

SISTEMATIZACION Y MÉTODOS ESTADÍSTICOS

 SAN JUAN BAUTISTA

UNIVERSIDAD PRIVADA



ESTUDIANTES:

- MARÍA LUCIA JACOBO ATUNCAR
- GAMBOA CANALES MARIPAZ
- ARIANA ABIGIAL VIDAL ROMUCHO
- FERNANDA GIANELLA CASTILLA SALVADOR
- SEBASTIAN PALOMINO ROJAS
- KRISTY STEFANY ALVAREZ PEVES

CARGAR PAQUETES Y IMPORTAR LA DATA



Cargar los paquetes

```
{r}
library(tidyverse)
library(here)
library(rio)
library(gtsummary)
library(car)
library(survival)
```

```
{r}
cirrosis_4 <- import(here("cirrosi.csv"))
```

Un vistazo a los datos

```
{r}
head(cirrosis_4)
```

A tibble: 6 × 21

id <dbl>	dias_seguimie... <dbl>	estado <chr>	medicamento <chr>	edad <dbl>
1	400	Fallecido	D_penicilamina	21464
2	4500	Censurado	D_penicilamina	20617
3	1012	Fallecido	D_penicilamina	25594
4	1925	Fallecido	D_penicilamina	19994
5	1504	Censurado_tras...	Placebo	13918
6	2503	Fallecido	Placebo	24201

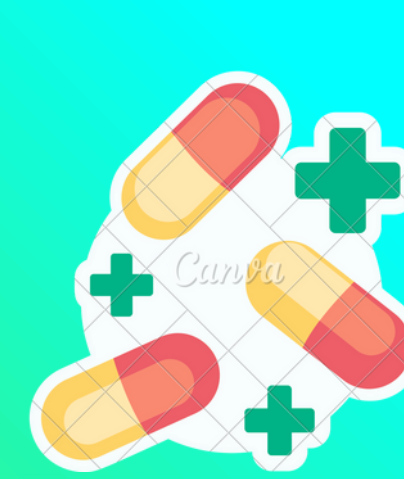
6 rows | 1-5 of 21 columns

pacientes total
de : 418

El análisis univariado

```
{r}
cirrosis_5 <- cirrosis_4 |>
  mutate(hipert = relevel(as.factor(dias_seguimiento), ref = "4500"), edades =
relevel(as.factor(medicamento), ref = "D_penicilamina"),
  desenlace = relevel(as.factor(etapa), ref = "Etapa 4")
) |>
na.omit()
```

Para obtener la tabla con los resultados del análisis univariado, se utiliza la función *tbl_uvregression()*, que permite generar tablas con las estimaciones de regresión logística para cada variable incluida



En esta tabla, los resultados se expresan como odds ratios no ajustados (OR) con sus respectivos intervalos de confianza al 95% y valores p

```
{r}
tabla_reg_log_univ <- cirrosis_5 |>
tbl_uvregression(
  include = c(dias_seguimiento, estado, medicamento, edad, sexo,ascitis, edema),
  y = desenlace,
  method = glm,
  method.args = list(family = binomial),
  exponentiate = TRUE,
  conf.int = TRUE,
  hide_n = TRUE,
  add_estimate_to_reference_rows = FALSE,
  pvalue_fun = ~ style_pvalue(.x, digits = 3),
  estimate_fun = ~ style_number(.x, digits = 2),
  label = list(
    dias_seguimiento ~ "dias post diagnostico",
    estado ~ "estado del paciente",
    medicamento ~ "medicamentos",
    edad ~ "dias x años",
    sexo ~ "Hombre/Mujer",
    ascitis ~ "Sí/No",
    edema ~ "estado de edema")
) |>
bold_labels() |>
bold_p(t = 0.05) |>
modify_header(estimate = "***OR no ajustado***", p.value = "***valor p**")
```

Characteristic	OR no ajustado	95% CI	Valor P
dias post diagnostico	1.00	1.00, 1.00	<0.001
estado del paciente			
Censurado	—	—	
Censurado_trasplante	0.44	0.16, 1.27	0.114
Fallecido	0.28	0.16, 0.48	<0.001
medicamentos			
D_penicilamina	—	—	
Placebo	0.92	0.56, 1.51	0.738
dias x años	1.00	1.00, 1.00	<0.001
Hombre/Mujer			
Hombre	—	—	
Mujer	1.42	0.67, 2.93	0.351
Sí/No			
No	—	—	
Sí	0.00		0.984
estado de edema			
Ausente	—	—	
Leve	0.44	0.19, 1.03	0.056
Severo	0.09	0.02, 0.28	<0.001
Abbreviations: CI = Confidence Interval, OR = Odds Ratio			

El análisis multivariado

Para el análisis de regresión logística multivariada, se aplicó una estrategia de selección automática de variables utilizando tres enfoques: eliminación hacia atrás (backward elimination), selección hacia adelante (forward selection) y selección paso a paso (stepwise selection).

