Trabajo Práctico

Barraza, Veronica y Maldonado, Kevin

Junio 2022

Índice

Planteo del problema

En este trabajo práctico se utilizará un conjunto de datos correspondiente a una encuesta con escala de tipo Likert (es decir, se pide al encuestado marcar un entero entre 1 y 5, donde 1 = Totalmente en desacuerdo y 5 = Totalmente de acuerdo). La encuesta consiste de 44 preguntas muy variadas, como "Disfruto de bailar" o "Creo que un desastre climatico podríia llegar a ser divertido". Para los individuos encuestados, se tienen también otras variables extra-encuesta que pueden ser de interés, como por ejemplo edad, géenero, religión, etc. El dataset contiene observaciones y 58 variables, el mismo dataset fue separado en un dataset de training y otro de testeo.

Para que la separación sea adecuada, los niveles de los factores en ambos subset tienen que ser iguales. Por lo tanto, para hacer esta separación utilizamos la libreria caret.

Caso de estudio

Supongamos que queremos modelar la respuesta de una de las preguntas del dataset en función de la edad y el género. En este caso: ¿Cuál sería el problema teórico de usar una regresión lineal para esto? ¿Cuál sería el problema de usar una regresión multinomial en este problema?

Para modelar una respuesta que toma valores enteros entre 1 a 5 necesitamos un modelo que cumpla con dos condiciones:

1. que devuelva valores discretos, que podamos *mapear* a estas cinco categorías,
2. a su vez que estos valores discretos tengan un orden que se corresponda con el orden natural de los valores de 1 a 5.

La regresión lineal nos da un modelo que toma valores reales no acotados; es naturalmente ordenado pero no es claro a priori cómo convertirlos en cinco categorías discretas. Por otro lado, la regresión multinomial nos provee de estas predicciones discretas que modelan la graduación de las respuestas de la pregunta, pero estas categorías no tienen el orden natural que buscamos.

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   country = col_character()
## )
## i Use `spec()` for the full column specifications.
```

Como hay dos preguntas repetidas (27 y 43): I think a natural disaster would be kind of exciting. Chequeamos que las respuestas sean iguales,pero dado que las respuestas son diferentes las excluimos del análisis. Veamos algunas relaciones entre las variables demográficas. También analizamosos histogramas de frecuencias de las variables factores relacionadas con las preguntas (se presentan algunos en la fig. 2).

```
## [1] FALSE
```

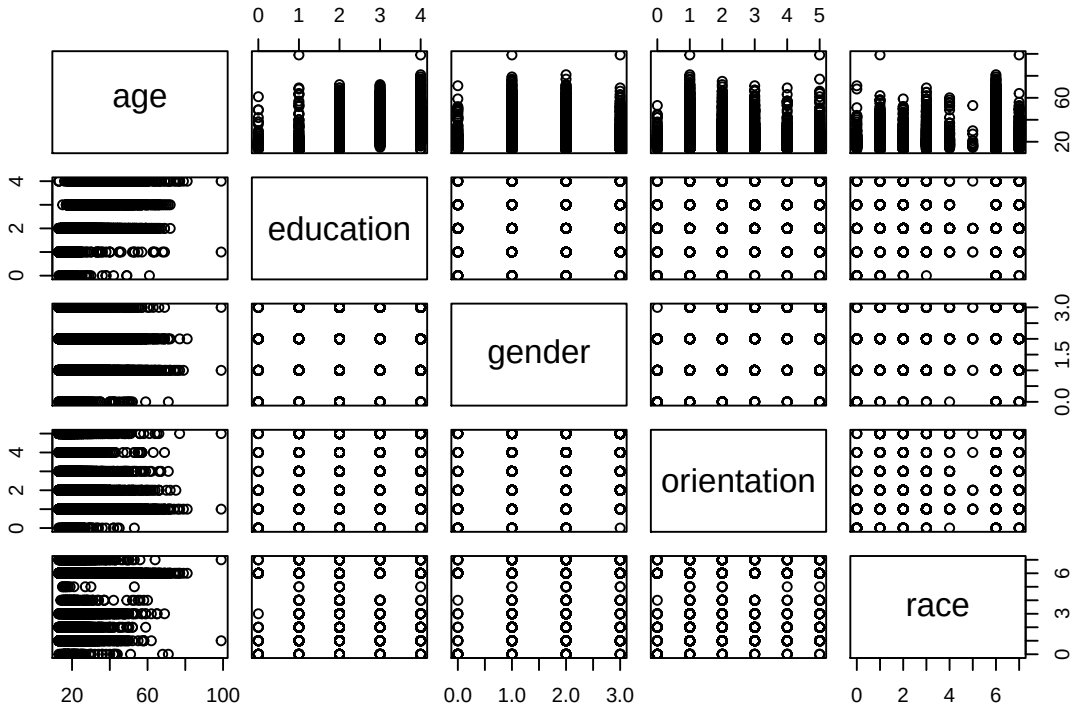

Figura 1: Paiplor entre algunas de las variables del dataset

Regresión ordinal

La regresión ordinal intenta cumplir ambos objetivos. El objetivo es: si tenemos variables explicativas x y una variable respuesta y, vamos a modelar $P(y \leq i|x)$ para cada $i = 1, \ldots, 4$. Es fácil ver que estos valores nos permiten calcular $P(y = i|x)$ para cada $i = 1, \ldots, 5$. Basta con que estas probabilidades sean crecientes en i. Vamos a modelarlas como lineales en x, con un término independiente creciente en i que nos asegura el orden de las probabilidades que necesitamos, y con una función de link (monótona) con imagen en el intervalo $(0, 1)$. Esto es, $P(y \leq i|x) = s(x^t\beta + \theta_i)$, con $\theta_1 < \cdots < \theta_4$ y s función de link: por ejemplo, la función sigmoidea (modelo *logit*) o la función de distribución acumulada de la normal estándar (modelo *probit*).

Vamos a trabajar con la pregunta 37: *I have played a lot of video games*. Es natural pensar que podría estar correlacionar con la edad.

Veamos qué distribución tienen las variables $Q37$ y *age*.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Intentemos ver ahora si hay alguna relación entre las variables. Grafiquemos el promedio de la respuesta 37 en función de la edad.

