

SEGUNDO TRABAJO
INTRODUCCIÓN A LA ANALÍTICA
MÓDULO 2

ELABORADO POR:
SANTIAGO FRANCO VALENCIA
DAVID JARAMILLO CALLE
DUVAN CAMILO MANRIQUE VARGAS
VALENTINA VANEGAS CASTAÑO

DOCENTE
MAURICIO ALEJANDRO MAZO LOPERA

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE MEDELLÍN
FACULTAD DE CIENCIAS

2022

Contents

Punto 2	2
Variables mas relevantes	2
Significancia en los parametros	3
cross validation para escoger el mejor lambda.	3
Coeficientes significativos.	4
Modelos	6
Punto 3	8
Regresión local	8
Punto 5	10
Actividad 1	10
Actividad 2	16
Datos	17
Análisis descriptivo	17
Modelo	20

Table 1: Tiempo de ejecución

	user.self	sys.self	elapsed
txt	29.39	0.59	30.08
RData	1.17	0.17	1.34

Punto 2

Cargue la base de datos en R, guardela como .RData y luego carguela nuevamente. ¿Cuál fue la reducción en tamaño del archivo?

Al guardar la base de datos con formato .RData el tamaño pasa de 188 a 89 MB, reduciéndose en un poco más de la mitad. Además, como se muestra en la siguiente tabla, el tiempo de ejecución o de carga usando .RData resulta ser mucho más eficiente y rápido. En este sentido, al trabajar con bases de datos de gran tamaño es recomendable usar esta herramienta.

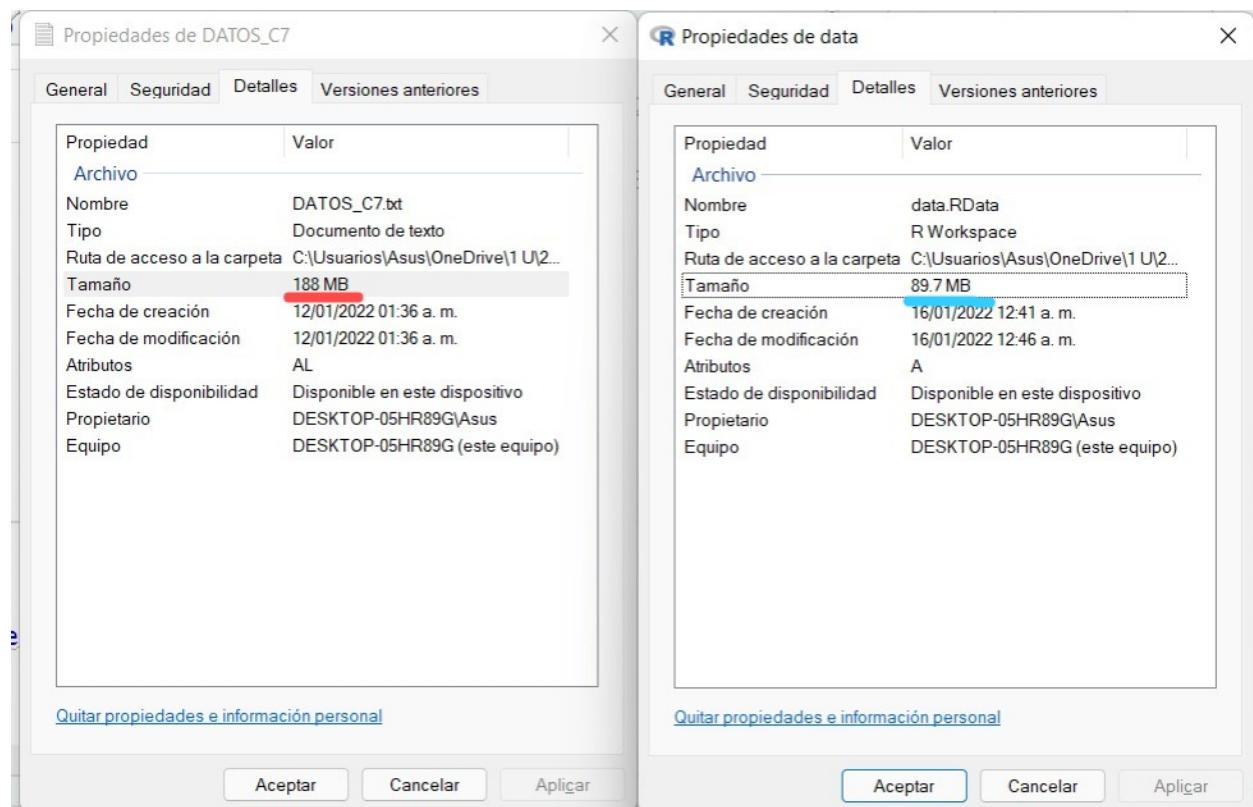


Figure 1: Tamaño de las bases de datos

Variables mas relevantes

Realice un análisis para seleccionar las variables más relevantes para explicar Y.

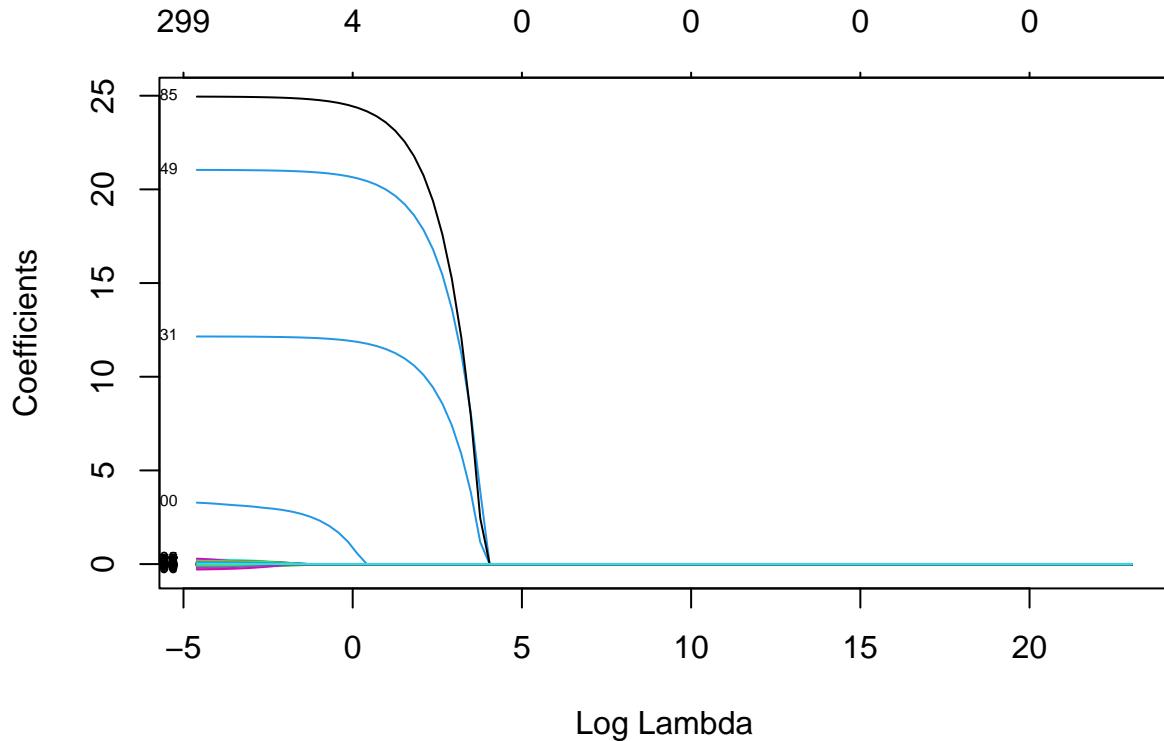
La base de datos contiene **89717** observaciones y **311** variables, por tanto como método para seleccionar las covariables que puedan explicar la variable independiente se usará la regresión **Lasso**.

Para la estimación del modelo Lasso se minimiza RSS_{Lasso} en función de los parámetros β_j .

$$RSS_{Lasso} = \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \sum_{j=1}^p |\beta_j|$$

Significancia en los parametros

Se puede observar que incluyendo el intercepto hay cuatro parámetros significativos.

