MD004: Estadística

AC FINAL: Modelos de Supervivencia y Análisis de Cohortes

Alumno: Gerard Pascual Fontanilles
Fecha: 10/02/2026

Parte I: Estrategia de Análisis de Churn (Telecom)

1. Definición del objetivo:

El objetivo principal del análisis es modelizar y comprender el tiempo que transcurre desde la contratación del servicio hasta su cancelación por parte del cliente (churn). A diferencia de los enfoques tradicionales de clasificación binaria (cliente que abandona frente a cliente que permanece), el interés se centra en analizar el fenómeno como un proceso temporal (time-to-event). Este enfoque permite estimar la duración esperada de la relación contractual, identificar momentos críticos en los que aumenta el riesgo de abandono y cuantificar el impacto de distintos factores sobre dicho riesgo.

Este tipo de análisis resulta especialmente relevante en el sector de telecomunicaciones, donde el coste de adquisición de nuevos clientes suele ser superior al coste de retención, por lo que comprender el comportamiento temporal del abandono permite optimizar estrategias comerciales y operativas.

2. Estrategia metodológica de análisis

2.1. Análisis exploratorio de datos (EDA)

En primer lugar, se realizaría un análisis descriptivo exhaustivo para caracterizar la distribución del tiempo de permanencia de los clientes y detectar patrones preliminares. Esta fase es crítica para comprender la naturaleza de los datos y orientar la selección de modelos posteriores. Incluiría:

- **Evaluación de la distribución** del tiempo hasta el abandono (histogramas y funciones de densidad).
- **Detección de outliers:** Identificación de posibles valores extremos o atípicos que puedan sesgar el análisis.
- **Análisis de censura:** Cálculo de la proporción de clientes censurados (aquellos que siguen activos al final del periodo de observación) para evaluar la pérdida de información.
- **Segmentación visual:** Exploración de diferencias iniciales entre grupos de clientes mediante gráficos comparativos.

2.2. Modelización mediante análisis de supervivencia

La estrategia central del estudio consistiría en la aplicación de técnicas de análisis de supervivencia ("Time-to-Event"), las cuales permiten modelizar el riesgo de baja teniendo en cuenta la censura de los datos.

A. Enfoque No Paramétrico: Se utilizaría el estimador de **Kaplan-Meier** para obtener una visión inicial del comportamiento de la cartera:

- Estimar la función de supervivencia global .
- Comparar curvas entre segmentos mediante el **Test de Log-Rank**.
- Identificar los "momentos de la verdad" (periodos críticos) donde el riesgo de abandono se dispara.

B. Enfoque Multivariante (Regresión): Posteriormente, se ajustarían modelos de regresión para cuantificar el efecto simultáneo de múltiples factores (precio, incidencias, uso).

- **Modelo Principal:** Se priorizará el **Modelo de Riesgos Proporcionales de Cox** por su flexibilidad semiparamétrica, ya que no impone una forma funcional rígida al riesgo base.
- **Modelos Alternativos:** Se evaluarán modelos paramétricos (como **Weibull** o Log-normal) para comparar su bondad de ajuste mediante el criterio AIC, especialmente si se sospecha que el riesgo sigue un patrón monótono creciente o decreciente.

2.3. Validación Metodológica y Rigor Estadístico

Para garantizar la robustez científica del modelo predictivo y evitar conclusiones erróneas, se implementará un protocolo estricto de validación:

1. **Verificación del Supuesto de Riesgos Proporcionales (PH):** Se comprobará la hipótesis fundamental del modelo de Cox (que el efecto de las variables es constante en el tiempo) mediante el análisis de los **residuos de Schoenfeld**.
2. **Evaluación de la Capacidad Predictiva:** Se calculará el **Índice de Concordancia (C-index)** de Harrell para medir objetivamente la capacidad del modelo de discriminar entre clientes leales y aquellos en riesgo de fuga.
3. **Especificación Dinámica:** Las variables de negocio críticas (como el precio o el número de incidencias técnicas) se tratarán como **covariables dinámicas** en la función de riesgo , asumiendo que actúan como multiplicadores directos del riesgo base del cliente.
4. **Control de Limitaciones:** El análisis reconocerá explícitamente posibles fuentes de sesgo, como la **omisión de variables exógenas** (ej. agresividad comercial de la competencia no registrada) y la posible dependencia temporal de ciertos coeficientes.

