

BANK MARKETING CLASSIFICATION TASK

Dataset relacionado con campañas de marketing realizadas por una institución bancaria portuguesa mediante llamadas telefónicas para ofrecer depósitos a plazo.

Yassin Ettijani
Ana Gil
Wenya Zhong



VARIABLES

age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
30	unemployed	married	primary	no	1787	no	no	cellular	19	oct	79	1	-1	0	unknown	no
33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	339	4	failure	no
35	management	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	330	1	failure	no
30	management	married	tertiary	no	1476	yes	yes	unknown	3	jun	199	4	-1	0	unknown	no
59	blue-collar	married	secondary	no	0	yes	no	unknown	5	may	226	1	-1	0	unknown	no
35	management	single	tertiary	no	747	no	no	cellular	23	feb	141	2	176	3	failure	no

VARIABLE A ESTUDIAR: y

¿Se ha suscrito el cliente a un depósito a plazo?

Toma los valores "yes" o "no", variable binaria.

Problema (1/2)



Este es João. João es uno de los clientes de nuestro banco . ¿Querrá João suscribirse a un depósito a plazo?

Problema

¿Qué modelo de regresión podemos utilizar?

Regresión Lineal

- La variable Target Y_i sigue una distribución Normal.
- Las Y_i deben ser independientes

Regresión Logística

- La variable Target tiene naturaleza dicotómica.
- Muestra de gran tamaño

Regresión Poisson

- La variable Target se trata de un conteo de eventos en un determinado tiempo.
- Independencia de eventos

SELECCIÓN DEL MODELO



MODELO 1: CON TODAS LAS VARIABLES Y CON INTERACCIONES

**Salen demasiadas
interacciones.**



**Demasiado coste
computacional.**

NO NOS SIRVE

MODELO 2: CON TODAS LAS VARIABLES

Deviance: D = 2173.7

Bondad del modelo: $R^2 = 0.3272513$

AIC = 2259.7

Mejora un
32.73% el
modelo nulo

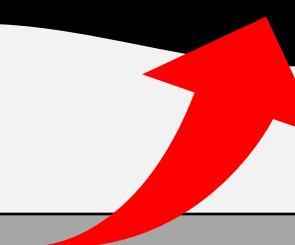
MODELO 3: USANDO STEPWISE

Deviance: D = 2195.5

Bondad del modelo: $R^2 = 0.320489$

AIC = 2249.5

Mejora un
32.04% el
modelo nulo



Entre dos modelos con similar Deviance elegimos el que tiene menor AIC (métrica que tiene en cuenta el Deviance, pero también el número de variables). La métrica AIC indica que merece la pena excluir algunas variables.

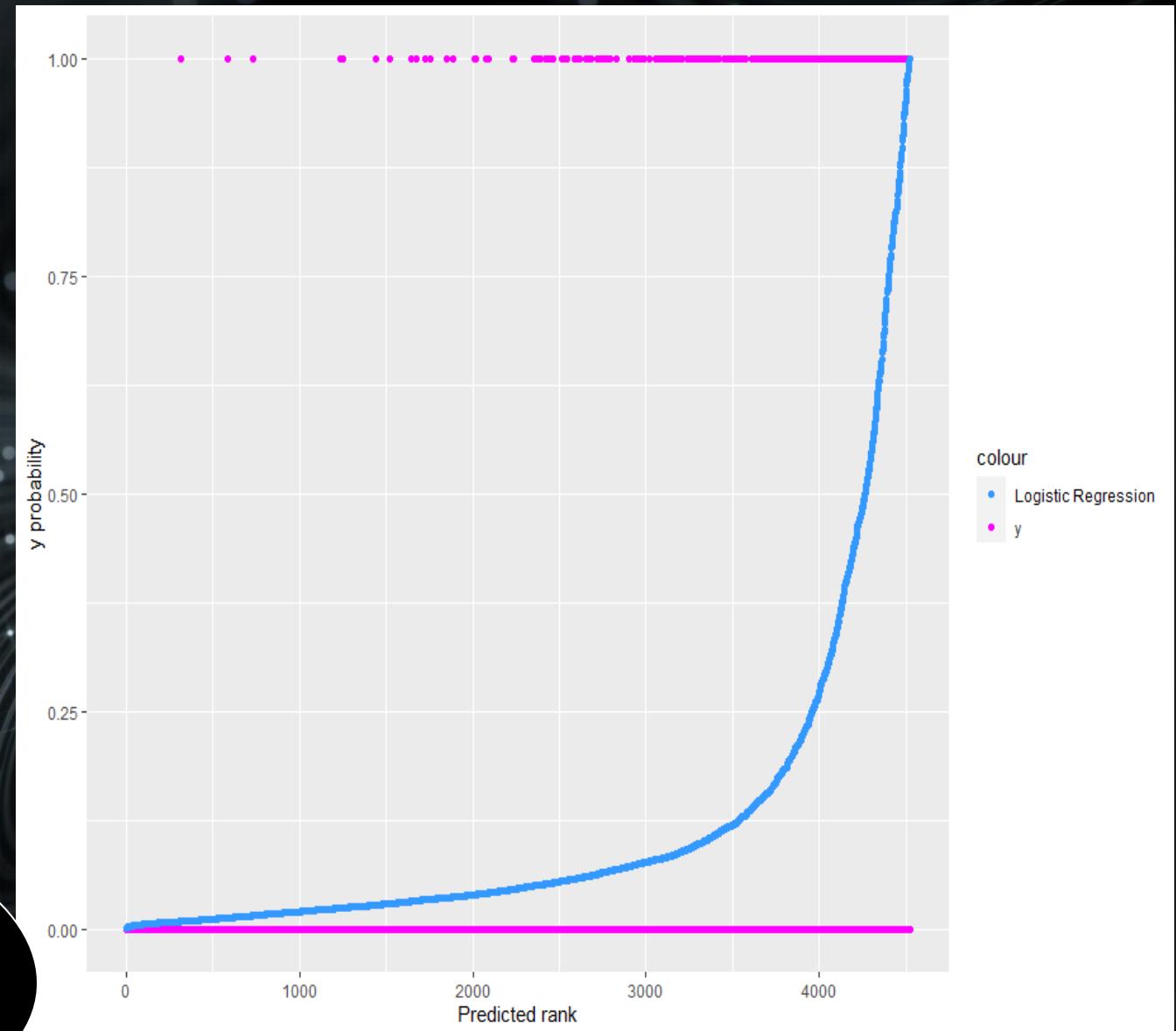
AJUSTE GRÁFICO DEL MODELO

Variable dependiente: y

Variables independientes:

- marital
- education
- housing
- loan
- contact
- day
- month
- duration
- campaign
- potcome

¡Solo el 25.41% de las predicciones están lejos del 0 y del 1!



