Trabajo de Regresión en Introducción a Ciencia de Datos. Dataset ANACALT.

Danel Arias

2023-12-11

Descripción del trabajo

El trabajo consiste en la realización de un análisis exploratorio de datos (EDA) sobre un dataset y realizar regresión con diferentes métodos para un dataset. El dataset indicado es el dataset ANACALT (Analizing Categorical Data).

Descripción del dataset

Segun podemos leer en KEEL se trata de uno de los conjuntos de datos utilizados en el libro Analyzing Categorical Data de Jeffrey S. Simonoff, Springer-Verlag, Nueva York, 2003. Los datos contienen información sobre las decisiones adoptadas por un tribunal supremo.

Este contiene las siguientes variables:

- Actions_taken: Número de acciones tomadas por el tribunal supremo. [0-11]
- Liberal: Variable booleana [0-1] (No se especifica).
- Unconstitutional: Variable booleana [0-1] (No se especifica).
- Precedent_alteration: Variable booleana [0-1] (No se especifica).
- Unanimous: Variable booleana [0-1] (No se especifica).
- Year_of_decision: Año de la decisión. [1953-1988]
- Lower_court_disagreement: Variable booleana [0-1] (No se especifica).
- Log_exposure: Variable de output (No se especifica). [0-2.3]

En total el dataset tiene 8 variables y 4052 observaciones. Además, según la descripción en KEEL este dataset no contiene valores perdidos (NA's).

Preparación inicial del entorno

Carga de librerías

```
library(tidyverse)
library(ggplot2)
library(moments)
library(corrplot)
```

```
library(gridExtra)
library(kknn)
library(stats)
library(caret)
library(ggthemes)
```

Carga de datos

Cargamos los datos de ANACALT/ANACALT.dat. Al ser un fichero .dat debemos obviar las filas que no son datos, estas empiezan en '@'

```
anacalt <- read.table("ANACALT/ANACALT.dat", header = FALSE,
                      sep = ",", comment.char = "@")
```

Añadimos los nombres de las columnas. Los nombres de las columnas están en pima.dat. Y se encuentran en la línea tras '@inputs' así que lo leemos y lo añadimos.

```
nombres <- grep("^@inputs", readLines("ANACALT/ANACALT.dat"), value = TRUE) %>%
  # Eliminamos el @inputs
  str_remove("@inputs ") %>%
  #Eliminamos espacions
  str_remove_all(" ") %>%
  # Separamos por comas
  str_split(",") %>%
  # Convertimos a vector
  unlist()

# El nombre de la última variable se encuentra tras @outputs
nombre_output <- grep("^@outputs", readLines("ANACALT/ANACALT.dat"), value = TRUE) %>%
  # Eliminamos el @outputs
  str_remove("@outputs ")

# Añadimos el nombre de la última variable
nombres <- c(nombres, nombre_output)

# Añadimos los nombres al dataset
colnames(anacalt) <- nombres
```

Análisis exploratorio de datos

Primero echamos un vistazo a los datos son un `summary()`. Podemos ver que cuadra con la descripción proporcionada.

```
##  Actions_taken        Liberal       Unconstitutional  Precedent_alteration
##  Min.   : 0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median : 0.0000   Median :1.0000   Median :0.00000   Median :0.00000
##  Mean   : 0.1088   Mean   :0.5197   Mean   :0.07823   Mean   :0.02295
##  3rd Qu.: 0.0000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :11.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
```

```
##    Unanimous      Year_of_decision Lower_court_disagreement  Log_exposure
##  Min.   :0.0000   Min.   :1953     Min.   :0.0000           Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:1964     1st Qu.:0.0000           1st Qu.:2.080
##  Median :0.0000   Median :1973     Median :0.0000           Median :2.300
##  Mean   :0.3376   Mean   :1972     Mean   :0.2256           Mean   :2.034
##  3rd Qu.:1.0000   3rd Qu.:1981     3rd Qu.:0.0000           3rd Qu.:2.300
##  Max.   :1.0000   Max.   :1988     Max.   :1.0000           Max.   :2.300
```

Revisamos si tenemos NA's por si acaso y confirmamos lo que se nos indicaba en KEEL, no hay NA's.

```
colSums(is.na(anacalt))
```

```
##            Actions_taken                  Liberal         Unconstitutional
##                        0                        0                        0
##     Precedent_alteration                Unanimous         Year_of_decision
##                        0                        0                        0
## Lower_court_disagreement             Log_exposure
##                        0                        0
```

Análisis variable a variable

En este apartado analizaremos cada variable por separado.

Variable 'Actions_taken'

Comenzamos con la variable 'Actions_taken'. Sabemos que pertenece al rango [0-11]. Para más detalle, vemos un histograma en la Figura 1.

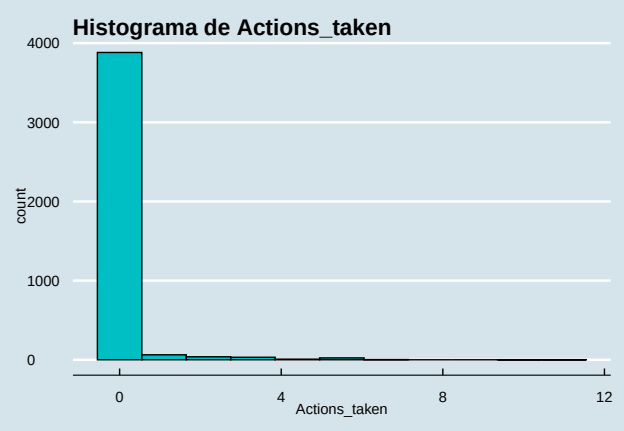

Figure 1: Histograma de Actions taken

Llama la atención el gran pico en 0, podríamos pensar que son valores perdidos, pero no podemos asegurarlo ya que es plausible que se hayan tomado 0 acciones.

Estudiamos también los cuantiles con un boxplot (Figura 2). Donde vemos que el 75% de los valores son 0, solo algunos valores son outliers.

```
##   0%  25%  50%  75% 100%
##    0    0    0    0   11
```

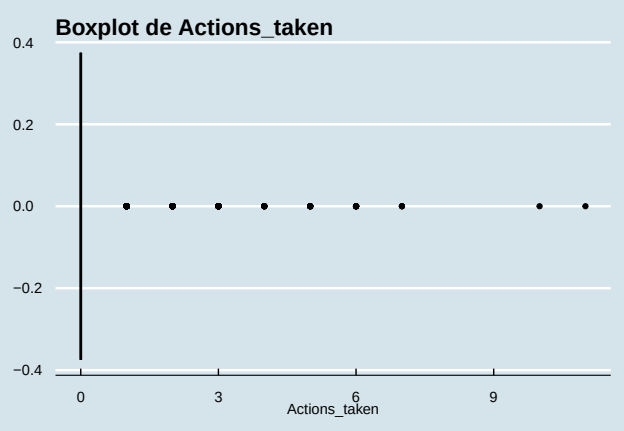

Figure 2: Boxplot de Actions taken

Estudiamos la forma de la variable Actions_taken. Para ello miramos la media, sd, skewness y kurtosis.

```
## [1] "Media: 0.1088"
```

```
## [1] "Deviación estándar: 0.6446"
```

Para estudiar la skewness y la kurtosis usaremos los tests de D'Agostino y Anscombe respectivamente.

```
## [1] "Skewness: 7.9659"
```

```
##
##  D'Agostino skewness test
##
## data:  anacalt$Actions_taken
## skew = 7.9659, z = 63.3294, p-value < 2.2e-16
## alternative hypothesis: data have a skewness
```

Dado que la skewness > 0 la distribución está sesgada a la derecha. Ademas el valor de p-value del test de D'agostino es menor que 0.05, por lo que podemos afirmar con seguridad que la distribución está sesgada a la derecha.

```
## [1] "Kurtosis: 80.5755"
```

```
##
##  Anscombe-Glynn kurtosis test
##
## data:  anacalt$Actions_taken
## kurt = 80.576, z = 39.005, p-value < 2.2e-16
## alternative hypothesis: kurtosis is not equal to 3
```

Dado que la kurtosis > 0 la distribución tiene colas pesadas. Además el valor de p-value del test de Anscombe es menor que 0.05, por lo que podemos afirmar con seguridad que la distribución no es normal. Podemos verlo mejor con un gráfico de densidad en la Figura 3.

Podemos ver como la distribución es significativamente densa en el valor 0 mientras que el resto de valores son muy poco probables.

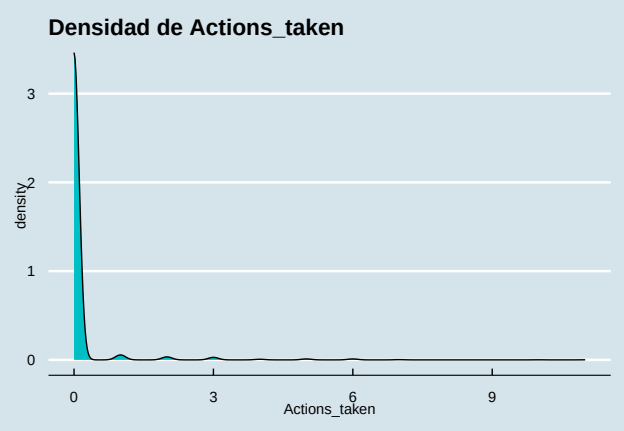

Figure 3: Densidad de Actions taken

Estudiamos la normalidad de la variable Actions_taken con un test de Shapiro-Wilk y el QQ-plot. Donde vemos que el valor de p-value es menor que 0.05, por lo que la variable Actions_taken no sigue una distribución normal. Lo vemos también el QQ-plot (Figura 4).

```
##
##  Shapiro-Wilk normality test
##
## data:  anacalt$Actions_taken
## W = 0.16301, p-value < 2.2e-16
```

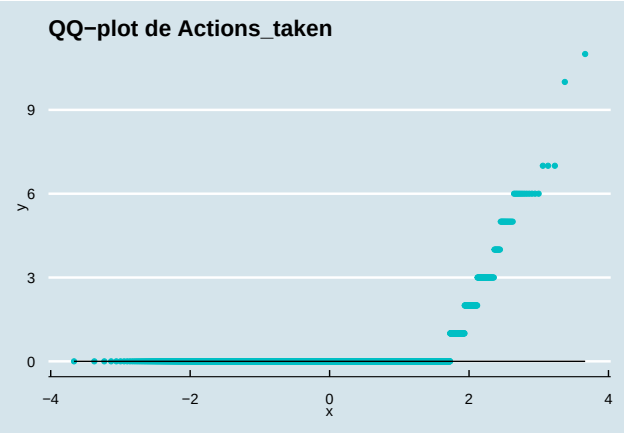

Figure 4: QQ-plot de Actions taken

Variables booleanas

Estudiamos la variable 'Liberal', 'Unconstitutional', 'Precedent_alteration', 'Unanimous' y 'Lower_court_disagreement' las cuales son variables booleanas de valores entre 0 y 1. Por lo que parece que son variables categóricas binarias. Transformamos las variables a factor con valores 'No' y 'Yes' y visualizamos las variables en barplots (Figura 5).

- Para la variable 'Liberal' vemos que la mayoría de los valores son 'Yes' pero no hay una gran diferencia.

- Para la variable 'Unconstitutional' vemos que la mayoría de los valores son 'No' y con una gran diferencia en relación a los 'Yes'.
- Para la variable 'Precedent_alteration' vemos que la mayoría de los valores son 'No' y con una gran diferencia en relación a los 'Yes'.
- Para la variable 'Unanimous' vemos que la mayoría de los valores son 'No' y con un diferencia significativa en relación a los 'Yes'.
- Para la variable 'Lower_court_disagreement' vemos que la mayoría de los valores son 'No' y con un diferencia significativa en relación a los 'Yes'.

