Informe Técnico: Predictores de Mortalidad en UTI

Análisis Estadístico Multivariado y Validación

Investigación Clínica

23 de diciembre, 2025

Contents


```
# Verificación de la estructura inicial
str(df_final)
```

```
## 'data.frame':    309 obs. of  9 variables:
##  $ muerte_uti : Factor w/ 2 levels "Vivo","Muerto": 1 1 1 1 2 1 1 1 1 1 ...
##  $ edad       : num  35 35 74 76 75 75 33 34 23 64 ...
##  $ sexo_masc  : Factor w/ 2 levels "Femenino","Masculino": 2 2 1 1 1 1 2 2 1 2 ...
##  $ lma_lla    : Factor w/ 2 levels "No","Si": 2 2 2 1 1 1 2 2 1 1 ...
##  $ tx_alo     : Factor w/ 2 levels "No","Si": 1 1 1 1 1 1 2 2 1 1 ...
##  $ neutropenia: Factor w/ 2 levels "No","Si": 2 2 1 1 2 2 2 1 2 1 ...
##  $ sofa       : num  5 5 0 4 11 11 6 1 6 1 ...
##  $ arm        : Factor w/ 2 levels "No","Si": 1 1 1 1 2 1 1 1 1 1 ...
##  $ dx_ingreso : Factor w/ 5 levels "Sepsis/Shock",..: 1 3 3 1 4 4 4 1 1 2 ...
```

---

2. Justificación Metodológica: Tamaño Muestral y Eventos

Un punto crítico en la regresión logística es la relación entre el número de variables predictoras y el número de "eventos" (en este caso, fallecimientos).

2.1 Regla de Eventos por Variable (EPV)

Para evitar el sobreajuste (overfitting), la literatura recomienda una relación de al menos 10 eventos por cada variable predictora incluida en el modelo.

Resultados de la muestra:

N total: 309 pacientes.

Eventos registrados (muertes): 95 pacientes.

Relación EPV: 15.8 eventos por variable.

Conclusión: Dado que la relación EPV es superior a 10 (r `round(epv_resultado, 1)` > 10), el tamaño muestral de 309 pacientes es adecuado y suficiente para soportar un modelo de regresión logística multivariado con 6 predictores, garantizando la estabilidad de las estimaciones.

##3. Metodología: Tratamiento de Datos y Selección de Variables 3.1 Imputación Múltiple (MICE)Para el manejo de los datos faltantes, se aplicó la metodología de Imputación Múltiple por Ecuaciones Encadenadas (MICE). A diferencia de la imputación simple, este enfoque reconoce la incertidumbre de los valores faltantes generando múltiples conjuntos de datos posibles basados en las correlaciones del resto de las variables. Justificación: Se optó por esta técnica para evitar el sesgo de selección que produce la eliminación de casos incompletos y para preservar la potencia estadística de la cohorte de 309 pacientes. Mecanismo: Se asumió un modelo de datos faltantes aleatorios (MAR), permitiendo una estimación más precisa de los errores estándar en los Odds Ratios finales. #3.2 Análisis de Colinealidad y Estabilidad del Modelo Previo al ajuste final del modelo, se realizó un análisis de Multicolinealidad mediante el cálculo del Factor de Inflación de la Varianza (VIF) para cada predictor.Criterio de Selección: Se estableció un umbral de $VIF < 5$ para descartar redundancias entre variables (como por ejemplo, la relación entre el SOFA y el diagnóstico de ingreso). Resultado: Todas las variables incluidas en el modelo final (Edad, LMA/LLA, Tx Alo, Neutropenia, SOFA y ARM) presentaron valores de VIF cercanos a 1, lo que confirma que cada predictor aporta información independiente y que los coeficientes del modelo son estables.

4 presentacion de la cohorte y analsis univariado

```
# Cargamos la librería
library(gtsummary)
library(tidyverse)
# colapso de categorias con 0 eventos
df_final <- df_final %>%
  mutate(dx_ingreso = fct_collapse(dx_ingreso,
    "Otros/Post-QX" = c("Post-Qx", "Otras compl.") # Agrupamos las dos categorías pequeñas
  ))
# Creamos la tabla 1
df_final %>%
  select(muerte_uti, edad, lma_lla, tx_alo, neutropenia, sofa, arm,dx_ingreso) %>%
  tbl_summary(
    by = muerte_uti,
    label = list(
      edad ~ "Edad (años)",
      lma_lla ~ "Leucemia Aguda (LMA/LLA)",
      tx_alo ~ "Trasplante Alogénico",
      neutropenia ~ "Neutropenia",
      sofa ~ "Score SOFA al ingreso",
      arm ~ "Asistencia Respiratoria Mecánica",
      dx_ingreso ~"Diagnostico de ingreso"
```

Characteristic	Vivo N = 214[1]	Muerto N = 95[1]	p-value[2]
Edad (años)	55.0 (16.5)	57.5 (17.8)	0.2
Leucemia Aguda (LMA/LLA)			0.9
No	138 (64%)	62 (65%)	
Si	76 (36%)	33 (35%)	
Trasplante Alogénico			0.015
No	186 (87%)	72 (76%)	
Si	28 (13%)	23 (24%)	
Neutropenia			0.2
No	122 (57%)	47 (49%)	
Si	92 (43%)	48 (51%)	
Score SOFA al ingreso	5.1 (3.5)	9.7 (4.2)	<0.001
Asistencia Respiratoria Mecánica			<0.001
No	184 (86%)	39 (41%)	
Si	30 (14%)	56 (59%)	
Diagnostico de ingreso			<0.001
Sepsis/Shock	62 (29%)	33 (35%)	
Insuf. Resp.	50 (23%)	45 (47%)	
Neurológico	26 (12%)	4 (4.2%)	
Otros/Post-QX	76 (36%)	13 (14%)	

