modelos a pesar de que en casi todos la regresión lineal es estadísticamente significativa al nivel 0.05 .*

Los modelos que mejor explican la variabilidad de los datos son los que tienen `indus`, `nox`, `rad` y `tax` respectivamente como variables predictoras, lo que era de esperarse debido a su nivel de correlación con la variable respuesta. Los modelos que usan la variable `chas` y la variable `rad` que son las variables categóricas solo muestran los valores correspondientes para su primer nivel después del nivel de referencia por lo que pareciera que no pasan la prueba de significancia de la regresión sin embargo tendríamos que ver que pasa en los demás niveles.

En la Figura 7 se muestran el ajuste de algunos de los modelos (entre ellos los que tienen R^2 más alta), el cual no parece ser muy bueno.

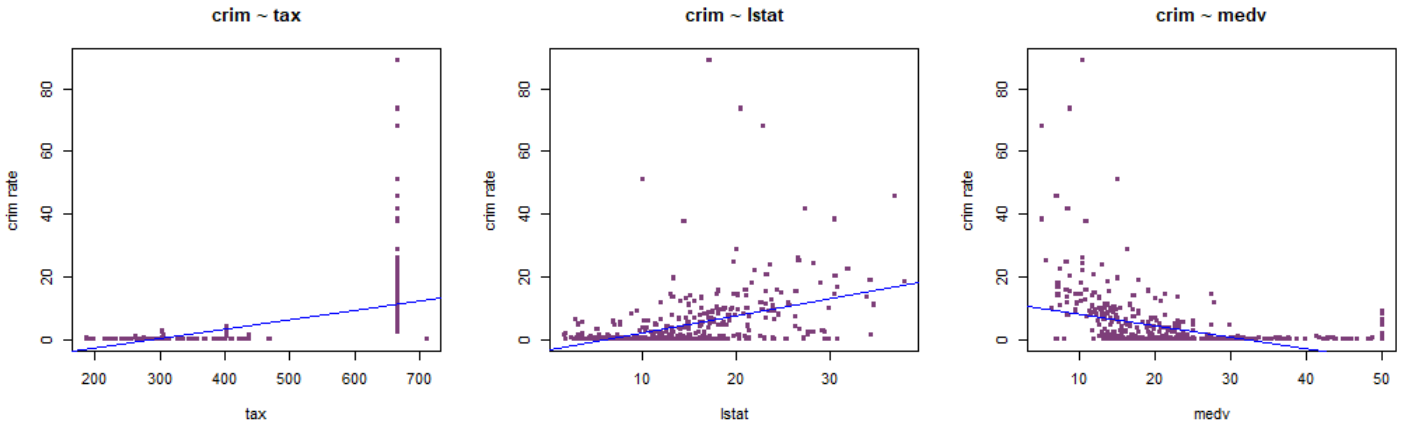

Figura 7: **Ajuste de los modelos de regresión lineal simple** *Las gráficas sugieren que algún reescalamiento de las variables podría mejorar el ajuste.*

Podemos intentar mejorar el ajuste en cada uno de los modelos aplicando la transformación logaritmo a la variable respuesta. Al hacer esto para cada modelo obtenemos los resultados que se resumen en el Cuadro 4. Podemos notar que el valor de la R^2 aumento la mayoría de los modelos y los que usan las variables `indus`, `nox`, `rad` y `tax` siguen siendo los que tienen el valor más alto y ahora solo para la varible asociada al segundo nivel de la variable categórica `chas` no se rechaza que β_1 sea cero. En la Figura 8 se nota gráficamente una mejoría en el ajuste de los modelos.

Variable predictora	R2	sigma_estimada	Intercepto	p_value_B0	beta_1	p_value_B1
zn	0.27	1.85	-0.24	0.01	-0.05	0.00
indus	0.53	1.48	-3.35	0.00	0.23	0.00
chas	0.00	2.16	-0.80	0.00	0.24	0.52
nox	0.62	1.33	-8.94	0.00	14.71	0.00
rm	0.09	2.06	5.16	0.00	-0.94	0.00
age	0.43	1.63	-4.25	0.00	0.05	0.00
dis	0.46	1.58	1.88	0.00	-0.70	0.00
rad	0.76	1.07	-3.57	0.00	0.82	0.01
tax	0.69	1.21	-5.12	0.00	0.01	0.00
ptratio	0.15	1.99	-7.96	0.00	0.39	0.00
black	0.23	1.90	3.26	0.00	-0.01	0.00
lstat	0.39	1.69	-3.18	0.00	0.19	0.00
medv	0.21	1.93	1.63	0.00	-0.11	0.00

Cuadro 4: **Resumen de los modelos con** $log($`crim`$)$ **como variable respuesta.**