Recordatorio

Un GLM pretende estudiar la relación entre una variable dependiente Y y p variables explicativas a partir de una *m.a.* con las siguientes características:

1. Las variables Y_1, \dots, Y_n siguen la misma distribución de la familia exponencial con la forma canónica y un sólo parámetro de interés: θ_i para cada i .
2. La relación entre $\eta_i = x_i^t \beta$ y $\mu_i = E[Y_i]$ está definida a partir de una función monótona y diferenciable, g , que denominamos "función link":

$$g(\mu_i) = x_i^t \beta$$

En los modelos con respuesta binaria, y_i puede ser 0 ó 1. Por lo tanto:

$$E[Y_i|x_i] = 1P(Y_i = 1|x_i) + 0P(Y_i = 0|x_i) = p_i$$

Considerando el modelo de regresión logística obtenido mediante Stepwise, clasificamos a un individuo de la siguiente manera:

- $y = \text{Yes}$ si $p_i > 0.5$
- $y = \text{No}$ en otro caso

		Clase real		Total
		No	Yes	
Clase predicha	No	3915	346	4261
	Yes	85	175	260
Total		4000	521	4521

- Tasa de error = $(85+346)/4521 = 0.09533289$ (baja)
- Precisión = $175/260 = 0.6730769$
 Exhaustividad = Sensibilidad = $175/521 = 0.3358925$
 Se pierden bastantes positivos.
- Especificidad = $3915/4000 = 0.97875$ (alta)
- Sin embargo, de los 521 que se suscriben, el 66.41% no son detectados.

Al cambiar el umbral a partir del cual se decide cómo clasificar a cada individuo, clasificamos de la siguiente manera:

- $y = \text{Yes}$ si $p_i > 0.2$
- $y = \text{No}$ en otro caso

		Clase real		Total
		No	Yes	
Clase predicha	No	3661	186	3847
	Yes	339	335	674
Total		4000	521	4521

- Tasa de error = $(339+186)/4521 = 0.1161248$ (aumenta, pero sigue siendo baja)
- Precisión = $335/674 = 0.4970326$ (disminuye)
 Exhaustividad = Sensibilidad = $335/521 = 0.6429942$ (aumenta)
 Se pierde precisión, pero se gana exhaustividad.
- Especificidad = $3661/4000 = 0.91525$ (disminuye, pero sigue siendo alta)
- De los 521 que se suscriben, solo el 35.7% no son detectados.

Probamos a cambiar una vez más el umbral, y clasificamos de la siguiente manera:

- $y = \text{Yes}$ si $p_i > 0.1$
- $y = \text{No}$ en otro caso

		Clase real		Total
		No	Yes	
Clase predicha	No	3238	90	3328
	Yes	762	431	1193
Total		4000	521	4521

- Tasa de error = $(762+90)/4521 = 0.1884539$ (aumenta considerablemente)

Para $p=0.1$ se produce un aumento considerable de errores.

Finalmente, nos quedamos con $p=0.2$.

EVALUACIÓN DEL ERROR

Considerando el umbral de clasificación de 0.2, podemos calcular distintas tasas de error:

MSE: Error Cuadrático Medio

$$\hat{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2$$

MAE: Error Absoluto Medio

$$MAE = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j|$$

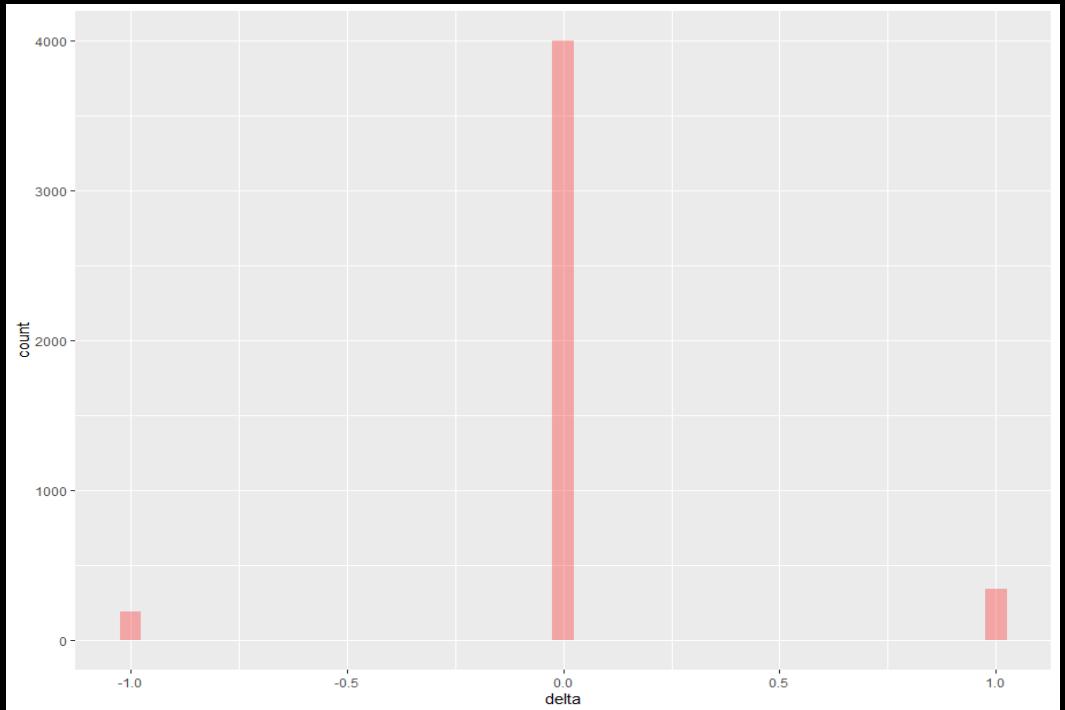
En nuestro caso, son iguales dado que la variable es binaria:

$$MSE = 0.1161248$$

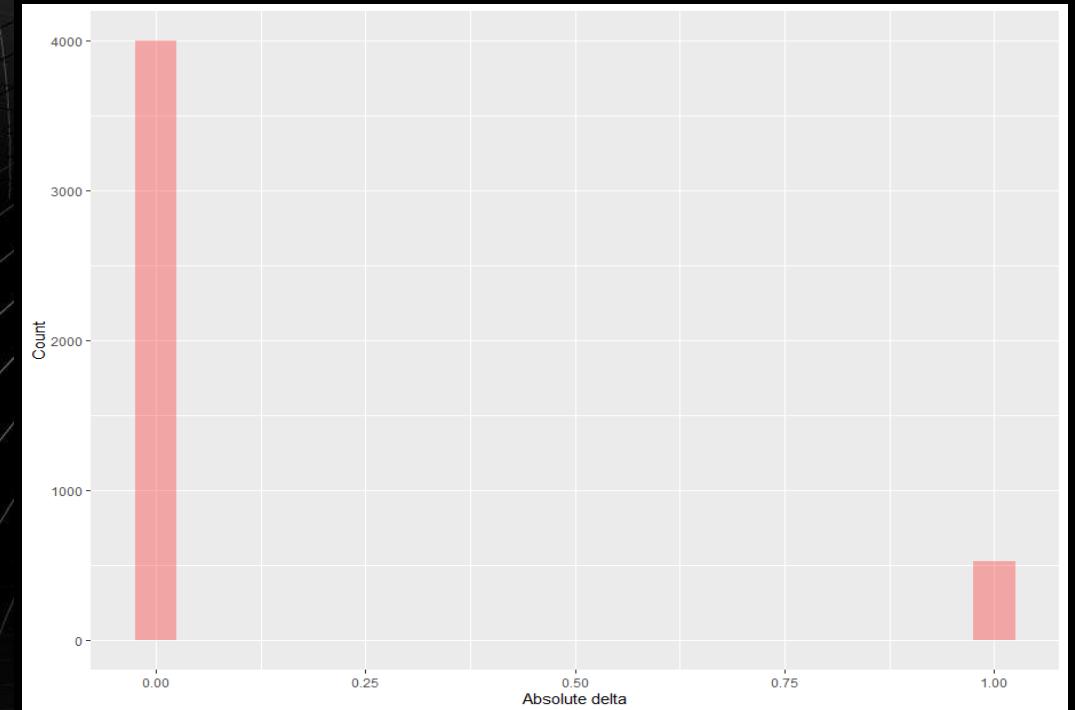
$$MAE = 0.1161248$$

ANÁLISIS GRÁFICO DEL ERROR

ERROR

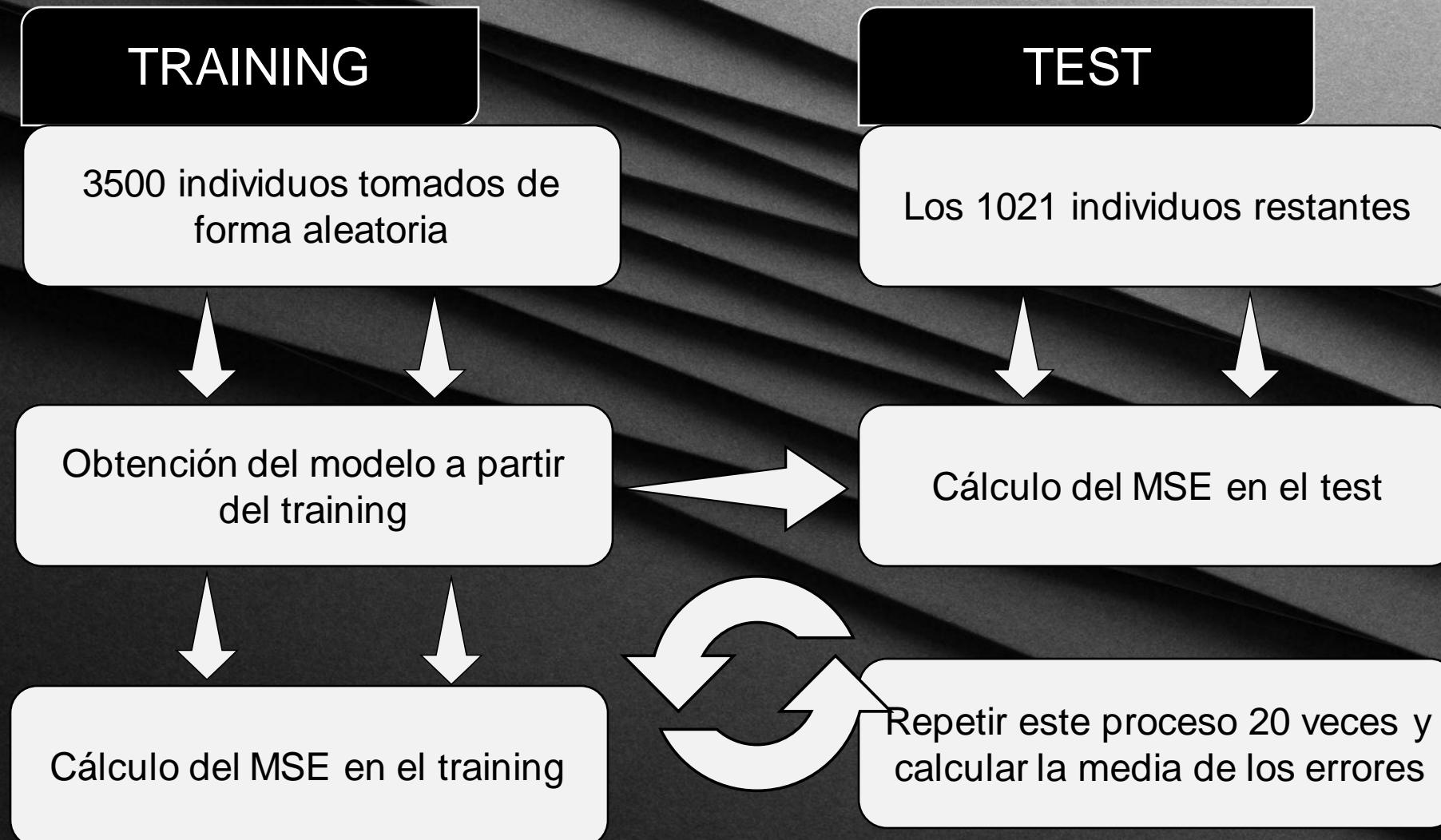


ERROR ABSOLUTO



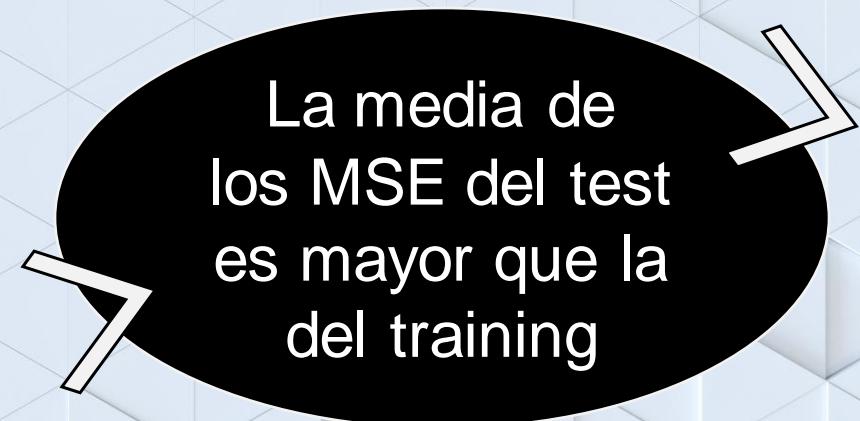
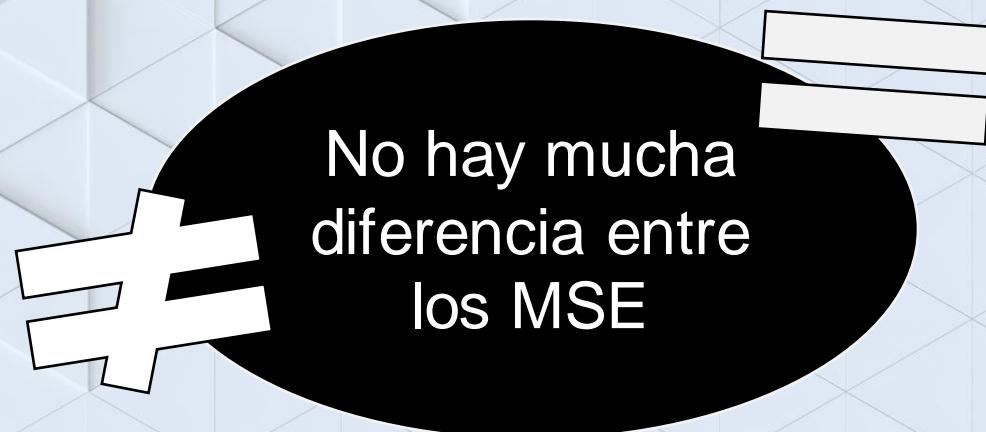
El error está centrado en 0 pero hay algunos outliers.