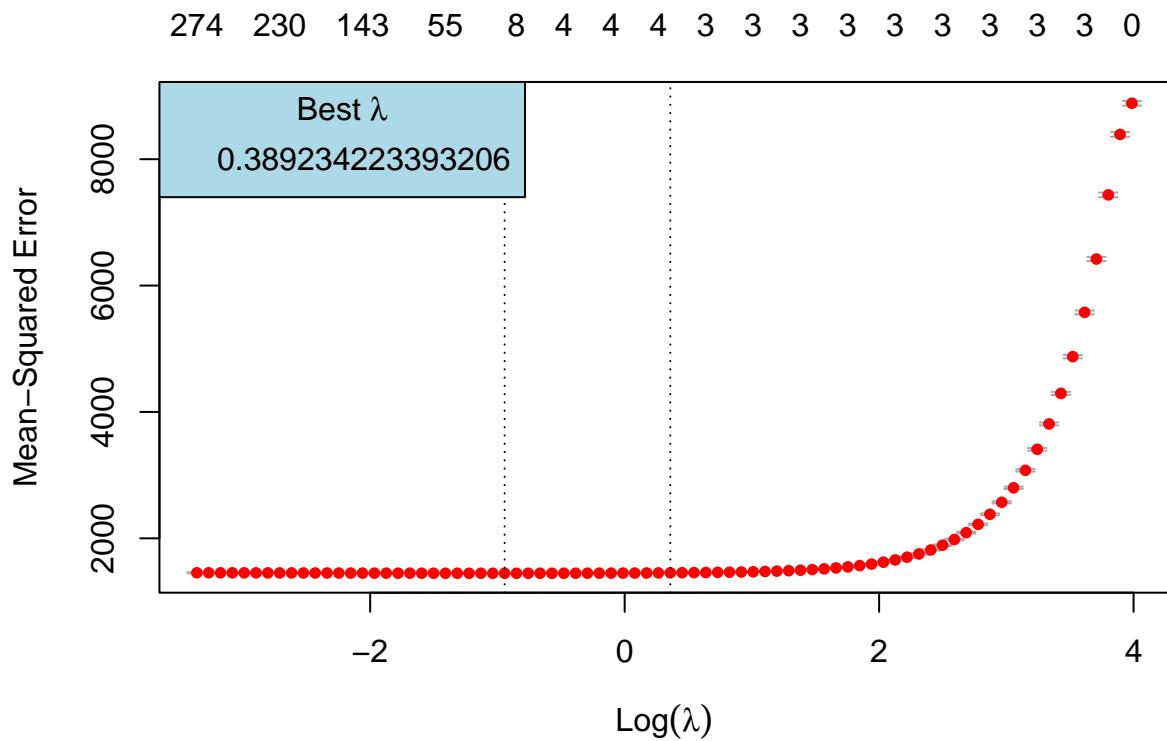


cross validation para escoger el mejor lambda.

Usando cross validation, se encuentra que, para este ejercicio, la estimación del parámetro de calibración es $\lambda = 0.3892$

Table 2: Coeficientes Regression Lasso

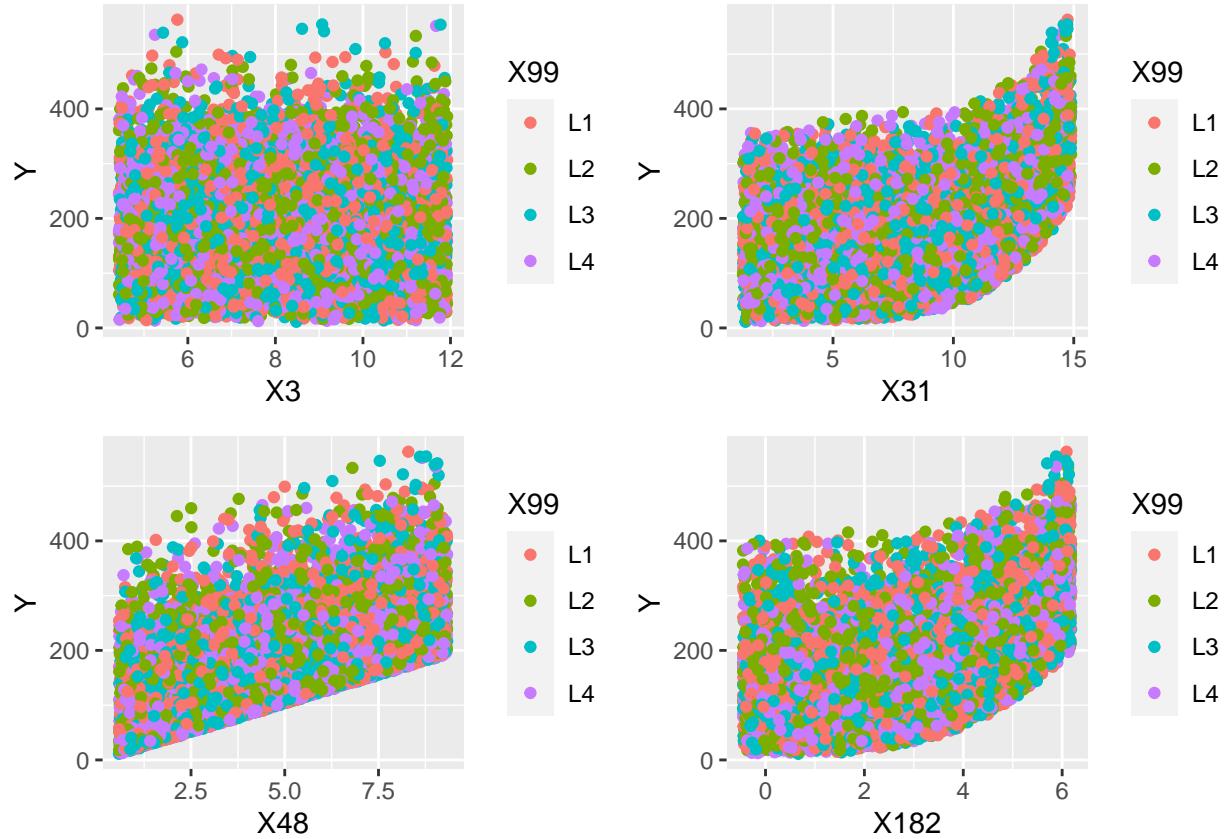
	Coef
(Intercept)	-83.855370
X31	12.051588
X48	20.891127
X99L2	2.290708
X182	24.755739



Coefficientes significativos.

Se encontró que las variables X31, X48, X182 y X99 resultaron significativas.

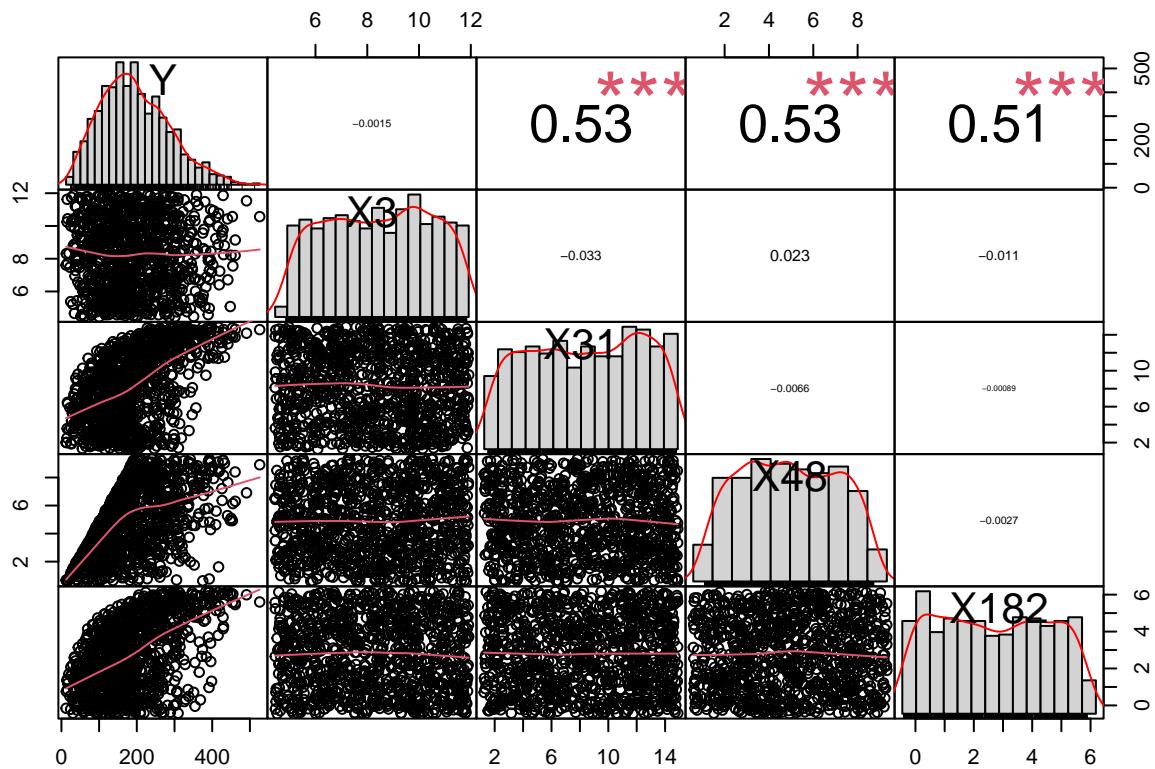
Selección de variables Grafique las variables más relevantes versus \mathbf{Y} y ajuste un modelo. ¿El comportamiento \mathbf{Y} es lineal con todas las variables explicativas?



El comportamiento de las variables se puede describir de la siguiente manera.