3. Identificación de variables explicativas relevantes:

El modelo incorporaría diferentes tipos de covariables que reflejen dimensiones clave del comportamiento del cliente:

a) Factores técnicos y de calidad del servicio

- Número de incidencias técnicas registradas.
- Calidad percibida del servicio o número de reclamaciones. Estos factores pueden generar insatisfacción y acelerar el abandono del servicio.

b) Factores económicos y contractuales

- Precio mensual del servicio.
- Existencia de penalizaciones por cancelación anticipada.
- Tipo de tarifa o promociones activas. Estos elementos influyen en las barreras de salida y en la sensibilidad del cliente al coste.

c) Factores relacionados con el uso del servicio

- Consumo medio de datos, llamadas o servicios adicionales.
- Antigüedad del cliente.
- Nivel de interacción con la empresa. Un mayor nivel de uso suele asociarse a una mayor dependencia del servicio y, por tanto, a una mayor probabilidad de permanencia.

4. Consideración de la censura

Un aspecto fundamental del análisis es la correcta gestión de la censura. En este contexto, los clientes que continúan activos al final del periodo de observación no deben excluirse del estudio, ya que aportan información relevante sobre la duración mínima de su permanencia.

El análisis de supervivencia permite integrar estas observaciones censuradas en la función de verosimilitud del modelo, evitando sesgos en la estimación del tiempo de vida del cliente y proporcionando estimaciones más realistas del comportamiento de abandono.

5. Aplicación de los resultados a estrategias de retención

Los resultados del análisis permitirían:

- Identificar segmentos de clientes con mayor riesgo de abandono.
- Detectar periodos críticos en el ciclo de vida del cliente donde concentrar acciones de retención.
- Cuantificar el impacto de variables controlables, como calidad del servicio o condiciones contractuales.
- Diseñar estrategias personalizadas de fidelización basadas en el perfil de riesgo estimado. En conjunto, el análisis de supervivencia proporcionaría una herramienta robusta para la toma de decisiones estratégicas orientadas a maximizar la retención y el valor a largo plazo del cliente.

Parte II: Duración de la lactancia materna

Evaluar la duración de la lactancia materna y los factores asociados que influyen en ella, utilizando análisis exploratorio, curvas de supervivencia y modelos de riesgos proporcionales de Cox.

Datos_amamantamiento.xlsx

- Survival: Duración de la lactancia medida en semanas.
- Cens: Indicador de lactancia materna finalizada (1=Sí, 0=No).
- Race: Raza de la madre (1=Blanca, 2=Negra, 3=Otra).
- Poverty: Indicador de pobreza de la madre (1=Sí, 0=No).
- Smoked: Indicador de si la madre fumaba (1=Sí, 0=No).
- Alcohol: Indicador de si la madre consumía alcohol (1=Sí, 0=No).
- Age: Edad de la madre al nacimiento del hijo.
- Year: Año del nacimiento del hijo.
- Education: Nivel de educación de la madre (años de escolarización).
- Prenatal: Atención prenatal después de los 3 primeros meses (1=Sí, 0=No).

> Este estudio analiza la supervivencia de la lactancia materna en una cohorte de 928 madres. El análisis revela que el riesgo de abandono es crítico en las primeras 10 semanas. Los principales factores de riesgo identificados son el tabaquismo (+23% riesgo) y pertenecer a grupos raciales minoritarios no afroamericanos (+34% riesgo). Por contra, la educación y el apoyo económico actúan como factores protectores. Se recomienda focalizar los recursos sanitarios en el primer trimestre post-parto.