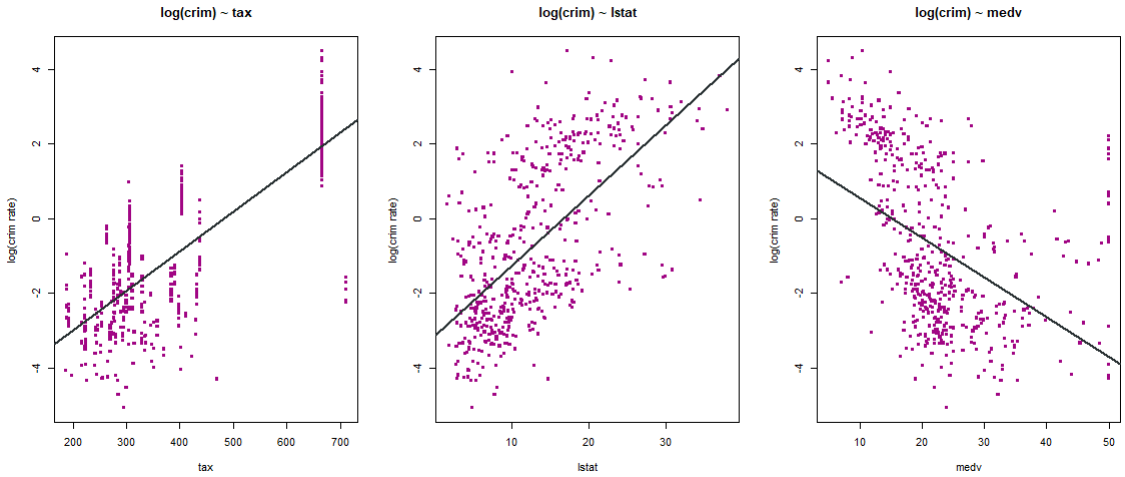

Figura 8: **Ajuste de los modelos con** *log*(`crim`) **como variable respuesta.** *Al igual que con el valor de* R^2 *podemos notar gráficamente una mejora en el ajuste con la transformación logaritmo de la variable respuesta.*

Para comprobar los supuestos del modelo de regresión lineal simple podemos observar las gráficas de los residuales de cada modelo. En la Figura 9 se observan las gráficas de comparación de cuantiles para algunos de los modelos y se nota que para el modelo original en todos hay problemas de ajuste especialmente el la cola de la distribución, lo cual nos indica que no se está cumpliendo el supuesto de distribución normal de los residuales. Al aplicar la transformación logaritmo a la variable respuesta se nota una mejora en el ajuste de las colas, sin embargo en el modelo de la variable `medv` el ajuste empeora en el centro de la distribución.

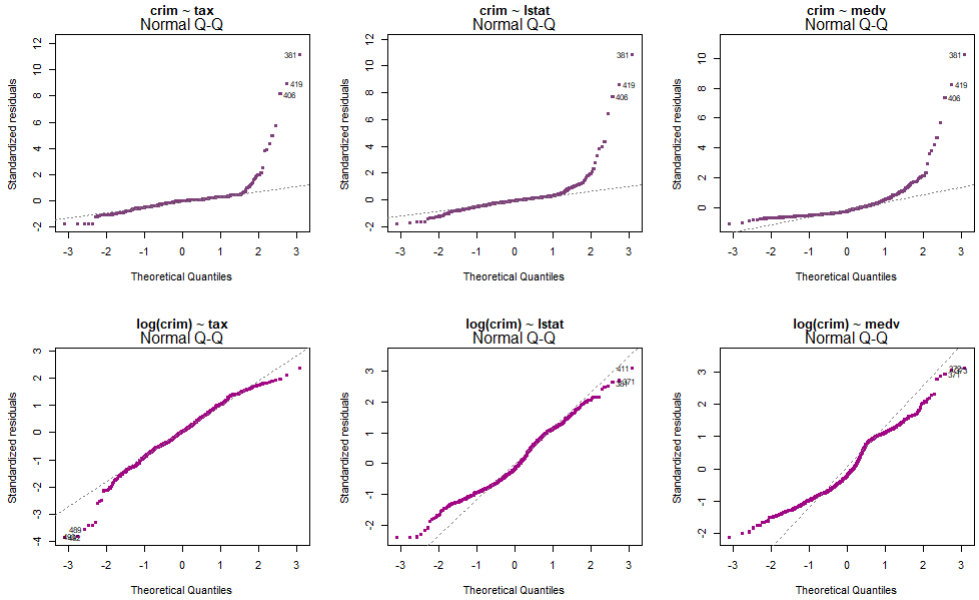

Figura 9: **Gráficas de comparación de cuantiles.***En general el ajuste en las colas mejora con la transformación logaritmo, aunque sigue sin ser totalmente satisfactorio.*

b) Ahora ajustaremos un modelo de regresión lineal múltiple en el que tomaremos en cuenta todas las variables de la base (numéricas y categóricas) en un mismo modelo. Para este modelo se estimaron: el intercepto, 11 parámetros correspondientes a las variables numéricas y 9 correspondientes a los diferentes niveles de las variables categóricas. En la Figura 10 se muestra el resumen de este modelo.

```
lm(formula = crim ~ ., data = Boston)

Residuals:
   Min     1Q Median     3Q    Max
 -9.91  -1.83  -0.27   0.93  74.82

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.28668    7.72082    2.76  0.00605 **
zn            0.03874    0.01966    1.97  0.04931 *
indus        -0.07884    0.08737   -0.90  0.36730
chas1        -0.75071    1.19561   -0.63  0.53038
nox         -10.81204    5.44114   -1.99  0.04747 *
rm            0.39765    0.62184    0.64  0.52281
age           0.00190    0.01817    0.10  0.91681
dis          -1.01633    0.29016   -3.50  0.00050 ***
rad2         -0.70404    2.03142   -0.35  0.72906
rad3          0.55521    1.85713    0.30  0.76510
rad4          0.20719    1.63888    0.13  0.89945
rad5          0.49125    1.66862    0.29  0.76858
rad6         -0.92578    2.01193   -0.46  0.64562
rad7          1.61448    2.17836    0.74  0.45897
rad8          1.60824    2.06982    0.78  0.43754
rad24        12.04502    2.44013    4.94  1.1e-06 ***
tax          -0.00312    0.00538   -0.58  0.56136
ptratio      -0.35118    0.20691   -1.70  0.09028 .
black        -0.00703    0.00369   -1.91  0.05708 .
lstat         0.12199    0.07680    1.59  0.11282
medv         -0.20533    0.06167   -3.33  0.00094 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.45 on 485 degrees of freedom
Multiple R-squared:  0.46,      Adjusted R-squared:  0.438
F-statistic: 20.7 on 20 and 485 DF,  p-value: <2e-16
```