- En X_3 no se observa ningún tipo de comportamiento.
- En X_{31} y X_{182} se observa un tipo de comportamiento polinomial.
- En X_{48} se observa comportamiento lineal.
- Por último, la variable categórica no parece incorporar algún comportamiento o patrón dentro de las combinaciones hechas.

Matriz de correlación Excluyendo a X_3 hay correlación entre las X_i y Y , ademas hay independencia entre las variables regresoras. Esto resulta ser apropiado para ajustes de modelos lineales.



Modelos

Ajuste un modelo polinómico y un modelo con funciones paso. Compare ambos modelos. ¿Cuál seleccionaría como el mejor modelo?

Modelo lineal En el modelo lineal se ajusta la siguiente ecuación

$$Y_t = \beta_1 \cdot X3_t + \beta_2 \cdot X31_t + \beta_3 \cdot X48_t + \beta_4 \cdot X182_t + \beta_5 \cdot X99L1t + \beta_6 \cdot X99L2t + \beta_7 \cdot X99L3t + \beta_8 \cdot X99L4t + \epsilon_t$$

	Estimate	Std. Error	t value	Pr(> t)
X3	-0.2399	0.05278	-4.546	5.486e-06
X31	12.17	0.02898	419.9	0
X48	21.05	0.04532	464.5	0
X99L1	-84.49	0.6078	-139	0
X99L2	-81.16	0.6101	-133	0
X99L3	-84.43	0.6077	-138.9	0
X99L4	-84.03	0.6093	-137.9	0
X182	24.9	0.05957	418	0

Table 4: Fitting linear model: $Y \sim -1 + X3 + X31 + X48 + X99 + X182$

Observations	Residual Std. Error	R^2	Adjusted R^2
112146	38.23	0.967	0.967

Modelo polinomico Considerando el análisis descriptivo y el comportamiento de los datos, para las variables $X31$ y $X182$ se considerarán para aplicarles polinomios de grado dos. En este sentido, la ecuación queda como sigue.

$$Y_t = \beta_1 \cdot X3t + \beta_2 \cdot X31t + \beta_3 \cdot X31_t^2 + \beta_4 \cdot X48t + \beta_5 \cdot X182t + \beta_6 \cdot X182_t^2 + \beta_7 \cdot X99_{L1t} + \beta_8 \cdot X99_{L2t} + \beta_9 \cdot X99_{L3t} + \beta_{10} \cdot X99_{L4t} + \epsilon_t$$

	Estimate	Std. Error	t value	Pr(> t)
poly(X31, degree = 2)1	16042	11.27	1423	0
poly(X31, degree = 2)2	9591	11.27	850.8	0
X48	20.99	0.01336	1571	0
X99L1	83.29	0.09478	878.8	0
X99L2	86.36	0.09456	913.3	0
X99L3	83.39	0.09396	887.5	0
X99L4	83.43	0.09437	884	0
poly(X182, degree = 2)1	16036	11.27	1422	0
poly(X182, degree = 2)2	7581	11.27	672.4	0

Table 6: Fitting linear model: $Y \sim -1 + \text{poly}(X31, \text{degree} = 2) + X48 + X99 + \text{poly}(X182, \text{degree} = 2)$

Observations	Residual Std. Error	R^2	Adjusted R^2
112146	11.27	0.9971	0.9971

modelo con funciones paso

$$Y_t = \sum_{i=1}^5 \beta_i bi(X3_t) + \sum_{i=6}^{11} \beta_i bi(X31_t) + \sum_{i=12}^{17} \beta_i bi(X48_t) + \sum_{i=18}^{24} \beta_i bi(X182_t) + \beta_{25} \cdot X99_{L1t} + \beta_{26} \cdot X99_{L2t} + \beta_{27} \cdot X99_{L3t} + \beta_{28} \cdot X99_{L4t} + \epsilon_t$$

	Estimate	Std. Error	t value	Pr(> t)
X3_cut(4.42,5.93]	26.96	0.323	83.48	0
X3_cut(5.93,7.43]	27.15	0.3219	84.34	0
X3_cut(7.43,8.93]	26.97	0.3237	83.29	0
X3_cut(8.93,10.4]	26.85	0.3249	82.65	0
X3_cut(10.4,11.9]	26.74	0.3248	82.35	0
X31_cut(3.53,5.81]	1.055	0.2337	4.516	6.316e-06
X31_cut(5.81,8.08]	5.447	0.2328	23.4	9.117e-121
X31_cut(8.08,10.3]	21.23	0.2322	91.42	0
X31_cut(10.3,12.6]	62.55	0.2333	268.1	0
X31_cut(12.6,14.9]	152	0.2328	652.9	0

Table 9: Metricas de los Modelos

	ECM	R2
Modelo Lineal	1447.5909	0.966996
Modelo Polinomial	126.3425	0.997130
Modelo Paso	514.1924	0.988481

	Estimate	Std. Error	t value	Pr(> t)
X48_cut(2.04,3.5]	30.39	0.234	129.9	0
X48_cut(3.5,4.96]	61.43	0.2334	263.1	0
X48_cut(4.96,6.41]	91.15	0.2347	388.4	0
X48_cut(6.41,7.87]	122.4	0.234	523.3	0
X48_cut(7.87,9.34]	153.1	0.2346	652.7	0
X182_cut(0.651,1.76]	1.444	0.2335	6.184	6.267e-10
X182_cut(1.76,2.86]	9.765	0.2336	41.8	0
X182_cut(2.86,3.97]	32.21	0.2334	138	0
X182_cut(3.97,5.08]	74.8	0.2338	319.9	0
X182_cut(5.08,6.19]	143.9	0.2337	616	0
X99L2	3.067	0.1909	16.06	5.272e-58
X99L3	0.06957	0.1912	0.3638	0.716
X99L4	0.1107	0.1911	0.5795	0.5623

Table 8: Fitting linear model: $Y \sim -1 + X3_cut + X31_cut + X48_cut + X182_cut + X99$

Observations	Residual Std. Error	R^2	Adjusted R^2
112146	22.59	0.9885	0.9885

Mejor Modelo Considerando el R^2 y ECM en el conjunto de test como métricas para escoger el mejor modelo, se puede concluir que para esta base de datos el **Modelo Polinomial** es el que consigue ajustarse mejor a los datos y tener una mejor predicción.

Punto 3

Usando la función **loess** (locally weighted smoothing), ajuste una curva a los datos de la base de datos **Wage** (Wage en función de Age).

Regresión local

La idea de la regresión local es construir en cada punto del conjunto de datos una regresión polinómica de grado bajo , con los valores de la variable explicativa situados cerca del punto cuya respuesta se está estimando. Lo que quiere decir que el polinomio se ajusta utilizando los mínimos cuadrados ponderados, dando más peso a los puntos cercanos al punto en cuestión. Un algoritmo simple consiste en:

- Tome una fracción $s = \frac{k}{n}$ de puntos x_i , alrededor de x_0 .

- Dele un peso $K_{i0} = K(x_i, x_0)$ a cada uno de los puntos vecinos de x_0 , de tal manera que los vecinos más cercanos tengan mayor peso que los más lejanos. En muchos casos se toma K_{i0} como la función tricubo.
- Estime los parámetros β_0 y β_1 minimizando la suma cuadrática ponderada:

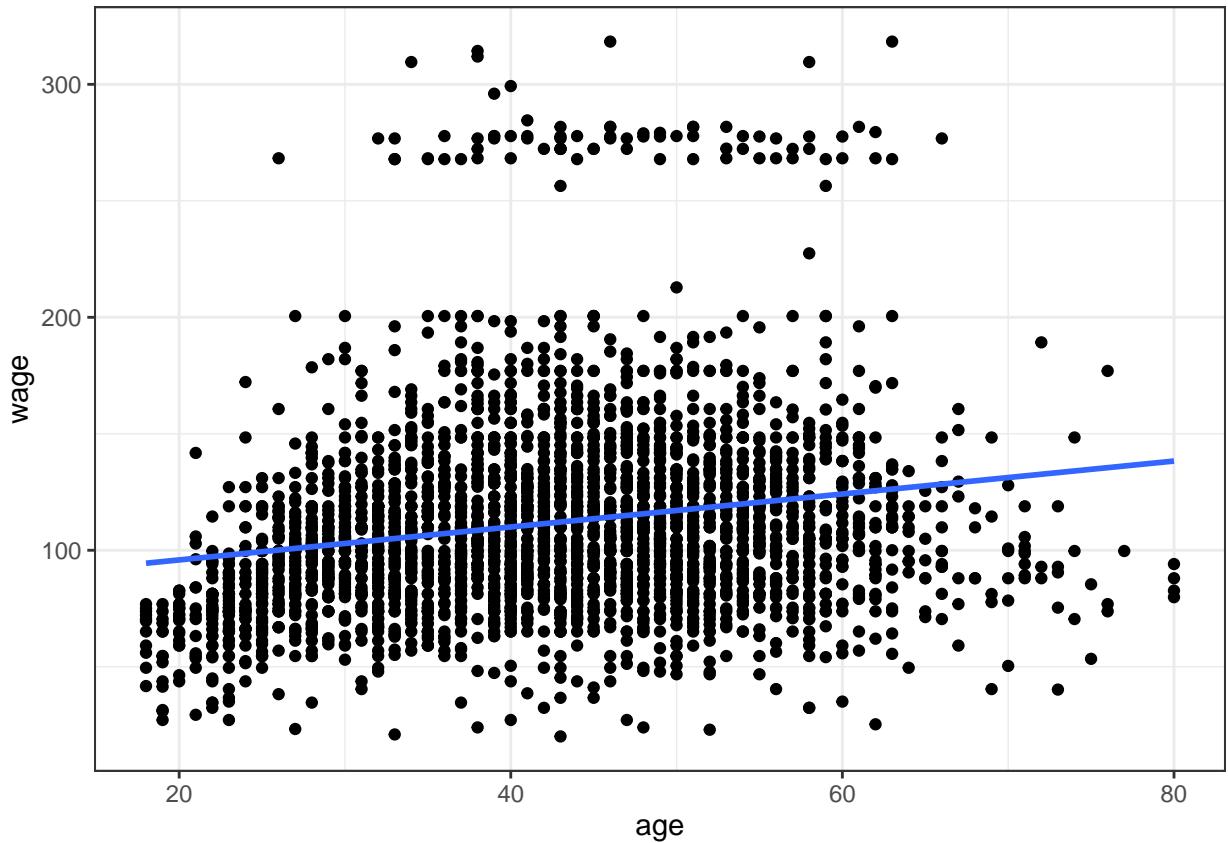
$$\sum_{i=1}^n K_{i0}(y_i - \beta_0 - \beta_1 x_i)^2$$

