UNIVERSIDAD Finis Terrae | FACULTAD DE ECONOMÍA Y NEGOCIOS

ECO5008 Modelos predictivos

Clase 02 - Modelos de Supervivencia y Weibull aplicados a Salud

Sebastián Egaña Santibáñez

Enlaces del profesor

https://segana.netlify.app

https://github.com/sebaegana

https://www.linkedin.com/in/sebastian-egana-santibanez/

Clase 02

Conceptos iniciales

Introducción a supervivencia

Cuando hablamos de supervivencia en términos de predicciones nos referimos al esfuerzo de modelar el tiempo que transcurre hacía cierto evento. En el caso del área de salud, podemos pensar en el tiempo que transcurre hasta remisión, alta o rehospitalización. Un área intermedia de aplicación, corresponde a la de ingeniería de la confiabilidad, sirviendo para poder determinar tiempo entre fallas o entre reparaciones.

Elementos claves

- La variable relevante para ser analizada corresponde al tiempo hasta el evento, considerando $T > 0$
- Censura: corresponde a evento que por alguna u otra razón no puede ser observados de manera completa; por ejemplo, si estamos analizando el tiempo antes del alta de los pacientes existirán casos de paciente que aún se encuentran hospitalizados pero que dado el momento de selección del estudio debemos **censurar** en base al valor que posee al momento del estudio. Este tipo de observación solo **añade** a la probabilidad de que el evento ocurra posterior a lo observado y debe ser marcada para incorporarla de manera distinta al análisis.

Tipos de censura (diagrama)

Leyenda

- `-` tiempo observado
- `[` inicio
- `]` fin
- `|` corte/fin
- `*` evento observado
- `?` evento en el intervalo

```
1) Evento observado (no censurado)
   Sujeto A: [------*------]            → observamos T = t

2) Censura a derecha (right censoring)
   Sujeto B: [------------|  ]          → sabemos T > C
   (no ocurrió antes del corte)

3) Censura a izquierda (left censoring)
   Sujeto C:     *-----]                → sabemos T <= L
   (ocurrió antes del 1er registro)

4) Censura por intervalo (interval censoring)
   Sujeto D: [---?---]                  → sabemos L < T <= R
   (entre dos visitas y relacionada con inspecciones)

5) Censura administrativa (Tipo I, corte común de calendario)
   Sujeto E: [-------|]                 → típicamente "a derecha"
```

```
6) Censura Tipo II (por número de eventos)
   Cohorte: el estudio termina cuando ocurre el k-ésimo evento
```

Preguntas

¿Considere en los casos de gestión para los cuales esto podría ser valioso?

Distribución de Weibull

Se utiliza para modelos los tiempos hasta un evento cuando la tasa de eventos cambia de forma monótona (crece o decrece).

Parámetros.

k (forma): indica cómo cambia el riesgo en el tiempo.

- $k > 1$: riesgo creciente (cada vez más probable el evento).
- $k = 1$: riesgo constante (caso exponencial).
- $k < 1$: riesgo decreciente (más probable al inicio).

η (escala): fija la escala temporal (en las mismas unidades que t).

Se puede calcular con covariables (otras ariables relevantes), pero para este caso solo utilizaremos la función sin otras variables.

Fórmulas clave (Weibull)

Sea $T \sim \text{Weibull}(k, \eta)$ con $k > 0$ (forma) y $\eta > 0$ (escala), $t > 0$.

- **PDF (Weibull) :** $f(t) = \frac{k}{\eta}\left(\frac{t}{\eta}\right)^{k-1} \exp[-(t/\eta)^k]$
- **CDF:** $F(t) = 1 - \exp\ [-(t/\eta)^k]$
- **Supervivencia:** $S(t) = \exp[-(t/\eta)^k]$
- **Riesgo (hazard):** $h(t) = \dfrac{k}{\eta}\ \left(\frac{t}{\eta}\right)^{k-1}$
- **Riesgo acumulado:** $H(t) = -\ln S(t) = (t/\eta)^k$
- **Percentil p :** $t_p = \eta\,[-\ln(1-p)]^{1/k}$
- **Mediana:** $t_{0.5} = \eta\,(\ln 2)^{1/k}$
- **Media:** $\mathbb{E}[T] = \eta\,\Gamma\Big(1 + \frac{1}{k}\Big)$

En la práctica, el software (R en este caso) te devuelve k, η, la media, la mediana y los intervalos sin que tengas que calcularlos a mano.