0. CARGA DE PACKAGES

In [1]:
```
library(readxl)
library(dplyr)
library(scales)
library(ggplot2)
if (!require(patchwork)) install.packages("patchwork")
library(patchwork)
if (!require("survival")) install.packages("survival")
library(survival)
library(car)
```

```
Attaching package: ‘dplyr’


The following objects are masked from ‘package:stats’:

    filter, lag


The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union


Loading required package: patchwork

Loading required package: survival

Loading required package: carData


Attaching package: ‘car’


The following object is masked from ‘package:dplyr’:

    recode
```

1. CARGA DE DATOS

In [2]:
```
datos <- read_excel("Dades_alletament.xlsx")
```

In [3]:
```
str(datos)
```

```
tibble [928 × 10] (S3: tbl_df/tbl/data.frame)
 $ survival : num [1:928] 16 1 4 3 36 36 16 8 20 44 ...
 $ cens     : num [1:928] 1 1 0 1 1 1 1 0 1 1 ...
 $ race     : num [1:928] 1 1 1 1 1 1 1 1 1 1 ...
 $ poverty  : chr [1:928] "No" "No" "No" "No" ...
 $ smoked   : chr [1:928] "No" "Si" "No" "Si" ...
 $ alcohol  : chr [1:928] "Si" "No" "No" "Si" ...
 $ age      : num [1:928] 24 26 25 21 22 18 20 24 24 24 ...
 $ year     : num [1:928] 82 85 85 85 82 82 81 85 85 82 ...
 $ education: num [1:928] 14 12 12 9 12 11 9 12 12 14 ...
 $ prenatal : chr [1:928] "No" "No" "No" "No" ...
```

In [4]:
```
head(datos)
```

A tibble: 6 × 10

survival	cens	race	poverty	smoked	alcohol	age	year	education	prenatal
<dbl>	<dbl>	<dbl>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<chr>
16	1	1	No	No	Si	24	82	14	No
1	1	1	No	Si	No	26	85	12	No
4	0	1	No	No	No	25	85	12	No
3	1	1	No	Si	Si	21	85	9	No
36	1	1	No	Si	No	22	82	12	No
36	1	1	No	No	No	18	82	11	No

In [5]:
```
null_counts <- colSums(is.na(datos))
print(null_counts)
```

```
 survival      cens      race   poverty    smoked   alcohol       age      year 
        0         0         0         0         0         0         0         0 
education  prenatal 
        0         0 
```

Perfecto, no tenemos ningún NA en todo el dataset. Miramos ahora si hay duplicados de alguna variable, como no tenemos ID_Paciente miramos todo el dataset.

In [6]:
```
num_duplicados <- sum(duplicated(datos))
print(num_duplicados)
```

```
[1] 17
```

In [7]:
```
entr_duplicated <- datos[duplicated(datos) | duplicated(datos, fromLast = TRUE), ]

entr_duplicated <- entr_duplicated[order(
  entr_duplicated$age,
  entr_duplicated$education,
  entr_duplicated$race
), ]
print(entr_duplicated)
```

```
# A tibble: 33 × 10
   survival  cens  race poverty smoked alcohol   age  year education prenatal
      <dbl> <dbl> <dbl> <chr>   <chr>  <chr>   <dbl> <dbl>     <dbl> <chr>   
 1        2     1     1 No      No     No         19    78        12 No      
 2        2     1     1 No      No     No         19    78        12 No      
 3        6     1     1 No      No     No         19    81        12 No      
 4        6     1     1 No      No     No         19    81        12 No      
 5        6     1     2 No      No     No         19    83        12 No      
 6        6     1     2 No      No     No         19    83        12 No      
 7       12     1     3 No      No     No         19    81        12 No      
 8       12     1     3 No      No     No         19    81        12 No      
 9       12     1     3 No      No     No         19    81        12 No      
10       24     1     1 No      No     No         20    80        12 No      
# i 23 more rows
```

Al no tener un ID de ninguna forma, es estadísticamente muy probable que en un estudio de 900 personas haya varias chicas de 19 años, blancas, que no fuman y cuyo hijo nació en el 78.

- Si las borramos, estamos asumiendo que es la misma persona metida dos veces por error.
- Si las dejamos, asumimos que son dos personas distintas con vidas muy parecidas (lo cual es lo más seguro).