```
##                    age meanAnswer
## age         1.0000000 -0.6667442
## meanAnswer -0.6667442  1.0000000
```

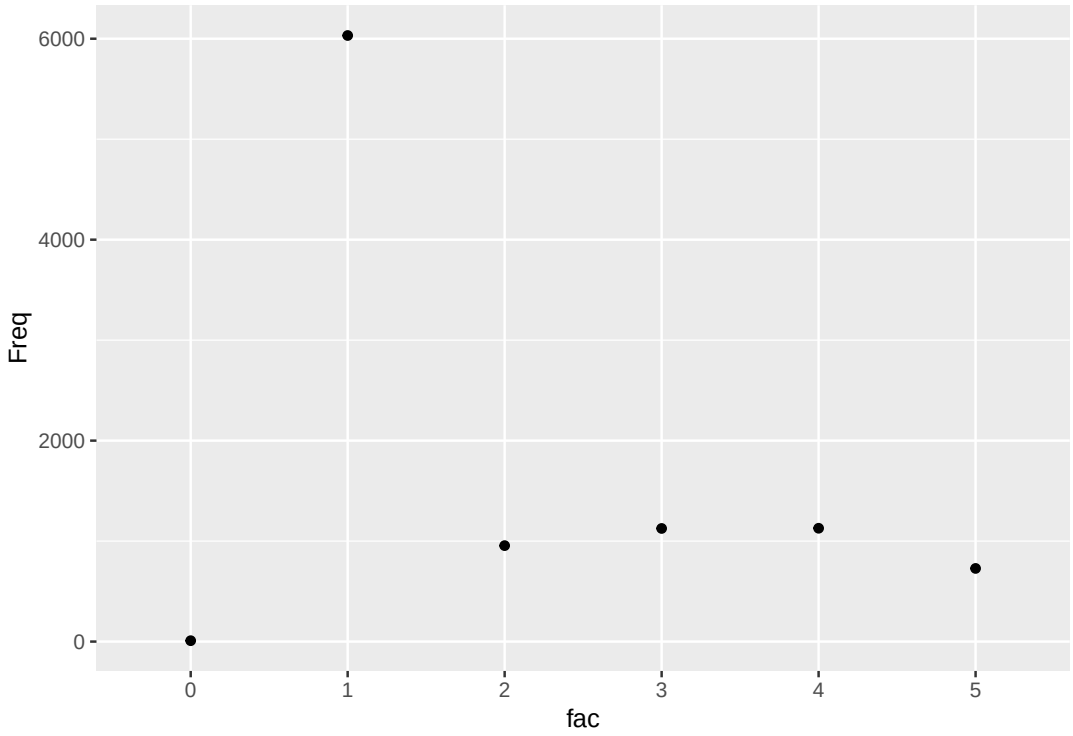

Figura 2: Frecuencias de algunas clases de factores

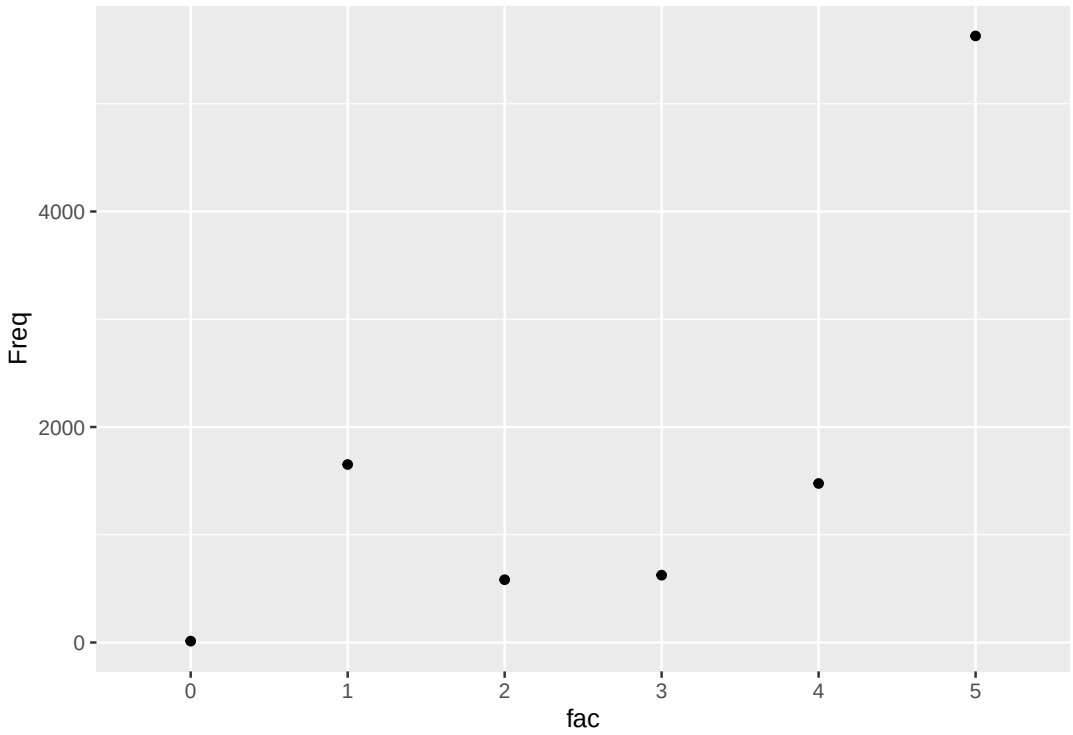