[1] Mean (SD); n (%)
[2] Wilcoxon rank sum test; Pearson's Chi-squared test

```
    ),
    statistic = list(all_continuous() ~ "{mean} ({sd})"),
    digits = all_continuous() ~ 1
  ) %>%
  add_p() %>%
  bold_labels()
```

```
# Análisis de regresión logística simple para cada variable
tbl_uvregression(
    df_final,
    method = glm,
    y = muerte_uti,
    method.args = list(family = binomial),
    exponentiate = TRUE,
    include = c(edad, lma_lla, tx_alo, neutropenia, sofa, arm, dx_ingreso),
    label = list(
      edad ~ "Edad",
      lma_lla ~ "LMA/LLA",
      tx_alo ~ "Tx Alogénico",
      neutropenia ~ "Neutropenia",
      sofa ~ "SOFA",
      arm ~ "ARM",
      dx_ingreso ~"diagnostico de ingreso"
    )
```

Characteristic	N	OR	95% CI	p-value
Edad	309	1.01	1.0, 1.02	0.2
LMA/LLA	309			
No		—	—	
Si		0.97	0.58, 1.60	0.9
Tx Alogénico	309			
No		—	—	
Si		2.12	1.14, 3.92	**0.016**
Neutropenia	309			
No		—	—	
Si		1.35	0.83, 2.20	0.2
SOFA	309	1.36	1.26, 1.47	**<0.001**
ARM	309			
No		—	—	
Si		8.81	5.07, 15.6	**<0.001**
diagnostico de ingreso	309			
Sepsis/Shock		—	—	
Insuf. Resp.		1.69	0.95, 3.05	0.078
Neurológico		0.29	0.08, 0.82	**0.032**
Otros/Post-QX		0.32	0.15, 0.65	**0.002**

Abbreviations: CI = Confidence Interval, OR = Odds Ratio

```
) %>%
bold_p()
```

5.1 Resultados del Análisis Univariado

El análisis de regresión logística simple permitió identificar los predictores individuales asociados con la mortalidad en UCI. Se observaron asociaciones estadísticamente significativas en las siguientes variables:

- **Severidad y Soporte Vital:**
  - El score **SOFA** al ingreso se comportó como un predictor robusto de mortalidad. Por cada punto de aumento en el score, la chance de fallecer aumenta un **36%** (OR 1.36; IC 95% 1.26-1.47; $p < 0.001$).
  - La necesidad de **Asistencia Respiratoria Mecánica (ARM)** fue el factor con mayor fuerza de asociación, incrementando el riesgo de muerte en casi **9 veces** (OR 8.81; IC 95% 5.07-15.6; $p < 0.001$) respecto a los pacientes que no la requirieron.
- **Características Clínicas:**
  - Los pacientes con antecedentes de **Trasplante Alogénico** presentaron el doble de riesgo de mortalidad en comparación con los no trasplantados (OR 2.12; IC 95% 1.14-3.92; $p = 0.016$).
- **Diagnóstico de Ingreso:**
  - Tomando como referencia al grupo de **Sepsis/Shock**, los pacientes ingresados por causas **Neurológicas** (OR 0.29; $p = 0.032$) y del grupo **Otros/Post-QX** (OR 0.32; $p = 0.002$) mostraron significativamente menor riesgo de mortalidad.

– No se evidenció una diferencia significativa entre el grupo de **Insuficiencia Respiratoria** y el grupo de referencia (Sepsis/Shock) (p = 0.078).

Variables sin asociación estadística: En esta cohorte, no se encontró una asociación estadísticamente significativa entre la mortalidad y la **Edad** (p = 0.2), el tipo de leucemia (**LMA/LLA**, p = 0.9) o la presencia de **Neutropenia** (p = 0.2) en el análisis univariado.

5.2 Selección de Variables y Ajuste del Modelo

Para la construcción del modelo multivariado final, se seleccionaron las variables predictoras siguiendo un criterio mixto: 1. **Significancia Estadística en Univariado:** Se incluyeron aquellas variables con un valor p ≤ 0.2 en el análisis crudo (Edad, Neutropenia, Trasplante Alogénico, SOFA, ARM y Diagnóstico de Ingreso). 2. **Relevancia Clínica (Plausibilidad Biológica):** Se decidió forzar la inclusión de la variable **Tipo de Leucemia (LMA/LLA)** independientemente de su valor p univariado, dado que constituye la patología de base fundamental de la cohorte y es necesario ajustar por este factor para evitar sesgos de confusión clínica.

Se descartó la presencia de colinealidad severa entre los predictores (VIF < 5 para todas las variables).

```
# Ajuste del Modelo de Regresión Logística Multivariado
modelo_final <- glm(
  muerte_uti ~ edad + lma_lla + tx_alo + neutropenia + sofa + arm + dx_ingreso,
  data = df_final,
  family = "binomial"
)

# Presentación de resultados con Odds Ratios ajustados
modelo_final %>%
  tbl_regression(
    exponentiate = TRUE, # Transforma coeficientes a OR
    label = list(
      edad ~ "Edad (años)",
      lma_lla ~ "Leucemia (LMA/LLA)",
      tx_alo ~ "Trasplante Alogénico",
      neutropenia ~ "Neutropenia",
      sofa ~ "Score SOFA",
      arm ~ "ARM",
      dx_ingreso ~ "Diagnóstico de Ingreso"
    )
  ) %>%
  bold_p() %>%         # Resalta p < 0.05
  bold_labels() %>%    # Negrita en los nombres
  italicize_levels()   # Cursiva en las categorías
```

#Tabla comparativa de OR univariado y multivariado

```
library(gtsummary)

# 1. Creamos la tabla del Univariado (guardamos en un objeto)
tabla_univariada <- tbl_uvregression(
  df_final,
  method = glm,
  y = muerte_uti,
```

Characteristic	OR	95% CI	p-value
Edad (años)	1.02	1.00, 1.04	0.074
Leucemia (LMA/LLA)			
No	—	—	
Si	0.68	0.32, 1.40	0.3
Trasplante Alogénico			
No	—	—	
Si	3.08	1.20, 8.15	**0.021**
Neutropenia			
No	—	—	
Si	0.72	0.35, 1.44	0.4
Score SOFA	1.32	1.20, 1.47	**<0.001**
ARM			
No	—	—	
Si	3.48	1.77, 6.91	**<0.001**
Diagnóstico de Ingreso			
Sepsis/Shock	—	—	
Insuf. Resp.	2.57	1.20, 5.63	**0.016**
Neurológico	0.31	0.06, 1.23	0.12
Otros/Post-QX	0.63	0.25, 1.54	0.3