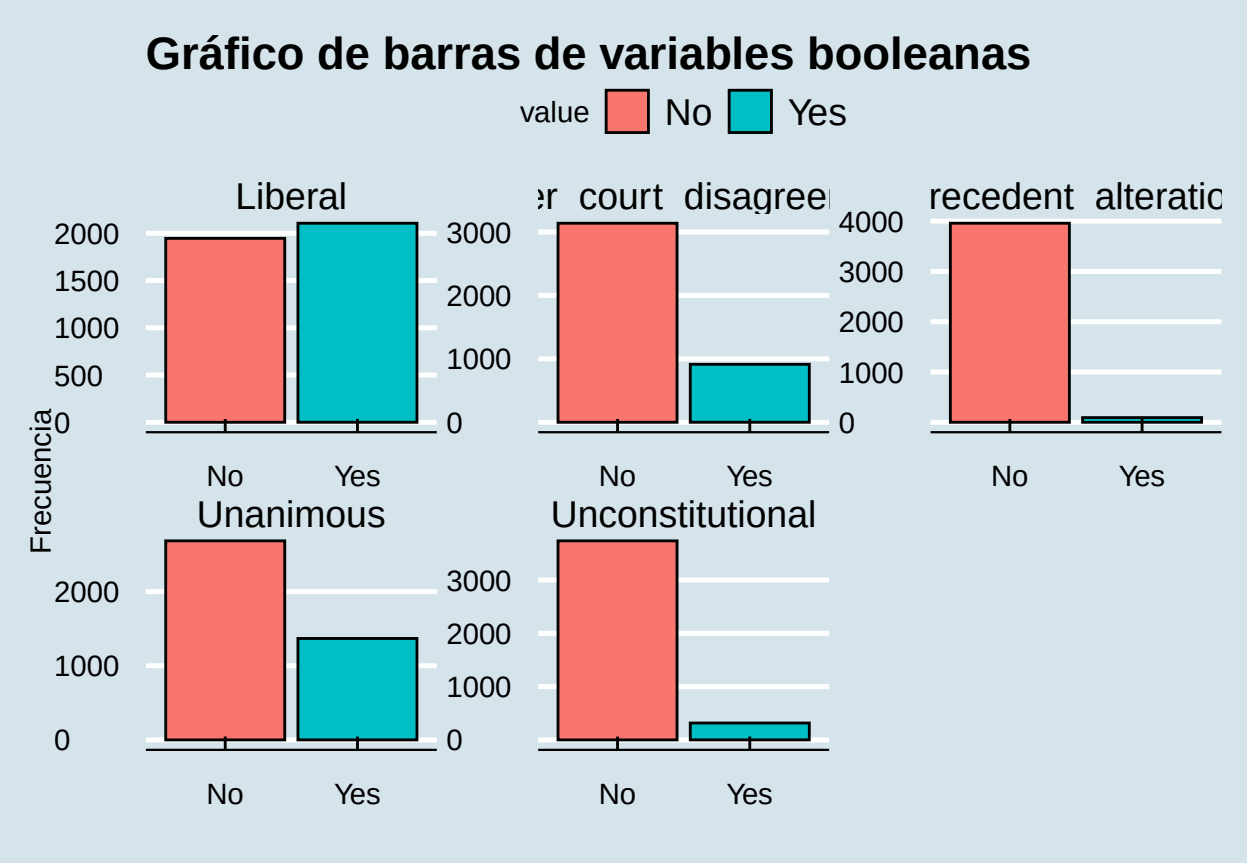

Figure 5: Gráficos de barras de las variables booleanas

Variable 'Year_of_decision'

Estudiamos la variable 'Year_of_decision'. Sabemos que pertenece al rango [1953-1988]. Para más detalle, vemos un histograma en la Figura (6).

Estudiamos también los cuantiles con un boxplot (Figura 7). Donde vemos que los valores parecen estar bastante repartidos con algo de mayor frecuencia hacia el rango superior.

```
##   0%  25%  50%  75% 100%
## 1953 1964 1973 1981 1988
```

Estudiamos para más detalle la forma de la variable. Usamos la media, sd, skewness y kurtosis.

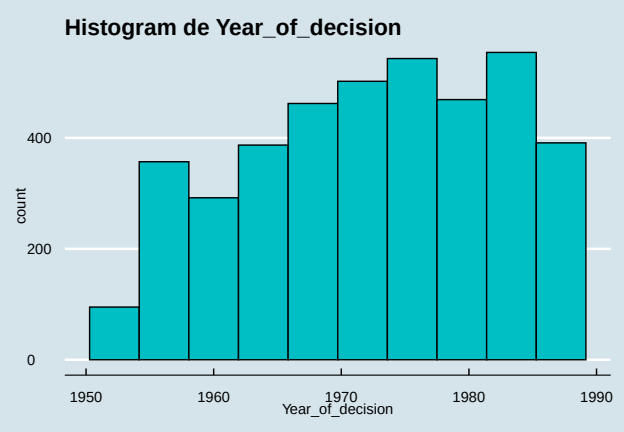

Figure 6: Histograma de Year of decision

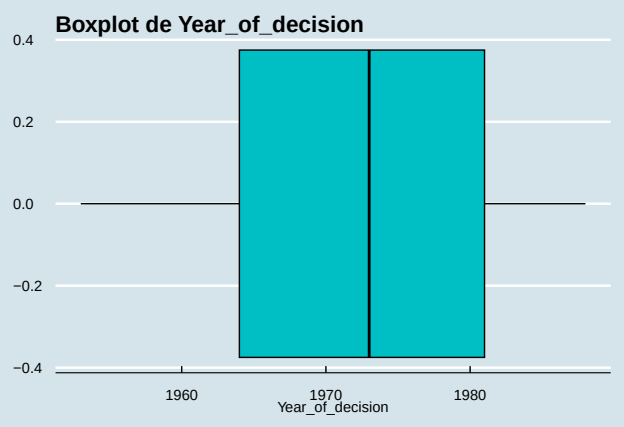

Figure 7: Boxplot de Year of decision

```
## [1] "Media: 1972.3485"
```

```
## [1] "Deviación estándar: 9.851"
```

Para estudiar la skewness y la kurtosis usaremos los tests de D'Agostino y Anscombe respectivamente.

```
## [1] "Skewness: -0.1706"
```

```
##
##  D'Agostino skewness test
##
## data:  anacalt$Year_of_decision
## skew = -0.17057, z = -4.40935, p-value = 1.037e-05
## alternative hypothesis: data have a skewness
```

Dados que skewness < 0 la distribución está sesgada a la izquierda Además el valor de p-value del test de D'agostino es menor que 0.05, por lo que podemos afirmar con seguridad que la distribución está sesgada.

```
## [1] "Kurtosis: 1.8948"
```

```
##
##  Anscombe-Glynn kurtosis test
##
## data:  anacalt$Year_of_decision
## kurt = 1.8948, z = -41.0356, p-value < 2.2e-16
## alternative hypothesis: kurtosis is not equal to 3
```

Dado que kurtosis > 0 la distribución tiene colas pesadas. Además el valor de p-value del test de Anscombe es menor que 0.05, por lo que podemos afirmar con seguridad que la distribución tiene colas pesadas. Podemos verlo mejor con un gráfico de densidad 8.

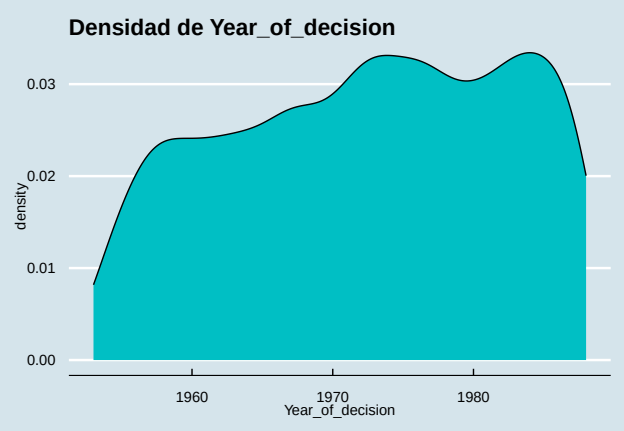

Figure 8: Densidad de Year of decision

Estudiamos la normalidad de la variable Year_of_decision con un test de Shapiro-Wilk y el QQ-plot. Donde vemos que el valor de p-value es menor que 0.05, por lo que la variable Year_of_decision no sigue una distribución normal. Lo vemos también el QQ-plot (Figura 9) donde podemos ver como tiene una forma de 'S' muy aplastada, por lo que no es normal.

```
##
##  Shapiro-Wilk normality test
##
## data:  anacalt$Year_of_decision
## W = 0.95709, p-value < 2.2e-16
```

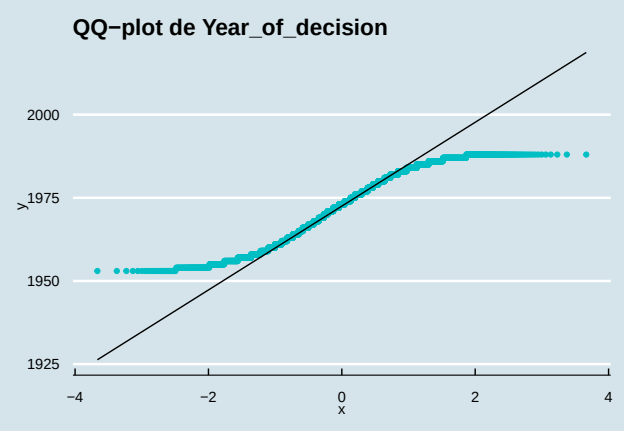

Figure 9: QQ-plot de Year of decision

Variable de output "Log_exposure"

Estudiamos ahora la variable de output "Log_exposure". Sabemos que pertenece al rango [0-2.3]. Por lo que es una variable continua. Para más detalle, vemos un histograma en la Figura (10). Donde vemos que hay un pequeño detalle en el valor 0 y a partir de 0.5 va creciendo casi de manera exponencial.

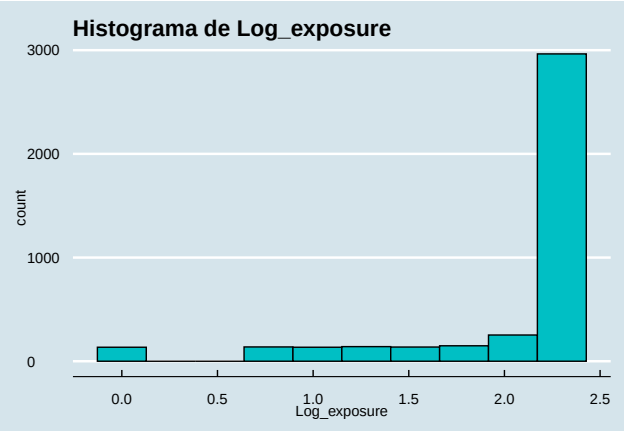

Figure 10: Histograma de Log exposure

Estudiaremos también los cuantiles con un boxplot (Figura 11). Donde vemos que el 50% de los valores superiores es exactamente 2.3, esto se visualiza perfectamente en el boxplot.

```
##   0%  25%  50%  75% 100%
## 0.00 2.08 2.30 2.30 2.30
```

Estudiamos para más detalle la forma de la variable. Usamos la media, sd, skewness y kurtosis.