Figura 3: Frecuencias de algunas clases de factores

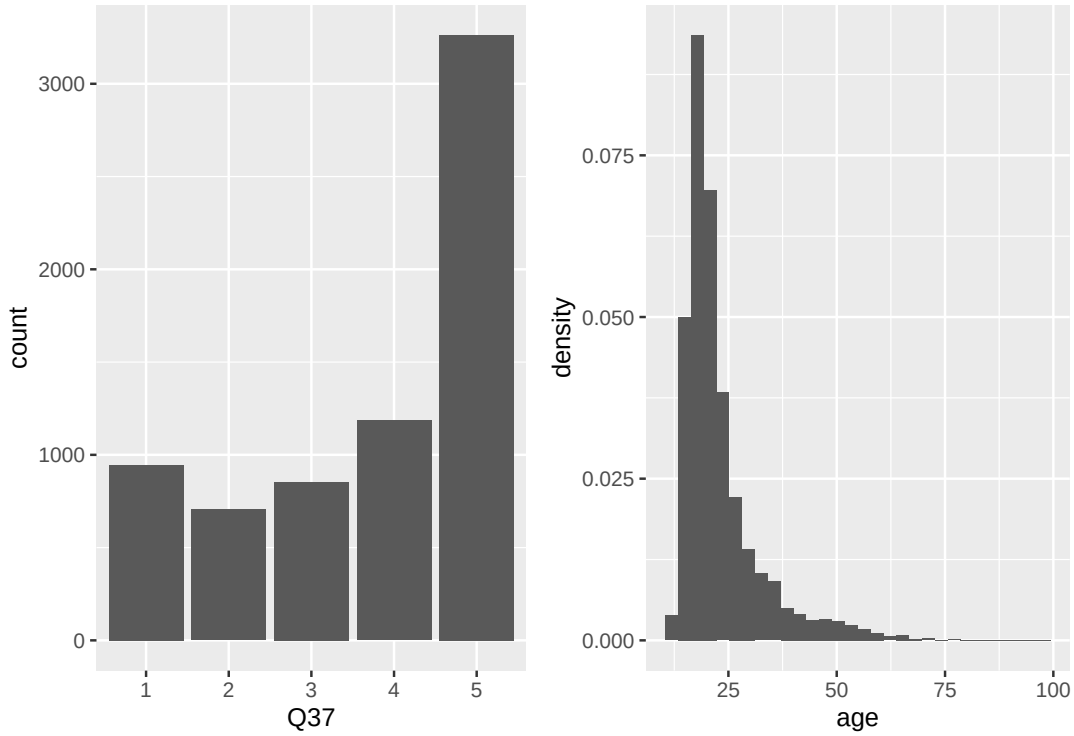

Figura 4: Distribución de las variables Q37 y age

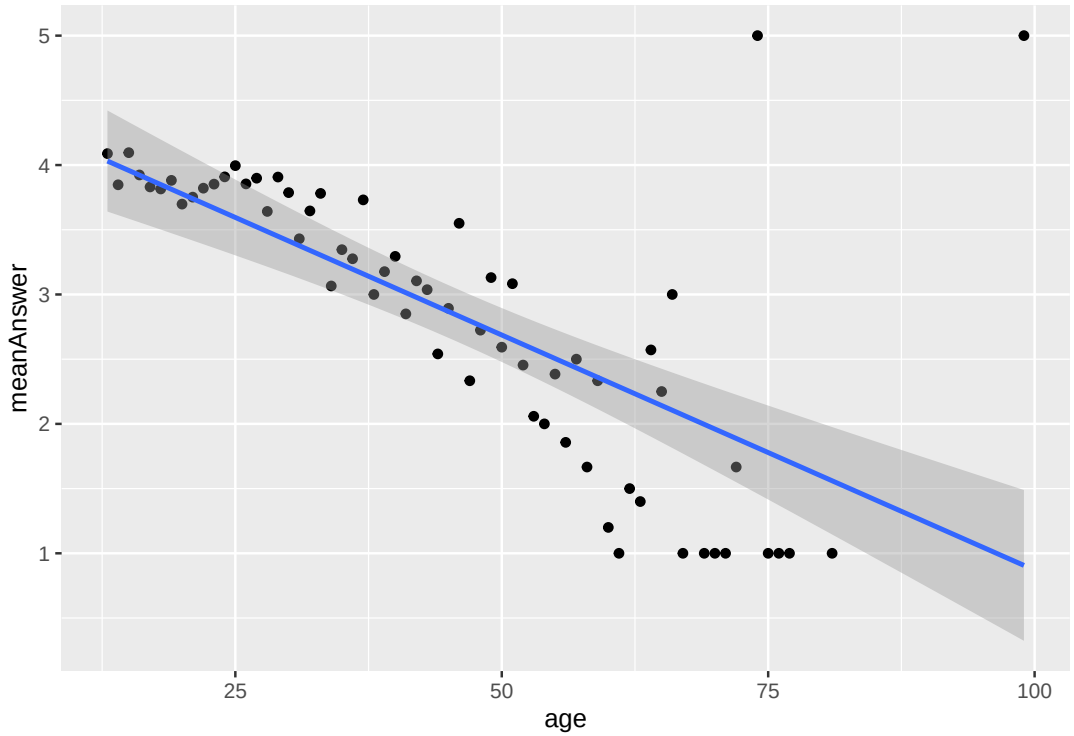

Figura 5: Promedio de la respuesta 37 en función de la edad