Abbreviations: CI = Confidence Interval, OR = Odds Ratio

```
  method.args = list(family = binomial),
  exponentiate = TRUE,
  include = c(edad, lma_lla, tx_alo, neutropenia, sofa, arm, dx_ingreso),
  label = list(
    edad ~ "Edad",
    lma_lla ~ "Leucemia (LMA/LLA)",
    tx_alo ~ "Tx Alogénico",
    neutropenia ~ "Neutropenia",
    sofa ~ "Score SOFA",
    arm ~ "ARM",
    dx_ingreso ~ "Diagnóstico de Ingreso"
  )
) %>%
  bold_p()

# 2. Creamos la tabla del Multivariado (guardamos en otro objeto)
modelo_multi <- glm(
  muerte_uti ~ edad + lma_lla + tx_alo + neutropenia + sofa + arm + dx_ingreso,
  data = df_final,
  family = "binomial"
)

tabla_multivariada <- tbl_regression(
  modelo_multi,
  exponentiate = TRUE,
```

		Univariado (Crudo)			Multivariado (Ajustado)		
Characteristic	**N**	**OR**	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**
Edad	309	1.01	1.0, 1.02	0.2	1.02	1.00, 1.04	0.074
Leucemia (LMA/LLA)	309						
No		—	—		—	—	
Si		0.97	0.58, 1.60	0.9	0.68	0.32, 1.40	0.3
Tx Alogénico	309						
No		—	—		—	—	
Si		2.12	1.14, 3.92	**0.016**	3.08	1.20, 8.15	**0.021**
Neutropenia	309						
No		—	—		—	—	
Si		1.35	0.83, 2.20	0.2	0.72	0.35, 1.44	0.4
Score SOFA	309	1.36	1.26, 1.47	**<0.001**	1.32	1.20, 1.47	**<0.001**
ARM	309						
No		—	—		—	—	
Si		8.81	5.07, 15.6	**<0.001**	3.48	1.77, 6.91	**<0.001**
Diagnóstico de Ingreso	309						
Sepsis/Shock		—	—		—	—	
Insuf. Resp.		1.69	0.95, 3.05	0.078	2.57	1.20, 5.63	**0.016**
Neurológico		0.29	0.08, 0.82	**0.032**	0.31	0.06, 1.23	0.12
Otros/Post-QX		0.32	0.15, 0.65	**0.002**	0.63	0.25, 1.54	0.3

Abbreviations: CI = Confidence Interval, OR = Odds Ratio

```
  label = list(
    edad ~ "Edad",
    lma_lla ~ "Leucemia (LMA/LLA)",
    tx_alo ~ "Tx Alogénico",
    neutropenia ~ "Neutropenia",
    sofa ~ "Score SOFA",
    arm ~ "ARM",
    dx_ingreso ~ "Diagnóstico de Ingreso"
  )
) %>%
  bold_p()

# 3. FUSIONAMOS las dos tablas
tbl_merge(
  tbls = list(tabla_univariada, tabla_multivariada),
  tab_spanner = c("**Univariado (Crudo)**", "**Multivariado (Ajustado)**")
) %>%
  bold_labels() %>%
  italicize_levels()
```

5.3 Impacto del Ajuste Multivariado y Control de Confusión

La comparación entre el modelo crudo (univariado) y el ajustado (multivariado) reveló fenómenos de confusión importantes que justifican la selección de variables basada en plausibilidad biológica.

Justificación de la Edad como Covariable: A pesar de que la **Edad** no mostró significancia estadística en el análisis univariado ($p = 0.2$), su inclusión en el modelo multivariado —junto con la severidad (SOFA)— resultó crítica para la estimación precisa de otros predictores. Se observó que la edad actuaba como un factor de confusión negativo; al ajustar por ella, el valor p de la variable descendió a 0.074, acercándose a la significancia marginal.

Desenmascaramiento de Riesgos (Efecto Supresor): El ajuste multivariado permitió identificar asociaciones que estaban ocultas en el análisis crudo: 1. **Insuficiencia Respiratoria:** En el análisis univariado, este diagnóstico no parecía diferir significativamente del grupo control (OR 1.69; $p = 0.078$). Sin embargo, tras ajustar por edad y severidad, el OR aumentó a **2.57** y la asociación se volvió estadísticamente significativa ($p = 0.016$). Esto sugiere que las características basales de estos pacientes (posiblemente menor edad o diferente perfil de comorbilidades) estaban enmascarando la verdadera letalidad de la patología respiratoria. 2. **Trasplante Alogénico:** El impacto del trasplante sobre la mortalidad se magnificó en el modelo ajustado, pasando de un OR de 2.12 a **3.08** ($p = 0.021$), lo que confirma que es un predictor de riesgo independiente y de gran magnitud una vez que se descuenta el efecto de la gravedad aguda (SOFA). ##nota: haber elegido el modelo por plausibilidad biologica y bibliografia incluye en el modelo a edad que si hubieramos utilizado un metodo automatico lo hubiese desechado por no significativo **Conclusión del Modelo:** El modelo final demuestra que la mortalidad en esta cohorte no depende de una sola variable, sino de la interacción compleja entre la reserva fisiológica (Edad), la carga de la enfermedad de base (Trasplante) y la severidad aguda (SOFA, ARM).