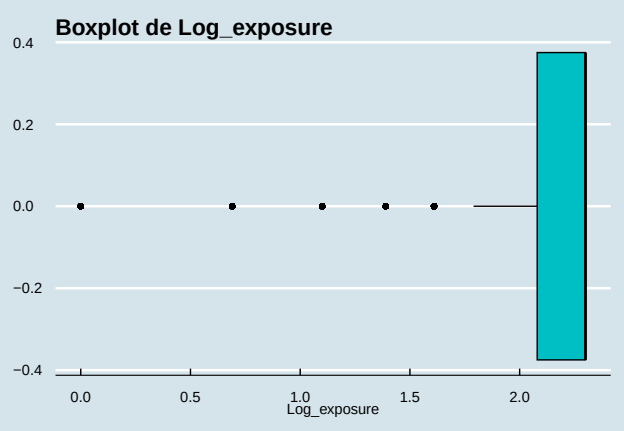

Figure 11: Boxplot de Log exposure

```
## [1] "Media: 2.0337"
```

```
## [1] "Deviación estándar: 0.5495"
```

Para estudiar la skewness y la kurtosis usaremos los tests de D'Agostino y Anscombe respectivamente.

```
## [1] "Skewness: -2.3454"
```

```
##
##  D'Agostino skewness test
##
## data:  anacalt$Log_exposure
## skew = -2.3454, z = -37.8307, p-value < 2.2e-16
## alternative hypothesis: data have a skewness
```

Dado que skewness < 0 la distribución está sesgada a la izquierda Además el valor de p-value del test de D'agostino es menor que 0.05, por lo que podemos afirmar con seguridad que la distribución está sesgada.

```
## [1] "Kurtosis: 7.8682"
```

```
##
##  Anscombe-Glynn kurtosis test
##
## data:  anacalt$Log_exposure
## kurt = 7.8682, z = 21.0267, p-value < 2.2e-16
## alternative hypothesis: kurtosis is not equal to 3
```

Dado que kurtosis > 0 la distribución tiene colas pesadas. Además el valor de p-value del test de Anscombe es menor que 0.05, por lo que podemos afirmar con seguridad que la distribución tiene colas pesadas. Podemos verlo mejor con un gráfico de densidad 12.

Se puede ver como la mayoría de valores se concentran sobre el 2.3.

Estudiamos también la normalidad de la variable Log_exposure con un test de Shapiro-Wilk y el QQ-plot. Donde vemos que el valor de p-value es menor que 0.05, por lo que la variable Log_exposure no sigue una distribución normal. Lo vemos también el QQ-plot (Figura 13) donde vemos bastante claramente que la variable Log_exposure no sigue una distribución normal.

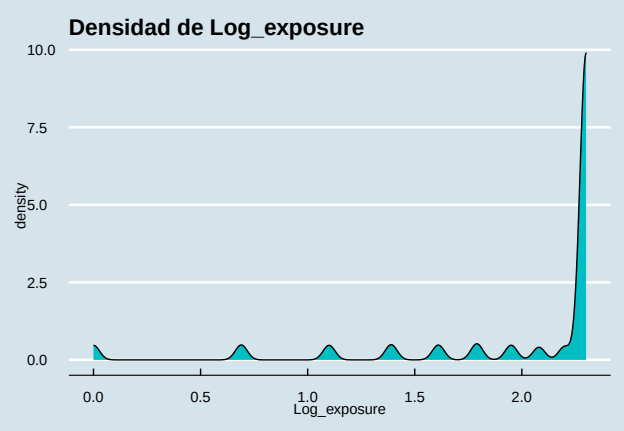

Figure 12: Densidad de Log exposure

```
##
##  Shapiro-Wilk normality test
##
## data:  anacalt$Log_exposure
## W = 0.55506, p-value < 2.2e-16
```

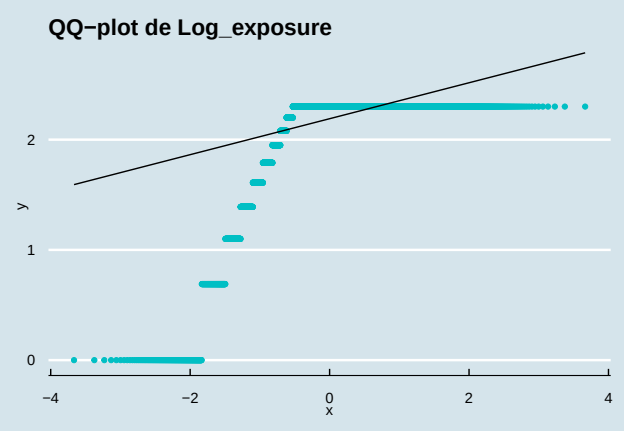

Figure 13: QQ-plot de Log exposure

Análisis bivariante

Tras el análisis de cada variable por separado, haremos un análisis bivariable, donde veremos la relación entre las variables.

Correlación entre variables numéricas

Para ver la correlación entre las variables numéricas usaremos el método de Spearman pues las variables no siguen una distribución normal (ver Figura 14).

Podemos observar como la variable de input 'Year_of_decision' tiene una correlación negativa con la variable de output 'Log_exposure', lo que indica que a mayor año de decisión, menor es el valor de 'Log_exposure'. Podemos verlo mejor con un scatterplot (Figura 15).

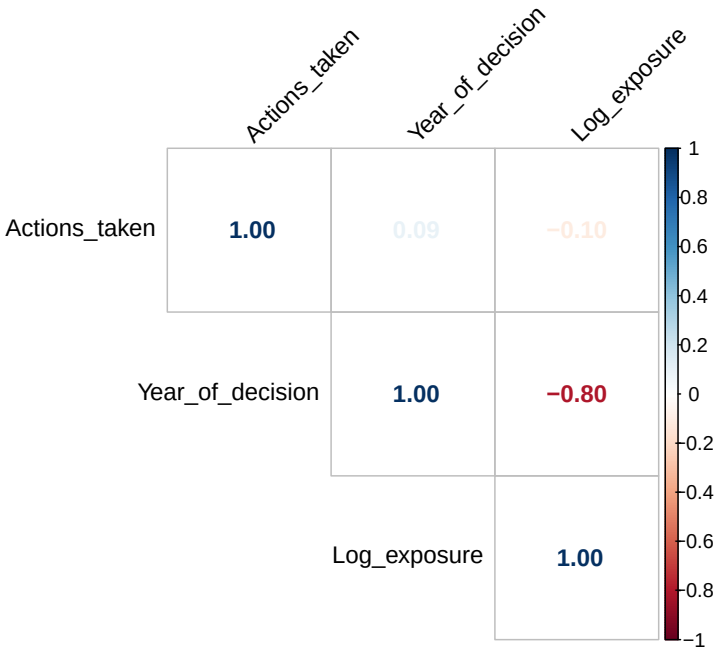

Figure 14: Correlación entre variables numéricas

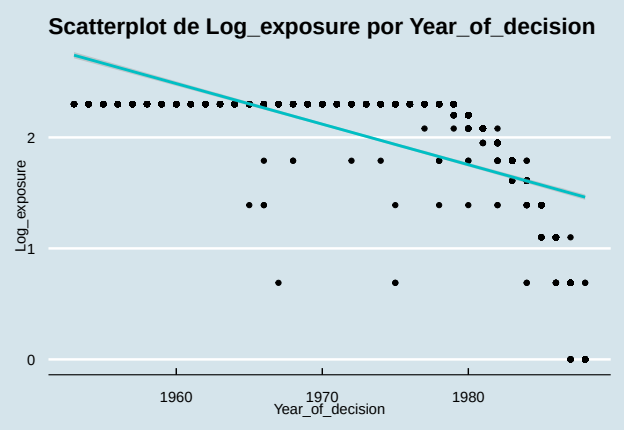

Figure 15: Scatterplot de Log exposure por Year of decision

Podemos observar como los valores previos a 1960 son prácticamente todos 2.3 y a partir de 1960 se empiezan a descender hasta 0. Por lo tanto la relación entre ambas variables es negativa y podemos verlo de mejor manera con una media por años.

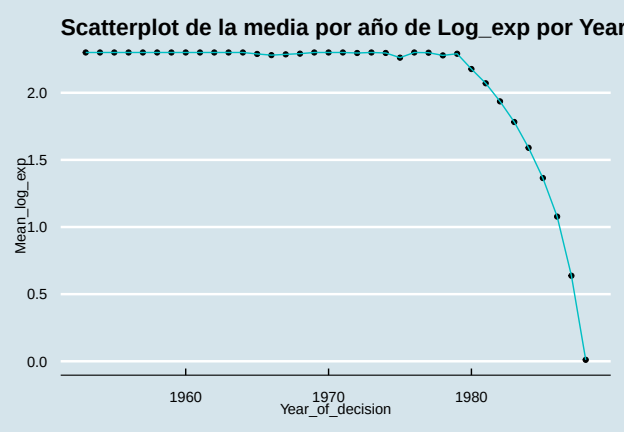

Figure 16: Scatterplot de la media por año de Log exp por Year of decision

Podemos observar como la media de Log_exposure se mantiene sobre 2.3 hasta finales de los 70 y va claramente descendiendo a medida que aumenta el año de decisión.

Relación entre variables categóricas

Para ver la relación entre las variables categóricas usaremos tablas de contingencia y gráficos de mosaico. Primero vemos una tabla de contingencia que compara cada variable en las Figuras 17 y 18.

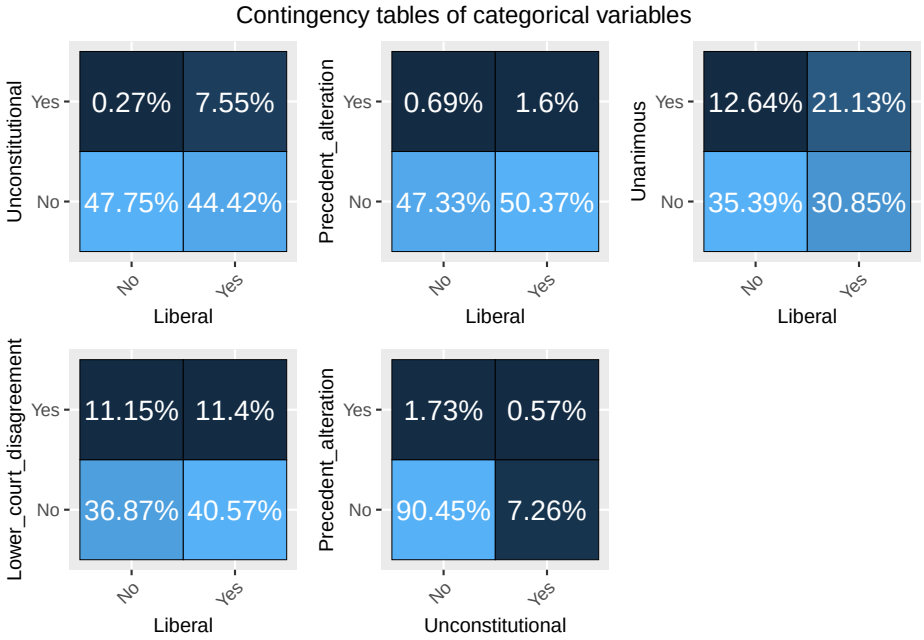

Figure 17: Tablas de contingencia de las variables categóricas (1)

Observarndo las tablas de contingencia podemos ver como algunas variables tienen una relación clara. Resalta que el 90% de los casos son 'No''Unconstitutional' y 'No' 'Precedent_alteration', por lo que parece que no