```
## `geom_smooth()` using formula 'y ~ x'
```

Del gráfico se desprende que, en efecto, parecen estar correlacionadas: el promedio de la respuesta can con la edad -que es la intuición que uno puede tener de antemano- y esa caída no parece estar muy lejos de ser lineal, al menos en las edades < 60. Hacia el final los datos son más ruidosos (mayor dispersión), que bien puede deberse a la menor cantidad de encuestados en ese rango etario (ver gráfico anterior). En particular, la coeficiente de correlación entre ambas variables es aproximadamente -0.67, lo que indica un cierto grado de asociación lineal negativa entre las mismas.

Aplicación del modelo de regresión ordinal para predecir Q en función de la edad.

A continuación, usamos el comando **polr** del paquete **MASS** para estimar un modelo de regresión ordinal para estas variables. polr usa la interfaz de estándar en R para especificar un modelo de regresión con resultado seguido de predictores. También especificamos Hess=TRUE para que el modelo devuelva la matriz de información observada de la optimización (llamada Hessian) que se usa para obtener errores estándar.

```
## Call:
## MASS::polr(formula = Q37 ~ age, data = pollTrain, Hess = TRUE)
##
## Coefficients:
##        Value Std. Error t value
## age -0.04205   0.002457  -17.11
##
## Intercepts:
##     Value    Std. Error t value
## 1|2  -2.8725   0.0708   -40.5826
## 2|3  -2.1640   0.0658   -32.9111
## 3|4  -1.5514   0.0630   -24.6447
## 4|5  -0.8339   0.0611   -13.6391
##
## Residual Deviance: 19416.40
## AIC: 19426.40
```

A continuación, vemos la tabla resumen de los coeficientes de salida de la regresión que incluye el valor de cada coeficiente, los errores estándar y el valor t, que es simplemente la relación entre el coeficiente y su error estándar. A continuación, vemos las estimaciones de las interseptos, que a veces se denominan puntos de corte. Indican dónde se corta la variable latente para formar los grupos que observamos en nuestros dataset:

Tabla 1. Resumen de la regresión ordinal

	Value	Std. Error	t value
age	-0.04205	0.002457	-17.11

Tabla 2. Estimaciones de las intersepto

	Value	Std. Error	t value
1	-2.8725	0.0708	-40.5826
2	-2.1640	0.0658	-32.9111
3	-1.5514	0.0630	-24.6447
4	-0.8339	0.0611	-13.6391

En el gráfico podemos observar como varía la probabilidad estimada de cada posible valor de respueta, en función de la edad. Hay un fenómeno notable: la relación de la que hablábamos antes es clara para los valores extremos (1

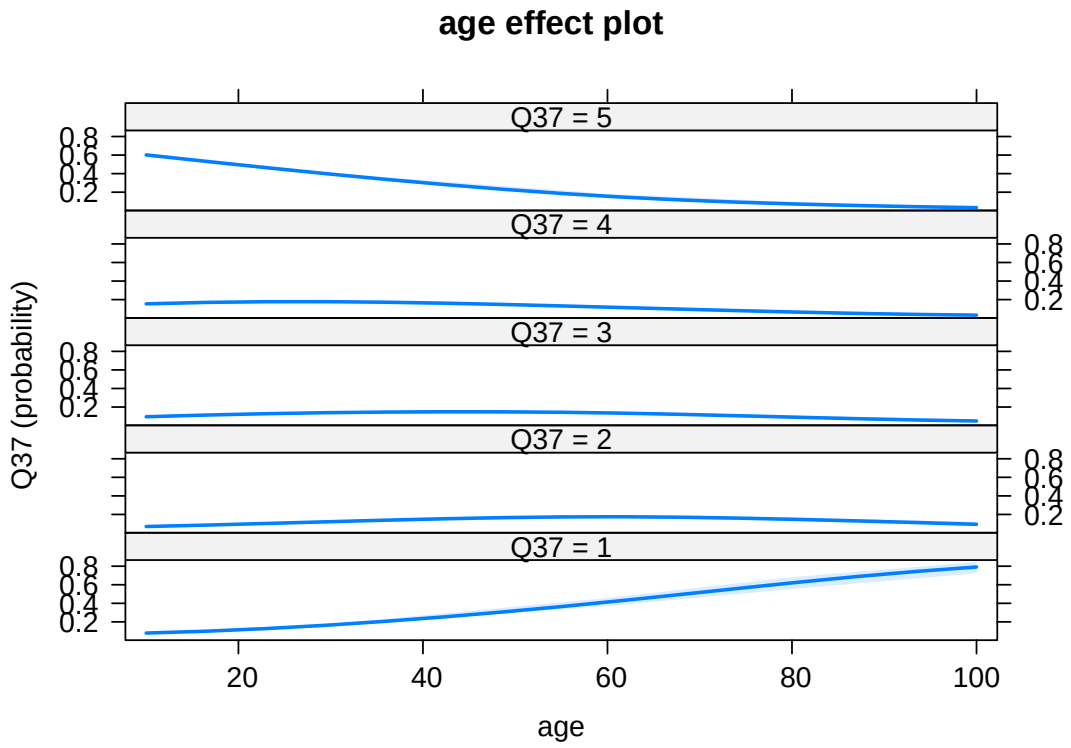

Figura 6: Regresión ordinal

y 5), pero no parece ser tan fuerte para los valores intermedios. De alguna manera, el modelo está reconociendo que estos valores límite se comportan distinto. Hemos mencionado este comportamiento en clase, y se ha mencionado en la literatura también. Tal distribución podría ser problemático para el modelo.

Estimación de la probabilidad de que a una persona de 25 años esté al menos de acuerdo con la frase "me gustan las armas"

A continuación, veamos otro ejemplo con la pregunta Q9. Vamos a estimar la probabilidad de que a una persona de 25 años esté al menos de acuerdo con la frase "me gustan las armas" utilkizando una regresión ordinal de la misma forma que hicimos en los parrafos anteriores. Encontramos que probabilidad de al menos de acuerdo con la frase es 0.33.