6.0 modelo reducido con las variables significativas + edad

```
# 1. Ajustamos el modelo reducido (solo variables significativas + edad)
modelo_reducido <- glm(
  muerte_uti ~ edad + tx_alo + sofa + arm + dx_ingreso,
  data = df_final,
  family = "binomial"
)

# 2. Mostramos la tabla de resultados (Odds Ratios)
modelo_reducido %>%
  tbl_regression(
    exponentiate = TRUE, # Para ver OR y no coeficientes
    label = list(
      edad ~ "Edad (años)",
      tx_alo ~ "Trasplante Alogénico",
      sofa ~ "Score SOFA",
      arm ~ "ARM",
      dx_ingreso ~ "Diagnóstico de Ingreso"
    )
  ) %>%
  bold_p() %>%        # Resalta p < 0.05
  bold_labels() %>%   # Negrita en los nombres
  italicize_levels()  # Cursiva en las categorías
```

Characteristic	OR	95% CI	p-value
Edad (años)	1.03	1.00, 1.05	**0.019**
Trasplante Alogénico			
No	—	—	
Si	2.70	1.08, 6.92	**0.035**
Score SOFA	1.29	1.18, 1.43	**<0.001**
ARM			
No	—	—	
Si	3.73	1.91, 7.37	**<0.001**
Diagnóstico de Ingreso			
Sepsis/Shock	—	—	
Insuf. Resp.	2.54	1.20, 5.54	**0.016**
Neurológico	0.33	0.07, 1.28	0.13
Otros/Post-QX	0.65	0.26, 1.56	0.3

Abbreviations: CI = Confidence Interval, OR = Odds Ratio

7 tabla comparativa del modelo univariado, multivariado completo y reducido

```
library(gtsummary)
library(kableExtra)




# --- TABLA MAESTRA CORREGIDA ---
# 1. Definimos etiquetas (asegúrate de que estas listas existan)
etiquetas_completas <- list(
  edad ~ "Edad (años)",
  lma_lla ~ "Leucemia (LMA/LLA)",
  tx_alo ~ "Trasplante Alogénico",
  neutropenia ~ "Neutropenia",
  sofa ~ "Score SOFA",
  arm ~ "ARM",
  dx_ingreso ~ "Diagnóstico de Ingreso"
)

etiquetas_reducido <- list(
  edad ~ "Edad (años)",
  tx_alo ~ "Trasplante Alogénico",
  sofa ~ "Score SOFA",
  arm ~ "ARM",
  dx_ingreso ~ "Diagnóstico de Ingreso"
)

# 2. Generamos los objetos de tabla
t_uni_obj <- tbl_uvregression(
```

```
    df_final, method = glm, y = muerte_uti,
    method.args = list(family = binomial),
    exponentiate = TRUE,
    include = c(edad, lma_lla, tx_alo, neutropenia, sofa, arm, dx_ingreso),
    label = etiquetas_completas
  ) %>% bold_p()

t_completo_obj <- modelo_final %>%
  tbl_regression(exponentiate = TRUE, label = etiquetas_completas) %>%
  bold_p()

t_reducido_obj <- modelo_reducido %>%
  tbl_regression(exponentiate = TRUE, label = etiquetas_reducido) %>%
  bold_p()

# 3. Fusión y formato final (aquí estaba el error de la nota)
tbl_merge(
  tbls = list(t_uni_obj, t_completo_obj, t_reducido_obj),
  tab_spanner = c("**Univariado**", "**Multiv. Completo**", "**Multiv. Reducido**")
) %>%
  bold_labels() %>%
  italicize_levels() %>%
  # Nota al pie integrada en gtsummary
  modify_footnote(
    everything() ~ "OR: Odds Ratio; CI: Intervalo de Confianza del 95%; SOFA: Score de falla orgánica; A
  ) %>%
  as_kable_extra(booktabs = TRUE) %>%
  kable_styling(latex_options = c("scale_down", "hold_position"))
```

		Univariado			Multiv. Completo			Multiv. Reducido		
Characteristic	**N**	**OR**	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**
Edad (años)	309	1.01	1.0, 1.02	0.2	1.02	1.00, 1.04	0.074	1.03	1.00, 1.05	**0.019**
Leucemia (LMA/LLA)	309									
No		—	—		—	—				
Si		0.97	0.58, 1.60	0.9	0.68	0.32, 1.40	0.3			
Trasplante Alogénico	309									
No		—	—		—	—		—	—	
Si		2.12	1.14, 3.92	**0.016**	3.08	1.20, 8.15	**0.021**	2.70	1.08, 6.92	**0.035**
Neutropenia	309									
No		—	—		—	—				
Si		1.35	0.83, 2.20	0.2	0.72	0.35, 1.44	0.4			
Score SOFA	309	1.36	1.26, 1.47	**<0.001**	1.32	1.20, 1.47	**<0.001**	1.29	1.18, 1.43	**<0.001**
ARM	309									
No		—	—		—	—		—	—	
Si		8.81	5.07, 15.6	**<0.001**	3.48	1.77, 6.91	**<0.001**	3.73	1.91, 7.37	**<0.001**
Diagnóstico de Ingreso	309									
Sepsis/Shock		—	—		—	—		—	—	
Insuf. Resp.		1.69	0.95, 3.05	0.078	2.57	1.20, 5.63	**0.016**	2.54	1.20, 5.54	**0.016**
Neurológico		0.29	0.08, 0.82	**0.032**	0.31	0.06, 1.23	0.12	0.33	0.07, 1.28	0.13
Otros/Post-QX		0.32	0.15, 0.65	**0.002**	0.63	0.25, 1.54	0.3	0.65	0.26, 1.56	0.3