Figura 10: **Resumen del modelo de regresión lineal múltiple**

Podemos notar que a diferencia de los modelos de regresión lineal simple (sin transformación) en los que solo los parámetros del segundo nivel de ambas variables categóricas no eran significativos y los 11 correspondientes a las variables continuas sí lo eran, ahora tenemos que solo 4 parámetros correspondientes a las variables continuas y 1 correspondiente al último nivel de una variable categórica son significativos y el resto son estadísticamente cero. Esto nos dice que en el modelo de regresión lineal múltiple, las variables predictoras no influyen tanto como lo hicieran por separado.

Podemos ahora probar con la transformación logaritmo para la variable respuesta al igual que lo hicimos con los modelos simples. El resumen de este modelo se muestra en la Figura 11 y en él podemos ver que el número de parámetros significativos aumenta considerablemente al igual que la R^2 ajustada pues pasa de 0.46 a 0.88.

```
lm(formula = log(crim) ~ ., data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3069 -0.4780 -0.0474  0.4554  2.4843

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.07e+00   8.81e-01   -3.49  0.00053 ***
zn          -1.13e-02   2.24e-03   -5.05  6.2e-07 ***
indus        1.09e-02   9.97e-03    1.09  0.27531
chas1       -1.53e-01   1.36e-01   -1.12  0.26280
nox          3.96e+00   6.21e-01    6.39  4.0e-10 ***
rm          -5.12e-02   7.09e-02   -0.72  0.47065
age          6.19e-03   2.07e-03    2.99  0.00295 **
dis         -2.85e-02   3.31e-02   -0.86  0.38890
rad2         1.66e-01   2.32e-01    0.71  0.47532
rad3         6.54e-01   2.12e-01    3.09  0.00213 **
rad4         1.29e+00   1.87e-01    6.92  1.4e-11 ***
rad5         9.21e-01   1.90e-01    4.84  1.8e-06 ***
rad6         7.24e-01   2.29e-01    3.15  0.00172 **
rad7         1.48e+00   2.48e-01    5.96  5.0e-09 ***
rad8         1.78e+00   2.36e-01    7.53  2.4e-13 ***
rad24        3.78e+00   2.78e-01   13.57  < 2e-16 ***
tax          6.98e-05   6.13e-04    0.11  0.90938
ptratio     -8.56e-02   2.36e-02   -3.63  0.00032 ***
black       -1.49e-03   4.21e-04   -3.54  0.00043 ***
lstat        2.78e-02   8.76e-03    3.17  0.00163 **
medv         6.38e-03   7.03e-03    0.91  0.36484
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.735 on 485 degrees of freedom
Multiple R-squared:  0.889,     Adjusted R-squared:  0.884
F-statistic:  194 on 20 and 485 DF,  p-value: <2e-16
```

Figura 11: **Resumen del modelo con** *log*(`crim`) **como variable respuesta.**

Para verificar los supuestos del modelo podemos ver las gráficas de los residuales en la Figura 12 donde en el modelo sin transformación notamos que aunque la media de los residuales es cercana a cero, la varianza parece ser creciente y en la gráfica de comparación de cuantiles la cola de la distribución se aleja bastante de la normal. Al aplicar la tranformación el ajuste mejora en ambos aspectos, sobre todo en los cuantiles, que resultan ser muy similares a los de la normal incluso en las colas.