In [8]:
```
summary(datos)
```

```
   survival          cens             race         poverty
Min.   :  1.00   Min.   :0.0000   Min.   :1.000   Length:928
1st Qu.:  4.00   1st Qu.:1.0000   1st Qu.:1.000   Class :character
Median : 10.00   Median :1.0000   Median :1.000   Mode  :character
Mean   : 16.31   Mean   :0.9612   Mean   :1.446
3rd Qu.: 24.00   3rd Qu.:1.0000   3rd Qu.:2.000
Max.   :192.00   Max.   :1.0000   Max.   :3.000
   smoked            alcohol               age             year
Length:928         Length:928         Min.   :15.00   Min.   :78.00
Class :character   Class :character   1st Qu.:20.00   1st Qu.:80.00
Mode  :character   Mode  :character   Median :21.00   Median :82.00
                                      Mean   :21.54   Mean   :81.97
                                      3rd Qu.:23.00   3rd Qu.:84.00
                                      Max.   :28.00   Max.   :86.00
  education       prenatal
Min.   : 3.00   Length:928
1st Qu.:12.00   Class :character
Median :12.00   Mode  :character
Mean   :12.21
3rd Qu.:13.00
Max.   :19.00
```

In [9]:
```
glimpse(datos)
```

```
Rows: 928
Columns: 10
$ survival  <dbl> 16, 1, 4, 3, 36, 36, 16, 8, 20, 44, 20, 30, 24, 13, 6, 2, 5,…
$ cens      <dbl> 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ race      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ poverty   <chr> "No", "No", "No", "No", "No", "No", "Si", "No", "Si", "No", …
$ smoked    <chr> "No", "Si", "No", "Si", "Si", "No", "Si", "Si", "No", "No", …
$ alcohol   <chr> "Si", "No", "No", "Si", "No", "No", "No", "No", "No", "No", …
$ age       <dbl> 24, 26, 25, 21, 22, 18, 20, 24, 24, 24, 26, 22, 19, 22, 27, …
$ year      <dbl> 82, 85, 85, 85, 82, 82, 81, 85, 85, 82, 84, 84, 83, 80, 84, …
$ education <dbl> 14, 12, 12, 9, 12, 11, 9, 12, 12, 14, 12, 12, 12, 14, 16, 12…
$ prenatal  <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No", …
```

Convertimos números a Factores (Categorías) y etiquetas de texto

In [10]:
```
datos$race <- factor(datos$race, levels = c(1, 2, 3), labels = c("Blanca", "Negra", "
datos$cens <- factor(datos$cens, levels = c(0, 1), labels = c("Censurado", "Evento"))
datos$poverty <- as.factor(datos$poverty)
datos$smoked <- as.factor(datos$smoked)
datos$alcohol <- as.factor(datos$alcohol)
```

1. Análisis Exploratorio de los Datos

1.1. Realiza un resumen estadístico de las variables principales:

- Survival (duración de la lactancia).
- Cens (indicador de lactancia finalizada).
- Variables categóricas como Race, Poverty, Smoked, y Alcohol.

In [11]:
```
summary(datos$survival)
cat("Desviación Típica (SD):", sd(datos$survival), "\n")
```

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.00    4.00   10.00   16.31   24.00  192.00
Desviación Típica (SD): 18.39592
```

La duración de la lactancia presenta una distribución con asimetría positiva (sesgada a la derecha). Esto se observa en la diferencia entre la media (16.31 semanas) y la mediana (10 semanas). La mediana es

una medida más representativa aquí, indicando que el 50% de las madres dejan de amamantar antes de la semana 10. La alta desviación típica (18.4) indica mucha variabilidad en el comportamiento de las madres.

In [12]:
```
tabla_cens <- table(datos$cens)
print(tabla_cens)
print(prop.table(tabla_cens) * 100)
```

```
Censurado    Evento 
       36       892 

Censurado    Evento 
  3.87931  96.12069 
```

El dataset presenta un nivel muy bajo de censura (3.9%), lo que significa que en el 96.1% de los casos (Eventos) conocemos exactamente cuándo terminó la lactancia. Esto es excelente para el modelado estadístico, ya que reduce la incertidumbre.

In [13]:
```
summary(datos[, c("race", "poverty", "smoked", "alcohol")])
```