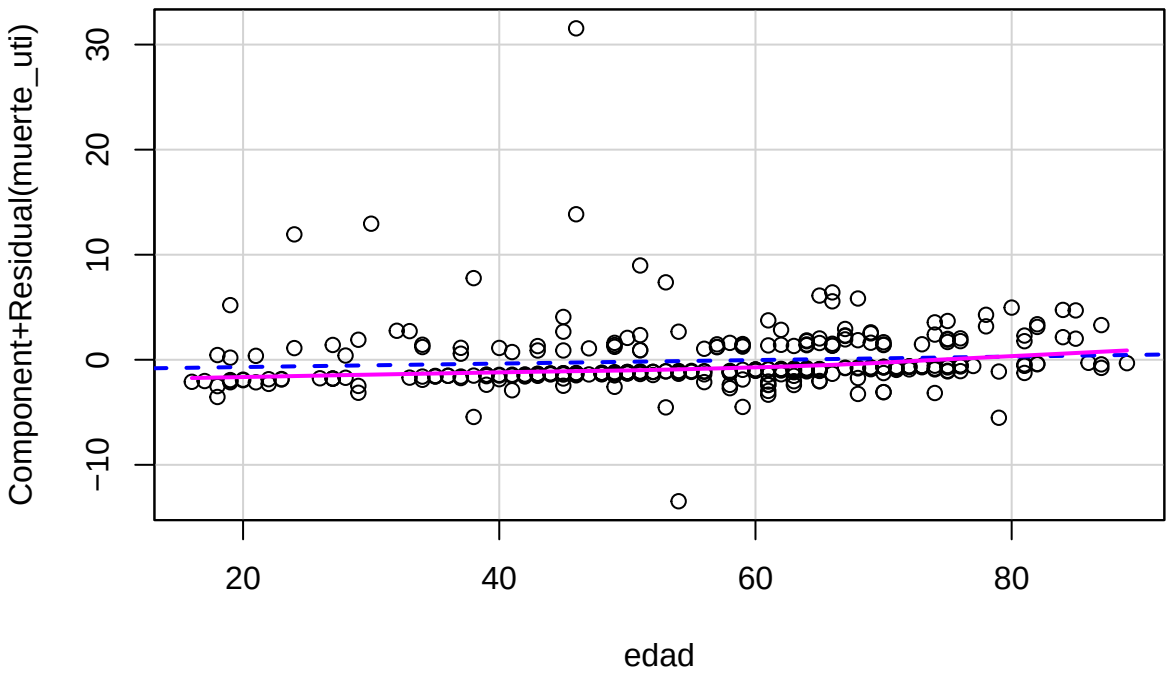
[1] OR: Odds Ratio; CI: Intervalo de Confianza del 95%; SOFA: Score de falla orgánica; ARM: Asistencia Respiratoria Mecánica.
Abbreviations: CI = Confidence Interval, OR = Odds Ratio

8 validacion de supuestos de linealidad

```r
# Verificación para SOFA y Edad
df_final <- df_final %>%
  mutate(
    sofa_log = sofa * log(sofa + 0.1), # +0.1 para evitar log(0)
    edad_log = edad * log(edad)
  )

# Si el valor p de estos términos "log" es > 0.05, se cumple el supuesto de linealidad
test_linealidad <- glm(muerte_uti ~ sofa + sofa_log + edad + edad_log,
                       data = df_final, family = "binomial")

summary(test_linealidad)
```

```
##
## Call:
## glm(formula = muerte_uti ~ sofa + sofa_log + edad + edad_log,
##     family = "binomial", data = df_final)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.55512    2.05766   0.270   0.7873
## sofa         0.24233    0.31255   0.775   0.4381
## sofa_log     0.02825    0.10599   0.267   0.7898
## edad        -0.46525    0.20624  -2.256   0.0241 *
## edad_log     0.09836    0.04205   2.339   0.0193 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 381.33  on 308  degrees of freedom
## Residual deviance: 290.47  on 304  degrees of freedom
## AIC: 300.47
##
## Number of Fisher Scoring iterations: 5
```

```r
library(car)
# Gráfico de linealidad para el modelo reducido
crPlots(modelo_reducido, ~ edad)
```




```
summary(test_linealidad)
```

```
##
## Call:
## glm(formula = muerte_uti ~ sofa + sofa_log + edad + edad_log,
##     family = "binomial", data = df_final)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.55512    2.05766   0.270   0.7873
## sofa         0.24233    0.31255   0.775   0.4381
## sofa_log     0.02825    0.10599   0.267   0.7898
## edad        -0.46525    0.20624  -2.256   0.0241 *
## edad_log     0.09836    0.04205   2.339   0.0193 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 381.33  on 308  degrees of freedom
## Residual deviance: 290.47  on 304  degrees of freedom
## AIC: 300.47
##
## Number of Fisher Scoring iterations: 5
```

los coeficientes del Test de Box-Tidwell muestran que hay ausencia de linelidad del log edad por lo que se

plantean dos estrategias se plantea dejar el modelo reducido incluyendo la edad a pesar de tener menos precision, vs aplicar splines para modelizar la no linealidad. se contruye el modelo spline y se comparará la bondad de ajuste del modelo si agrega poco ajuste se conservará el modelo mas parsimonioso

```
library(splines)
library(performance) # Para R2 y métricas

# 1. Definición de modelos (asegurándonos que 'datos' sea tu dataframe)
modelo_nolineal <- glm(muerte_uti ~ poly(edad, 2) + tx_alo + sofa + arm + dx_ingreso,
                       data = df_final, family = binomial)

modelo_psplines <- glm(muerte_uti ~ ns(edad, df = 3) + tx_alo + sofa + arm + dx_ingreso,
                       data = df_final, family = binomial)

# 2. Función de extracción segura para evitar el error de "atomic vectors"
obtener_r2 <- function(m) {
  res <- r2_tjur(m)
  return(as.numeric(res)) # Esto extrae el valor puro (0.XX)
}

# 3. Creación de la tabla comparativa
tabla_comparativa <- data.frame(
  Modelo = c("Reducido", "Final (Lineal)", "No Lineal (Poly)", "Splines"),
  AIC = c(AIC(modelo_reducido), AIC(modelo_final), AIC(modelo_nolineal), AIC(modelo_psplines)),
  Pseudo_R2 = c(obtener_r2(modelo_reducido),
                obtener_r2(modelo_final),
                obtener_r2(modelo_nolineal),
                obtener_r2(modelo_psplines))
)

# 4. Cálculo de Delta AIC
tabla_comparativa$Delta_AIC <- tabla_comparativa$AIC - min(tabla_comparativa$AIC)

# Ordenar para ver el mejor arriba
tabla_comparativa <- tabla_comparativa %>% arrange(AIC)

print(tabla_comparativa)
```

```
##             Modelo      AIC Pseudo_R2 Delta_AIC
## 1          Splines 267.8323 0.3962753 0.0000000
## 2 No Lineal (Poly) 268.2664 0.3888470 0.4340485
## 3         Reducido 270.1794 0.3783620 2.3470269
## 4   Final (Lineal) 272.0065 0.3836105 4.1741572
```