- El valor ajustado en x_0 y que sirve para elaborar la curva es:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

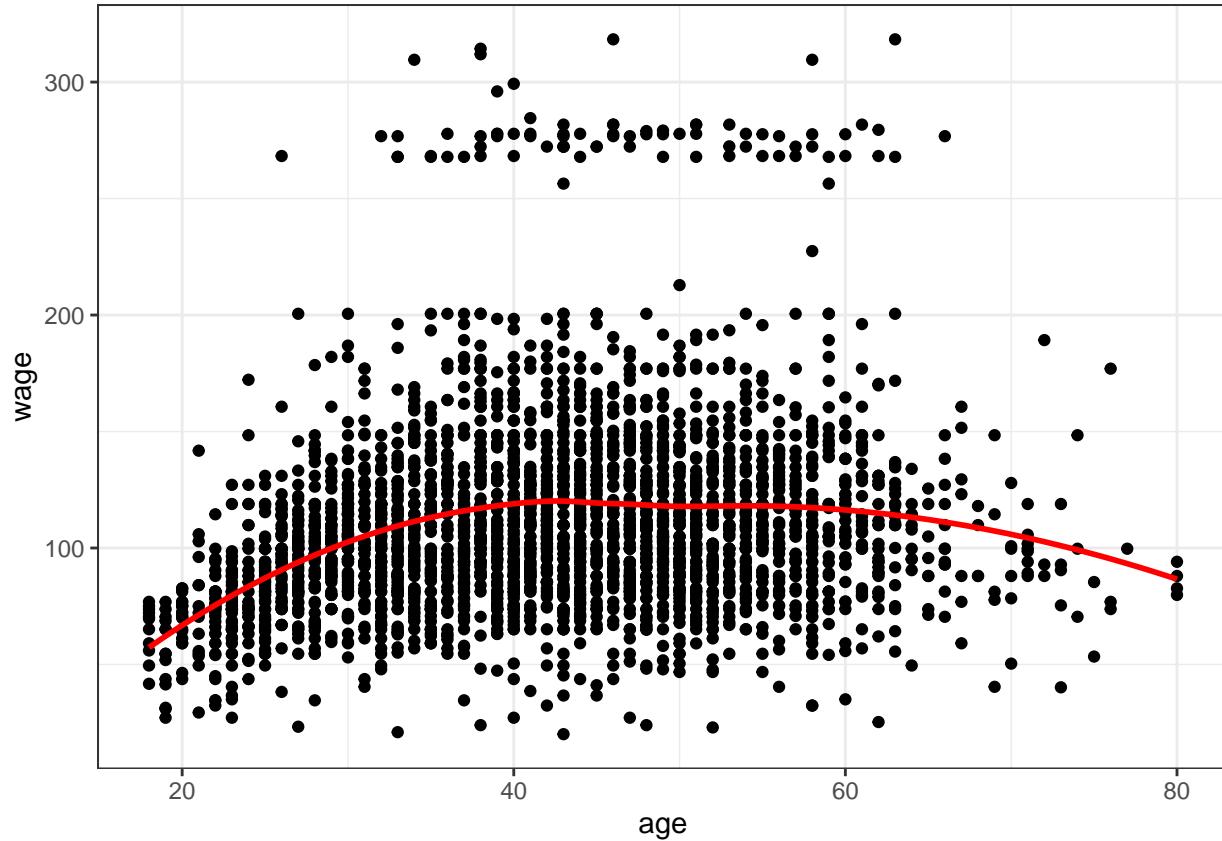
Wage según Age

Es entonces de nuestro interés construir una curva que pueda explicar de manera confiable la relación que puede existir entre la variable “Age” y “Wage”.



Observando los datos, es razonable concluir que una función lineal no es la más indicada a la hora de explicar el comportamiento que tiene el “wage” basandnos en “age”. Parece ser más adecuado ajustar una función más flexible que no sea lineal.

Teniendo esto en cuenta, sumado a la metodología antes explicada. Se tiene como una buena alternativa para explicar la relación antes mencionada ajustar una regresión local sobre los dato como se muestra a continuación.



De esta manera se puede obtener una curva no lineal mucho más flexible para poder explicar como se comporta el salario (Wage) a partir de la edad (Age) de una persona en específico. Aún así es claro que se observan varios puntos extremos que no se llegan a explicar muy bien con la curva construida utilizando locally weightd smoothing, por lo que, eventualmente, sería ideal tratar de alguna manera con estos outliers.

Punto 5

Actividad 1

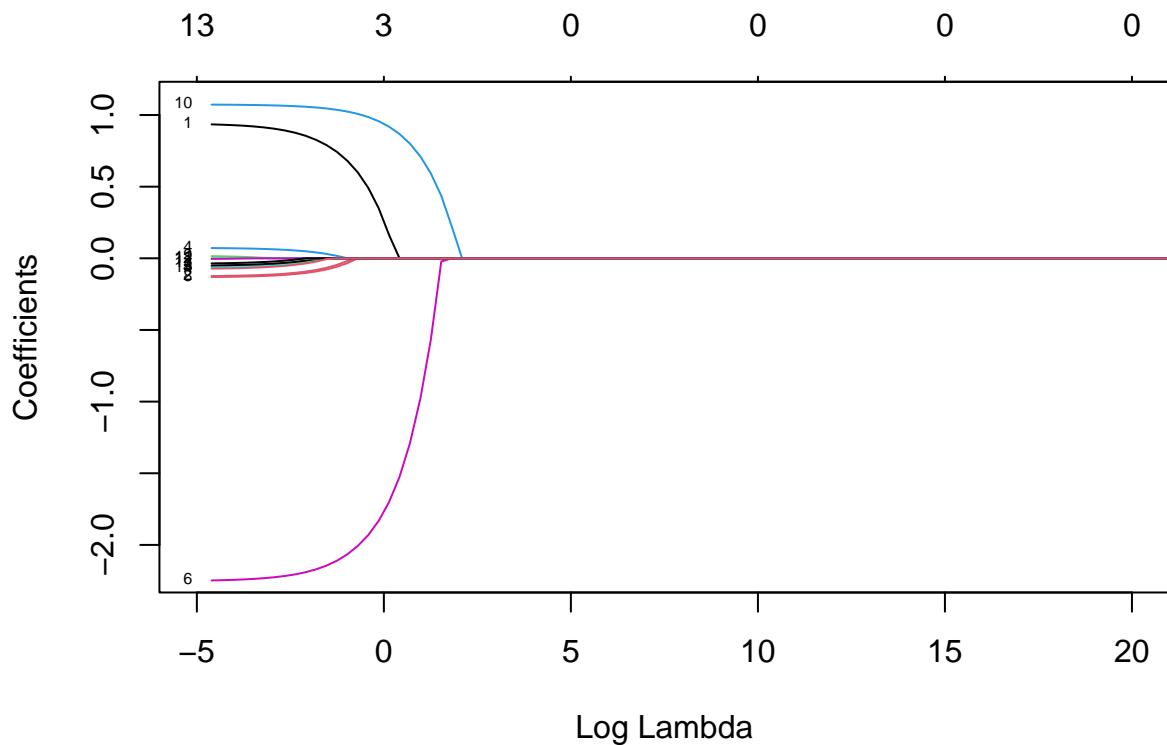
Las primeras observaciones de la base de datos son:

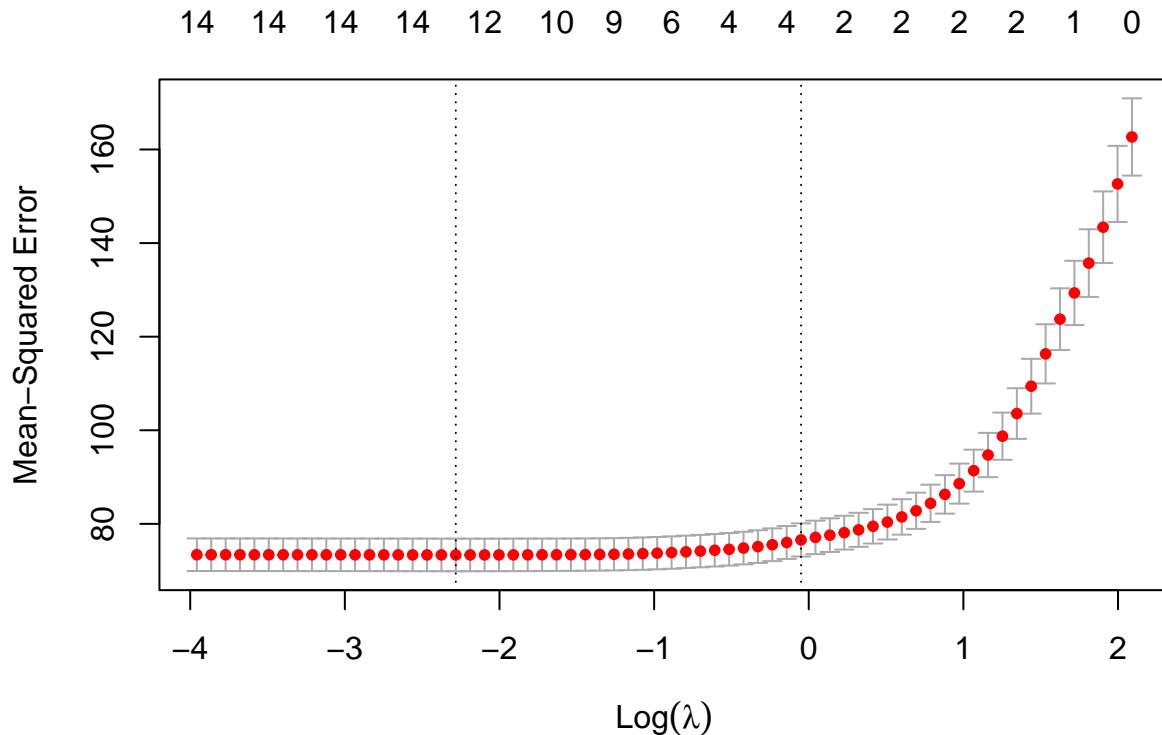
Y	X1	X2	X3	X4	X5	X6	X7	X8	X9
11.19162	-1.08114	9.1861	-0.0937	2.8957	1.5078	4.34157	5.2942	5.9344	11.1092
22.60943	-1.28549	8.6449	6.1533	0.1205	3.0802	5.24621	6.8235	5.1159	13.2634
29.71620	0.85479	6.6983	5.8730	10.6547	10.0062	4.93574	3.3363	3.6772	10.0689
10.33344	0.80922	13.0474	8.6096	3.6979	10.2647	6.25590	11.2681	7.4279	11.2968
28.84943	-0.26611	9.3007	0.5411	11.7838	8.2244	2.88766	2.3122	2.6440	8.9499
29.63682	1.31320	3.2235	0.6024	3.2813	3.9245	4.97632	5.0332	3.6060	10.6468

X10	X11	X12	X13	X14
8.46643	11.1092	1.9148	5.2417	11.0799
17.64208	13.2634	-2.3003	8.4040	7.0935
29.27286	10.0689	-2.7403	8.8594	5.5088
9.52291	11.2968	0.4165	4.7300	5.6324
19.12384	8.9499	8.7977	9.6763	7.7176
17.81611	10.6468	0.1442	7.1415	6.5541

LASSO Regression

```
## [1] 15 100
```





```

## [1] 0.1020261

## [1] 68.13816

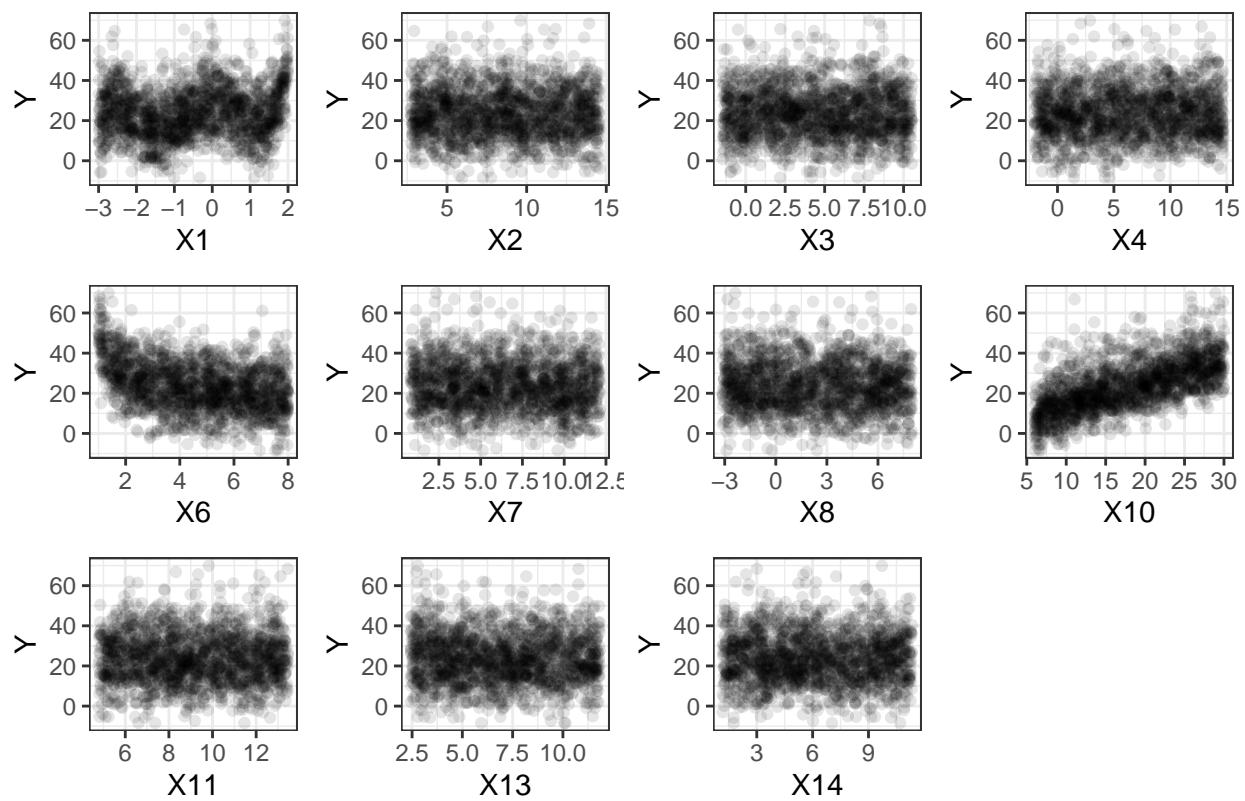
## (Intercept)          X1          X2          X3          X4          X5
## 16.980966399  0.870227421 -0.098835616 -0.033630082  0.053293669 -0.026638028
##          X6          X7          X8          X9          X10         X11
## -2.201192460 -0.022431096 -0.107541378  0.000000000  1.060619812  0.000000000
##          X12         X13         X14
## 0.000000000 -0.006393338 -0.042563546

## (Intercept)          X1          X2          X3          X4          X5
## 16.980966399  0.870227421 -0.098835616 -0.033630082  0.053293669 -0.026638028
##          X6          X7          X8          X10         X13         X14
## -2.201192460 -0.022431096 -0.107541378  1.060619812 -0.006393338 -0.042563546

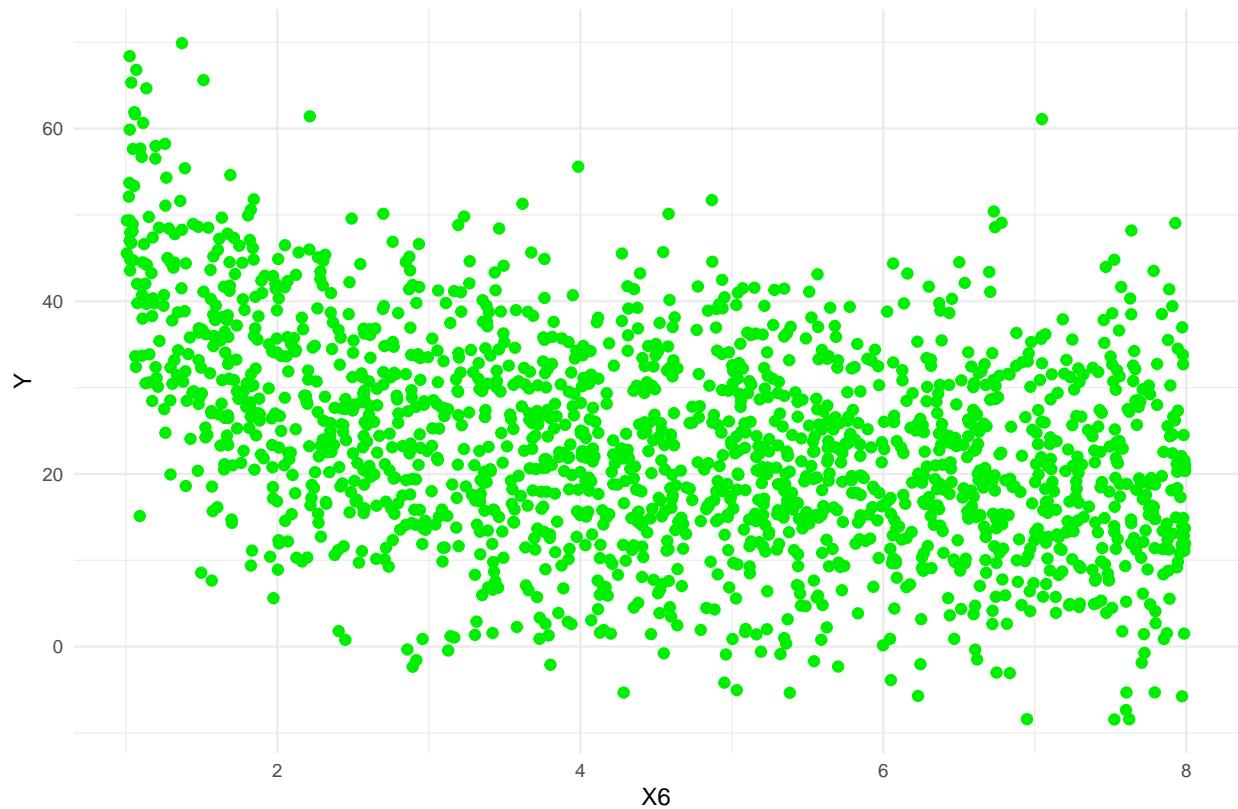
```