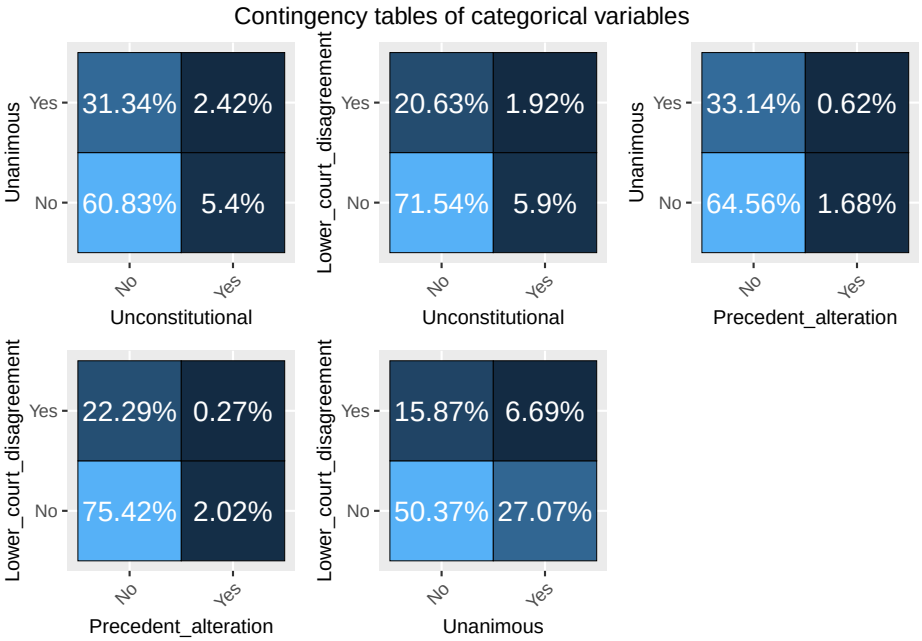

Figure 18: Tablas de contingencia de las variables categóricas (2)

hay casos inconstitucionales sin que haya una "alteración precedente". Además vemos que el 75% de los casos son 'NO' 'Precedent_alteration' y 'No' 'Lower_court_disagreement', por lo que parece que no hay muchos casos de alteración precedente sin que haya un desacuerdo de la corte inferior. Por último, vemos que el 71% de los casos son 'No' 'Lower_court_disagreement' y 'No' 'Unconstitutional', por lo que parece que no hay muchos casos de desacuerdo de la corte inferior sin que haya un caso inconstitucional.

Relación con la variable de output

Analizaremos la relación de las variables con la variable de output. Para ello, usaremos boxplots (Figura 19) para las variables categóricas y scatterplots (Figura 20) para las variables numéricas.

Al igual que lo estudiado en el apartado de correlación, vemos en la Figura 20 que Year_of_decisión parece tener una relación cuadrática con Log_exposure.

Modelo de regresión

Metodología de trabajo

Para la realización del modelo de regresión lineal, he decidido usar la técnica de backward elimination. Esta técnica consiste en ir eliminando variables del modelo hasta que todas las variables que quedan son significativas. Una vez se consiga un modelo satisfactorio a este nivel, se estudiará la posibilidad de añadir interacciones entre variables. Y tras esto se estudiará la posibilidad de añadir variables polinómicas.

Preparación del dataset

Dado que el modelo de regresión lineal necesita de variables numéricas, vamos a convertir las variables categóricas en numéricas. Como son variables binarias, convertimos los valores 'Yes' en 1 y los valores 'No' en 0.

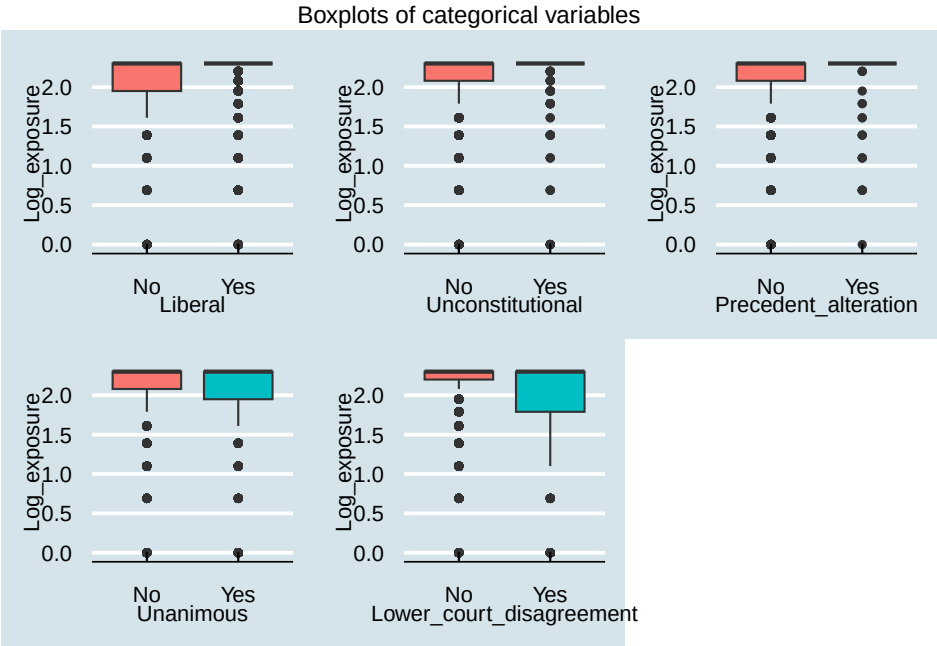

Figure 19: Boxplots de las variables categóricas

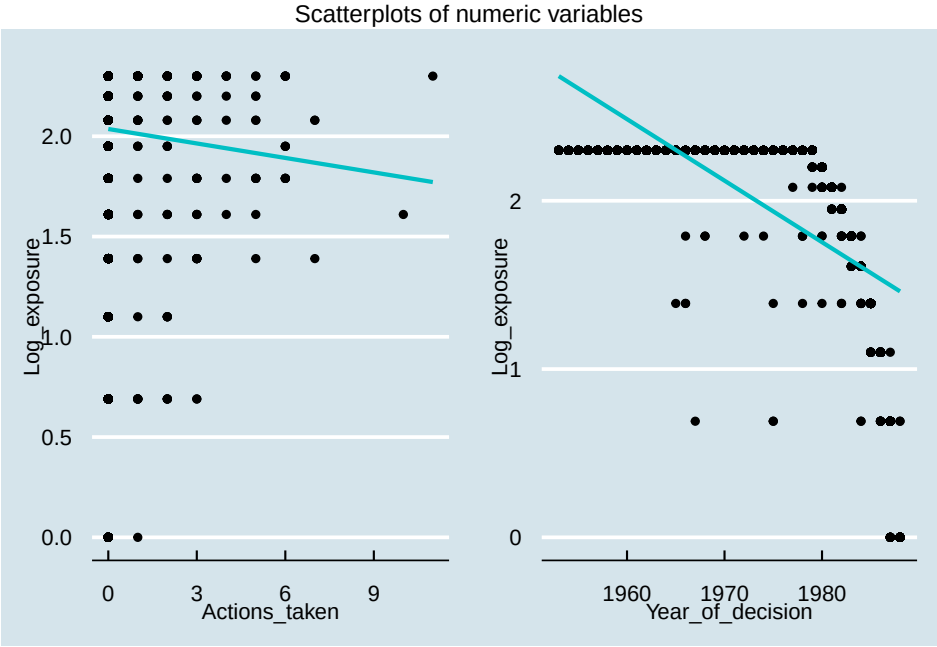

Figure 20: Scatterplots de las variables numéricas

```
##  Actions_taken        Liberal       Unconstitutional  Precedent_alteration
##  Min.   : 0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median : 0.0000   Median :1.0000   Median :0.00000   Median :0.00000
##  Mean   : 0.1088   Mean   :0.5197   Mean   :0.07823   Mean   :0.02295
##  3rd Qu.: 0.0000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :11.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##    Unanimous      Year_of_decision Lower_court_disagreement  Log_exposure
##  Min.   :0.0000   Min.   :1953     Min.   :0.0000           Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:1964     1st Qu.:0.0000           1st Qu.:2.080
##  Median :0.0000   Median :1973     Median :0.0000           Median :2.300
##  Mean   :0.3376   Mean   :1972     Mean   :0.2256           Mean   :2.034
##  3rd Qu.:1.0000   3rd Qu.:1981     3rd Qu.:0.0000           3rd Qu.:2.300
##  Max.   :1.0000   Max.   :1988     Max.   :1.0000           Max.   :2.300
```

Además, normalizamos los datos y separmos el conjunto de datos en train y test (80% y 20% respectivamente).

```
test_size <- 0.2
train_size <- 1 - test_size

set.seed(123)

# Normalizamos los datos
pObj <- preProcess(num_anacalt, method=c('range'))
num_anacalt <- predict(pObj, num_anacalt)

# Separamos el conjunto de datos en train y test
train_index <- createDataPartition(num_anacalt$Log_exposure, p = train_size, list = FALSE)
train <- num_anacalt[train_index, ] %>% dplyr::select(-Log_exposure)
test <- num_anacalt[-train_index, ] %>% dplyr::select(-Log_exposure)
train_lbl <- num_anacalt[train_index, ]$Log_exposure
test_lbl <- num_anacalt[-train_index, ]$Log_exposure
```

Modelo de regresión lineal

Modelo inicial

Primero, vamos a crear un modelo inicial con todas las variables. Para ello, usaremos la función lm(). Comenzamos con un modelo inicial con todas las variables.

```
##
## Call:
## lm(formula = train_lbl ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69106 -0.08625  0.03199  0.13065  0.25765
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.2105886  0.0085834 141.038  < 2e-16 ***
## Actions_taken             0.1279428  0.0539554   2.371   0.0178 *
```

```
## Liberal                  -0.0269254  0.0067701  -3.977 7.13e-05 ***
## Unconstitutional          0.0616941  0.0122485   5.037 4.99e-07 ***
## Precedent_alteration      0.0037307  0.0228494   0.163   0.8703
## Unanimous                 0.0005367  0.0067669   0.079   0.9368
## Year_of_decision         -0.5687389  0.0115819 -49.106  < 2e-16 ***
## Lower_court_disagreement -0.0188179  0.0075786  -2.483   0.0131 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1796 on 3235 degrees of freedom
## Multiple R-squared:  0.4387, Adjusted R-squared:  0.4375
## F-statistic: 361.2 on 7 and 3235 DF,  p-value: < 2.2e-16
```

Vemos que el modelo inicial tiene un R^2 de 0.438 y que la variable 'Unanimous' no es significativa. Por lo que la eliminamos, tras esto el modelo mantiene un R^2 ajustado de 0.438 y la variable 'Precedent_alterations' no se marca como significativa. Por lo que la eliminamos también.

```
##
## Call:
## lm(formula = train_lbl ~ . - Unanimous - Precedent_alteration,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69134 -0.08647  0.03295  0.13044  0.25740
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.210791   0.008402 144.100  < 2e-16 ***
## Actions_taken             0.127832   0.053907   2.371   0.0178 *
## Liberal                  -0.026830   0.006668  -4.024 5.86e-05 ***
## Unconstitutional          0.061857   0.012150   5.091 3.76e-07 ***
## Year_of_decision         -0.568723   0.011573 -49.142  < 2e-16 ***
## Lower_court_disagreement -0.018885   0.007563  -2.497   0.0126 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1796 on 3237 degrees of freedom
## Multiple R-squared:  0.4387, Adjusted R-squared:  0.4379
## F-statistic:   506 on 5 and 3237 DF,  p-value: < 2.2e-16
```