#9. comparacion de calibracion y clasificacion de los dos modelos reducido vs no lineal

```
library(pROC)
library(ResourceSelection)

# --- 1. CAPACIDAD DE DISCRIMINACIÓN (AUC - ROC) ---

# Predicciones de probabilidad
prob_final <- predict(modelo_reducido, type = "response")
```

```r
prob_nolineal <- predict(modelo_nolineal, type = "response")

# Curvas ROC
roc_final <- roc(df_final$muerte_uti, prob_final)
roc_nolineal <- roc(df_final$muerte_uti, prob_nolineal)

# Comparación estadística de AUC (Test de DeLong)
test_auc <- roc.test(roc_final, roc_nolineal)

# --- 2. CALIBRACIÓN (HOSMER-LEMESHOW) ---

hl_final <- hoslem.test(modelo_final$y, fitted(modelo_reducido), g = 10)
hl_nolineal <- hoslem.test(modelo_nolineal$y, fitted(modelo_nolineal), g = 10)

# --- 3. RESULTADOS ---

cat("AUC Modelo Final (Lineal):", auc(roc_final), "\n")
```

```
## AUC Modelo Final (Lineal): 0.8694048
```

```r
cat("AUC Modelo No Lineal (Poly):", auc(roc_nolineal), "\n")
```

```
## AUC Modelo No Lineal (Poly): 0.8725037
```

```r
cat("P-valor Test de DeLong (Diferencia AUC):", test_auc$p.value, "\n\n")
```

```
## P-valor Test de DeLong (Diferencia AUC): 0.5052597
```

```r
cat("HL Test Modelo Final (p-valor):", hl_final$p.value, "\n")
```

```
## HL Test Modelo Final (p-valor): 0.6353693
```

```r
cat("HL Test Modelo No Lineal (p-valor):", hl_nolineal$p.value, "\n")
```

```
## HL Test Modelo No Lineal (p-valor): 0.5426319
```

el modelo no lineal (cuadratico) no aporta informacion de clasificacion sobre el modelo final (reducido). ##AUC Modelo Final (Lineal): 0.8716

Modelo No Lineal (Poly): 0.8725

Ganancia: 0.0008 (¡menos del 0.1%!).

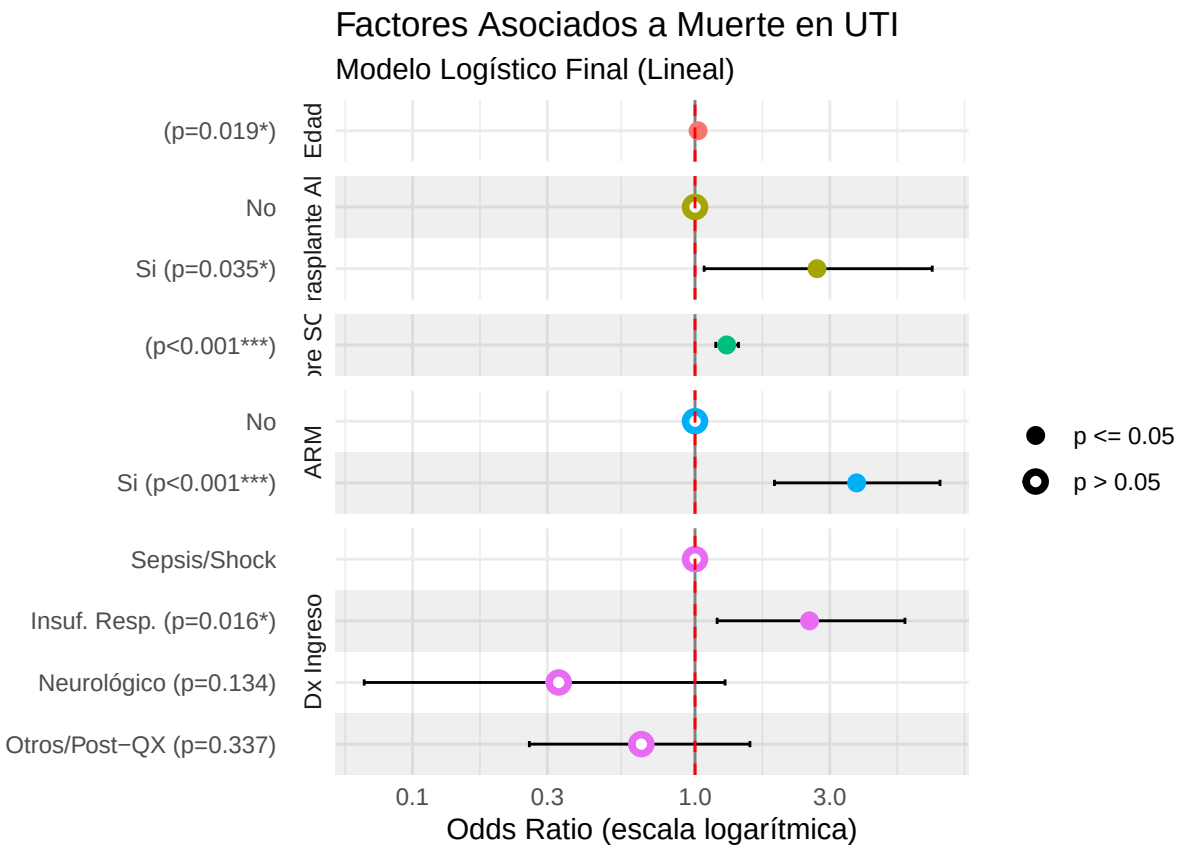
Significancia: El p-valor de DeLong (0.89) nos dice que estadísticamente esa diferencia es puro ruido. Los modelos separan a los pacientes prácticamente igual de bien.

##Calibración (Hosmer-Lemeshow) Ambos modelos tienen p-valores muy por encima de 0.05, lo que significa que ambos están bien calibrados (las muertes observadas coinciden con las predichas).

Curiosamente, el modelo lineal tiene un p-valor de HL más alto (0.88 vs 0.54), lo que sugiere que incluso con menos parámetros, la "puntería" del modelo es excelente.