```
predictions.q9 = predict(MASS::polr(Q9 ~ age, data = pollTrain),
                         newdata = data.frame(age = 25),
                         type="probs")
prob = predictions.q9["4"] + predictions.q9["5"]
print(paste("Prob de al menos De acuerdo", prob))
```

```
## [1] "Prob de al menos De acuerdo 0.338387160374882"
```

Para la pregunta Q (Q37) definir la siguiente función de pérdida

La función de pérdida presenta la siguiente forma:

$L(y, \hat{y}) = \frac{1}{n} \sum_{x_n=i}^{n} |y_i \hat{y}|$

donde y_i es la respuesta del individuo i a la pregunta Q y $\hat{y}$ es la correspondiente predicción, notar que son números enteros entre 1 y 5.

A continuación, se muestra la implementación de la función de pérdida:

```r
l1_loss = function(y1, y2) {
  return(sum(abs(y1-y2)) / length(y1))
}
```

Implementar un modelo lineal que prediga la respuesta a la pregunta Q en función de la edad.

Este modelo tendrá predicciones $\hat{y}$ que pertenecen a toda la recta real. Para hacerlo comparable con el modelo de regresión ordinal, tomaremos como predicción final al número entero entre 1 y 5 más cercano a y_i

El comando básico es *lm* (linear models). El primer argumento de este comando es una fórmula $y \sim x$ en la que se especifica cuál es la variable respuesta o dependiente (y) y cuál es la variable regresora o independiente (x). El segundo argumento, llamado data especifica cuál es el fichero en el que se encuentran las variables. El resultado lo guardamos en un objeto llamado regresion. Este objeto es una lista que contiene toda la información relevante sobre el análisis. Mediante el comando summary obtenemos un resumen de los principales resultados:

En este ejemplo la ecuación de la recta de mínimos cuadrados es:

$\hat{y} = 4.52 - 0.03 * age$

El coeficiente de determinación (es decir, el coeficiente de correlación al cuadrado) mide la bondad del ajuste de la recta a los datos. A partir de la salida anterior, vemos que su valor en este caso es Multiple R-squared: 0.04.

```
##
## Call:
## lm(formula = Q37 ~ age, data = pollTrain %>% mutate(Q37 = as.numeric(levels(Q37))[Q37]))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0834 -0.9465  0.4299  1.1484  3.8592
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.528181   0.046873   96.61   <2e-16 ***
## age         -0.034216   0.001883  -18.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.433 on 6949 degrees of freedom
## Multiple R-squared:  0.04535,    Adjusted R-squared:  0.04521
## F-statistic: 330.1 on 1 and 6949 DF,  p-value: < 2.2e-16
```

Comparación del modelo de regresión ordinal y lineal Comparar el valor de la pérdida L para el modelo de regresión ordinal y el modelo de regresión lineal (modificado) del item anterior, aplicando ambos. Decidir cuál de los dos es preferible. Para esto entrenamos los modelos con el set de entrenamiento y los evaluamos con el set de testeo.

Como se observa en la siguiente figura, vemos que la función de pérdida es menor para la regresión lineal que para la regresión ordinal.

Comparación entre múltiples modelos Ahora vamos a probar al menos 5 modelos modelos que le parezca que tengan sentido, agregando nuevas variables, interacciones, probando otros algoritmos, etc, intentando minimizar la pérdida L. Para la misma usamos:

1. Regresión ordinal
2. Regresión lineal

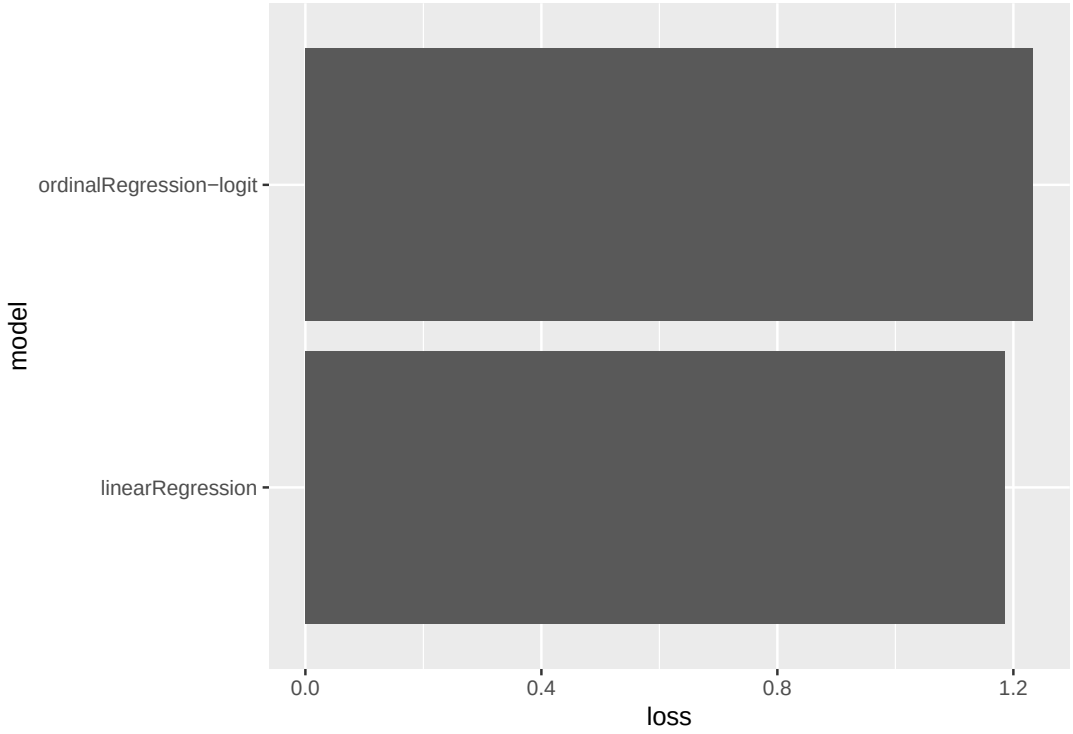

Figura 7: Función de perdida: Regresión ordinal y lineal

3. Regresión ordinal con el método probit.
4. Regresión múltinomial
5. Regresión ordinal con dos variables (edad y educación)

```
## # weights:  15 (8 variable)
## initial  value 11187.202929
## iter  10 value 9719.884585
## final  value 9695.160847
## converged