Ejemplo simplificado

Tenamos los siguientes datos de altas pacientes con neumonía y queremos predecir el tiempo promedio de hospitalización en días:

```
tiempo <- c(2,3,1,5,7,8,9,3,4,6,2,10,11,7,5,4,12,8,9,6)
alta   <- c(1,1,1,1,1,0,0,1,1,1,1,0,0,1,1,1,0,1,1,1)
```

Recordar que acá la censura corresponde a los pacientes que no han sido dados de alta (alta = 0).

Table 1: Conteos: total, eventos (DELTA=1), censuras (DELTA=0)

Total	Eventos_DELTA1	Censuras_DELTA0	Proporcion_evento
20	15	5	0.75

Table 2: Resumen global de TCC (días)

n	min	mean	sd	median	max
20	1	6.1	3.161	6	12

Table 3: Resumen de TCC por estado DELTA (días)

DELTA	n	min	mean	sd	median	max
0	5	8	10.0	1.581	10	12
1	15	1	4.8	2.366	5	9

Veamos el análisis de Weibull para estos datos:

Table 4: Parámetros Weibull 2P por caso

Caso	shape (k)	scale ( )
Sin censura	2.2424	5.4240
Con censura	1.5848	7.8569

Table 5: Mediana, media y percentiles t_p (p=0.1, 0.5, 0.9)

Caso	Mediana	Media	p=0.1	p=0.5	p=0.9
Sin censura	4.6061	4.8041	1.9883	4.6061	7.8678
Con censura	6.2347	7.0506	1.8993	6.2347	13.2985

Actividad aplicada

Recuerden, seguimos siendo un equipo de analistas. Tenemos datos relacionados con tiempo de cirugías cardíacas en donde nuestro interés se relaciona con estimar el tiempo de la operación hasta que un paciente muere.

Estimación de Weibull

EDA

Veamos algunas estadísticas descriptivas:

Table 6: Conteos: total, eventos (DELTA=1), censuras (DELTA=0)

Total	Eventos_DELTA1	Censuras_DELTA0	Proporcion_evento
145	139	6	0.959

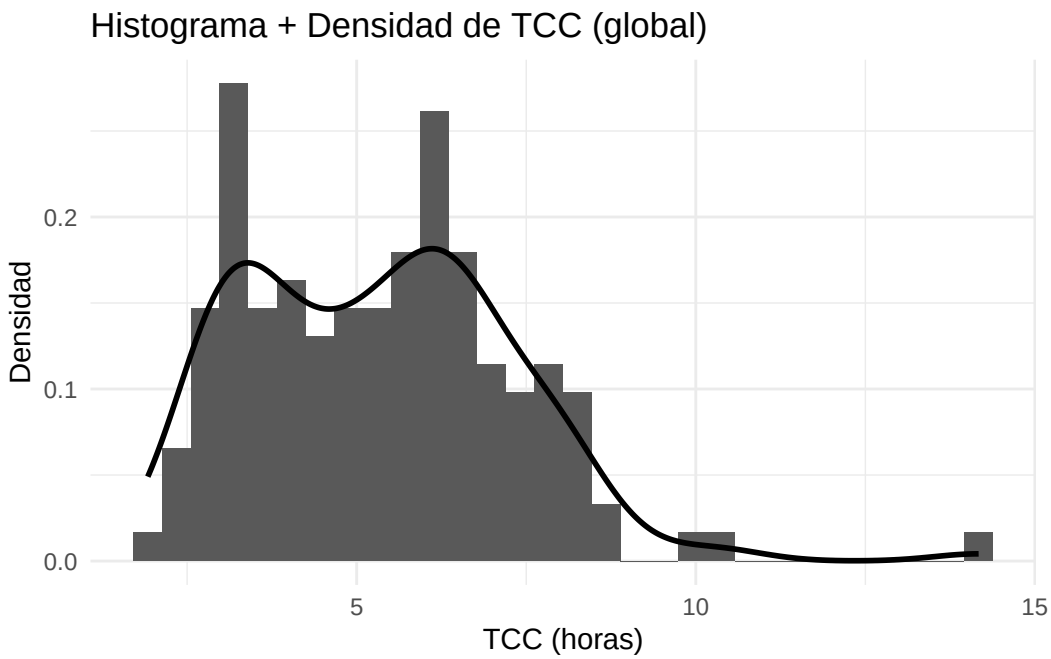
Table 7: Resumen global de TCC (horas)