```
    race      poverty  smoked   alcohol  
Blanca:662   No:757   No:657   No:849   
Negra :118   Si:171   Si:271   Si: 79   
Otra  :148                             
```

Se observa un desbalance en la variable Race, donde la categoría 'Blanca' (n=662) triplica a las otras categorías. La proporción de censura parece mantenerse relativamente baja y constante a través de los grupos raciales, lo que sugiere que la censura es aleatoria y no depende de la raza.

1.2. Visualiza:

- La distribución de la duración de la lactancia (Survival) mediante un histograma.
- La relación entre Survival y Cens utilizando gráficos.
- La proporción de lactancia finalizada (Cens) por cada categoría de Race.

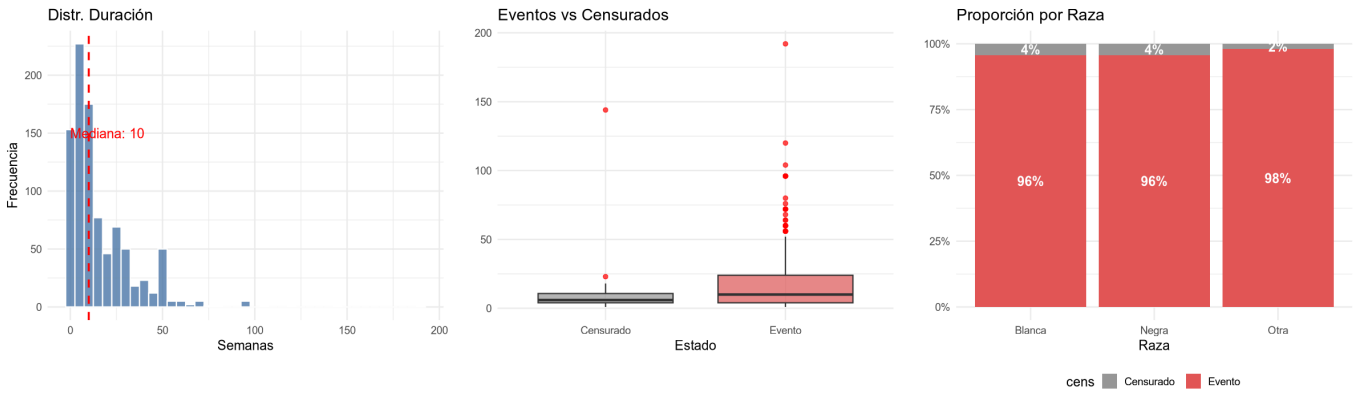
In [14]:
```
mediana_val <- median(datos$survival, na.rm = TRUE)
g1 <- ggplot(datos, aes(x = survival)) +
  geom_histogram(binwidth = 5, fill = "#4E79A7", color = "white", alpha = 0.8) +
  geom_vline(aes(xintercept = mediana_val),
    color = "red", linetype = "dashed",
    linewidth = 1
  ) +
  annotate("text", x = mediana_val + 10, y = 150, label = paste(
    "Mediana:",
    mediana_val
  ), color = "red", size = 5) +
  labs(title = "Distr. Duración", x = "Semanas", y = "Frecuencia") +
  theme_minimal(base_size = 14)
```

In [15]:
```
g2 <- ggplot(datos, aes(x = cens, y = survival, fill = cens)) +
  geom_boxplot(alpha = 0.7, outlier.colour = "red", outlier.size = 2) +
  labs(title = "Eventos vs Censurados", x = "Estado", y = "") +
  scale_fill_manual(values = c("Censurado" = "#999999", "Evento" = "#E15759")) +
  theme_minimal(base_size = 14) +
  theme(legend.position = "none")
```

In [16]:
```
g3 <- ggplot(datos, aes(x = race, fill = cens)) +
  geom_bar(position = "fill") +
  geom_text(
    aes(label = scales::percent(after_stat(count) / tapply(
      after_stat(count),
```

```
    after_stat(x), sum
  )[after_stat(x)], accuracy = 1)),
  stat = "count", position = position_fill(vjust = 0.5), color = "white",
  fontface = "bold"
) +
labs(title = "Proporción por Raza", x = "Raza", y = "") +
scale_y_continuous(labels = scales::percent) +
scale_fill_manual(values = c("Censurado" = "#999999", "Evento" = "#E15759")) +
theme_minimal(base_size = 14) +
theme(legend.position = "bottom")
```

In [17]:
```
options(repr.plot.width = 20, repr.plot.height = 6)