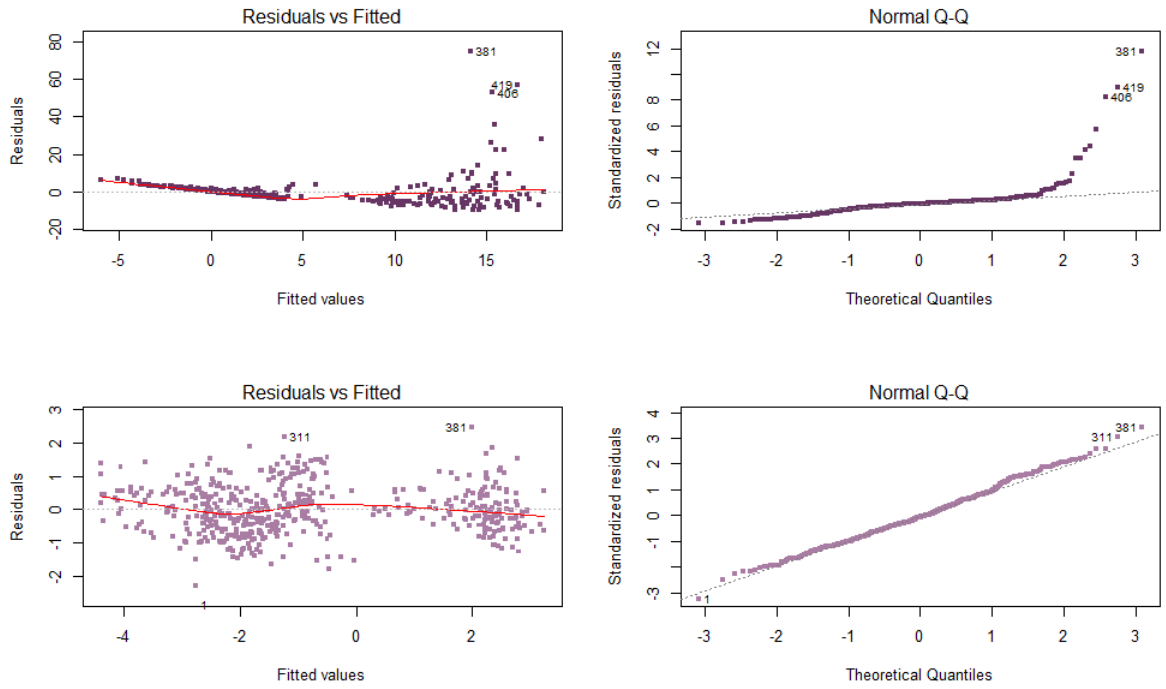

Figura 12: **Residuales del modelo de regresión lineal múltiple con y sin transformación.**

c) Para comparar los resultados obtenidos en **a)** y **b)** graficamos el valor de los parámetros estimados para cada variable predictora en ambos modelos (el simple y el múltiple) y para el modelo con las variables originales y el que usó la $log($`crim`$)$ como variable respuesta.

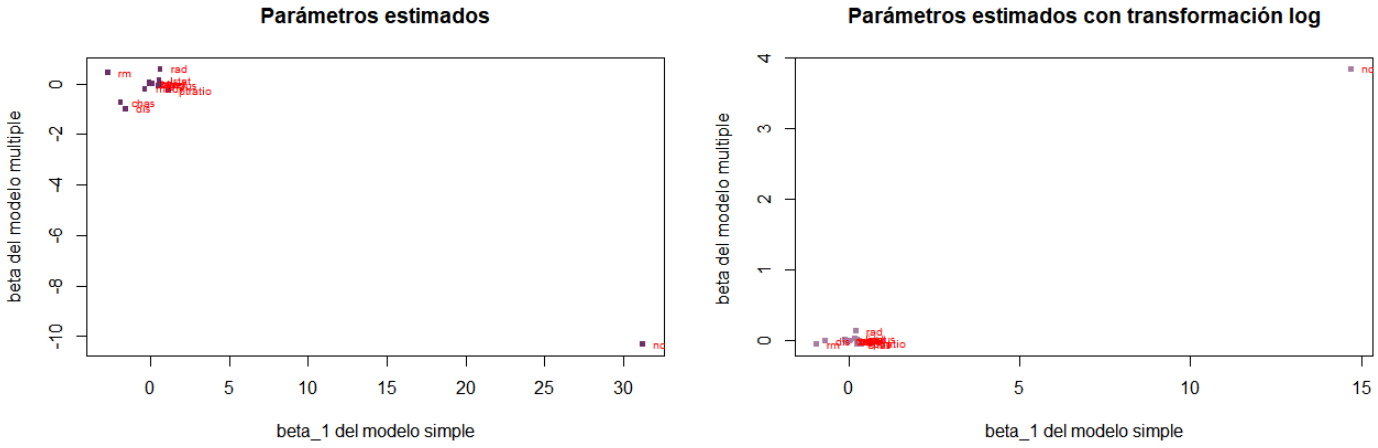

Figura 13: **Parámetros estimados en el modelo simple vs en el modelo múltiple.**

Podemos notar que los parámetros con valores cercanos a cero tienen valores estimados cercanos para el modelo simple y el modelo múltiple, pero el de la variables `nox` es el que más difieren. El parámetro asociado a `nox` es considerablemente más alto en el modelo simple (tiene más influencia que en el modelo múltiple).

d) Para verificar si hay dependencia no lineal entre las variables predictoras y la respuesta ajustamos un modelo de la forma:

$$crim = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$

para cada variable continua de la base. El resumen del ajuste de estos modelos se encuentra en el Cuadro 5

	R^2	$\sigma_{estimada}$	Intercepto	β_1	p-value β_1	β_2	p-value β_2	β_3	p-value β_3
zn	0.06	8.37	4.85	-0.33	0.00	0.01	0.09	-0.00	0.23
indus	0.26	7.42	3.66	-1.97	0.00	0.25	0.00	-0.01	0.00
nox	0.30	7.23	233.09	-1279.37	0.00	2248.54	0.00	-1245.70	0.00
rm	0.07	8.33	112.62	-39.15	0.21	4.55	0.36	-0.17	0.51
age	0.17	7.84	-2.55	0.27	0.14	-0.01	0.05	0.00	0.01
dis	0.28	7.33	30.05	-15.55	0.00	2.45	0.00	-0.12	0.00
tax	0.37	6.85	19.18	-0.15	0.11	0.00	0.14	-0.00	0.24
ptratio	0.11	8.12	477.18	-82.36	0.00	4.64	0.00	-0.08	0.01
black	0.15	7.95	18.26	-0.08	0.14	0.00	0.47	-0.00	0.54
lstat	0.22	7.63	1.20	-0.45	0.33	0.06	0.06	-0.00	0.13
medv	0.42	6.57	53.17	-5.09	0.00	0.16	0.00	-0.00	0.00