ESTIMACIÓN DEL ERROR MEDIANTE VALIDACIÓN CRUZADA



Tras llevar a cabo este proceso, obtenemos que la media de los errores cuadráticos medios obtenidos son:

- MSE del training = 0.1149286
- MSE del test = 0.1172870



MODELO ANOVA

```
> modANOVA1 <- aov(y.bin~marital + education + housing + loan +
+                     contact + month +
+                     data$duration_category + data$day_category + data$campaign_category + poutcome,data = data)
> summary(modANOVA1)

Df Sum Sq Mean Sq F value    Pr(>F)
marital          2   1.9   0.97  12.795 2.88e-06 ***
education        3   1.2   0.40   5.248  0.0013 **
housing          1   4.5   4.50  59.334 1.63e-14 ***
loan             1   2.1   2.06  27.212 1.90e-07 ***
contact          2   6.1   3.03  40.021 < 2e-16 ***
month            11  19.2   1.74  22.972 < 2e-16 ***
data$duration_category  2   63.4   31.69 417.994 < 2e-16 ***
data$day_category     2   0.5   0.26   3.435  0.0323 *
data$campaign_category 1   0.2   0.19   2.490  0.1147
poutcome          3   21.4   7.12  93.884 < 2e-16 ***
Residuals       4492 340.6   0.08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Al ser duration, day y campaign variables cuantativas, hemos transformado esas variables en variables categóricas para hacer el ANOVA.

Variable más significativa !

Duration_category	Rango=[4-3025]s
Corta	0s-180s
Media	181s-480s
Larga	480s-Infs

MODELO ANOVA

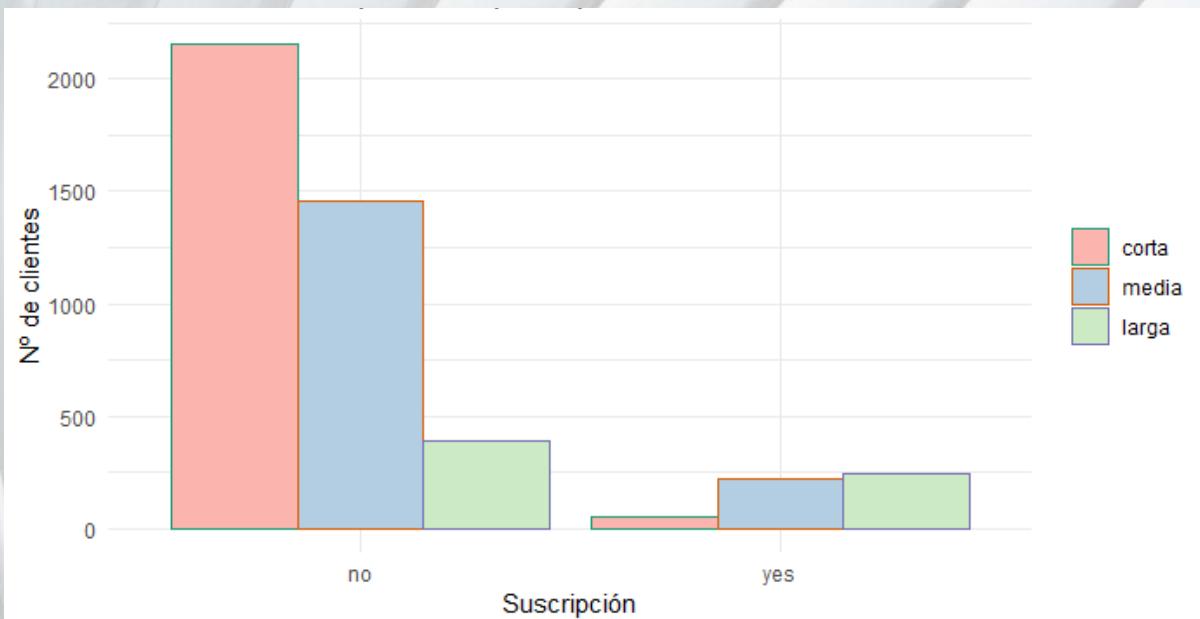
MODELO DEFINITIVO

```
> modelo_def1 <- aov(y.bin~duration_category,data = data)
> summary(modelo_def1)

   Df Sum Sq Mean Sq F value Pr(>F)
duration_category     2    63.3    31.66   359.8 <2e-16 ***
Residuals            4518   397.6    0.09
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

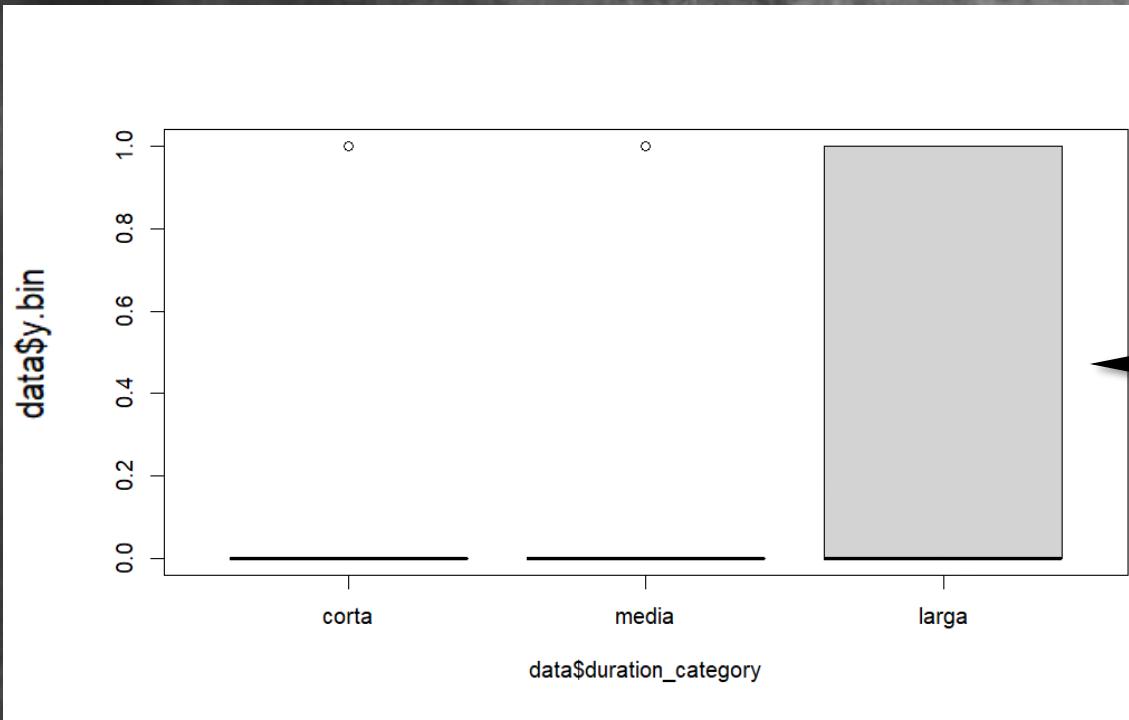
Podemos ver que cuando la duración es larga, parece que el número de clientes suscritos es mayor.