Para hacer la selección de variables se utilizó la Regresión LASSO, y los coeficientes que son diferentes de 0 son los asociados a estas variables X_1 , X_2 , X_3 , X_4 , X_5 , X_6 , X_7 , X_8 , X_{10} , X_{13} , X_{14} y por tanto estas variables son las más relevantes para explicar Y

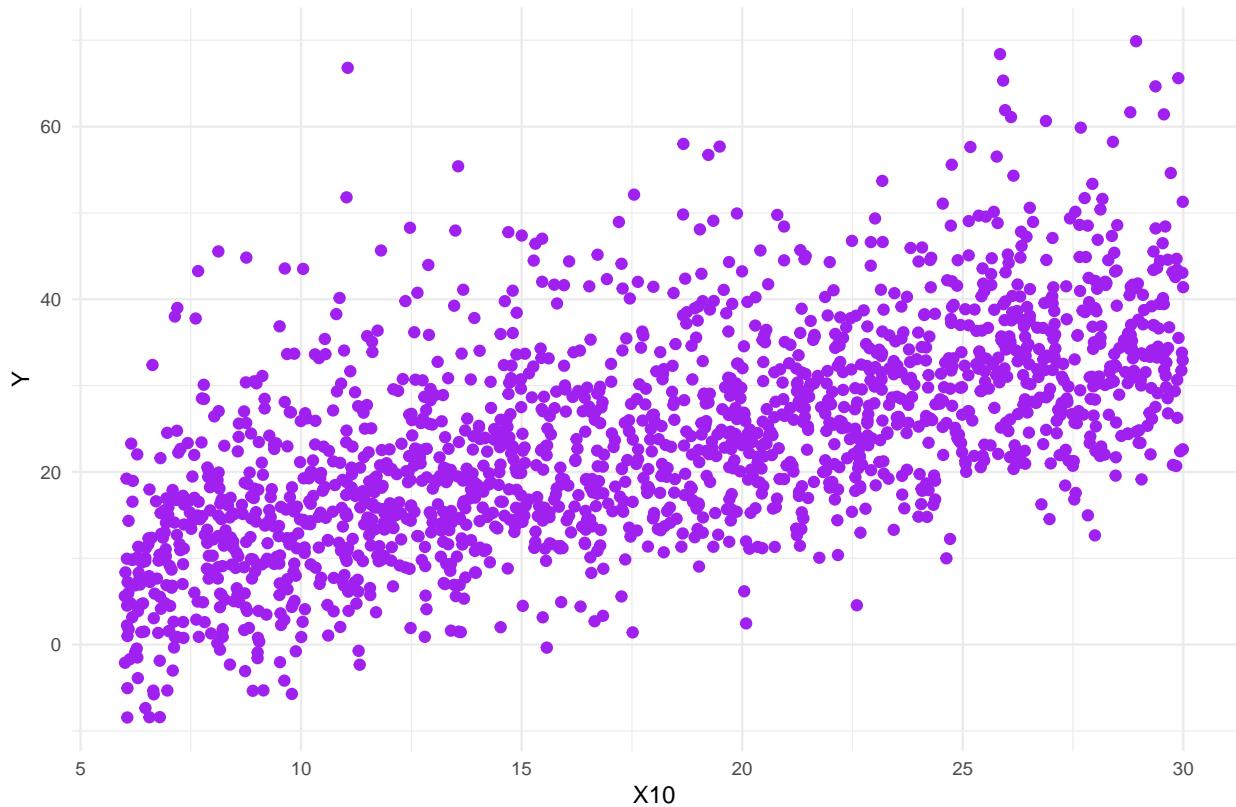
Relación entre Y y las variables



Relación entre la variable Y y la variable X6



Relación entre la variable Y y la variable X10



```

## GAMLSS-RS iteration 1: Global Deviance = 12495.73
## GAMLSS-RS iteration 2: Global Deviance = 12490.55
## GAMLSS-RS iteration 3: Global Deviance = 12490.46
## GAMLSS-RS iteration 4: Global Deviance = 12490.46
## GAMLSS-RS iteration 5: Global Deviance = 12490.46

## ****
## Family: c("NO", "Normal")
##
## Call: gammss(formula = y ~ X1 + X2 + X3 + X4 + X5 + X6 +
##   X7 + X8 + X10, sigma.formula = ~X6 + X10, nu.formula = X13,
##   tau.formula = X14, family = NO(mu.link = "identity",
##   sigma.link = "identity"), data = basea)
##
## Fitting method: RS()
##
## -----
## Mu link function: identity
## Mu Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.12656   1.15887 13.916 < 2e-16 ***
## X1          0.93853   0.13578  6.912 6.66e-12 ***
## X2         -0.13423   0.05685 -2.361  0.0183 *
## X3         -0.05157   0.05674 -0.909  0.3636
## X4          0.07819   0.03992  1.959  0.0503 .

```

```

## X5      -0.07069   0.06950  -1.017   0.3093
## X6     -2.01865   0.10510 -19.207 < 2e-16 ***
## X7     -0.04560   0.06012  -0.759   0.4482
## X8     -0.10510   0.06088  -1.726   0.0845 .
## X10    1.08187   0.02813  38.466 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## -----
## Sigma link function: identity
## Sigma Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.044412  0.504372 19.915 < 2e-16 ***
## X6        -0.400010  0.065924 -6.068 1.59e-09 ***
## X10       0.006649  0.019731  0.337   0.736
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## -----
## No. of observations in the fit: 1767
## Degrees of Freedom for the fit: 13
## Residual Deg. of Freedom: 1754
## at cycle: 5
##
## Global Deviance: 12490.46
## AIC: 12516.46
## SBC: 12587.66
## ****
## (Intercept)          X1          X2          X3          X4          X5
## 16.12656323  0.93852557 -0.13423349 -0.05156573  0.07818569 -0.07068867
## X6          X7          X8          X10
## -2.01865279 -0.04560176 -0.10509898  1.08187350
##
## (Intercept)          X6          X10
## 10.044411997 -0.400009991  0.006648753

```

Actividad 2

Día a día cientos de accidentes automovilísticos ocurren en las ciudades más congestionadas vialmente del mundo. Esto es un problema de altísima importancia pues estos accidentes generan aún más problemas de movilidad y, aún más preocupante, en muchas ocasiones pueden dejar una gran cantidad de heridos o, incluso, muertos como resultado.

Por lo tanto, es más que comprensible querer entender como se relacionan diferentes características del día a día con respecto a la cantidad de accidentes que se pueden presentar con el fin de implementar medidas que puedan, de una manera u otra, intentar minimizar la cantidad de accidentes de manera general.

Con esto en mente, tenemos en mente ajustar un modelo lineal que nos permita, como ya se mencionó, observar como una selección de variables afecta al número de accidentes, en este caso, de una cierta ciudad en el mundo.