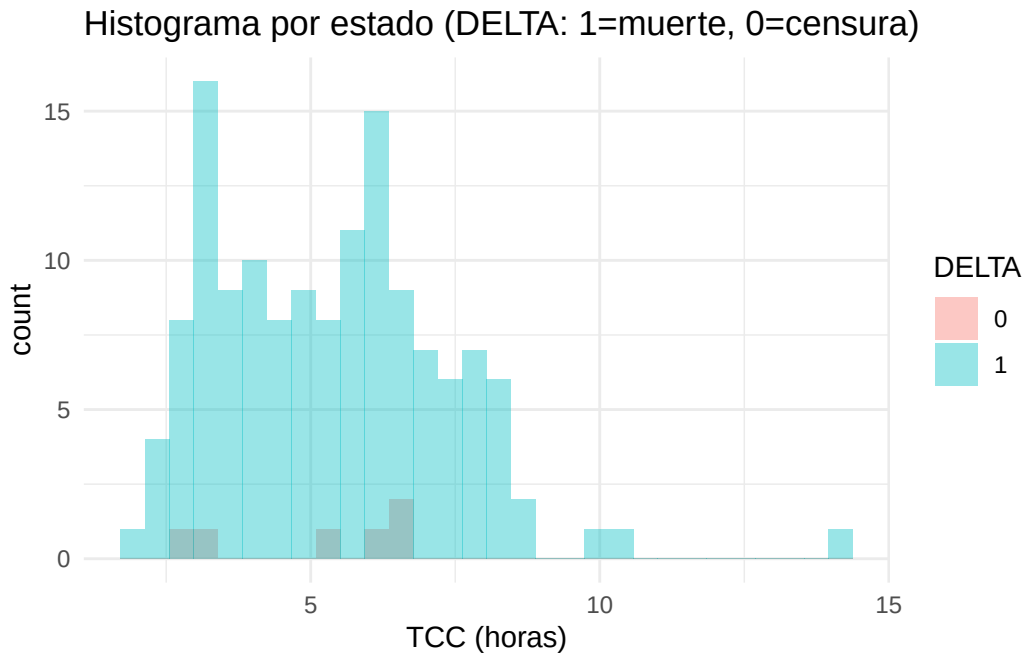
n	min	mean	sd	median	max
145	1.92	5.329	1.979	5.25	14.17

Table 8: Resumen de TCC por estado DELTA (horas)