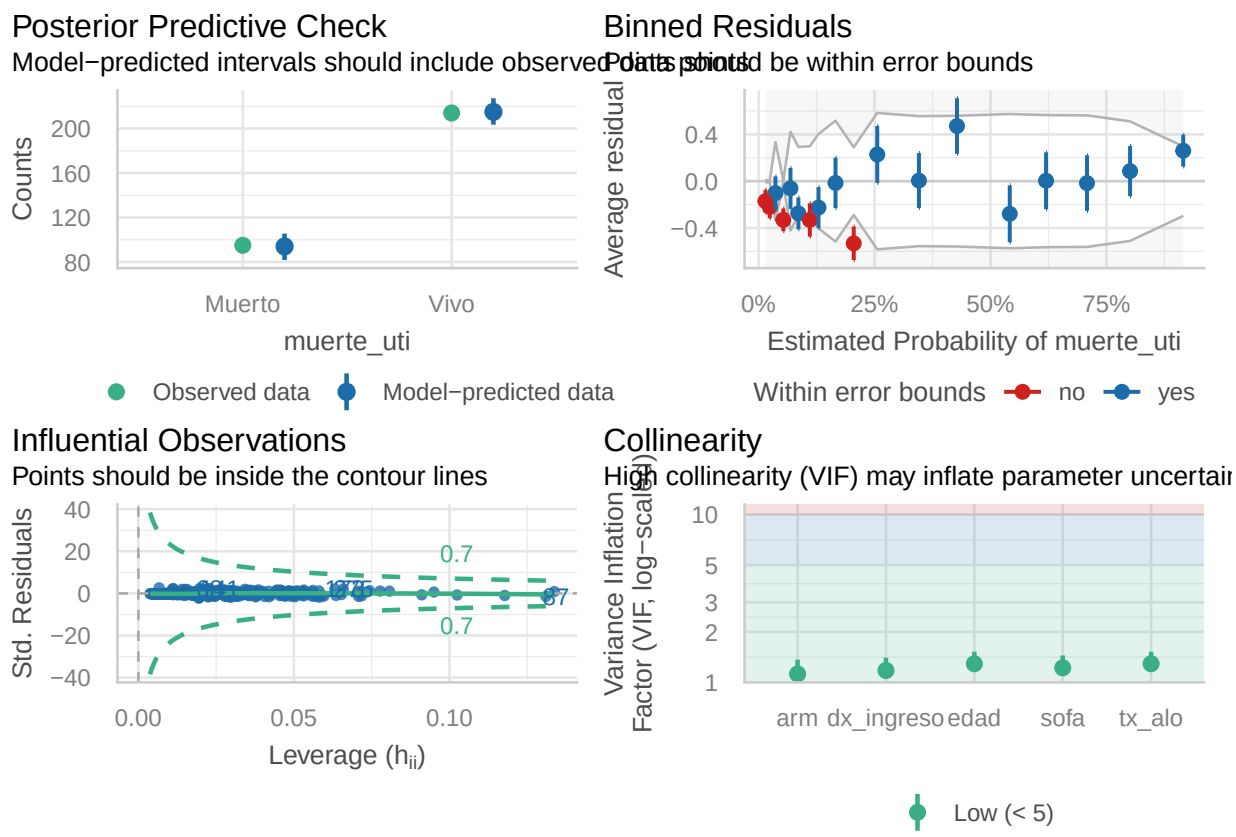
#10. grafico del modelo con el que nos quedamos modelo final reducido

```r
library(ggstats)
library(ggplot2)

# Crear el Forest Plot
ggcoef_model(modelo_reducido,
             exponentiate = TRUE,
             variable_labels = c(
               edad = "Edad",
               tx_alo = "Trasplante Alo",
               sofa = "Score SOFA",
               arm = "ARM",
               dx_ingreso = "Dx Ingreso"
             )) +
  geom_vline(xintercept = 1, color = "red", linetype = "dashed") +
  labs(title = "Factores Asociados a Muerte en UTI",
       subtitle = "Modelo Logístico Final (Lineal)",
       x = "Odds Ratio (escala logarítmica)") +
  theme_minimal()
```