Duración del último contacto /suscripción



Hipótesis de ANOVA

BOXPLOT



- Parece que la duración corta y media tienen proporciones alineadas y diferentes a la duración larga
- Valores atípicos en la duración corta y media
- Homocedasticidad: la duración larga rompe con la posible dinámica de igualdad de varianzas

NORMALIDAD

SHAPIRO TEST

H_0 : La distribución es normal

H_1 : La distribución no es normal

Todos los p-value < 0.05

```
> tapply(data$y.bin, data$duration_category, shapiro.test)  
$corta
```

Shapiro-Wilk normality test

```
data: X[[i]]  
W = 0.14133, p-value < 2.2e-16
```

```
$media
```

Shapiro-Wilk normality test

```
data: X[[i]]  
W = 0.40122, p-value < 2.2e-16
```

```
$larga
```

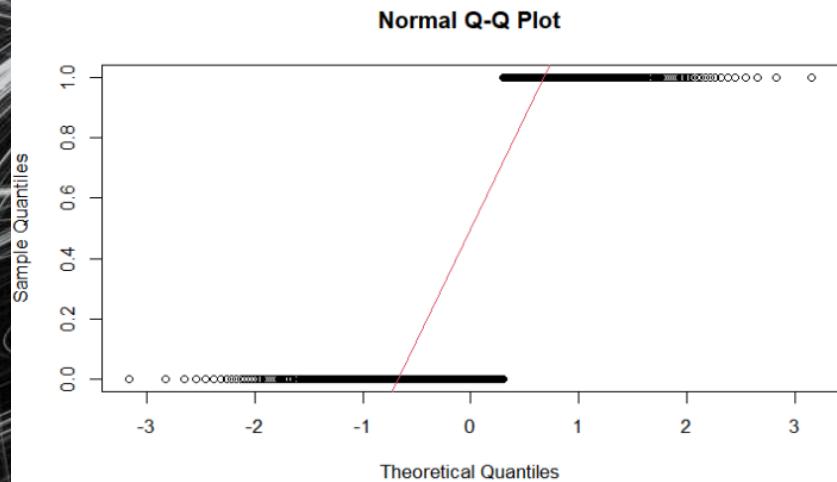
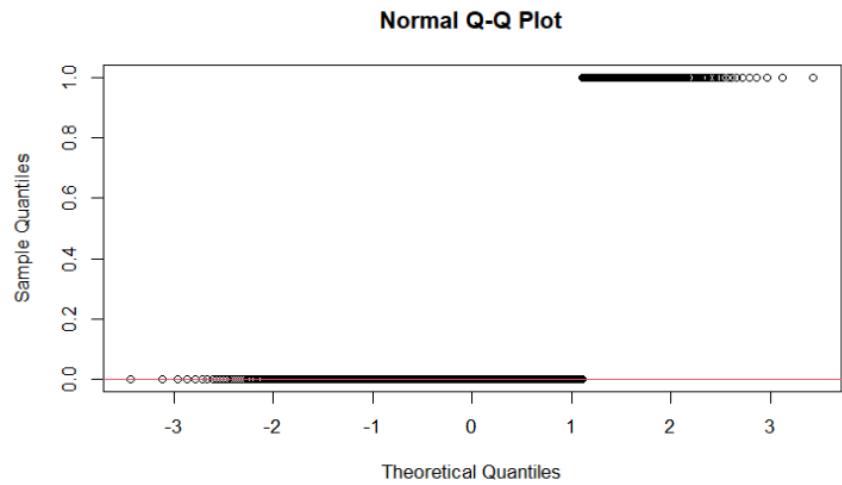
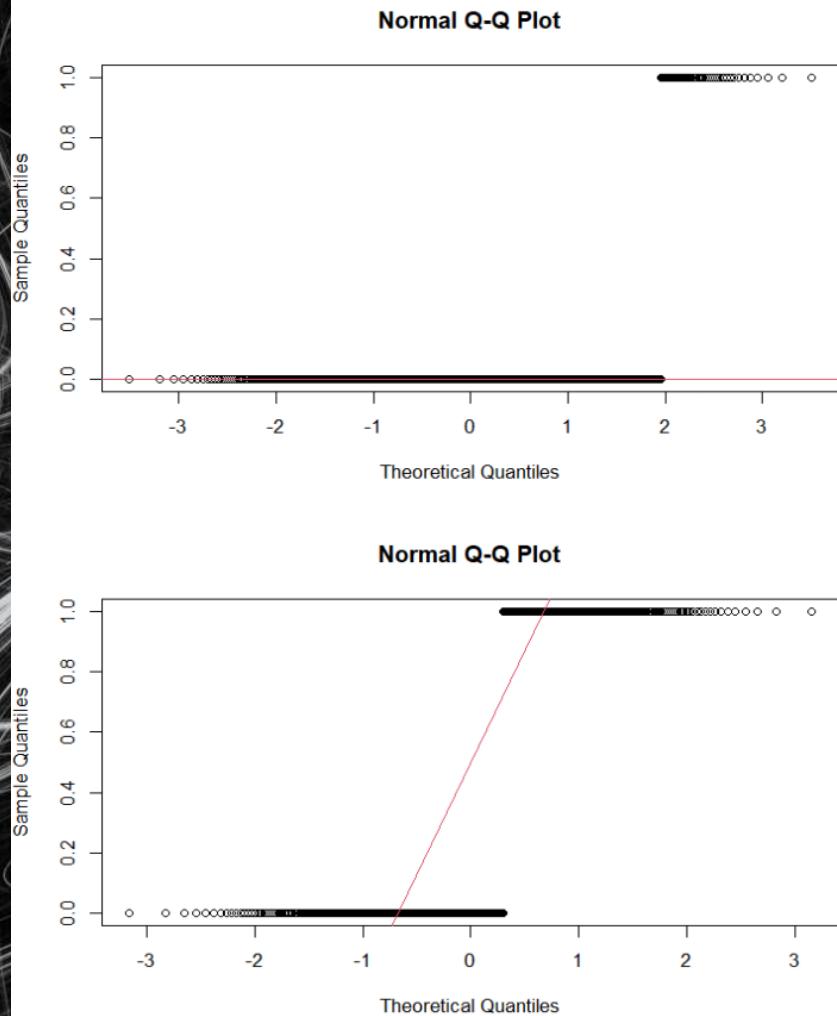
Shapiro-Wilk normality test

```
data: X[[i]]  
W = 0.61596, p-value < 2.2e-16
```

Rechazamos normalidad para todos los grupos.