Tras quitar estas 2 variables vemos que el modelo se mantiene en un R^2 de 0.438 y que la variable 'Actions_taken' es la menos significativa, por lo que probamos a eliminarla. Por lo que vamos a eliminarla del modelo pero esto empeora el R^2 ajustado a 0.437, por lo que la dejamos en el modelo. Observamos como se ajusta este modelo a los datos en la Figura 21. Calculamos también el MSE (Mean Squared Error) con el conjunto de test.

```
## [1] "MSE: 0.032"
```

Interacciones

Ahora vamos a estudiar la posibilidad de añadir interacciones entre variables. Para esto, nos remitimos a la correlación estudiada anteriormente, y vemos que las variables no parecen tener casi ninguna correlación

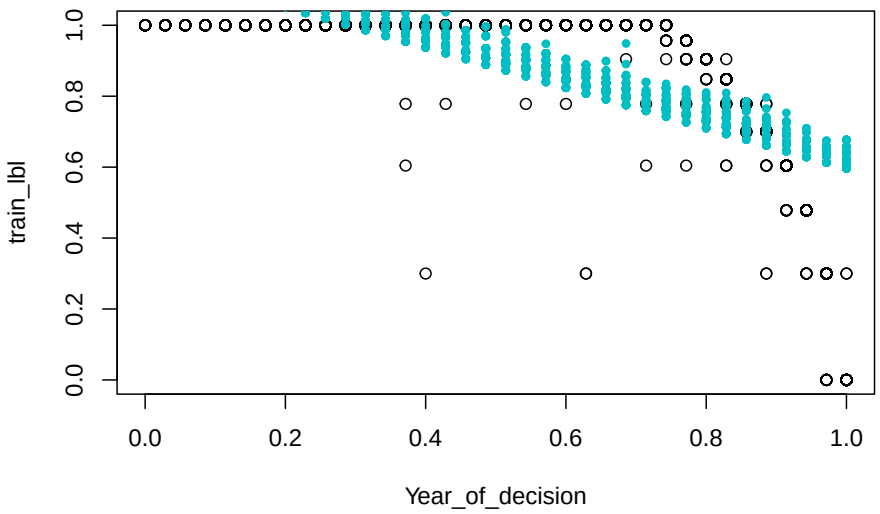

Figure 21: Ajuste del modelo inicial

entre ellas. Por lo que vamos a probar a añadir interacciones entre todas las variables e ir podando las menos significativas.

```
## [1] "R^2 ajustado: 0.4451"
```

Se puede observar como el R^2 ajustado ha pasado de 0.438 a 0.4451. Por lo que el modelo con las interacciones promete, pero eliminaremos las variables que no son significativas, o sea la interacción a 4 bandas y probamos solo con las interacciones a 3 niveles.

```
## [1] "R^2 ajustado: 0.4453"
```

```
##                                                    Estimate
## (Intercept)                                     1.222227152
## Actions_taken                                  -0.424696336
## Unconstitutional                                0.097722195
## Year_of_decision                               -0.587533673
## Liberal                                        -0.074078764
## Lower_court_disagreement                        0.077502654
## Actions_taken:Unconstitutional                 -1.884254283
## Actions_taken:Year_of_decision                  0.755180790
## Unconstitutional:Year_of_decision              -0.005863182
## Unconstitutional:Liberal                       -0.048060108
## Year_of_decision:Liberal                        0.095116855
## Actions_taken:Liberal                          -0.042882875
## Actions_taken:Lower_court_disagreement          2.331755870
## Liberal:Lower_court_disagreement               -0.026274525
## Unconstitutional:Lower_court_disagreement       0.155219155
## Year_of_decision:Lower_court_disagreement      -0.147042930
## Actions_taken:Unconstitutional:Year_of_decision 4.184823226
```

```
## Unconstitutional:Year_of_decision:Liberal                   0.013559126
## Actions_taken:Liberal:Lower_court_disagreement             -0.211468828
## Unconstitutional:Liberal:Lower_court_disagreement          -0.083889139
## Actions_taken:Year_of_decision:Lower_court_disagreement    -2.790219053
## Year_of_decision:Liberal:Lower_court_disagreement           0.006733162
## Unconstitutional:Year_of_decision:Lower_court_disagreement -0.105290464
##                                                            Std. Error
## (Intercept)                                                0.01252294
## Actions_taken                                              0.25934947
## Unconstitutional                                           0.27026690
## Year_of_decision                                           0.01931272
## Liberal                                                    0.01628697
## Lower_court_disagreement                                   0.02841556
## Actions_taken:Unconstitutional                             4.18542929
## Actions_taken:Year_of_decision                             0.34653312
## Unconstitutional:Year_of_decision                          0.42367882
## Unconstitutional:Liberal                                   0.27259225
## Year_of_decision:Liberal                                   0.02695230
## Actions_taken:Liberal                                      0.12410800
## Actions_taken:Lower_court_disagreement                     0.92111619
## Liberal:Lower_court_disagreement                           0.03666421
## Unconstitutional:Lower_court_disagreement                  0.18605166
## Year_of_decision:Lower_court_disagreement                  0.04054240
## Actions_taken:Unconstitutional:Year_of_decision            7.88027062
## Unconstitutional:Year_of_decision:Liberal                  0.42812659
## Actions_taken:Liberal:Lower_court_disagreement             0.33691317
## Unconstitutional:Liberal:Lower_court_disagreement          0.16402210
## Actions_taken:Year_of_decision:Lower_court_disagreement    1.10497084
## Year_of_decision:Liberal:Lower_court_disagreement          0.05615805
## Unconstitutional:Year_of_decision:Lower_court_disagreement 0.12114873
##                                                                t value
## (Intercept)                                                97.59903668
## Actions_taken                                              -1.63754462
## Unconstitutional                                            0.36157663
## Year_of_decision                                          -30.42210996
## Liberal                                                    -4.54834609
## Lower_court_disagreement                                    2.72747225
## Actions_taken:Unconstitutional                             -0.45019379
## Actions_taken:Year_of_decision                              2.17924564
## Unconstitutional:Year_of_decision                          -0.01383874
## Unconstitutional:Liberal                                   -0.17630768
## Year_of_decision:Liberal                                    3.52908150
## Actions_taken:Liberal                                      -0.34552870
## Actions_taken:Lower_court_disagreement                      2.53144597
## Liberal:Lower_court_disagreement                           -0.71662588
## Unconstitutional:Lower_court_disagreement                   0.83427988
## Year_of_decision:Lower_court_disagreement                  -3.62689296
## Actions_taken:Unconstitutional:Year_of_decision             0.53105070
## Unconstitutional:Year_of_decision:Liberal                   0.03167083
## Actions_taken:Liberal:Lower_court_disagreement             -0.62766566
## Unconstitutional:Liberal:Lower_court_disagreement          -0.51145021
## Actions_taken:Year_of_decision:Lower_court_disagreement    -2.52515175
## Year_of_decision:Liberal:Lower_court_disagreement           0.11989665
## Unconstitutional:Year_of_decision:Lower_court_disagreement -0.86910082
```

```
##                                                          Pr(>|t|)
## (Intercept)                                           0.000000e+00
## Actions_taken                                         1.016144e-01
## Unconstitutional                                      7.176922e-01
## Year_of_decision                                     6.628540e-179
## Liberal                                               5.606414e-06
## Lower_court_disagreement                              6.416775e-03
## Actions_taken:Unconstitutional                        6.526010e-01
## Actions_taken:Year_of_decision                        2.938567e-02
## Unconstitutional:Year_of_decision                     9.889595e-01
## Unconstitutional:Liberal                              8.600633e-01
## Year_of_decision:Liberal                              4.228398e-04
## Actions_taken:Liberal                                 7.297195e-01
## Actions_taken:Lower_court_disagreement                1.140653e-02
## Liberal:Lower_court_disagreement                      4.736569e-01
## Unconstitutional:Lower_court_disagreement             4.041852e-01
## Year_of_decision:Lower_court_disagreement             2.913007e-04
## Actions_taken:Unconstitutional:Year_of_decision       5.954203e-01
## Unconstitutional:Year_of_decision:Liberal             9.747365e-01
## Actions_taken:Liberal:Lower_court_disagreement        5.302675e-01
## Unconstitutional:Liberal:Lower_court_disagreement     6.090709e-01
## Actions_taken:Year_of_decision:Lower_court_disagreement 1.161246e-02
## Year_of_decision:Liberal:Lower_court_disagreement     9.045725e-01
## Unconstitutional:Year_of_decision:Lower_court_disagreement 3.848568e-01
```

Vemos que el R^2 ajustado ha pasado de 0.4451 a 0.4453 y la interacción entre Actions_taken, Year_of_decision y Lower_court_disagreeement tiene un p-valor aceptable. Por lo que seguimos puliendo este modelo, eliminando todas las viendo que las interacciones de 3 variables que no son significativas. Así probamos un modelo con las interacciones entre 2 variables, mantiendo la interacción de 3 identificada.

```
##
## Call:
## lm(formula = train_lbl ~ Actions_taken * Year_of_decision * Lower_court_disagreement +
##     Actions_taken * Unconstitutional + Unconstitutional * Year_of_decision +
##     Unconstitutional * Liberal + Year_of_decision * Liberal +
##     Actions_taken * Liberal + Liberal * Lower_court_disagreement +
##     Unconstitutional * Lower_court_disagreement, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70247 -0.08747  0.03472  0.13051  0.27312
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                         1.22161    0.01185
## Actions_taken                                      -0.41609    0.25823
## Year_of_decision                                   -0.58687    0.01806
## Lower_court_disagreement                            0.08041    0.02074
## Unconstitutional                                    0.11521    0.06398
## Liberal                                            -0.07419    0.01471
## Actions_taken:Year_of_decision                      0.76640    0.34572
## Actions_taken:Lower_court_disagreement              2.04525    0.80142
## Year_of_decision:Lower_court_disagreement           -0.14998    0.02734
## Actions_taken:Unconstitutional                      0.34137    0.46449
```

```
## Year_of_decision:Unconstitutional                          -0.01677    0.05274
## Unconstitutional:Liberal                                   -0.05301    0.05575
## Year_of_decision:Liberal                                    0.09649    0.02359
## Actions_taken:Liberal                                      -0.07199    0.11528
## Lower_court_disagreement:Liberal                           -0.02527    0.01607
## Lower_court_disagreement:Unconstitutional                   0.01452    0.02826
## Actions_taken:Year_of_decision:Lower_court_disagreement -2.50821    1.01169
##                                                            t value Pr(>|t|)
## (Intercept)                                                103.095  < 2e-16 ***
## Actions_taken                                               -1.611 0.107215
## Year_of_decision                                           -32.494  < 2e-16 ***
## Lower_court_disagreement                                     3.878 0.000108 ***
## Unconstitutional                                             1.801 0.071828 .
## Liberal                                                     -5.044 4.82e-07 ***
## Actions_taken:Year_of_decision                               2.217 0.026705 *
## Actions_taken:Lower_court_disagreement                       2.552 0.010755 *
## Year_of_decision:Lower_court_disagreement                   -5.485 4.44e-08 ***
## Actions_taken:Unconstitutional                               0.735 0.462429
## Year_of_decision:Unconstitutional                           -0.318 0.750568
## Unconstitutional:Liberal                                    -0.951 0.341771
## Year_of_decision:Liberal                                     4.090 4.41e-05 ***
## Actions_taken:Liberal                                       -0.625 0.532323
## Lower_court_disagreement:Liberal                            -1.572 0.116079
## Lower_court_disagreement:Unconstitutional                    0.514 0.607414
## Actions_taken:Year_of_decision:Lower_court_disagreement  -2.479 0.013218 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1782 on 3226 degrees of freedom
## Multiple R-squared:  0.4488, Adjusted R-squared:  0.4461
## F-statistic: 164.2 on 16 and 3226 DF,  p-value: < 2.2e-16
```