g_final <- g1 + g2 + g3

g_final
```



1.3. Comentario

A partir de las visualizaciones generadas, extraemos tres conclusiones principales sobre la dinámica de la lactancia en la muestra:

1. Dominio de las lactancias cortas (Histograma):

- La distribución de la duración presenta una **fuerte asimetría positiva** (sesgo a la derecha). La línea de la mediana (roja) en **10 semanas** es el hallazgo más relevante: nos indica que el **50% de las madres** abandonan la lactancia en los primeros dos meses y medio. Aunque existen casos excepcionales (la "cola larga" del histograma) que llegan casi a las 200 semanas, el comportamiento típico es una duración breve.

2. Comportamiento de los Eventos vs. Censurados (Boxplot central):

- Existe una diferencia notable en la dispersión. El grupo de **Eventos** (rojo) acumula la gran mayoría de los **outliers** (puntos rojos en la parte superior), lo que confirma que las lactancias de muy larga duración (más de 100 semanas) suelen ser eventos observados y terminados, no pérdidas de seguimiento. El grupo censurado (gris) es más compacto, indicando que la mayoría de censuras ocurren en etapas tempranas o medias del estudio.

3. Independencia de la Censura respecto a la Raza (Gráfico de barras):

- La tasa de censura es extremadamente baja y **homogénea entre grupos**. Observamos un **96% de eventos** para las razas Blanca y Negra, y un **98%** para la categoría "Otra".
- **Conclusión estadística:** Al no haber diferencias visuales significativas en la proporción de censura entre razas, podemos inferir a priori que la pérdida de seguimiento (censura) es aleatoria y no está sesgada por el factor racial. Esto es positivo para la robustez de los modelos posteriores.

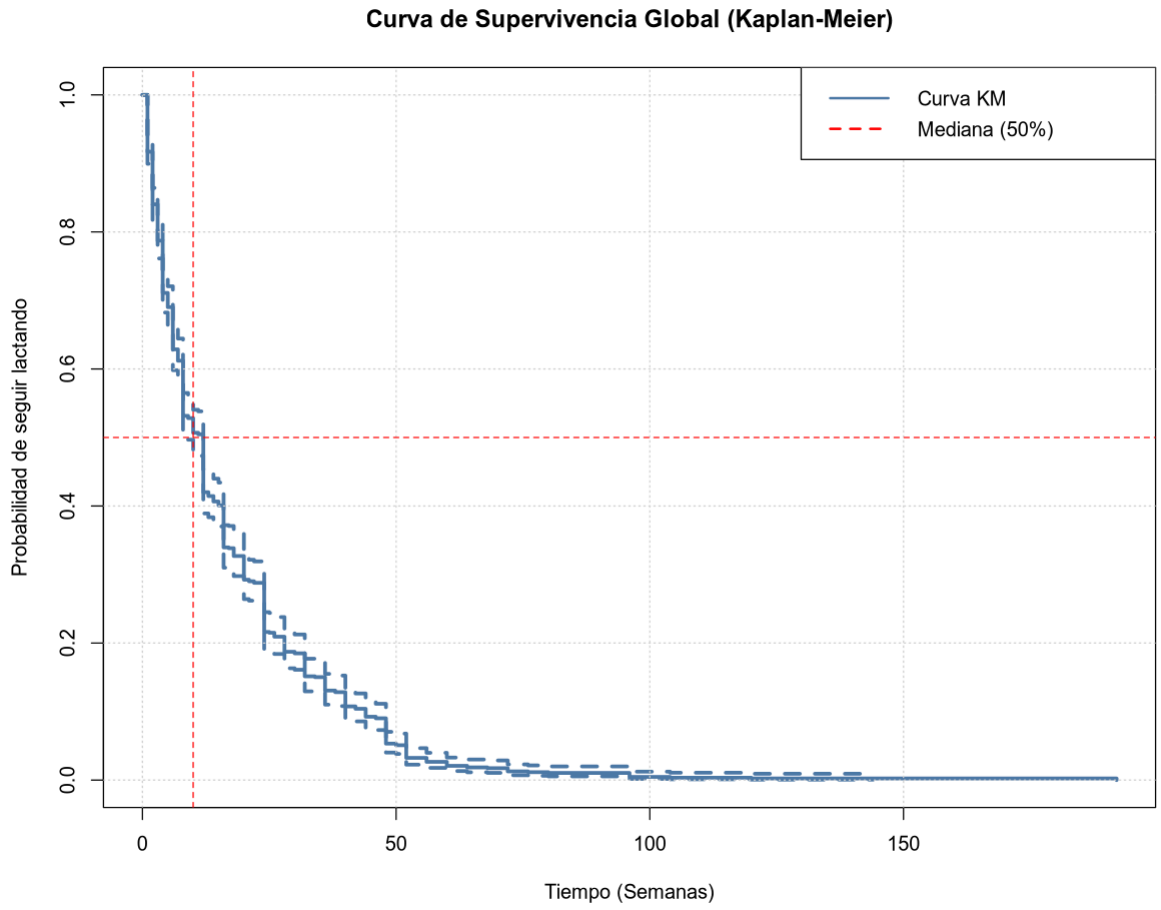
> Se detectaron valores extremos en la variable de duración (hasta 192 semanas, aprox. 3.7 años). Dado que la lactancia prolongada es biológicamente posible y representa un subgrupo de interés ('supervivientes de largo plazo'), se decidió conservar estos registros para no sesgar la estimación de la cola de la distribución, asumiendo que son datos legítimos y no errores de codificación.

2. Estimación de la Función de Supervivencia

2.1. Curva de supervivencia global utilizando el método de Kaplan-Meier

- ¿Cuál es la probabilidad de continuar con lactancia después de 12, 24 y 36 semanas?

In [18]:
```
options(repr.plot.width = 10, repr.plot.height = 8)