Datos

Para esto, contamos una base de datos de dicha ciudad que registra la cantidad de accidentes que se presentan en la misma es 728 días diferentes. Además de un grupo de características de dicho día de las que se quiere conocer su efecto sobre la variable respuesta (# de accidentes), las cuales son:

- X1: # de vehículos que transitan por día.
- X2: LLueve (1) o No llueve (0).
- X3: Día de la semana.

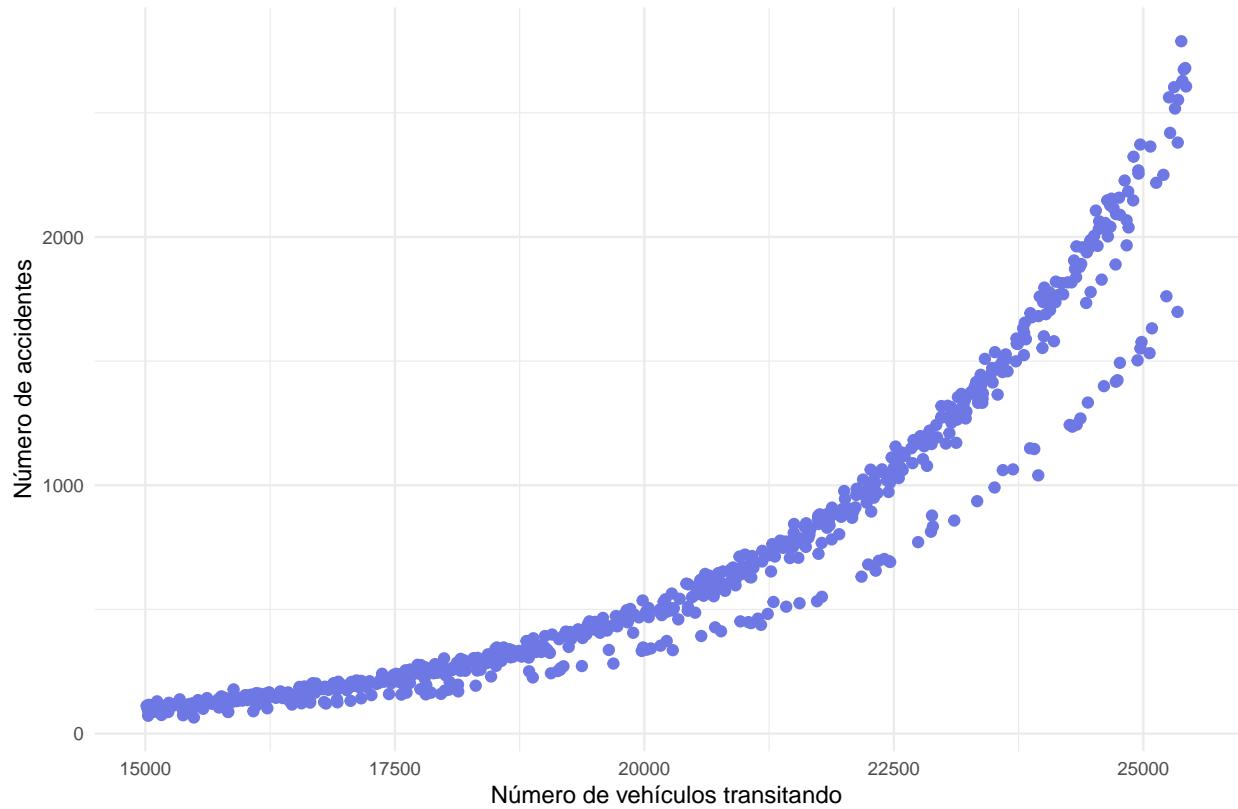
Para el caso del día de la semana, tendremos usar variables indicadoras. Es decir, variables que valdrán 1 si el día en cuestión corresponde a dicho día de la semana y 0 en cualquier otro caso.

Y	X1	X2	X3
1112	22548	1	LUN
399	19075	1	MAR
103	15425	0	MIE
502	19859	0	JUE
181	16774	0	VIE
280	18223	0	SAB

Análisis descriptivo

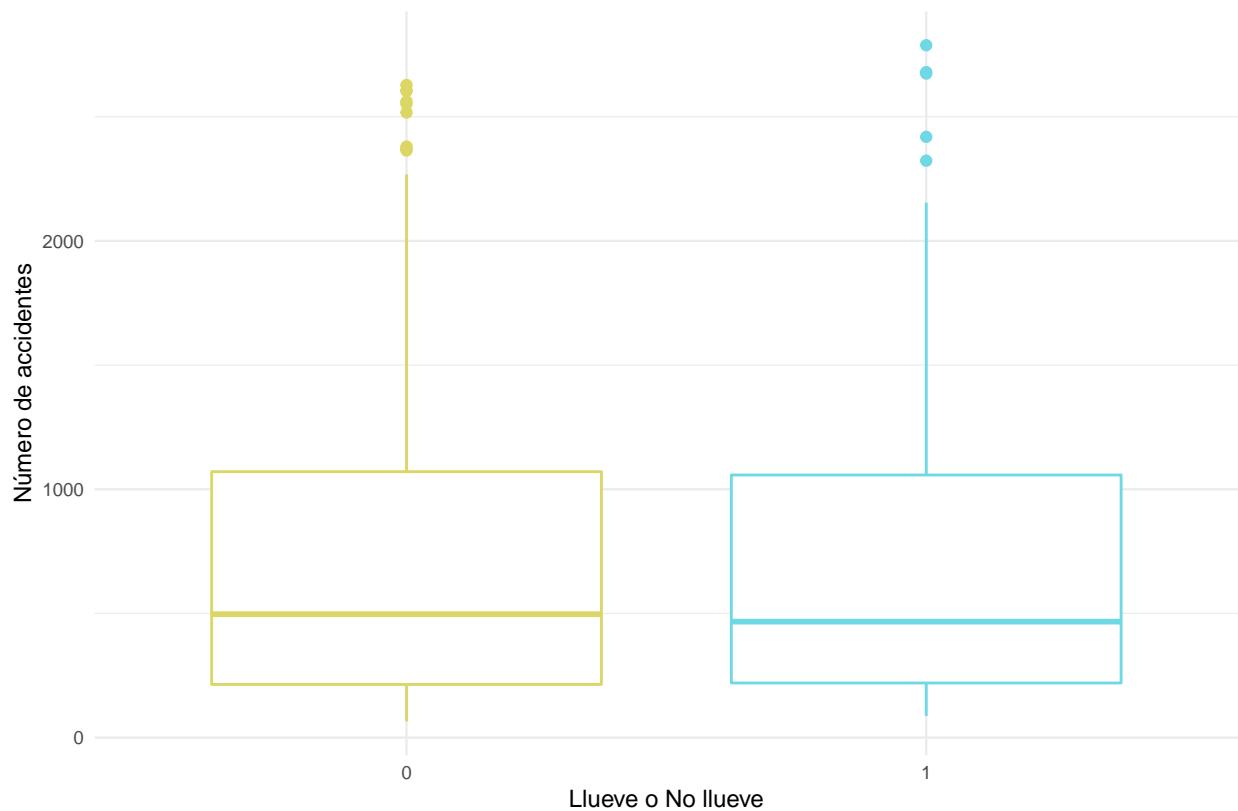
Antes de proceder a la construcción del modelo ya mencionado; es de interés realizar un breve análisis con el fin de observar, gráficamente, como se comporta nuestra variable respuesta en relación las variables independientes.