NORMALIDAD

Q-Q PLOT



NORMALIDAD DE LOS RESIDUOS

SHAPIRO TEST

```
> shapiro.test(res1)

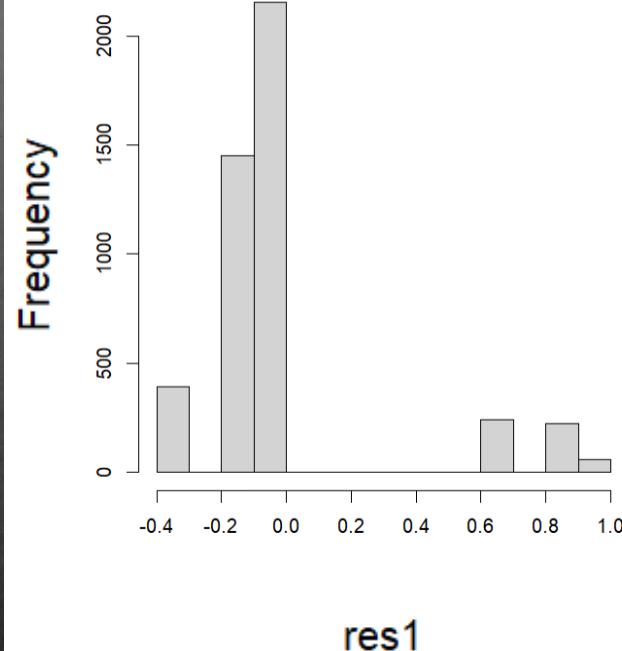
Shapiro-Wilk normality test

data: res1
W = 0.65466, p-value < 2.2e-16
```

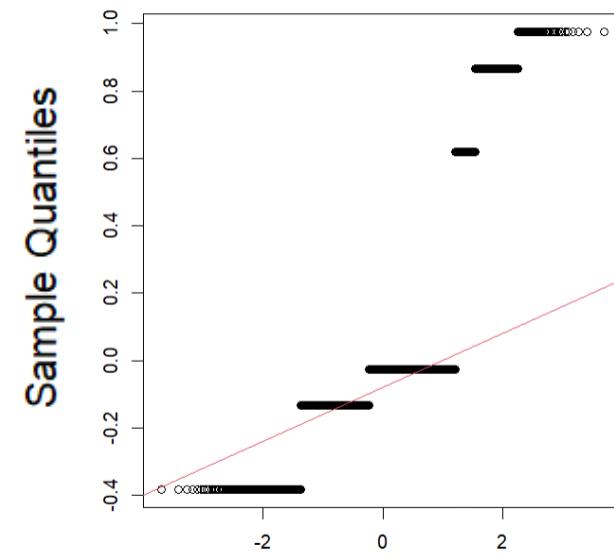


Como $p\text{-valor} < 0.05 \Rightarrow$
Rechazamos normalidad de
los residuos.

Histogram of res1



Normal Q-Q Plot



Theoretical Quantiles

HOMOCEDASTICIDAD

Test de Barlett

H_0 : Igualdad de varianzas
 H_1 : Al menos una es diferente

```
> bartlett.test(data$y.bin~data$duration_category)

Bartlett test of homogeneity of variances

data: data$y.bin by data$duration_category
Bartlett's K-squared = 1728, df = 2, p-value < 2.2e-16
```

Tenemos que p-value < 0.05

Rechazamos igualdad de varianzas.

HOMOCEDASTICIDAD

Patrones en los residuos

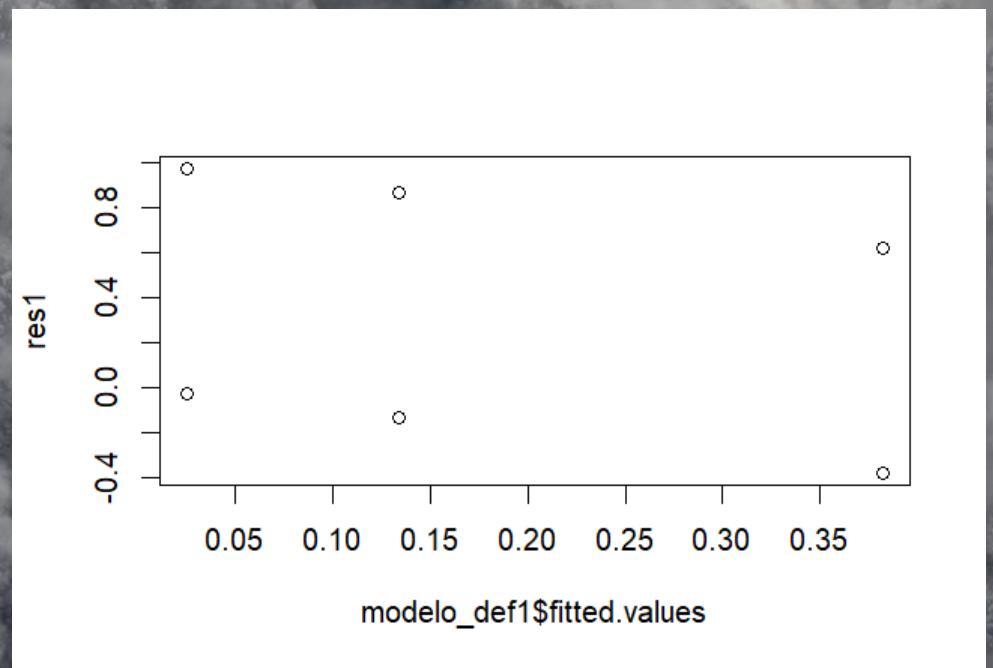
```
> summary(lm(res1~data$duration_category))

Call:
lm(formula = res1 ~ data$duration_category)

Residuals:
    Min      1Q   Median      3Q     Max 
-0.09785 -0.09785 -0.02403 -0.02403  0.92533 

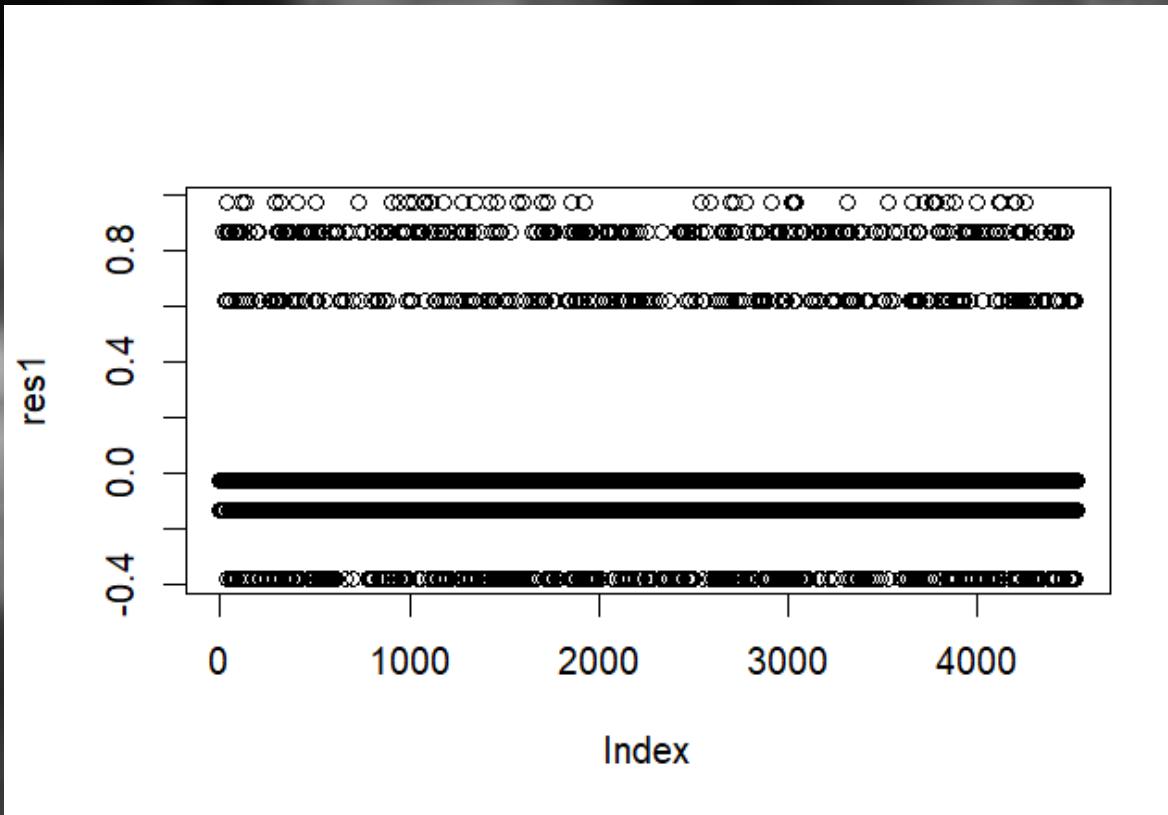
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.024676  0.004024  6.133 9.38e-10 ***
data$duration_categorymedia 0.090997  0.006126 14.854 < 2e-16 ***
data$duration_categorylarga 0.211385  0.008541 24.751 < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1892 on 4518 degrees of freedom
Multiple R-squared:  0.1291,    Adjusted R-squared:  0.1287 
F-statistic: 334.9 on 2 and 4518 DF,  p-value: < 2.2e-16
```