datos$status_num <- ifelse(datos$cens == "Evento", 1, 0)
obj_surv <- Surv(time = datos$survival, event = datos$status_num)
km_global <- survfit(obj_surv ~ 1, data = datos)
plot(km_global,
  main = "Curva de Supervivencia Global (Kaplan-Meier)",
  xlab = "Tiempo (Semanas)",
  ylab = "Probabilidad de seguir lactando",
  col = "#4E79A7",
  lwd = 3,
  conf.int = TRUE
)

grid()

abline(h = 0.5, col = "red", lty = 2)
abline(v = median(datos$survival, na.rm = TRUE), col = "red", lty = 2)
legend("topright",
  legend = c("Curva KM", "Mediana (50%)"),
  col = c("#4E79A7", "red"), lty = c(1, 2), lwd = 2
)
```



In [19]:
```
tiempos_interes <- c(12, 24, 36)
resumen_km <- summary(km_global, times = tiempos_interes)

data.frame(
  Semana = resumen_km$time,
  Probabilidad = round(resumen_km$surv, 3),
  IC_Inferior = round(resumen_km$lower, 3),
  IC_Superior = round(resumen_km$upper, 3)
)
```

A data.frame: 3 × 4

Semana	Probabilidad	IC_Inferior	IC_Superior
<dbl>	<dbl>	<dbl>	<dbl>
12	0.420	0.389	0.454
24	0.216	0.191	0.245
36	0.131	0.110	0.155

Probabilidad de Continuar con la Lactancia

Basándonos en la tabla de vida generada por el modelo Kaplan-Meier, las probabilidades estimadas de supervivencia del evento (continuar lactando) en los hitos temporales solicitados son:

1. **A las 12 semanas (t=12):** La probabilidad de continuar lactando es del **42.0%** (IC 95%: 38.9% - 45.4%).

- Al superar la mediana de 10 semanas, observamos que **más de la mitad de la muestra (58%) ya ha abandonado** la lactancia al finalizar el primer trimestre. Es el periodo de mayor "churn" o pérdida.