Número de vehículos transitando Vs número de accidentes



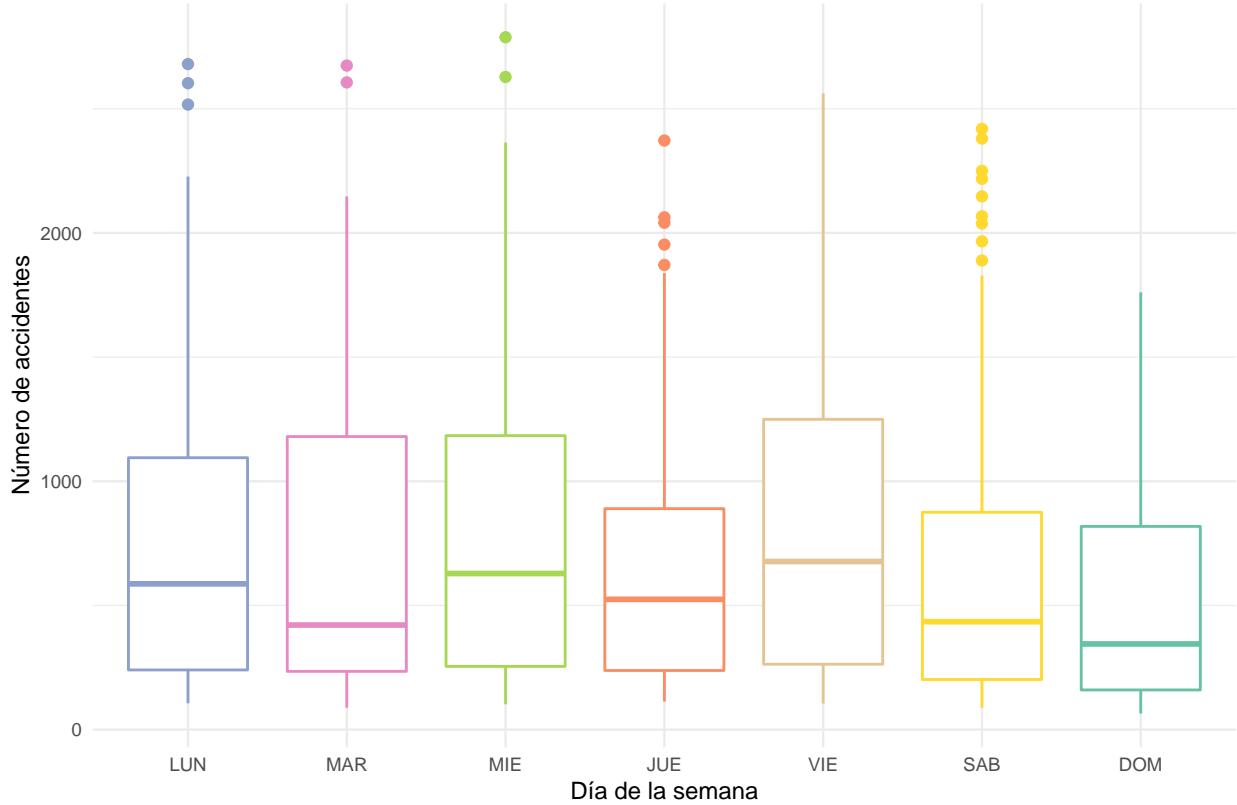
Como era de esperarse, es evidente que existe una relación no lineal (posiblemente cuadrática o exponencial) y positiva entre el número de vehículos que transitan en cierto día y el número de accidentes que se tienen en ese mismo día. Esto quiere decir que a más vehículos hay transitando, mayor es la cantidad de accidentes que se presentan.

Número de accidentes si llueve o no



A pesar de que nuestra intuición podría llevarnos a pensar que en días lluviosos se presentan más accidentes de tránsito, la anterior figura nos está indicando que, en promedio, sin importar si llueve o no, se presentan el mismo número de accidentes de tránsito.

Número de accidentes según el día de la semana



Finalmente, de manera visual, no parece que el efecto del día de la semana en el que se esté, sea significativo para la cantidad de accidentes que se presentan en dicho día. Sin embargo, este efecto podría estar enmascarado por un posible interacción con otro de los factores o que la escala del gráfico no permita observar claramente la diferencia entre los distintos días de la semana.

Modelo

Ahora, procedemos a ajustar el modelo ya mencionado con anterioridad. Inmediatamente, al tratarse del número de accidentes en un día, parece más que lógico y razonable utilizar una distribución Poisson para ajustar los datos de interés. Por lo tanto, vamos a haer uso de la metodología GAMLSS (Generalized additive models for location, scale and shape), en esta se modelan cada uno de los parámetros de la distribución. Dado que usaremos una Poisson, sólo hay un parámetro por modelar de la siguiente forma:

$$Y \sim Pois(\lambda)$$

$$g_1(\lambda) = \beta_0 + f_1(X_1) + f_2(X_2) + \dots + f_3(X_p)$$

Donde g_1 es una función link para el parámetro en cuestión. En sete caso, dado que la distribución Poisson explica conteos, y estos sólo pueden ser mayor o iguales a 0, parece razonable utilizar la función logarítmica como función link.

```
## GAMLSS-RS iteration 1: Global Deviance = 6561.99
## GAMLSS-RS iteration 2: Global Deviance = 6561.99
```

```

## ****
## Family: c("PO", "Poisson")
##
## Call: gammLSS(formula = Y ~ poly(X1, 2) + X2 + X3, family = "PO",
##               data = df)
##
## Fitting method: RS()
##
## -----
## Mu link function: log
## Mu Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.855799  0.004524 1294.320 < 2e-16 ***
## poly(X1, 2)1 24.544451  0.060229  407.521 < 2e-16 ***
## poly(X1, 2)2  0.061449  0.045350    1.355   0.176
## X2          0.012738  0.003116    4.089  4.83e-05 ***
## X3JUE        0.390531  0.005677   68.796 < 2e-16 ***
## X3LUN        0.395928  0.005504   71.938 < 2e-16 ***
## X3MAR        0.393911  0.005565   70.784 < 2e-16 ***
## X3MIE        0.397331  0.005473   72.597 < 2e-16 ***
## X3SAB        0.310996  0.005672   54.834 < 2e-16 ***
## X3VIE        0.392859  0.005469   71.829 < 2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## -----
## No. of observations in the fit: 728
## Degrees of Freedom for the fit: 10
##             Residual Deg. of Freedom: 718
##                               at cycle: 2
##
## Global Deviance:      6561.99
## AIC:                  6581.99
## SBC:                  6627.893
## ****

```

A pesar de nuestras sospechas iniciales y lo que se vió en el análisis descriptivo, parece que la componente cuadrática para la variable “número de vehículos transitando” no es significativa, por lo que esta será retirada y sólo se mantendrá la componente lineal.

```

## GAMLSS-RS iteration 1: Global Deviance = 6563.825
## GAMLSS-RS iteration 2: Global Deviance = 6563.825

## ****
## Family: c("PO", "Poisson")
##
## Call: gammLSS(formula = Y ~ X1 + X2 + X3, family = "PO", data = df)
##
## Fitting method: RS()
##
## -----
## Mu link function: log
## Mu Coefficients:

```

```

##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.779e-01  1.404e-02 -26.921 < 2e-16 ***
## X1           3.093e-04  5.881e-07 525.857 < 2e-16 ***
## X2           1.276e-02  3.117e-03   4.092 4.76e-05 ***
## X3JUE        3.899e-01  5.659e-03  68.899 < 2e-16 ***
## X3LUN        3.956e-01  5.500e-03  71.937 < 2e-16 ***
## X3MAR        3.937e-01  5.563e-03  70.773 < 2e-16 ***
## X3MIE        3.969e-01  5.462e-03  72.660 < 2e-16 ***
## X3SAB        3.110e-01  5.672e-03  54.831 < 2e-16 ***
## X3VIE        3.923e-01  5.456e-03  71.907 < 2e-16 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## -----
## No. of observations in the fit: 728
## Degrees of Freedom for the fit: 9
##       Residual Deg. of Freedom: 719
##                           at cycle: 2
##
## Global Deviance:    6563.825
## AIC:                6581.825
## SBC:                6623.137
## ****

```

Ahora todos los efectos incluidos son significativos y, además, obtenemos un AIC menor al que teníamos para el modelo que incluía la componente cuadrática. Por tanto, es este último modelo con el que nos quedaremos.

Recordemos que estamos modelando el logaritmo del parámetro λ , por lo tanto los coeficientes se interpretan basados en este. Por ejemplo, para el efecto debido al número de vehículos transitando, obtenemos una estimación de 0.0003093; este quiere decir que, si las demás variables se mantienen constantes, un aumento de un vehículo en transito trae consigo un aumento promedio de 0.0003093 en el logaritmo del número de accidentes. Otra forma de interpretar este coeficiente, si calculamos $e^{0.0003093} = 1.0003$, es al decir que, en promedio, un aumento de una unidad en los vehículos que transitan, genera aproximadamente, un aumento del 0.03% en el número de accidentes. De manera similar, se puede observar como, en un día lluvioso se aumenta el logaritmo del número de accidentes en un 0.01276 en promedio o, si calculamos $e^{0.01276} = 1.0128$, en un día lluvioso aumentan el número de accidentes en aproximadamente 1.28% en promedio. Finalmente, dado que el día domingo es el nivel de referencia, las estimaciones para los demás días de la semana nos dejan concluir que, si es cualquier día de la semana, excepto por el sábado, se aumenta el logaritmo del número de accidentes en, aproximadamente, 0.39 en promedio con respecto al domingo. Además, en este caso el intercepto estimado no tiene una interpretación real, pues es absurdo pensar en que transiten 0 vehículos.

A partir de esto, es directo concluir que los días en los que menos accidentes se presentan en promedio son los domingos en los que no llueve.

Finalmente entonces, el modelo predictivo estimado se ve de la siguiente forma:

$$\begin{aligned}
\log(\lambda) = & -0.3779 + 0.0003093 \times \text{Num de vehiculos} + 0.01276 \times \text{Llueve} \\
& + 0.389 \times \text{Jueves} + 0.3956 \times \text{Lunes} + 0.3937 \times \text{Martes} + 0.3969 \times \text{Miercoles} \\
& + 0.311 \times \text{S\u00e1bado} + 0.3923 \times \text{Viernes}
\end{aligned}$$