1

Ajuste del modelo inicial Ajustamos un modelo de regresión logística binaria que incluya todas las variables candidatas

```
{r}  
var_modelo = glm(  
  desenlace ~ dias_seguimiento + edad + bilirrubina + colesterol + albumina +  
  fosfatasa_alcalina + trigliceridos,  
  data = cirrosis_5,  
  family = binomial(link = "logit"))
```

2

Realizamos la selección de variables usando la técnica Eliminación hacia atrás (Backward elimination)

```
{r}  
multi_backward <- var_modelo |>  
  step(direction = "backward", trace = FALSE)
```

3

Realizamos la selección de variables usando la técnica Selección hacia adelante (Forward selection).

```
{r}  
multi_forward <- var_modelo |>  
  step(direction = "forward", trace = FALSE)
```

4

Realizamos la selección de variables usando la técnica Selección paso a paso (Stepwise selection).

```
{r}  
multi_stepwise <- var_modelo |>  
  step(direction = "both", trace = FALSE)
```



Estimados el AIC para los modelos.

Podemos visualizar el AIC y cuáles variables han sido seleccionadas en cada modelo, usando la función `summary`.

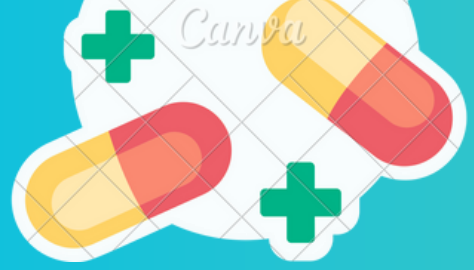
```
{r}  
summary(multi_backward)
```

```
{r}  
summary(multi_forward)
```

```
{r}  
summary(multi_stepwise)
```

```
{r}  
summary(multi_backward)  
  
Call:  
glm(formula = desenlace ~ dias_seguimiento + edad + bilirrubina +  
      colesterol + albumina, family = binomial(link = "logit"),  
      data = cirrosis_5)  
  
Coefficients:  
              Estimate Std. Error z value Pr(>|z|)  
(Intercept)  -5.006e+00  1.803e+00  -2.777  0.005488 **  
dias_seguimiento  4.674e-04  1.636e-04   2.856  0.004289 **  
edad            -6.360e-05  3.901e-05  -1.630  0.103046 .  
bilirrubina      -7.280e-02  4.093e-02  -1.779  0.075260 .  
colesterol        2.303e-03  9.098e-04   2.531  0.011368 *  
albumina         1.542e+00  4.344e-01   3.551  0.000384 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
Null deviance: 354.06  on 275  degrees of freedom  
Residual deviance: 286.53  on 270  degrees of freedom  
AIC: 298.53  
  
Number of Fisher scoring iterations: 4
```





Estimados el AIC para los modelos.

Podemos visualizar el AIC y cuáles variables han sido seleccionadas en cada modelo, usando la función `summary`.

```
{r}  
summary(multi_forward)
```

Call:

```
glm(formula = desenlace ~ dias_seguimiento + edad + bilirrubina +  
     colesterol + albumina + fosfatasa_alcalina + trigliceridos,  
     family = binomial(link = "logit"), data = cirrosis_5)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-5.218e+00	1.846e+00	-2.826	0.004713	**
dias_seguimiento	4.469e-04	1.677e-04	2.664	0.007720	**
edad	-6.326e-05	3.911e-05	-1.617	0.105777	
bilirrubina	-7.840e-02	4.429e-02	-1.770	0.076709	.
colesterol	2.224e-03	9.150e-04	2.431	0.015059	*
albumina	1.588e+00	4.456e-01	3.562	0.000367	***
fosfatasa_alcalina	4.170e-05	7.786e-05	0.536	0.592206	
trigliceridos	4.117e-04	2.559e-03	0.161	0.872183	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 354.06 on 275 degrees of freedom
Residual deviance: 286.19 on 268 degrees of freedom
AIC: 302.19

Number of Fisher Scoring iterations: 4





Estimados el AIC para los modelos.

Podemos visualizar el AIC y cuáles variables han sido seleccionadas en cada modelo, usando la función `summary`.

```
{r}
summary(multi_stepwise)

Call:
glm(formula = desenlace ~ dias_seguimiento + edad + bilirrubina +
     colesterol + albumina, family = binomial(link = "logit"),
     data = cirrosis_5)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.006e+00  1.803e+00  -2.777  0.005488 **
dias_seguimiento  4.674e-04  1.636e-04   2.856  0.004289 **
edad          -6.360e-05  3.901e-05  -1.630  0.103046
bilirrubina    -7.280e-02  4.093e-02  -1.779  0.075260 .
colesterol     2.303e-03  9.098e-04   2.531  0.011368 *
albumina       1.542e+00  4.344e-01   3.551  0.000384 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 354.06  on 275  degrees of freedom
Residual deviance: 286.53  on 270  degrees of freedom
AIC: 298.53

Number of Fisher Scoring iterations: 4
```



Evaluación de colinealidad

Finalmente, evaluamos la colinealidad usando la función `check_collinearity()` del paquete `performance`.