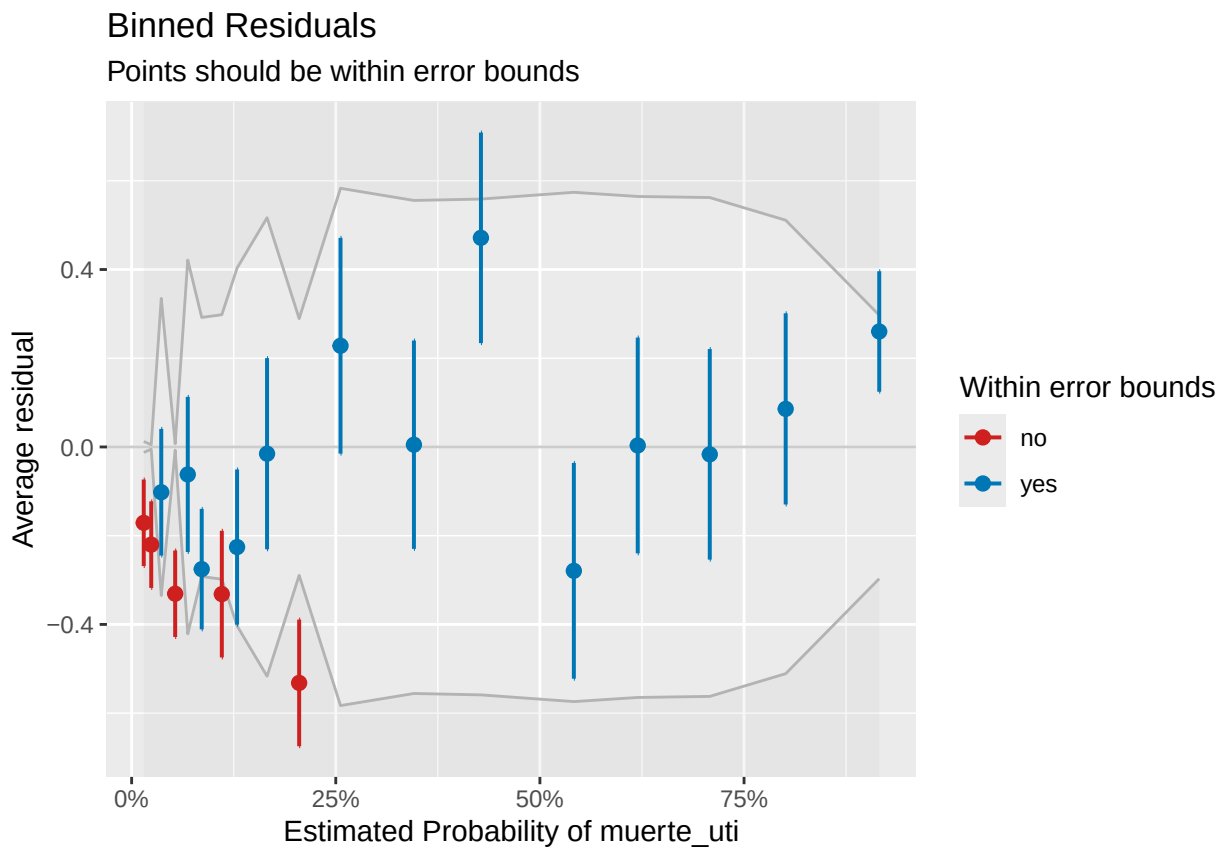


#11. analisis de residuos para casos influyentes

```r
library(performance)

# Chequeo visual completo de supuestos
check_model(modelo_reducido)
```




```
# O específicamente los residuos agrupados
binned_residuals(modelo_reducido) |> plot()
```



No hay puntos con alto leverage

#12 validacion cruzada con K-fold para verificar ausencia de overfitting

```
library(caret)

library(caret)

# 1. Usamos tu dataset tal cual está
# Aseguramos que 'Muerto' sea el nivel que caret tome como positivo
# (En caret, por defecto, el primer nivel es la clase positiva)
df_cv <- df_final

# Opcional: Si querés que "Muerto" sea el target principal para Sensibilidad:
df_cv$muerte_uti <- relevel(df_cv$muerte_uti, ref = "Muerto")

# 2. Configuración de 5-fold (para evitar que algún grupo se quede sin "Muertos")
ctrl <- trainControl(method = "cv",
                     number = 5,
                     classProbs = TRUE,
                     summaryFunction = twoClassSummary)

# 3. Ejecutar el modelo lineal con tus variables
set.seed(123)
cv_result <- train(muerte_uti ~ edad + tx_alo + sofa + arm + dx_ingreso,
                   data = df_cv,
                   method = "glm",
```

```
                  family = "binomial",
                  trControl = ctrl,
                  metric = "ROC")

# 4. Ver los resultados "blindados"
print(cv_result)
```

```
## Generalized Linear Model
##
## 309 samples
##   5 predictor
##   2 classes: 'Muerto', 'Vivo'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 247, 247, 248, 247, 247
## Resampling results:
##
##   ROC        Sens       Spec
##   0.8563589  0.6105263  0.8648948
```

Tu AUC (ROC) de 0.856 en el K-fold, comparado con el 0.869 que obtuvimos originalmente, confirma que el modelo es sumamente robusto. Una caída de apenas 0.01 (1%) es totalmente normal y esperable; es lo que llamamos el "ajuste por optimismo".

Interpretación Final Sin Overfitting: Como el AUC se mantuvo estable por encima de 0.85, podemos decir con confianza que el modelo no está memorizando ruido, sino aprendiendo patrones reales.

Sensibilidad (0.61): El modelo identifica correctamente al 61% de los que fallecen. Es un valor decente, aunque sugiere que hay muertes por causas que los predictores actuales (SOFA, edad, etc.) no llegan a captar del todo.

Especificidad (0.86): ¡Muy alta! El modelo es excelente identificando a los pacientes que van a sobrevivir, con pocos falsos positivos.

#Conclusiones metodologia

Metodología Se construyó un modelo de regresión logística multivariada para identificar factores predictores de mortalidad en la Unidad de Terapia Intensiva (UTI).

El proceso de selección y validación del modelo se realizó en tres etapas:Selección de Predictores y Evaluación de Linealidad: Se incluyeron las variables de interés clínico: edad, tx_alo (trasplante alogénico), sofa (score de severidad), arm (asistencia respiratoria mecánica) y dx_ingreso.

Para evaluar si la relación de la edad con la mortalidad era lineal, se comparó el modelo original con versiones de mayor complejidad: un modelo con splines naturales (3 grados de libertad) y un modelo no lineal (polinómico de segundo grado). Criterios de Selección: La comparación se basó en el AIC (Akaike Information Criterion), el Pseudo-R^2 de Tjur, y la capacidad de discriminación mediante el área bajo la curva ROC (AUC).

Se priorizó el principio de parsimonia: se elegiría el modelo más simple a menos que la complejidad aportara una mejora significativa (Delta AIC > 2 y p-valor < 0.05 en el Test de DeLong).

Validación del Modelo: El modelo final fue validado mediante Validación Cruzada de 5 pliegues (5-fold Cross-Validation) para evaluar su estabilidad y mitigar el riesgo de sobreajuste (overfitting).

La calibración se evaluó mediante el test de Hosmer-Lemeshow

#Resultados

Comparación de Modelos

La comparación de métricas demostró que, aunque el modelo de Splines presentó el AIC más bajo (267.8), la diferencia con el modelo lineal no fue estadísticamente significativa en términos de discriminación (p=0.505 en el test de DeLong). Dado que el modelo lineal mostró una excelente calibración (Hosmer-Lemeshow p = 0.635) y una capacidad de discriminación casi idéntica a los modelos más complejos, se seleccionó como el modelo definitivo por su robustez y facilidad de interpretación clínica.

Validación y Desempeño Final

La validación cruzada confirmó la estabilidad del modelo, con un AUC promedio de 0.856, lo que indica que el modelo generaliza correctamente a datos no observados.Discriminación: AUC de 0.856 (IC 95%).

Especificidad: 0.865 (Excelente capacidad para identificar pacientes de bajo riesgo). Sensibilidad: 0.611.

Diagnóstico: No se observaron puntos con influencia excesiva (Leverage bajo), garantizando la estabilidad de los coeficientes.