## Call:
## nnet::multinom(formula = Q37 ~ age, data = pollTrain)
##
## Coefficients:
##   (Intercept)         age
## 2   0.6175133 -0.03543688
## 3   0.7876543 -0.03458729
## 4   1.5012549 -0.05131756
## 5   2.7932453 -0.06397082
##
## Std. Errors:
##   (Intercept)         age
## 2   0.1301079 0.004811365
## 3   0.1225468 0.004482865
## 4   0.1186714 0.004510802
## 5   0.1003502 0.003715125
##
## Residual Deviance: 19390.32
## AIC: 19406.32
```

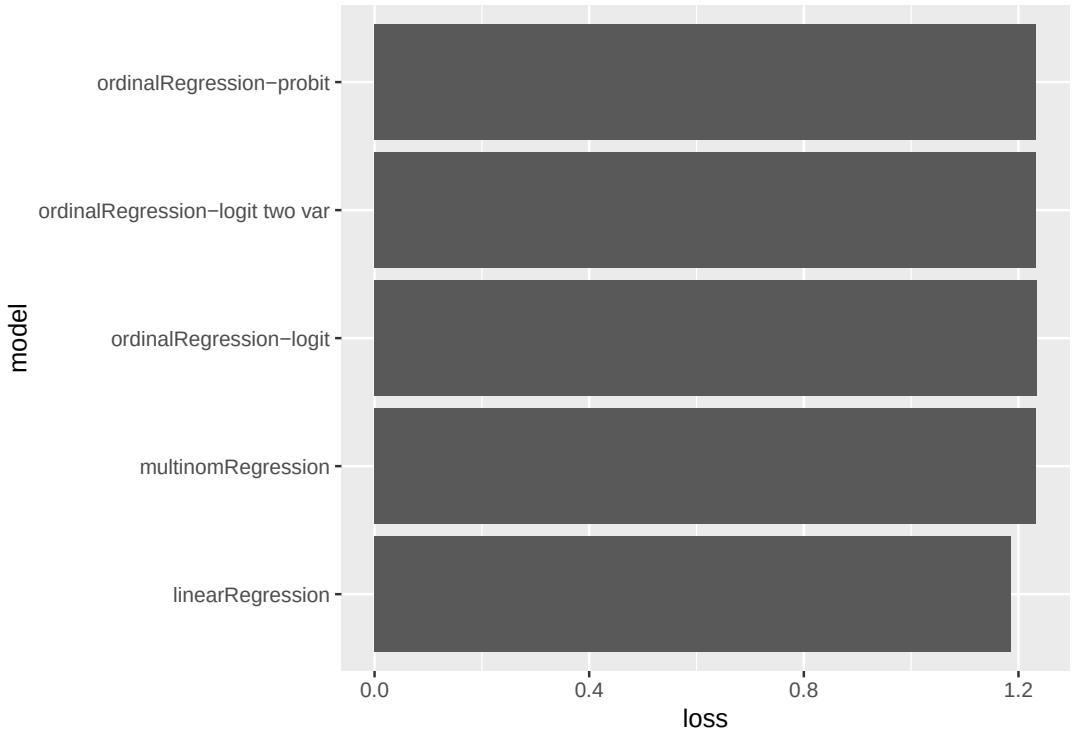

Figura 8: Función de perdida

Regresión lógistica La Regresión Logística Simple, desarrollada por David Cox en 1958, es un método de regresión que permite estimar la probabilidad de una variable cualitativa binaria en función de una variable cuantitativa. Una de las principales aplicaciones de la regresión logística es la de clasificación binaria, en el que las observaciones se clasifican en un grupo u otro dependiendo del valor que tome la variable empleada como predictor.

Consideremos ahora el problema de regresión logística que consiste en predecir si una persona tiene un titulo universitario o no (ver variable education, categorías 3 y 4). Para esto vamos a tomar como covariables engant y age.Tomaremos como pérdida la exactitud, es decir, la proporción de predicciones correctas.

En primer lugar, miremos cuantas instancias tenemos de cada variable con o sin título:

Antes de implementar el modelo, miremos un brevemente los datos, por ejemplo gráfiquemos boxplot para la variable de interes en función de la edad:

Ahora si, apliquemos el modelo de regresión lógistica:

1. El coeficiente estimado para la intersección es el valor esperado del logaritmo de odds: -4.026.