Cuadro 5: **Resumen de los modelos con transformaciones no lineales.**

Podemos notar que para algunas variables, como `indus` y `nox`, los dos parámetros extras que añadimos sí son significativos, es decir, estas varibles sí tienen una relación no lineal con la varible respuesta; mientras que variables como `rm` no rechazan la prueba para sus parámetros β_2 y β_3, es decir, estos son estadísticamente cero, por lo que su relación con la variable respuesta no indica ser de mayor grado. Por otro lado, en la variable `lstat`, β_2 no es significativa a un nivel de 0.05 pero sí lo es a un nivel de 0.10, sin embargo β_3 ya no es significativa ni siquiera a un nivel de 0.10, por lo que con este nivel de significancia podríamos decir que el cuadrado de `lstat` es influyente en la respuesta pero su cubo ya no lo es. Sucede lo contrario en el caso de la variable `age` pues β_2 es significativa a un nivel de 0.05 pero no al nivel 0.01 y en cambio β_3 sí lo es en ambos niveles, es decir, hay más evidencia de que el cubo de la edad sea significativo a que lo sea su cuadrado.

2. Ejercicio 2

La base de datos con la que trabajamos en este ejercicio es `"June_13_data.csv"` la cual tiene 23137 observaciones de 14 variables relacionadas con incidentes automovilísticos registrados entre 2014 y 2019. Contiene información sobre el mes, la hora del accidente y las condiciones climatológicas presentes así como las condiciones del camino donde sucedió.

En este ejercicio consideraremos solo las siguientes 5 variables de la base:

- `Crash_Score`: Grado del accidente. Se calcula tomando en cuenta el número de heridos y muertos y el número de vehículos involucrados, entre otros factores.
- `Time_of_Day`: Bloque en el que se encuentra la hora en la que sucedió el accidente. Se dividen las 24 horas del día en 6 bloques de 4 horas consecutivas empezando por medianoche, por lo que si el accidente ocurrió entre las 20 horas y medianoche, se registra con el número 6.
- `Rd_Feature`: Caracterísitica especial del camino donde ocurrió el accidente (NONE, INTERSECTION, RAMP, DRIVEAWAY, OTHER).
- `Rd_Congiguration`: Diseño del camino (TWO-WAY-PROTECTED-MEDIAN, TWO-WAY-UNPROTECTED-MEDIAN, TWO-WAY-NO-MEDIAN, ONE-WAY, UNKNOWN).

- `Traffic_Control`: Tipo de señales sobre el camino para control el tráfico (SIGNAL, STOP-SIGN, YIELD, NONE, OTHER).

Para tratar de explicar la variable `Crash_Score` se ajustará un modelo de regresión lineal múltiple con las variables `Time_of_Day`, `Rd_Feature`, `Rd_Congiguration` y `Rd_Congiguration` como variables predictoras, las cuales son todas categóricas, por lo que en `R` se tratarán como factores. Nos referiremos de aquí en adelante a este modelo como Modelo 1. El resumen de este primer modelo se muestra en la Figura 14 y en este podemos ver que el p-value de la regresión es muy cercano a cero, lo que nos dice que la regresión sí es significativa, sin embargo solo hay 9 niveles de las variables categóricas que son significativas al nivel 0.05.

```
Call:
lm(formula = Crash_Score ~ ., data = june_13)

Residuals:
   Min     1Q Median     3Q    Max
 -7.21  -3.00  -0.88   2.00  46.86

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                                5.52141    0.20638   26.75  < 2e-16 ***
Time_of_Day2                               0.43487    0.18342    2.37   0.0178 *
Time_of_Day3                               0.69017    0.16210    4.26  2.1e-05 ***
Time_of_Day4                               0.63991    0.15854    4.04  5.4e-05 ***
Time_of_Day5                               0.79720    0.15891    5.02  5.3e-07 ***
Time_of_Day6                               0.43740    0.17374    2.52   0.0118 *
Rd_FeatureINTERSECTION                     0.36437    0.12138    3.00   0.0027 **
Rd_FeatureNONE                             0.00434    0.09761    0.04   0.9646
Rd_FeatureOTHER                           -0.02473    0.28425   -0.09   0.9307
Rd_FeatureRAMP                            -0.31118    0.19748   -1.58   0.1151
Rd_ConfigurationTWO-WAY-NO-MEDIAN          0.03328    0.12791    0.26   0.7947
Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN   0.44770    0.14630    3.06   0.0022 **
Rd_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN 0.21234    0.13415    1.58   0.1135
Rd_ConfigurationUNKNOWN                    0.10654    0.58227    0.18   0.8548
Traffic_ControlOTHER                       0.31450    0.28866    1.09   0.2759
Traffic_ControlSIGNAL                      0.47734    0.08546    5.59  2.4e-08 ***
Traffic_ControlSTOP-SIGN                   0.42616    0.10716    3.98  7.0e-05 ***
Traffic_ControlYIELD                       0.26069    0.27754    0.94   0.3476
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.26 on 23119 degrees of freedom
Multiple R-squared:  0.0105,    Adjusted R-squared:  0.00975
F-statistic: 14.4 on 17 and 23119 DF,  p-value: <2e-16
```

Figura 14: Resumen del Modelo 1.