Vemos que el R^2 ajustado ha pasado de 0.4453 a 0.4461, Por lo que seguimos podando, en este caso, eliminamos las interacciones introducidas que no son significativas. Eso significa que solo se mantiene Year_of_decision*Liberal y la interacción de a 3.

```
##
## Call:
## lm(formula = train_lbl ~ Unconstitutional + Actions_taken * Year_of_decision *
##     Lower_court_disagreement + Year_of_decision * Liberal, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71048 -0.08862  0.03759  0.12735  0.26340
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                         1.22436    0.01172
## Unconstitutional                                    0.06006    0.01207
## Actions_taken                                      -0.42142    0.25000
## Year_of_decision                                   -0.58618    0.01803
## Lower_court_disagreement                            0.06325    0.01737
## Liberal                                            -0.07759    0.01432
## Actions_taken:Year_of_decision                      0.72580    0.34082
```

```
## Actions_taken:Lower_court_disagreement                    1.90388    0.78997
## Year_of_decision:Lower_court_disagreement                -0.14095    0.02668
## Year_of_decision:Liberal                                  0.08983    0.02285
## Actions_taken:Year_of_decision:Lower_court_disagreement -2.29843    0.98790
##                                                         t value Pr(>|t|)
## (Intercept)                                             104.460  < 2e-16 ***
## Unconstitutional                                          4.976 6.84e-07 ***
## Actions_taken                                            -1.686 0.091952 .
## Year_of_decision                                        -32.509  < 2e-16 ***
## Lower_court_disagreement                                  3.641 0.000276 ***
## Liberal                                                  -5.420 6.41e-08 ***
## Actions_taken:Year_of_decision                            2.130 0.033283 *
## Actions_taken:Lower_court_disagreement                    2.410 0.016005 *
## Year_of_decision:Lower_court_disagreement                 -5.282 1.36e-07 ***
## Year_of_decision:Liberal                                  3.932 8.59e-05 ***
## Actions_taken:Year_of_decision:Lower_court_disagreement  -2.327 0.020049 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1782 on 3232 degrees of freedom
## Multiple R-squared:  0.4481, Adjusted R-squared:  0.4464
## F-statistic: 262.4 on 10 and 3232 DF,  p-value: < 2.2e-16
```

Vemos que hemos conseguido un R^2 ajustado de 0.4464, mejorándolo. Si probamos a eliminar la interacción de 3 variables que es la menos significativa y sustituirla por las interacciones de a 2, vemos que el R^2 ajustado baja a 0.4456 por lo que lo dejamos con ella. Vemos como este modelo se ajusta a los datos en la Figura 22. Calculamos también el MSE (Mean Squared Error) con el conjunto de test.

```
## [1] "MSE: 0.0317"
```

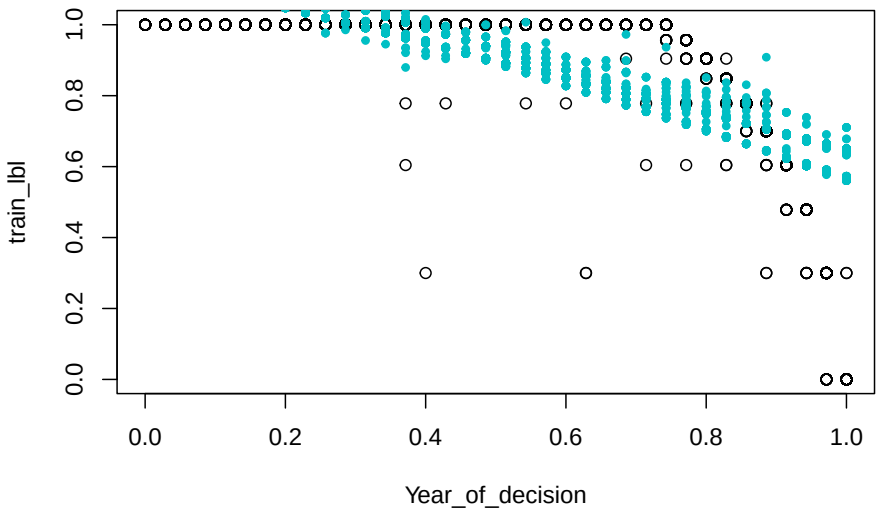

Figure 22: Ajuste del modelo con interacciones

Variables polinómicas

Ahora vamos a estudiar la posibilidad de añadir variables polinómicas. Para ello, estudiamos las variables en relación con la variable de output en la Figura 23.

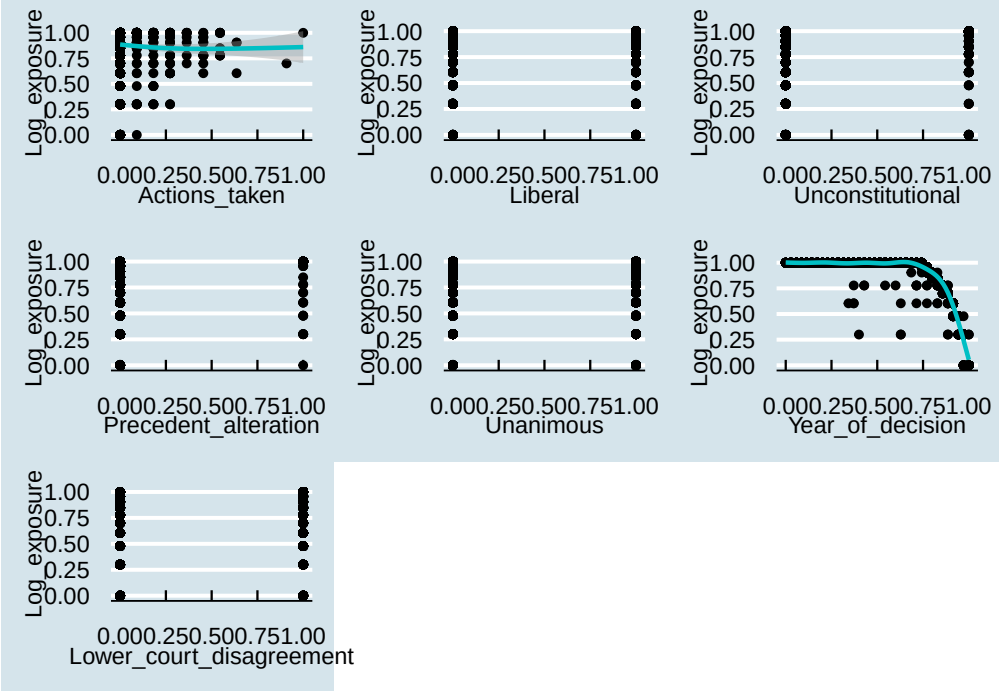

Figure 23: Relación de las variables con la salida

Observammos claramente que la variable 'Year_of_decision' tiene una relación cuadrática con la variable de output. Por lo que vamos a añadir una variable polinómica de grado 2 de 'Year_of_decision'.

```
##
## Call:
## lm(formula = train_lbl ~ Unconstitutional + Actions_taken * Year_of_decision *
##     Lower_court_disagreement + Year_of_decision * Liberal + I(Year_of_decision^2),
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66719 -0.07020 -0.00678  0.08474  0.28094
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                       0.8177403  0.0093725
## Unconstitutional                                 -0.0033972  0.0076467
## Actions_taken                                    -1.0586195  0.1575175
## Year_of_decision                                  1.4724720  0.0314164
## Lower_court_disagreement                          0.0193753  0.0109466
## Liberal                                          -0.0162062  0.0090481
## I(Year_of_decision^2)                            -1.9205788  0.0273326
## Actions_taken:Year_of_decision                    1.5329778  0.2146916
## Actions_taken:Lower_court_disagreement            0.9711665  0.4970884
```

```
## Year_of_decision:Lower_court_disagreement               -0.0373579  0.0168488
## Year_of_decision:Liberal                                 0.0006344  0.0144260
## Actions_taken:Year_of_decision:Lower_court_disagreement -1.2393915  0.6215945
##                                                        t value Pr(>|t|)
## (Intercept)                                             87.249  < 2e-16 ***
## Unconstitutional                                        -0.444   0.6569
## Actions_taken                                           -6.721 2.13e-11 ***
## Year_of_decision                                        46.870  < 2e-16 ***
## Lower_court_disagreement                                 1.770   0.0768 .
## Liberal                                                 -1.791   0.0734 .
## I(Year_of_decision^2)                                  -70.267  < 2e-16 ***
## Actions_taken:Year_of_decision                           7.140 1.14e-12 ***
## Actions_taken:Lower_court_disagreement                   1.954   0.0508 .
## Year_of_decision:Lower_court_disagreement               -2.217   0.0267 *
## Year_of_decision:Liberal                                 0.044   0.9649
## Actions_taken:Year_of_decision:Lower_court_disagreement  -1.994   0.0462 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1121 on 3231 degrees of freedom
## Multiple R-squared:  0.7817, Adjusted R-squared:  0.781
## F-statistic:  1052 on 11 and 3231 DF,  p-value: < 2.2e-16
```

Observamos que hemos pegado un salto significativo en el R^2 ajustado, pasando de 0.4464 a 0.781. Por lo que vamos a dejar el modelo con la variable polinómica. Tras esto, podemos probar a eliminar algunas de las variables que no son significativas ya que esta adición de la variable polinómica ha podido hacer que algunas variables que antes sí eran significativas ahora ya no. Es el caso de la interacción de 'Year_of_decision' y 'Liberal'. Al eliminarla el R^2 ajustado se mantiene igual, por lo que la eliminamos. Además, la variable 'Unconstitutional' ya no es significativa, por lo que probamos a eliminarla también

```
##
## Call:
## lm(formula = train_lbl ~ Liberal + Actions_taken * Year_of_decision *
##     Lower_court_disagreement + I(Year_of_decision^2), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66722 -0.07302 -0.00661  0.08465  0.28080
##
## Coefficients:
##                                                         Estimate Std. Error
## (Intercept)                                             0.817887   0.007750
## Liberal                                                -0.016344   0.004021
## Actions_taken                                          -1.058360   0.157464
## Year_of_decision                                        1.471208   0.029418
## Lower_court_disagreement                                0.019350   0.010943
## I(Year_of_decision^2)                                  -1.919218   0.027017
## Actions_taken:Year_of_decision                          1.533748   0.214590
## Actions_taken:Lower_court_disagreement                  0.975282   0.496850
## Year_of_decision:Lower_court_disagreement               -0.037395   0.016844
## Actions_taken:Year_of_decision:Lower_court_disagreement -1.244957   0.621225
##                                                        t value Pr(>|t|)
## (Intercept)                                            105.540  < 2e-16 ***
## Liberal                                                 -4.065 4.92e-05 ***
```

```
## Actions_taken                                          -6.721 2.12e-11 ***
## Year_of_decision                                       50.010  < 2e-16 ***
## Lower_court_disagreement                                1.768   0.0771 .
## I(Year_of_decision^2)                                 -71.037  < 2e-16 ***
## Actions_taken:Year_of_decision                          7.147 1.09e-12 ***
## Actions_taken:Lower_court_disagreement                  1.963   0.0497 *
## Year_of_decision:Lower_court_disagreement               -2.220   0.0265 *
## Actions_taken:Year_of_decision:Lower_court_disagreement -2.004   0.0451 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1121 on 3233 degrees of freedom
## Multiple R-squared:  0.7817, Adjusted R-squared:  0.7811
## F-statistic:  1286 on 9 and 3233 DF,  p-value: < 2.2e-16
```