```
{r}
performance::check_collinearity(multi_backward, ci = NULL)

# Check for Multicollinearity

Low Correlation
```

	Term	VIF	adj. VIF	Tolerance
	dias_seguimiento	1.22	1.11	0.82
	edad	1.04	1.02	0.96
	bilirrubina	1.53	1.24	0.65
	colesterol	1.37	1.17	0.73
	albumina	1.11	1.05	0.90

```
{r}
performance::check_collinearity(multi_forward, ci = NULL)

# Check for Multicollinearity

Low Correlation
```

	Term	VIF	adj. VIF	Tolerance
	dias_seguimiento	1.28	1.13	0.78
	edad	1.04	1.02	0.96
	bilirrubina	1.80	1.34	0.56
	colesterol	1.42	1.19	0.70
	albumina	1.16	1.08	0.86
	fosfatasa_alcalina	1.13	1.06	0.89
	trigliceridos	1.34	1.16	0.75

```
{r}
performance::check_collinearity(multi_stepwise, ci = NULL)

# Check for Multicollinearity

Low Correlation
```

	Term	VIF	adj. VIF	Tolerance
	dias_seguimiento	1.22	1.11	0.82
	edad	1.04	1.02	0.96
	bilirrubina	1.53	1.24	0.65
	colesterol	1.37	1.17	0.73
	albumina	1.11	1.05	0.90



Reporte del análisis univariado y multivariado



Como en las sesiones anteriores, reportaremos los resultados del modelo final de regresión logística.

Tabla para los resultados de la regresión univariado (no ajustado)

```
{r}
tabla_univ <- cirrosis_5 |>
tbl_uvregression(
  include = c(dias_seguimiento, estado, medicamento, edad, sexo, ascitis, edema),
  y = desenlace,
  method = glm,
  method.args = list(family = binomial),
  exponentiate = TRUE,
  conf.int = TRUE,
  hide_n = TRUE,
  add_estimate_to_reference_rows = FALSE,
  pvalue_fun = ~ style_pvalue(.x, digits = 3),
  estimate_fun = ~ style_number(.x, digits = 2),
  label = list(
    dias_seguimiento ~ "dias post diagnostico",
    estado ~ "estado del paciente",
    medicamento ~ "medicamentos",
    edad ~ "dias x años",
    sexo ~ "Hombre/Mujer",
    ascitis ~ "Sí/No",
    edema ~ "estado de edema")
) |>
bold_labels() |>
bold_p(t = 0.05) |>
modify_header(estimate = "***OR**", p.value = "***valor p**")
```

Tabla para los resultados de la regresión multivariable (ajustado)

```
{r}
tabla_multi <- glm(
  desenlace ~ dias_seguimiento + trigliceridos + bilirrubina + colesterol ,
  family = binomial(link = "logit"),
  data = cirrosis_5
) |>
tbl_regression(
  exponentiate = TRUE,
  conf.int = TRUE,
  pvalue_fun = ~ style_pvalue(.x, digits = 3),
  estimate_fun = ~ style_number(.x, digits = 2),
  label = list(
    dias_seguimiento ~ "dias post diagnostico",
    trigliceridos ~ "mg/dL",
    bilirrubina ~ "%",
    colesterol ~ "mg/dL"
  )
) |>
bold_labels() |>
bold_p(t = 0.05) |>
modify_header(estimate = "***OR**", p.value = "***valor p**")
```

Reporte del análisis univariado y multivariado

La tabla final la construimos usando la función `tbl_merge()`. De modo que la tabla del análisis univariado o no ajustado y multivariado o ajustado, se muestren lado a lado.

```
{r}
tabla_final <-
  tbl_merge(
    list(tabla_univ, tabla_multi),
    tab_spanner = c("**Univariado**", "**Multivariado**"))
```

Characteristic	Univariado			Multivariado		
	OR	95% CI	Valor P	OR	95% CI	Valor P
dias post diagnostico	1.00	1.00, 1.00	<0.001	1.00	1.00, 1.00	<0.001
estado del paciente						
Censurado	—	—				
Censurado_trasplante	0.44	0.16, 1.27	0.114			
Fallecido	0.28	0.16, 0.48	<0.001			
medicamentos						
D_penicilamina	—	—				
Placebo	0.92	0.56, 1.51	0.738			
dias x años	1.00	1.00, 1.00	<0.001			
Hombre/Mujer						
Hombre	—	—				
Mujer	1.42	0.67, 2.93	0.351			
Sí/No						
No	—	—				
Sí	0.00		0.984			

estado de edema						
Ausente	—	—				
Leve	0.44	0.19, 1.03	0.056			
Severo	0.09	0.02, 0.28	<0.001			
mg/dL		1.00	1.00, 1.01	0.788		
%		0.90	0.82, 0.98	0.017		
mg/dL		1.00	1.00, 1.00	0.005		
Abbreviations: CI = Confidence Interval, OR = Odds Ratio						





MUCHAS GRACIAS