Cuando los niveles no tienen un parámetro significativo, por ejemplo `Rd_FeatureNONE` nos dice que la media de este nivel es muy parecida a la media del nivel de referencia, por lo que podríamos no hacer distinción entre estos dos niveles.

Otro aspecto importante es que nuestra R^2 ajustada es muy pequeña, de 0.00975, es decir, la regresión solo explica un 0.975 % de la variabilidad de los datos, por lo que podemos intentar mejorar el ajuste.

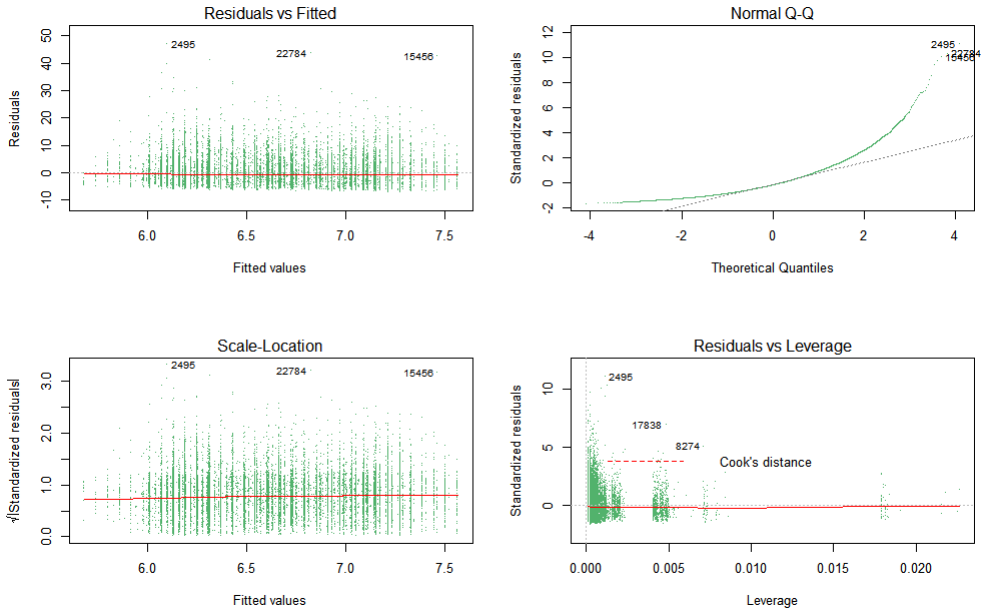

Figura 15: **Gráficas de residuales del Modelo 1.** *Aunque los residuales sí parecen tener media cero, la comparación con los cuantiles de la normal nos dice que no cumplen el supuesto de normalidad.*

Probamos otro modelo pero ahora transformando la variable respuesta. Usamos la función `boxcox` en `R` que nos regresa el valor de `lambda` que maximiza la log-verosimilitud. En este caso nos dio $\lambda = 0{,}263$, por lo que decidimos probar con $\texttt{Crash\_Score}^{0{,}263}$ (Modelo 2) y también con $log(\texttt{Crash\_Score})$ (Modelo 3). Estos dos modelos nos daban resultados muy similares: para el Modelo 2 la R^2 ajustada es de 0.0117 y para el Modelo 3 es de 0.0116 por lo que la diferencia entre estos ajustes es muy poca y en la Figura 16 se muestran los residuos de estos tres modelos, en los que parece mejorar el ajuste en la cola derecha pero en la cola izquierda sigue siendo insatisfactorio. Por facilidad de interpretación se eligió el Modelo 3.