```
##
## Call:
## glm(formula = hasTitle ~ engnat + age, family = "binomial", data = education.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.0726  -0.7359  -0.5979   0.9918   2.1027
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.026976   0.132735 -30.339  < 2e-16 ***
## engnat       0.426871   0.065800   6.487 8.73e-11 ***
## age          0.115819   0.003941  29.389  < 2e-16 ***
## ---
```

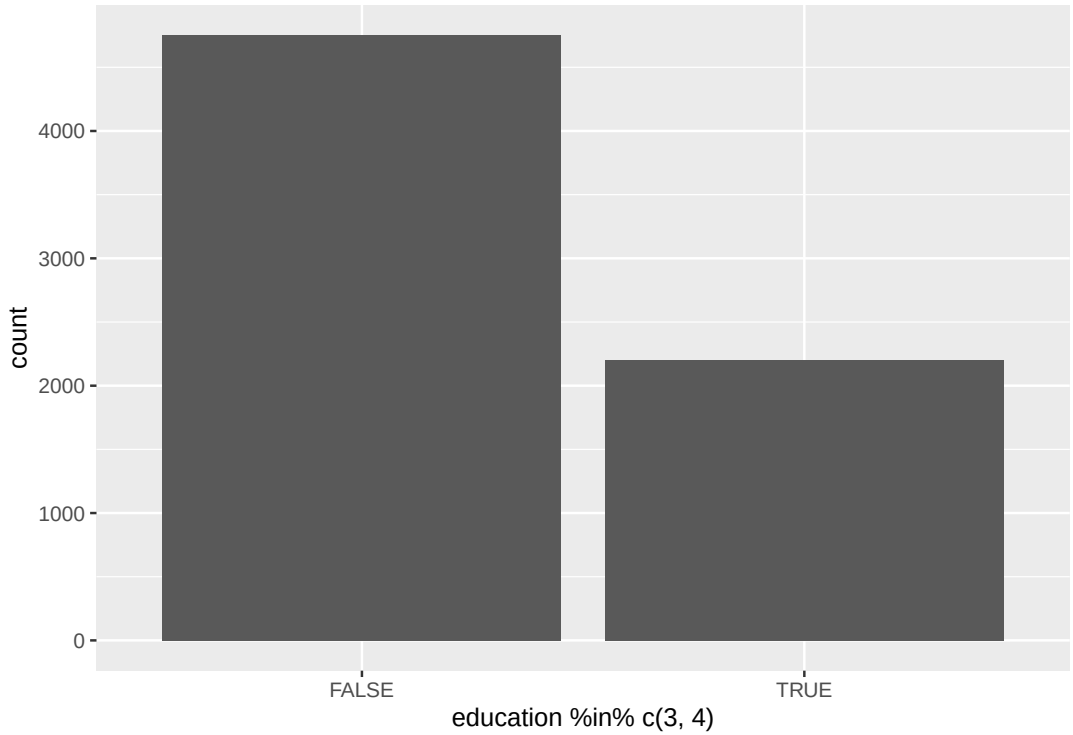

Figura 9: Frecuencia de casos con o sin título universitario

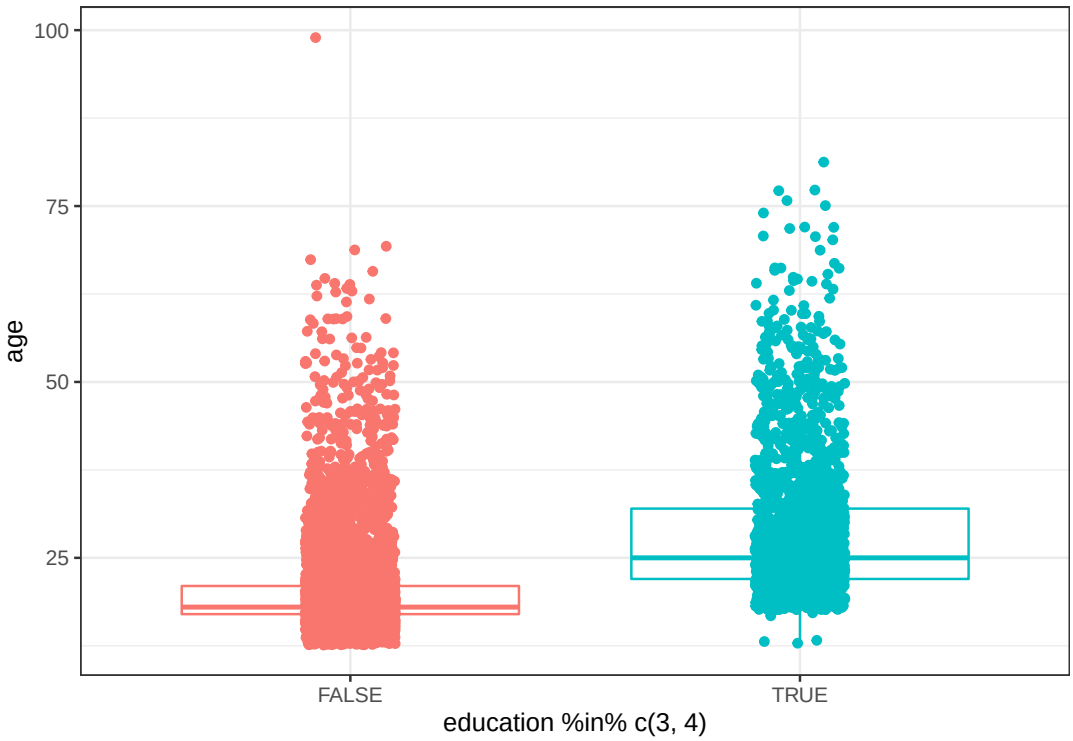

Figura 10: Box plot para la variable de interes en función de la edad

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8676.1  on 6950  degrees of freedom
## Residual deviance: 7399.6  on 6948  degrees of freedom
## AIC: 7405.6
##
## Number of Fisher Scoring iterations: 4
```

Para poder visualizar, vamos a recalcular el modelo pero solamente utilizando como variable independiente la edad:

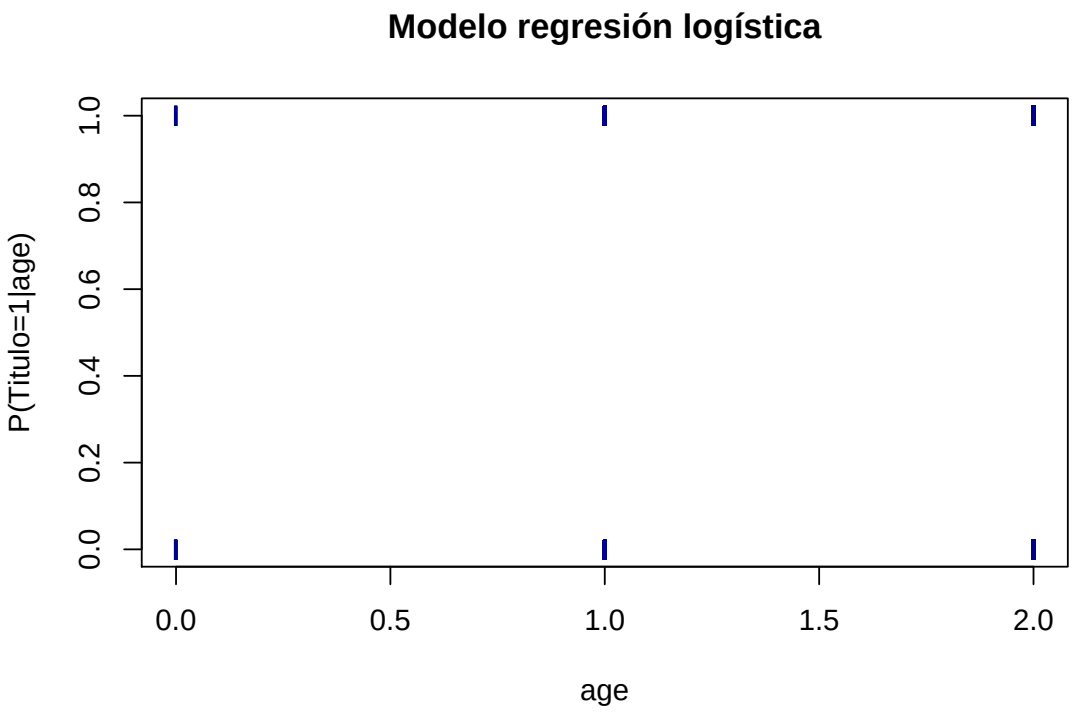

Figura 11: Regresión lógicstica en función de la edad

Posterior a ello, intentamos utilizar un tipo de Modelo de Selección de Variables: el de STEPWISE, el cual tiene como fin ayudar a seleccionar la mejor combinación de variables para así tener el menor AIC (se descarto esa idea porque tardaba mucho la corrida). Por lo cúal, probamos directamente incorporar todas las variables restantes observando que al menos en el dataset de training mejora la performance del modelo.

Analizar cómo varía la exactitud en función del tamaño de muestra.

En esta sección vamos a estimar la exactitud utilizando del modelo de regresión lógistica presentado anteriormente el cúal utiliza como variables independientes age y engnat para distintos tamaños de dataset de entrenamiento. Se observa que a medida que aumenta el tamaño muestran incrementa la exactitud, sin embargo se llega a un limite de 0.72. Esto nos estaria indicando que para mejorar la exactitud ya no alcanza con aumentar el tamaño mostral, pero se podrían incorporar más features o variables independientes.

```
## ============================================================================
```

Finalmente utilizamos una regresión beta para predecir cúal es el mínimo tamaño de muestra necesario para obtener una exactitud de 0.7. Se observa que no parece ser una buena aproximación (R-squared: 0.01121) para responder nuestra pregunta.

Figura 12: Regresión lógicstica en función de la edad

```
##
## Call:
## betareg(formula = accuracy ~ trainingSize, data = education.accuracies,
##     link = "loglog")
##
## Standardized weighted residuals 2:
##      Min       1Q   Median       3Q      Max
## -11.9389  -0.1580   0.0600   0.2622   1.5843
##
## Coefficients (mean model with loglog link):
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.0358212  0.0059798   173.2   <2e-16 ***
## trainingSize 0.0000156  0.0000104     1.5    0.134
##
## Phi coefficients (precision model with identity link):
##       Estimate Std. Error z value Pr(>|z|)
## (phi)   1887.9      188.7      10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Type of estimator: ML (maximum likelihood)
## Log-likelihood: 627.3 on 3 Df
## Pseudo R-squared: 0.01121
## Number of iterations: 102 (BFGS) + 3 (Fisher scoring)
```

Realizamos pruebas incorporando las demás variables demográficas, pero la exactitud no aumento por encima de 0.72. Dado que al rehalizar el split en el dataset de train y test no estratificamos por las variables de factor, no se encuentran todas las clases en ambos grupos. Se intento realizar una separación teniendo en cuenta esta caracteristica, sin emabrgo encontramos númersos grupos con solamente un elemento. Esto implica que hay instancias (filas) que cuentan con una combinación de las respuestas unica.

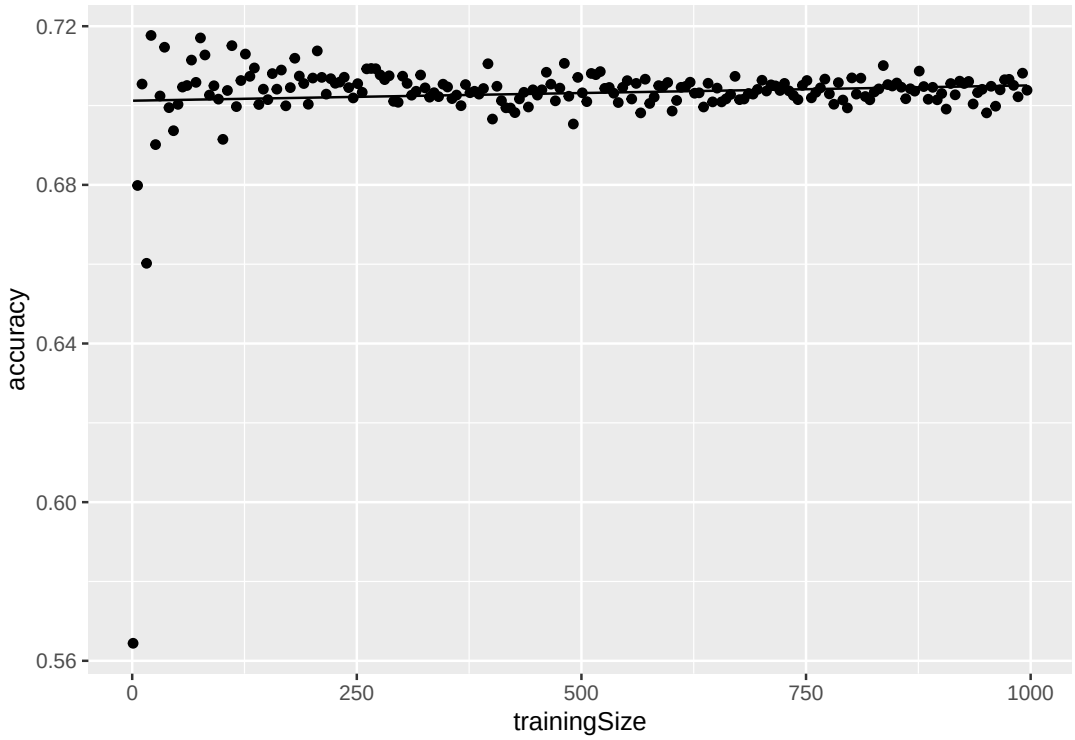

Figura 13: Exactitud en función de n