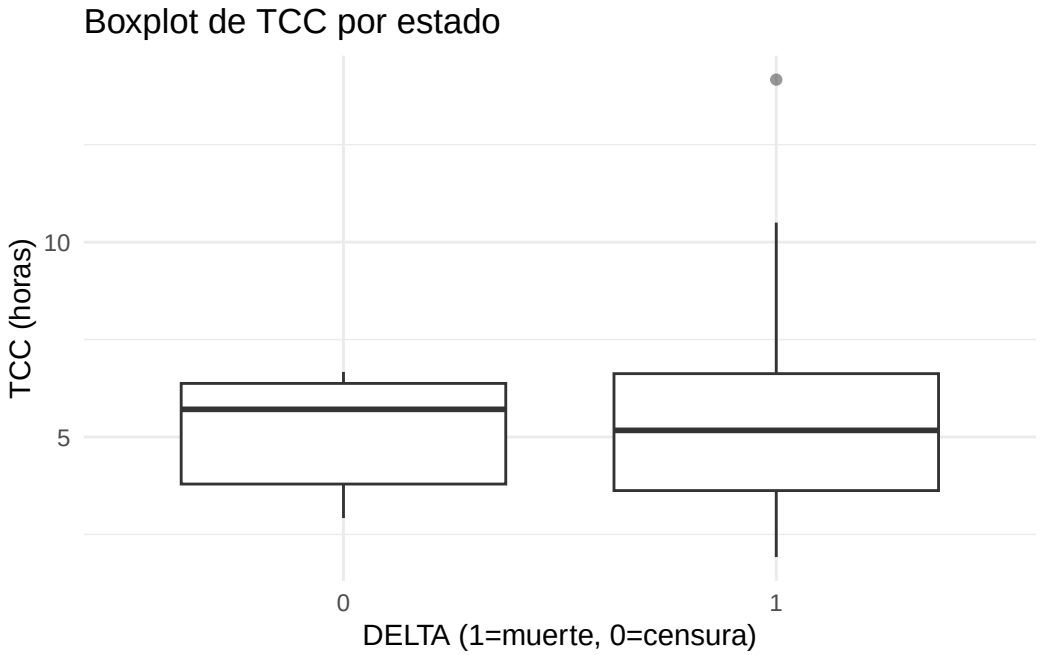
DELTA	n	min	mean	sd	median	max
0	6	2.92	5.127	1.644	5.71	6.67
1	139	1.92	5.338	1.997	5.17	14.17

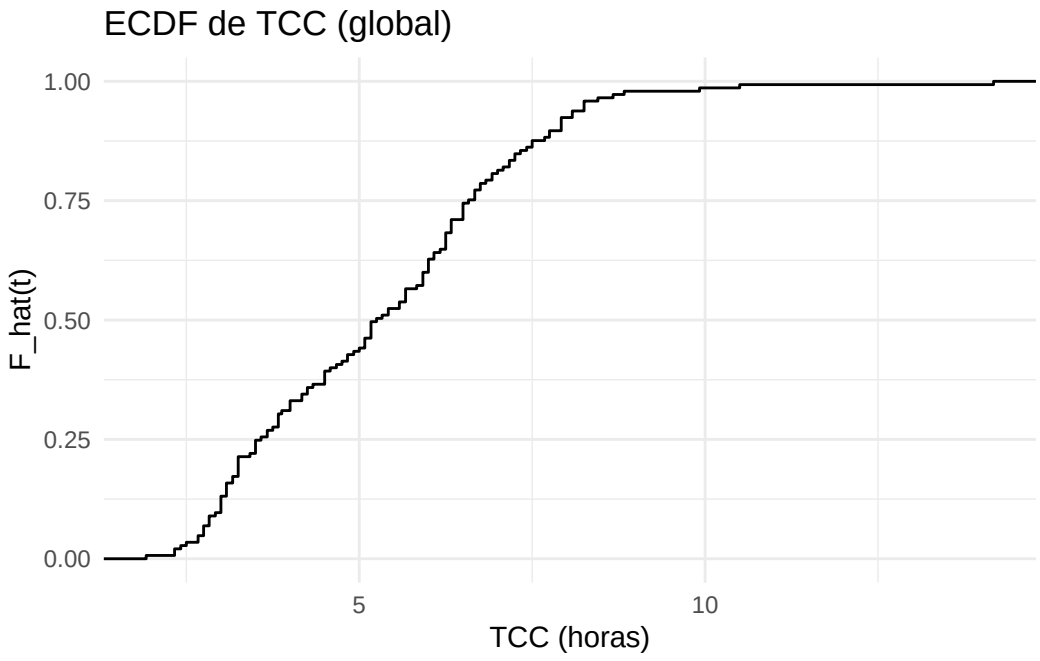
Consideremos realizar un histograma para ver la distribución de las horas de operación:





Otros gráficos relevantes:





Weibull 2P: cálculo con y sin censura

Table 9: Parámetros Weibull 2P por caso

Caso	shape (k)	scale ( )
Sin censura (DELTA==1)	2.8331	5.9906
Con censura (Surv)	2.8574	6.0694

Table 10: Funciones evaluadas en t = p10, p30, p50, p70, p90 del TCC

Caso	Funcion	t=3	t=3.84	t=5.25	t=6.33	t=7.85
Sin censura	PDF f(t)	0.115615	0.157600	0.186599	0.162547	0.090246
Sin censura	CDF F(t)	0.131477	0.246994	0.497462	0.689321	0.883794
Sin censura	Supervivencia S(t)	0.868523	0.753006	0.502538	0.310679	0.116206
Sin censura	Riesgo h(t)	0.133117	0.209294	0.371313	0.523198	0.776602
Sin censura	Riesgo acum. H(t)	0.140961	0.283682	0.688084	1.168995	2.152394
Con censura	PDF f(t)	0.111282	0.153514	0.185735	0.164822	0.094210
Con censura	CDF F(t)	0.124993	0.236878	0.483528	0.676210	0.875969
Con censura	Supervivencia S(t)	0.875007	0.763122	0.516472	0.323790	0.124031
Con censura	Riesgo h(t)	0.127179	0.201165	0.359622	0.509042	0.759569
Con censura	Riesgo acum. H(t)	0.133524	0.270337	0.660735	1.127662	2.087224

Table 11: Mediana, media y percentiles t_p (p=0.1, 0.5, 0.9)

Caso	Mediana	Media	p=0.1	p=0.5	p=0.9
Sin censura	5.2636	5.3367	2.7071	5.2636	8.0412
Con censura	5.3387	5.4087	2.7613	5.3387	8.1265

Para la interpretación de las tablas: Intuición rápida

- $S(t)$: probabilidad de seguir sin evento en el tiempo t.
- $F(t)$: probabilidad de haber tenido el evento antes del tiempo t.
- $h(t)$: tasa instantánea a la que cae $S(t)$; si $h(t)$ sube, $S(t)$ cae más rápido.
- $H(t)$: riesgo acumulado

Preguntas

- ¿Cuáles serían las aplicaciones de gestión de este análisis?

Una aplicación distinta

Se sabe que las maquinas de hemodiálisis son necesarias para poder mantener el tratamiento de los pacientes, requiriendo siempre un funcionamiento eficiente y continuo. Utilizando la distribución de Weibull podemos caracterizar la dinámica de fallas para dichas máquinas.

Tenemos los eventos de fallas para tres máquinas de hemodiálisis distintas:

Table 12: Fallas de máquinas

machine	Failure_no	Date_of_failure	TBF
M1	1	2014-01-29	1085.0
M1	2	2016-01-04	1904.0
M1	3	2016-10-19	609.0
M1	4	2016-10-24	10.5
M1	5	2017-05-29	605.5
M1	6	2020-03-06	2792.0
M1	7	2020-11-12	392.0
M1	8	NA	NA
M1	9	NA	NA
M1	10	NA	NA
M1	11	NA	NA
M2	1	2015-06-05	2450.0
M2	2	2017-05-29	2040.5
M2	3	2017-08-07	192.5
M2	4	2018-02-28	549.5
M2	5	2018-08-03	311.5
M2	6	2020-03-06	584.5
M2	7	2020-08-03	280.0
M2	8	NA	NA
M2	9	NA	NA
M2	10	NA	NA
M2	11	NA	NA
M3	1	2013-05-29	413.0
M3	2	2013-07-12	101.5
M3	3	2014-09-01	1130.5
M3	4	2015-06-05	763.0
M3	5	2015-11-09	427.0
M3	6	2017-08-07	1750.0
M3	7	2018-10-31	1228.5
M3	8	2019-08-09	780.5
M3	9	2020-03-06	560.0
M3	10	2020-07-29	255.5

machine	Failure_no	Date_of_failure	TBF
M3	11	2022-01-07	1456.0

Estimamos los parámetros de Weibull para cada maquína:

Table 13: Weibull 2P por componente (sin censura): parámetros y resúmenes

machine	shape (k)	scale ( )	Mediana	Media	p=0.1	p=0.5	p=0.9
M1	0.9318	1029.8916	694.9635	1064.0828	92.0256	694.9635	2520.746
M2	1.1167	957.6276	689.6906	919.5480	127.6425	689.6906	2020.991
M3	1.6215	899.2581	717.3276	805.2927	224.4637	717.3276	1504.082

Preguntas

- ¿Qué conclusiones podemos sacar de los números obtenidos?
- ¿Qué aplicaciones de negocio/gestión puede considerar?

Intervalo óptimo de mantención

Considerando que ya podemos caracterizar la dinámica de fallas para cada máquina, ¿podríamos hacer algún análisis para poder tener alguna estrageia que permita evitar la falla de la maquinaria? Acá debemos entender la diferencia entre falla correcta y falla preventiva:

- Falla correctiva: generar costos de reparación, pero a la vez costos relacionados con la interrupción operativa debido a la falla.
- Falla preventiva: debido a que puede ser planificada, se asume que genera solo los costos de la mantención preventiva (por ejemplo un overhaul)

No se tienen en cuenta costos de inspección, para simplificar el análisis.

Estrategia de mantención

La dinámica de mantención, cuando existe mantenimiento preventivo, considera la existencia de mantenimiento preventivo y mantenimiento correctivo que conviven de manera conjunta. Se debe entender que incluso con intervalos muy cortos de mantenciones preventivas nos enfrentamos igual a la posibilidad de eventos de mantención correctivos. Es debido a esto que la estrategia de mantención en dicho caso es una combinación de preventivo con correctivo que minimiza el costo de dicha estrategia de mantención.

El punto para optimizar la estrategia de mantención corresponde al punto donde se minimiza la siguiente función:

$$C(T) = \frac{C_{PM}}{T} + \frac{C_{CM}}{T}\lambda(T)$$

Donde

- C_{PM}: costo del mantenimiento preventivo
- C_{CM}: costo del mantenimiento correctivo
- T: intervalo de mantenimiento preventivo
- $\lambda(T)$: tasa de falla (hazard rate) al tiempo

Siendo $\lambda(T)$:

$$C(T) = \frac{k}{\eta}\left(\frac{T}{\eta}\right)^{k-1}$$

Para más información: Pardus Consulting

Table 14: Intervalo óptimo de mantenimiento preventivo (T*) por máquina con C_CM = 10×C_PM

	machine	k	eta	T_opt	Costo_h_min
T_opt	M1	0.9318	1029.8916	2059.7831	0.0194
T_opt1	M2	1.1167	957.6276	520.6800	0.0206
T_opt2	M3	1.6215	899.2581	196.1196	0.0134

Referencias

- Cavalcante, T., Ospina, R., Leiva, V., Cabezas, X., & Martin-Barreiro, C. (2023). Weibull regression and machine learning survival models: Methodology, comparison, and application to biomedical data related to cardiac surgery. Biology, 12(3), 442.
- Fenina, S., Jendoubi, S., & Bouchoucha, F. (2023). Failure rate estimation by Weibull distribution in a stochastic environment: application to the hemodialysis machine. Reliability: Theory & Applications, 18(3 (74)), 450-463.
- Nketiah, E. A. (2021). Parameter Estimation of the Weibull Distribution; Comparison of the Least-Squares Method and the Maximum Likelihood Estimation. Int. J. Adv. Eng. Res. Sci, 8, 210-224.