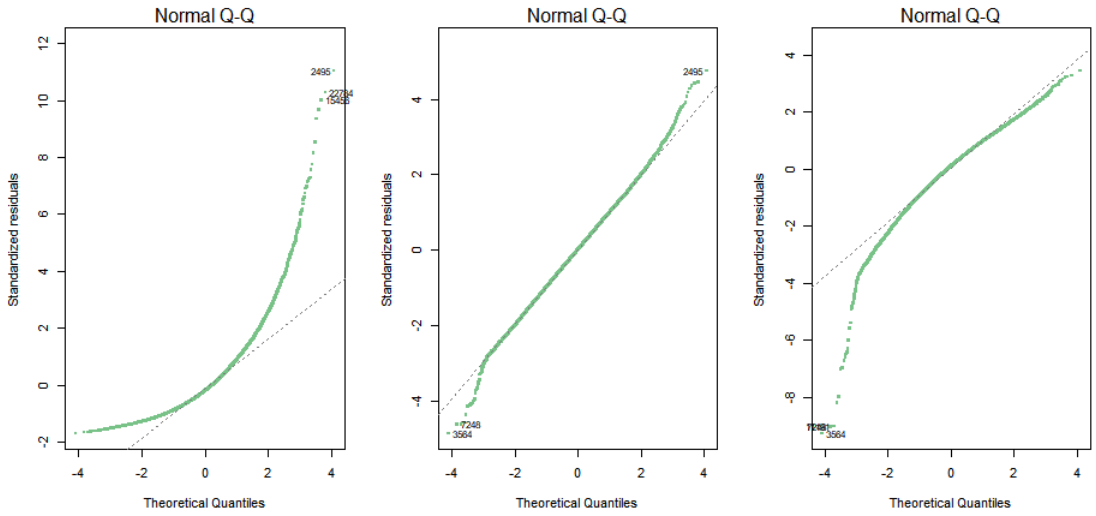

Figura 16: **Comparación de los residuales de los modelos 1,2 y 3.**

Usamos la función `drop1()` en `R` para verificar si todas nuestras variables predictoras son significativas.

```
Single term deletions

Model:
Crash_Score ~ Time_of_Day + Rd_Feature + Rd_Configuration + Traffic_Control
                  Df Sum of Sq    RSS   AIC F value  Pr(>F)
<none>                        419169 67060
Time_of_Day        5       669 419839 67087    7.38 6.3e-07 ***
Rd_Feature         4       480 419650 67079    6.62 2.5e-05 ***
Rd_Configuration   4       415 419585 67075    5.73 0.00013 ***
Traffic_Control    4       634 419803 67087    8.74 4.8e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 17: **Resultados de la función `drop1()` del modelo 3.** *Podemos ver que todas las variables predictoras son significativas al nivel 0.05, aunque no lo sean todos los niveles de cada una.*

```
Call:
lm(formula = log(Crash_Score) ~ ., data = june_13)

Residuals:
   Min     1Q Median     3Q    Max
-6.389 -0.404  0.065  0.475  2.341

Coefficients:
                                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  1.473903   0.033352   44.19  < 2e-16 ***
Time_of_Day2                                 0.103923   0.029642    3.51  0.00046 ***
Time_of_Day3                                 0.156862   0.026196    5.99  2.2e-09 ***
Time_of_Day4                                 0.153968   0.025621    6.01  1.9e-09 ***
Time_of_Day5                                 0.175348   0.025681    6.83  8.8e-12 ***
Time_of_Day6                                 0.115299   0.028078    4.11  4.0e-05 ***
Rd_FeatureINTERSECTION                       0.038666   0.019615    1.97  0.04871 *
Rd_FeatureNONE                              -0.011450   0.015774   -0.73  0.46793
Rd_FeatureOTHER                             -0.077927   0.045937   -1.70  0.08983 .
Rd_FeatureRAMP                              -0.058099   0.031914   -1.82  0.06870 .
Rd_ConfigurationTWO-WAY-NO-MEDIAN            0.000273   0.020671    0.01  0.98948
Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN     0.049231   0.023644    2.08  0.03734 *
Rd_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN   0.015852   0.021680    0.73  0.46468
Rd_ConfigurationUNKNOWN                      0.033878   0.094099    0.36  0.71883
Traffic_ControlOTHER                         0.039592   0.046649    0.85  0.39604
Traffic_ControlSIGNAL                        0.101778   0.013812    7.37  1.8e-13 ***
Traffic_ControlSTOP-SIGN                     0.085346   0.017317    4.93  8.4e-07 ***
Traffic_ControlYIELD                         0.058675   0.044852    1.31  0.19083
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.688 on 23119 degrees of freedom
Multiple R-squared:  0.0123,    Adjusted R-squared:  0.0116
F-statistic:   17 on 17 and 23119 DF,  p-value: <2e-16
```

Figura 18: **Resumen del Modelo 3**

Guiándonos del resumen del Modelo 3, decidimos agrupar las categorías NONE, OTHER y RAMP con la categoría de referencia de la variable `Rd_Feature` y dejar en otro grupo la categoría INTERSECTION de la misma variable. Para la variable `Rd_Configuration` se decidió agrupar TWO-WAY-NO-MEDIAN, TWO-WAY-UNPROTECTED-MEDIAN y UNKNOWN con la categoría de referencia. Por último, para la variable `Traffic_Control` se agruparon las categorías OTHER y YIELD con la de referencia.

Cada una de estas agrupaciones se hizo paso por paso, es decir, agrupamos las categorías de una variable y comparamos el modelo obtenido de esta forma con el modelo anterior (el que tenía todas las categorías por separado) y usando la función `anova()` comparamos estos dos modelos. Como el p-value de la prueba era mayor a 0.05 en cada caso, nos quedamos con el modelo con menos categorías. Al final llegamos a que el Modelo 6 reducido quedó de la siguiente forma:

```
Call:
lm(formula = log(Crash_Score) ~ ., data = june_13)

Residuals:
    Min      1Q  Median      3Q     Max
-6.382 -0.404   0.064   0.475   2.347