Vemos que nuestro R^2 ajustado ha incrementado casi imperceptiblemente, pero este es más sencillo de interpretar. Probamos también a eliminar la interacción a 3 niveles, ya que es la menos significativa, y vemos que el R^2 ajustado baja, por lo que la dejamos. A continuación, ya que hemos añadido una Year_of_decision^2 y ha funcionado tan bien, probamos a añadir Year_of_decision^3.

```
##
## Call:
## lm(formula = train_lbl ~ Liberal + Actions_taken * Year_of_decision *
##     Lower_court_disagreement + I(Year_of_decision^2) + I(Year_of_decision^3),
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70946 -0.04056  0.01221  0.04898  0.16210
##
## Coefficients:
##                                                         Estimate Std. Error
## (Intercept)                                             1.135067   0.005645
## Liberal                                                 0.003745   0.002230
## Actions_taken                                          -0.881614   0.086861
## Year_of_decision                                       -1.949215   0.042947
## Lower_court_disagreement                                0.007733   0.006036
## I(Year_of_decision^2)                                   6.069486   0.094062
## I(Year_of_decision^3)                                  -5.091218   0.059189
## Actions_taken:Year_of_decision                          1.104623   0.118444
## Actions_taken:Lower_court_disagreement                  0.520234   0.274048
## Year_of_decision:Lower_court_disagreement               -0.013596   0.009293
## Actions_taken:Year_of_decision:Lower_court_disagreement -0.619932   0.342662
##                                                        t value Pr(>|t|)
## (Intercept)                                            201.090   <2e-16 ***
## Liberal                                                  1.680   0.0931 .
## Actions_taken                                          -10.150   <2e-16 ***
## Year_of_decision                                       -45.387   <2e-16 ***
## Lower_court_disagreement                                 1.281   0.2003
## I(Year_of_decision^2)                                   64.527   <2e-16 ***
## I(Year_of_decision^3)                                  -86.016   <2e-16 ***
## Actions_taken:Year_of_decision                           9.326   <2e-16 ***
## Actions_taken:Lower_court_disagreement                   1.898   0.0577 .
## Year_of_decision:Lower_court_disagreement               -1.463   0.1436
```

```
## Actions_taken:Year_of_decision:Lower_court_disagreement  -1.809   0.0705 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06179 on 3232 degrees of freedom
## Multiple R-squared:  0.9336, Adjusted R-squared:  0.9334
## F-statistic:  4546 on 10 and 3232 DF,  p-value: < 2.2e-16
```

Vemos que sorprendentemente, hay un salto significativo a un R^2 ajustado de 0.9334. Vemos que nuestra interacción "estrella" ya no parece ser tan significativa, por lo que probamos a eliminarla y el R^2 ajustado se mantiene. Tras esto vemos que además las interacciones de 2 niveles de Actions_taken con Lower_court_disagreement y Year_of_decision con Lower_court_disagreement ya no son significativas, por lo que las eliminamos también.

```
##
## Call:
## lm(formula = train_lbl ~ Liberal + Lower_court_disagreement +
##     Actions_taken * Year_of_decision + I(Year_of_decision^2) +
##     I(Year_of_decision^3), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71123 -0.04089  0.01117  0.04807  0.15434
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.1364096  0.0055573 204.491   <2e-16 ***
## Liberal                         0.0038609  0.0022274   1.733   0.0831 .
## Lower_court_disagreement        0.0001877  0.0026076   0.072   0.9426
## Actions_taken                  -0.8359121  0.0816293 -10.240   <2e-16 ***
## Year_of_decision               -1.9527494  0.0429177 -45.500   <2e-16 ***
## I(Year_of_decision^2)           6.0744191  0.0940729  64.571   <2e-16 ***
## I(Year_of_decision^3)          -5.0961926  0.0591704 -86.127   <2e-16 ***
## Actions_taken:Year_of_decision  1.0488923  0.1089605   9.626   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06183 on 3235 degrees of freedom
## Multiple R-squared:  0.9335, Adjusted R-squared:  0.9334
## F-statistic:  6487 on 7 and 3235 DF,  p-value: < 2.2e-16
```

Se mantiene el R^2 ajustado, y vemos que la variable Lower_court_disagreement ya no es significativa, eso explica por que las interacciones tampoco lo eran, por lo que la eliminamos.

```
##
## Call:
## lm(formula = train_lbl ~ Liberal + Actions_taken * Year_of_decision +
##     I(Year_of_decision^2) + I(Year_of_decision^3), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71127 -0.04093  0.01113  0.04804  0.15433
##
```

```
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    1.136444   0.005536 205.284   <2e-16 ***
## Liberal                        0.003862   0.002227   1.734    0.083 .
## Actions_taken                 -0.836035   0.081599 -10.246   <2e-16 ***
## Year_of_decision              -1.952743   0.042911 -45.507   <2e-16 ***
## I(Year_of_decision^2)          6.074394   0.094058  64.581   <2e-16 ***
## I(Year_of_decision^3)         -5.096144   0.059158 -86.145   <2e-16 ***
## Actions_taken:Year_of_decision  1.049026   0.108928   9.630   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06182 on 3236 degrees of freedom
## Multiple R-squared:  0.9335, Adjusted R-squared:  0.9334
## F-statistic:  7571 on 6 and 3236 DF,  p-value: < 2.2e-16
```

Se mantiene el R^2 ajustado y vemos como parece que la variable Liberal también ha dejado de ser significativa, por lo que la eliminamos.

```
##
## Call:
## lm(formula = train_lbl ~ Actions_taken * Year_of_decision + I(Year_of_decision^2) +
##     I(Year_of_decision^3), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71318 -0.03920  0.00954  0.04965  0.15162
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    1.138531   0.005405 210.637   <2e-16 ***
## Actions_taken                 -0.835528   0.081624 -10.236   <2e-16 ***
## Year_of_decision              -1.947144   0.042803 -45.491   <2e-16 ***
## I(Year_of_decision^2)          6.057918   0.093606  64.717   <2e-16 ***
## I(Year_of_decision^3)         -5.085534   0.058858 -86.403   <2e-16 ***
## Actions_taken:Year_of_decision  1.049787   0.108961   9.635   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06183 on 3237 degrees of freedom
## Multiple R-squared:  0.9334, Adjusted R-squared:  0.9333
## F-statistic:  9079 on 5 and 3237 DF,  p-value: < 2.2e-16
```

Vemos que el modelo ha pasado de un R^2 ajustado de 0.9334 a 0.9333 por lo que, al ser un cambio tan pequeño y dado que el modelo pasa a ser más interpretable, vamos a dejar este modelo. Siguiendo con la misma lógica hasta ahora, probaremos a añaadir Year_of_decision^4.

```
##
## Call:
## lm(formula = train_lbl ~ Actions_taken * Year_of_decision + I(Year_of_decision^2) +
##     I(Year_of_decision^3) + I(Year_of_decision^4), data = train)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.68369 -0.01281  0.00122  0.01640  0.23451
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      0.941361   0.004426 212.680   <2e-16 ***
## Actions_taken                   -0.605521   0.051625 -11.729   <2e-16 ***
## Year_of_decision                 1.403190   0.055007  25.509   <2e-16 ***
## I(Year_of_decision^2)           -7.934629   0.208656 -38.027   <2e-16 ***
## I(Year_of_decision^3)           15.698688   0.299561  52.406   <2e-16 ***
## I(Year_of_decision^4)          -10.047861   0.143701 -69.922   <2e-16 ***
## Actions_taken:Year_of_decision   0.657709   0.069003   9.532   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03903 on 3236 degrees of freedom
## Multiple R-squared:  0.9735, Adjusted R-squared:  0.9734
## F-statistic: 1.981e+04 on 6 and 3236 DF,  p-value: < 2.2e-16
```

Vemos que el R^2 ajustado ha pasado de 0.9333 a 0.9734, por lo que conseguimos otro salto. Probamos por lo tanto con Year_of_decision^5.

```
##
## Call:
## lm(formula = train_lbl ~ Actions_taken * Year_of_decision + I(Year_of_decision^2) +
##     I(Year_of_decision^3) + I(Year_of_decision^4) + I(Year_of_decision^5),
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66620 -0.00303  0.00019  0.00269  0.26726
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      1.012997   0.004953 204.535  < 2e-16 ***
## Actions_taken                   -0.623318   0.047236 -13.196  < 2e-16 ***
## Year_of_decision                -0.433806   0.088766  -4.887 1.07e-06 ***
## I(Year_of_decision^2)            4.010880   0.512383   7.828 6.67e-15 ***
## I(Year_of_decision^3)          -14.633710   1.238105 -11.819  < 2e-16 ***
## I(Year_of_decision^4)           22.834309   1.315471  17.358  < 2e-16 ***
## I(Year_of_decision^5)          -12.762099   0.507999 -25.122  < 2e-16 ***
## Actions_taken:Year_of_decision   0.669193   0.063131  10.600  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03571 on 3235 degrees of freedom
## Multiple R-squared:  0.9778, Adjusted R-squared:  0.9778
## F-statistic: 2.037e+04 on 7 and 3235 DF,  p-value: < 2.2e-16
```

Vemos que incrementar el exponente sigue mejorando el modelo, llegando a R^2 de 0.9778 por lo que probamos con Year_of_decision^6 y vemos que el R^2 ajustado ha pasado de 0.9778 a 0.9779 por lo que no parece que tenga sentido seguir incrementando el exponente y nos quedamos con el modelo anterior.

Hemos conseguido un modelo que alcanza un R^2 ajustado de 0.9778. Lo que significa que el 97.78% de la variabilidad de la variable Log_exposure es explicada por nuestro modelo. Observamos como se ajusta este

modelo a los datos en la Figura 24. Calculamos también el MSE (Mean Squared Error) con el conjunto de test.

```
## [1] "MSE: 0.0011"
```

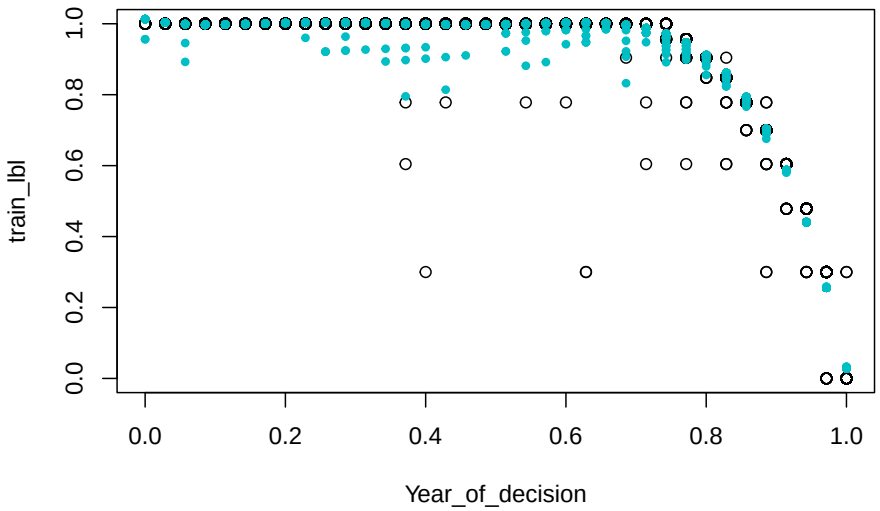

Figure 24: Ajuste del modelo polinómico

Validación cruzada (K-fold cross validation)

Para comprobar que nuestro modelo no está sobreajustado, vamos a realizar una validación cruzada con las 5 particiones que se indicaron junto al dataset. Por un lado usando el modelo de todas las variables y por otro usando el modelo polinómico que hemos obtenido `'Y ~ X1*X6 + I(X6^2) + I(X6^3) + I(X6^4) + I(X6^5)'`. Donde podemos observar como el MSE es mucho mejor en nuestro modelo que en el modelo con todas las variables.

```
## [1] "MSE del modelo con todas las variables en train: 0.1699"
```

```
## [1] "MSE del modelo con todas las variables en test: 0.171"
```

```
## [1] "MSE del modelo polinómico en train: 0.007"
```

```
## [1] "MSE del modelo polinómico en test: 0.0074"
```

Regresión usando K-Nearest Neighbors

Vamos a probar a realizar la regresión usando el algoritmo K-Nearest Neighbors. Para ello vamos a usar la librería `kknn`. Primero vamos a probar a usar todas las variables y luego vamos a probar a usar las variables más significativas que hemos obtenido con la regresión lineal.