No hay homocedasticidad.

INDEPENDENCIA DE LOS RESIDUOS



Parece que se observan patrones en los residuos.



¿No hay independencia de residuos?

INDEPENDENCIA DE LOS RESIDUOS

Durbin-Watson test

H_0 : No existe autocorrelación de primer orden en los residuos
 H_1 : Existe autocorrelación de primer orden en los residuos

```
> dwtest(modelo_def1)

Durbin-Watson test

data: modelo_def1
DW = 1.9836, p-value = 0.29
alternative hypothesis: true autocorrelation is greater than 0
```



Con un valor DW cercano a 2 y un p-value de 0.29, no hay suficiente evidencia para rechazar la hipótesis nula.

OUTIERS EN LOS RESIDUOS



Hay outliers en los residuos.

CONCLUSIONES



ES PREFERIBLE HACER LAS CAMPAÑAS DE MÁRKETING EN MARZO Y OCTUBRE, Y CONTACTAR A CLIENTES QUE TENGAN EDUCACIÓN TERCIARIA CUYA ANTERIOR CAMPAÑA HAYA TENIDO ÉXITO

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-2.7318547	0.3608341	-7.571	3.71e-14	***
maritalmarried	-0.5108288	0.1709685	-2.988	0.00281	**
maritalsingle	-0.3181354	0.1864058	-1.707	0.08788	.
educationsecondary	0.1268075	0.1827658	0.694	0.48779	
educationtertiary	0.3810998	0.1913291	1.992	0.04639	*
educationunknown	-0.2831229	0.3395018	-0.834	0.40432	
housingyes	-0.3434167	0.1329758	-2.583	0.00981	**
loanyes	-0.6430899	0.1981746	-3.245	0.00117	**
contacttelephone	-0.0077099	0.2217915	-0.035	0.97227	
contactunknown	-1.4577768	0.2253170	-6.470	9.81e-11	***
day	0.0162954	0.0081149	2.008	0.04463	*
monthaug	-0.3180448	0.2457796	-1.294	0.19566	
monthdec	0.2027593	0.6661682	0.304	0.76085	
monthfeb	0.1398131	0.2917392	0.479	0.63177	
monthjan	-1.0851644	0.3768278	-2.880	0.00398	**
monthjul	-0.7312482	0.2473493	-2.956	0.00311	**
monthjun	0.5600454	0.2975296	1.882	0.05979	.
monthmar	1.6262687	0.3813054	4.265	2.00e-05	***
monthmay	-0.4734989	0.2311010	-2.049	0.04047	*
monthnov	-0.8618761	0.2700920	-3.191	0.00142	**
monthoct	1.4102741	0.3255826	4.332	1.48e-05	***
monthsep	0.7558952	0.4089156	1.849	0.06452	.
duration	0.0041838	0.0001992	21.006	< 2e-16	***
campaign	-0.0722161	0.0281375	-2.567	0.01027	*
poutcomeother	0.4813847	0.2659176	1.810	0.07025	.
poutcomesuccess	2.4145453	0.2681035	9.006	< 2e-16	***
poutcomeunknown	-0.0931105	0.1847998	-0.504	0.61437	

> exp(modelo\$coefficients)

	(Intercept)	maritalmarried	maritalsingle	educationsecondary
	0.06509844	0.59999812	0.72750428	1.13519844
educationtertiary	1.46389370	educationunknown	0.75342716	housingyes
	contacttelephone	0.99231979	0.23275316	loanyes
	monthdec	1.22477760	monthfeb	0.70934257
	monthjan	1.15005882	monthjul	0.52566563
	monthjul	0.33784623	day	0.72757020
	monthaug	1.01642892	monthaug	0.72757020
	monthdec	0.23275316	monthjan	0.48130782
	monthfeb	1.15005882	monthjul	0.48130782
	monthjan	0.33784623	day	0.42236892
	monthjul	0.62281928	monthaug	0.42236892
	monthaug	1.00419258	duration	campaign
	duration	1.00419258	0.93032985	campaign
	campaign	0.93032985	0.91109284	poutcomeunknown
poutcomeother	1.61831378	11.18468362	poutcomesuccess	poutcomeunknown
poutcomesuccess				0.91109284
poutcomeunknown				

Next Steps

Planteamiento de nuevos
modelos predictivos teniendo
en cuenta nuevas variables

MODELOS

Utilizar técnicas de reducción
dimensional para tener un
conjunto de datos manejable