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                                1.4659     0.0245   59.93  < 2e-16 ***
Time_of_Day2                               0.1051     0.0296    3.55  0.00039 ***
Time_of_Day3                               0.1584     0.0262    6.05  1.5e-09 ***
Time_of_Day4                               0.1553     0.0256    6.07  1.3e-09 ***
Time_of_Day5                               0.1769     0.0257    6.89  5.7e-12 ***
Time_of_Day6                               0.1170     0.0281    4.17  3.1e-05 ***
Rd_FeatureINTERSECTION                     0.0596     0.0128    4.64  3.5e-06 ***
Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN   0.0417     0.0144    2.90  0.00373 **
Traffic_ControlSIGNAL                      0.0965     0.0129    7.49  7.3e-14 ***
Traffic_ControlSTOP-SIGN                   0.0807     0.0171    4.72  2.4e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.688 on 23127 degrees of freedom
Multiple R-squared:  0.0119,    Adjusted R-squared:  0.0116
F-statistic: 31.1 on 9 and 23127 DF,  p-value: <2e-16
```

Figura 19: **Resumen del modelo reducido.** *En este modelo todos los parámetros estimados son significativos y la R^2 ajustada es la misma que la del modelo completo (Modelo 3)*

Las gráficas de los residuales de este modelo se encuentran en la Figura 20 y en ellas podemos ver que la media de los residuales parece ser cero y podríamos decir que tienen varianza constante, pues no parecen seguir un patrón. Por otro lado, la comparación de cuantiles con la normal no es satisfactoria en la cola izquierda de la distribución, pero en el centro y la cola derecha se ve bastante bien y es mejor que lo que se observaba en el primer modelo.

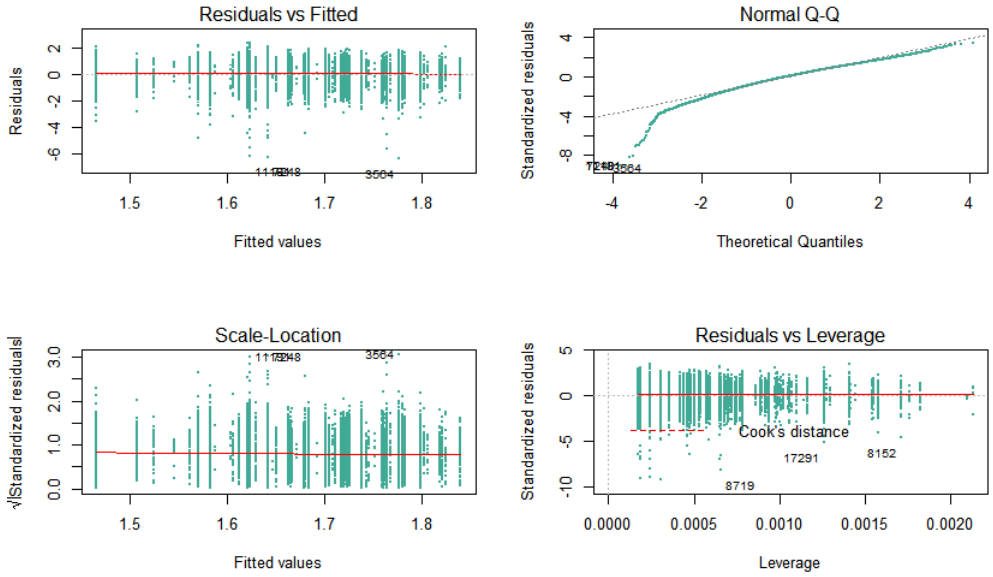

Figura 20: **Residuales del Modelo 6.**

Interpretación del modelo final (Modelo 6)

$$\begin{aligned} log(\hat{Crash_Score}) = \hat{\beta}_0 &+ \hat{\beta}_1(\text{Time of day 2}) + \hat{\beta}_2\text{Time of day 3}) + \hat{\beta}_3(\text{Time of day 4}) + \hat{\beta}_4(\text{Time of day 5}) + \hat{\beta}_5(\text{Time of day 6}) \\ &+ \hat{\beta}_6(\text{Rd Feat Intersection}) + \hat{\beta}_7(\text{RD Config Two way protected median}) \\ &+ \hat{\beta}_8(\text{Traffic Control Signal}) + \hat{\beta}_9(\text{Trafic Control Stop Sign}) \end{aligned}$$

donde cada varible asociada a las β_i, $i \in 1, ..., 9$ son variables que valen 1 si la observación pertenece a la categoría que indican y 0 en el otro caso. Cuando estas variables valen 0, quiere decir que la observación pertenece a la categoría de referencia de la variable correspondiente.

Por ejemplo, para una observación en la que el accidente ocurrió entre las 12am y las 4am (bloque 1 de `Time_of_Day`) en un camino que no tenía ninguna característica especial (categoría NONE de `Rd_Feature`), era de una dirección (categoría ONE-WAY de `Rd_Configuration`) y no tenía señales de tráfico (categoría NONE de `Traffic_Control`), entonces todas las variables asociadas a las β_i de 1 a 9 son 0, por lo que la estimación de log(`Crash_Score`) es un valor de $\hat{\beta}_0 = 1{,}47$. Si, por ejemplo, otra observación tiene estas mismas características excepto que había una señal de alto en el camino (categoría STOP SIGN de `Traffic_Control`) entonces el valor estimado de log(`Crash_Score`) es $\hat{\beta}_0 + \hat{\beta}_9 = 1{,}47 + 0{,}08 = 1{,}55$.

A pesar de que todas las variables que intervienen en este modelo son significativas, la regresión explica un porcentaje bajo de la variabilidad de los datos, por lo que el poder predictivo del mismo es poco. Una forma de mejorar esto podría ser ajustando modelos que tomen en cuenta las interacciones entre las variables.