Modelo con todas las variables

Probamos primeramente con todas las variables y con `k = 7` que es el valor por defecto.

```
modelknn <- kknn(train_lbl ~ ., train, test) # Por defecto k = 7
```

```
## [1] "MSE frente a test: 0.0022"
```

Si nos fijamos en el MSE el modelo de KNN parece prometedor, ya que parece tener menos error que el modelo linear de regresión. Probamos a ajustar el valor de K para ver si mejora el modelo, podemos ver el rendimiento en la Figura 25.

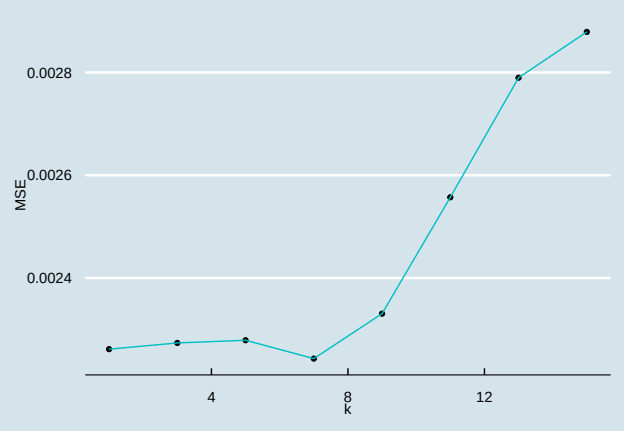

Figure 25: Ajuste del valor de K frente MSE

Podemos observar como el modelo con `k=7` es el que mejor resultado tiene. Ploteamos el modelo para ver cómo se ajusta a los datos (Figura 26).

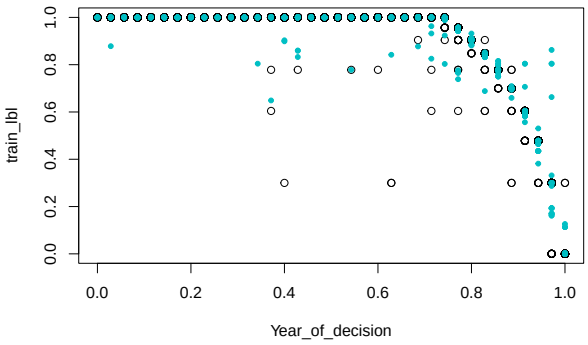

Figure 26: Modelo KNN con k=7

Modelo con la fórmula usada en regresión lineal

Podemos probar a ver si usando la misma fórmula que con la regresión lineal (`Log_exposure ~ Actions_taken*Year_of_decision + I(Year_of_decision^2) + I(Year_of_decision^3) + I(Year_of_decision^4) + I(Year_of_decision^5)`) mejoramos el modelo. Usamos `k = 7` que es el valor por defecto.

```
## [1] "MSE frente a test: 0.001"
```

Vemos que el error ha disminuido por lo que incluso en KNN parece que la fórmula polinómica tiene bastante éxito. Ploteamos el modelo para ver cómo se ajusta a los datos (Figura 27).

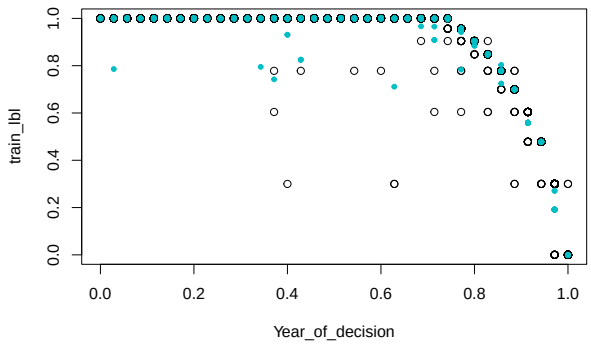

Figure 27: Modelo KNN con k=7 y fórmula customizada

Validación cruzada (K-fold cross validation)

Para asegurarnos de que esta disminución del error no es casualidad o un sobreajuste, vamos a realizar una validación cruzada.

```
## [1] "MSE del modelo con todas las variables en train: 0.0063"
```

```
## [1] "MSE del modelo con todas las variables en test: 0.0115"
```

```
## [1] "MSE del modelo con la fórmula polinómica en train: 0.0043"
```

```
## [1] "MSE del modelo con la fórmula polinómica en test: 0.0073"
```

Podemos observar como KNN suele sobreajustar, podemos intuirlo ya que tiene casi el doble de error en test que en train. Aun así, los datos de error son menores en general que los de regresión lineal.

Comparación de los modelos

Como paso final, vamos a comparar los modelos de regresión lineal y KNN con M5P. Los resultados de los correspondientes MSE de los modelos con todas las variables los tenemos almacenados en CSV's. Los cargamos y los comparamos.

```
## [1] "Summary de los datos de train:"
```

```
##       X              out_train_lm       out_train_kknn      out_train_m5p
##  Length:18          Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00
##  Class :character   1st Qu.:0.000e+00   1st Qu.:0.000e+00   1st Qu.:0.000e+00
##  Mode  :character   Median :5.000e+00   Median :2.000e+00   Median :3.000e+00
##                     Mean   :3.827e+08   Mean   :1.159e+08   Mean   :1.943e+08
##                     3rd Qu.:8.300e+01   3rd Qu.:2.200e+01   3rd Qu.:2.400e+01
##                     Max.   :4.826e+09   Max.   :1.561e+09   Max.   :2.559e+09
```

```
## [1] "Summary de los datos de test:"
```

```
##       X              out_test_lm        out_test_knn        out_test_m5p
##  Length:18          Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00
##  Class :character   1st Qu.:0.000e+00   1st Qu.:0.000e+00   1st Qu.:0.000e+00
##  Mode  :character   Median :5.000e+00   Median :6.000e+00   Median :3.000e+00
##                     Mean   :3.843e+08   Mean   :2.929e+08   Mean   :2.480e+08
##                     3rd Qu.:8.500e+01   3rd Qu.:5.300e+01   3rd Qu.:3.100e+01
##                     Max.   :4.844e+09   Max.   :3.846e+09   Max.   :3.158e+09
```

Para visualizarlo de mejor manera ploteamos los diferentes MSE para nuestro dataset (Figura 28)

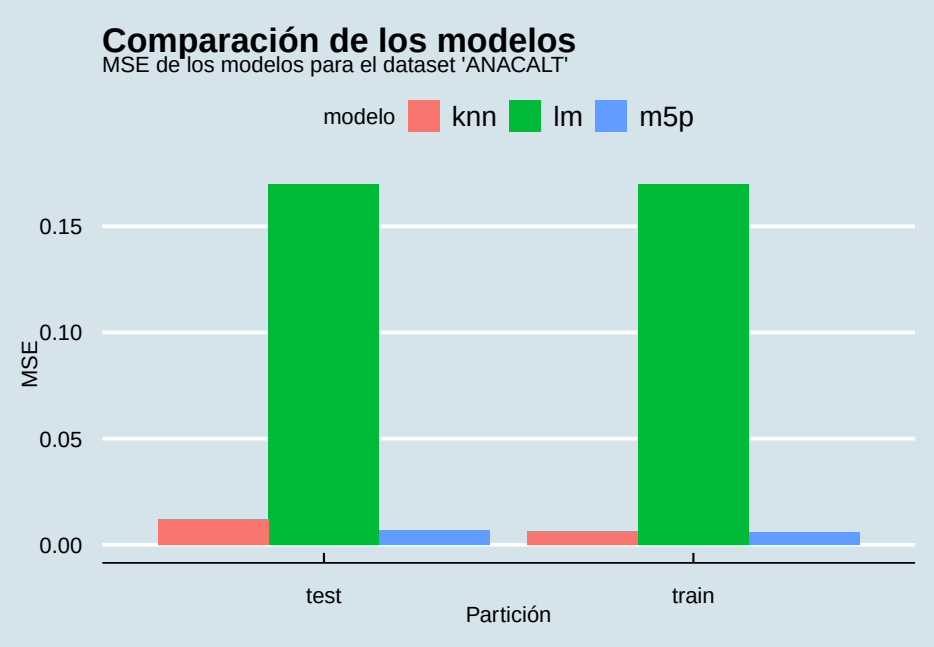

Figure 28: Comparación de los MSE de los modelos de regresión lineal, KNN y M5P para el dataset ANACALT

```
## List of 2
##  $ plot.title   :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 0.5
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ plot.subtitle:List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
```

```
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 0.5
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

De primeras, podemos observar como LM tiene bastante peor rendimiento en los modelos de todas las variables. Para hacer una comparación rigurosa, realizamos un test de Friedman.

```
##
##  Friedman rank sum test
##
## data:  .
## Friedman chi-squared = 8.4444, df = 2, p-value = 0.01467
```

Como el p-valor es menor que 0.05, podemos rechazar la hipótesis nula de que los modelos tienen el mismo rendimiento. Para saber cuáles son los modelos que tienen un rendimiento significativamente diferente, realizamos una comparación a pares con el post hoc de Holm.

```
##
##  Pairwise comparisons using Wilcoxon signed rank exact test
##
## data:  al_matrix and groups
##
##   1     2
## 2 0.580 -
## 3 0.081 0.108
##
## P value adjustment method: holm
```

Vemos que el p-valor entre el primer y el segundo algoritmo es 0.580, por lo que hay diferencias significativas entre ellos. O sea, parece haber diferencias significativas entre LM y KNN. El p-valor entre el primer y el tercer algoritmo es 0.108, por lo que también hay diferencias significativas entre ellos. O sea, parece haber diferencias significativas entre LM y M5'. Y si comparamos el segundo y el tercer algoritmo, el p-valor es 0.108, por lo que también hay diferencias significativas entre ellos. O sea, parece haber diferencias significativas entre KNN y M5'. Por lo tanto, hay diferencias significativas entre todos los algoritmos y dado que M5' es el que consigue el p-valor más pequeño, podemos decir que es el mejor algoritmo.

Anexo : Código de la práctica

Dado que esto se ha realizado en un archivo de RMarkdown, se puede ver el código de la práctica en el archivo `pima-cladificacion.Rmd` que se encuentra en el repositorio de GitHub. Click aquí