ACP

Explorar algoritmos de
Machine Learning para
mejorar el modelo predictivo

MACHINE LEARNING

Estudiar diferentes métricas
para evaluar los modelos
utilizados

MÉTRICAS

Explorar la posibilidad de
automatizar el entrenamiento
y evaluación del modelo

AUTOMATIZACIÓN

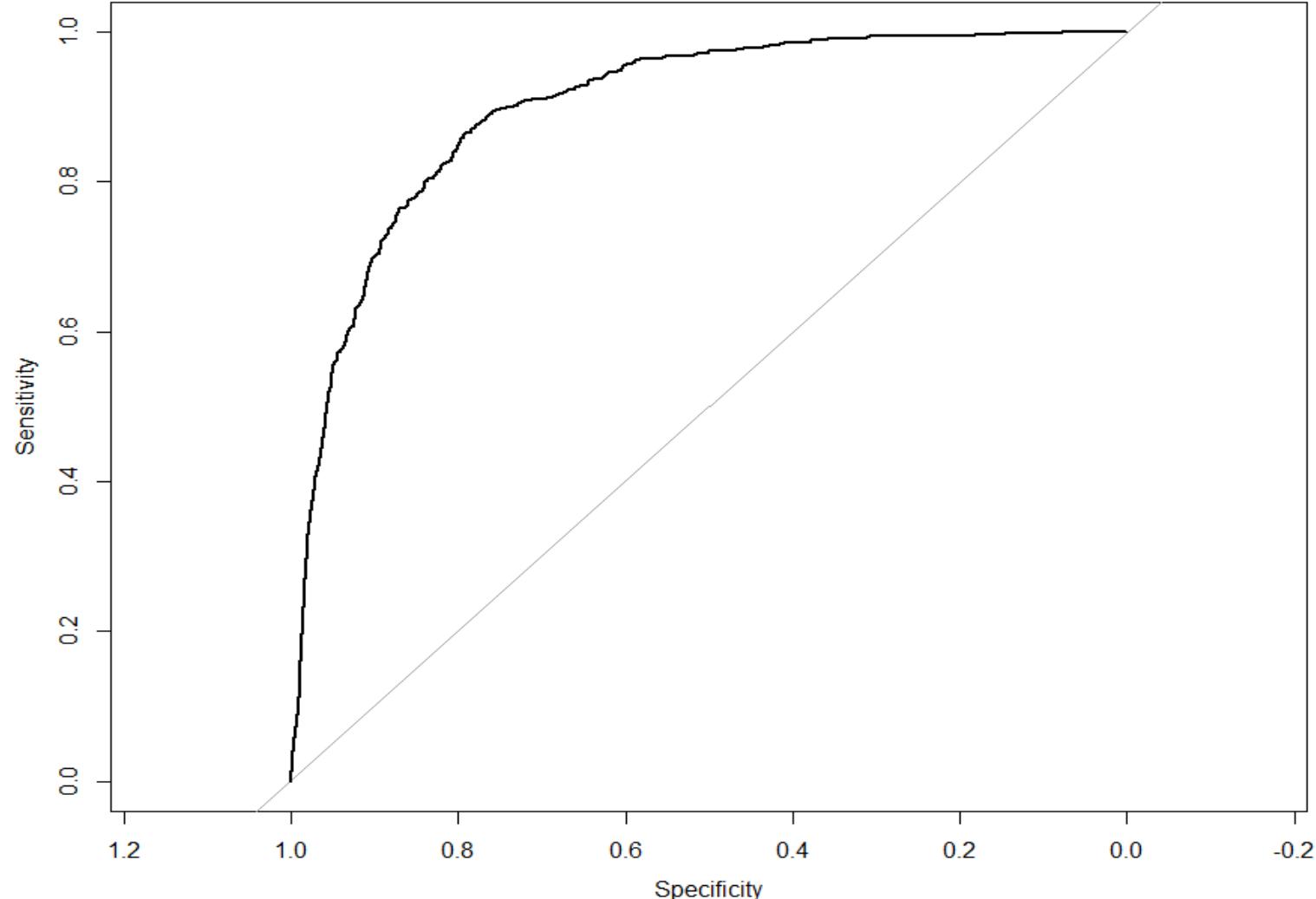
Explorar otros métodos de
procesamiento de datos no
lineales

MÉTODOS



BACKUP

Curva ROC



Repetiendo el cálculo de la sensibilidad y especificidad para diferentes valores de π en $[0,1]$ se puede dibujar la curva de ROC.

El mejor umbral para π será aquel que dé valores altos (cercaos al 1) de sensibilidad y de especificidad al mismo tiempo.

Cálculo de IC por diferentes métodos

IC clásico

```
> t.test(y_cm~p_cm, p.adjust.method = "none")$conf.int  
[1] -0.12573071 -0.09062088  
attr("conf.level")  
[1] 0.95  
> t.test(y_cl~p_cl, p.adjust.method = "none")$conf.int  
[1] -0.3951870 -0.3180469  
attr("conf.level")  
[1] 0.95  
> t.test(y_ml~p_ml, p.adjust.method = "none")$conf.int  
[1] -0.2897824 -0.2071000  
attr("conf.level")  
[1] 0.95
```

Hay diferencia de proporciones entre los grupos ya que el 0 no está en el IC.

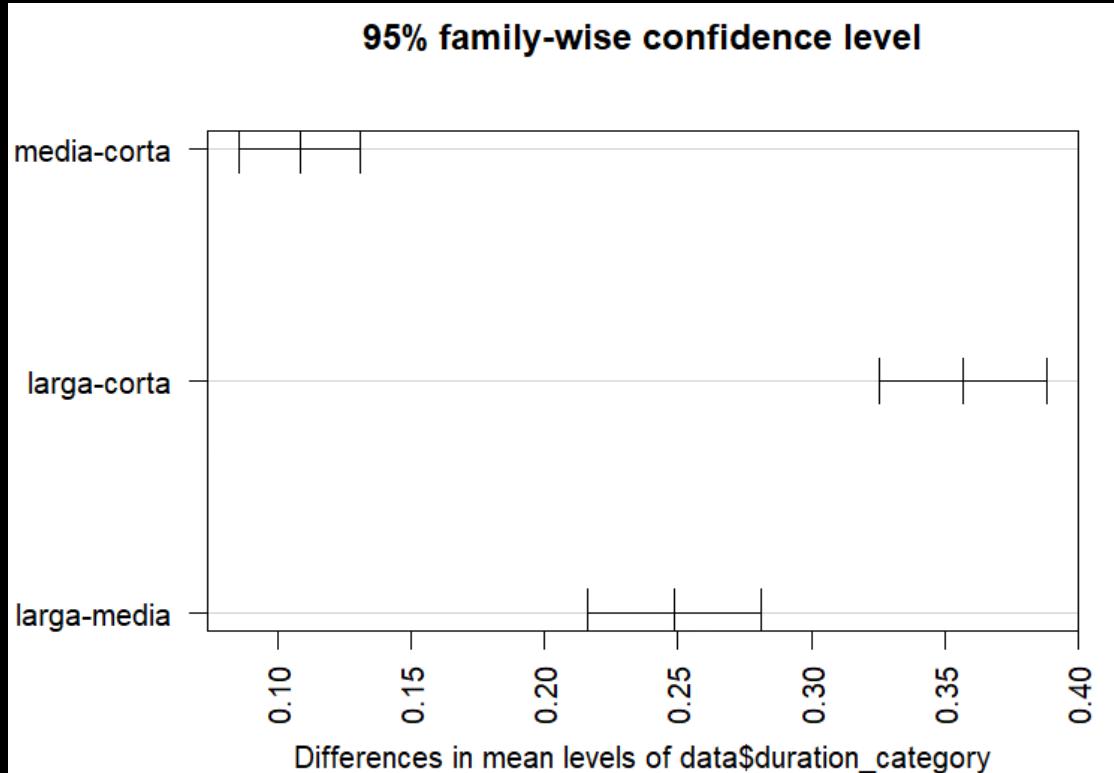
Cálculo de IC por algunos métodos

IC Tukey

```
> TukeyHSD(modelo_def1)
Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = y.bin ~ duration_category, data = data)

$duration_category
      diff      lwr      upr p adj
media-corta 0.1081758 0.08565931 0.1306923 0
larga-corta 0.3566170 0.32522668 0.3880073 0
larga-media 0.2484412 0.21596122 0.2809212 0
```



Hay diferencia de proporciones entre los grupos porque el 0 no está en el IC y los p ajustados=0.

Ajuste de p-valores

Ajustar los p-valores por Bonferroni

```
> pairwise.t.test(data$y.bin, data$duration_category, p.adjust.method="bonfe")
  Pairwise comparisons using t tests with pooled SD
data: data$y.bin and data$duration_category
  corta   media
media <2e-16 -
larga <2e-16 <2e-16
  P value adjustment method: bonferroni
```



Hay diferencia de proporciones entre todos los pares de grupos de la variable duración: corta-media, corta-larga y media-larga.