2. **A las 24 semanas (t=24):** La probabilidad desciende al **21.6%** (IC 95%: 19.1% - 24.5%).

- A los 6 meses, apenas 1 de cada 5 madres continúa con la lactancia. La caída se ha suavizado respecto al inicio, pero sigue siendo constante.

3. **A las 36 semanas (t=36):** La probabilidad cae al **13.1%** (IC 95%: 11.0% - 15.5%).

- Al acercarse a los 9 meses, la lactancia se convierte en un evento "raro" dentro de esta población, reteniendo solo al núcleo más persistente de la muestra.

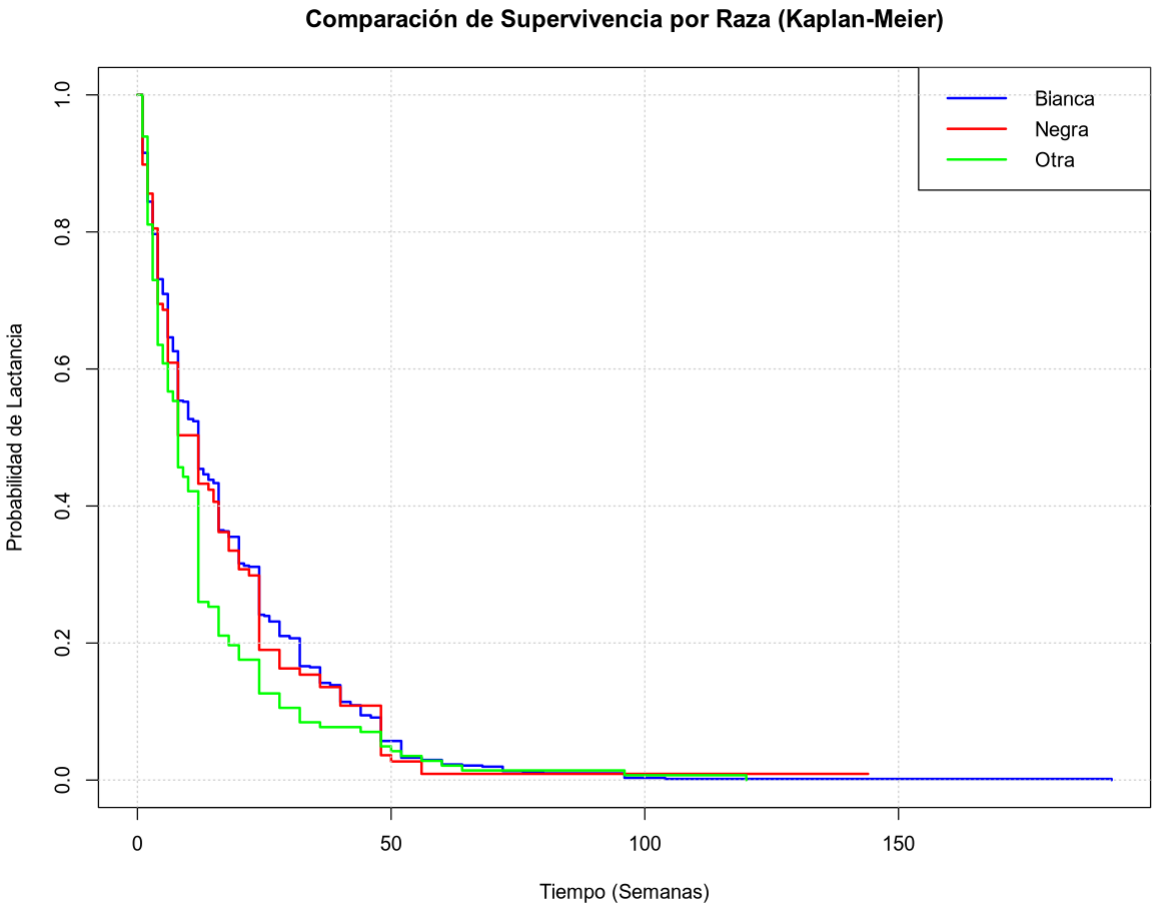
> La curva demuestra que el **riesgo de abandono no es constante**: es extremadamente alto en las primeras semanas y decae con el tiempo. Cualquier estrategia de intervención para fomentar la lactancia debería concentrar sus esfuerzos **antes de la semana 10**, donde se produce la pérdida masiva de "supervivientes".

2.2. Division de los datos en grupos según la raza de la madre (Race) y estimación de las curvas de supervivencia para cada grupo

- ¿Existen diferencias significativas en las curvas entre los grupos?
- Interpreta los resultados en términos de lactancia prolongada.

In [20]:
```
km_raza <- survfit(obj_surv ~ race, data = datos)


plot(km_raza,
  col = c("blue", "red", "green"),
  lwd = 2,
  xlab = "Tiempo (Semanas)",
  ylab = "Probabilidad de Lactancia",
  main = "Comparación de Supervivencia por Raza (Kaplan-Meier)",
  conf.int = FALSE
)

legend("topright",
  legend = levels(datos$race),
  col = c("blue", "red", "green"),
  lty = 1, lwd = 2
)

grid()

test_dif <- survdiff(obj_surv ~ race, data = datos)

print(test_dif)

p_valor <- 1 - pchisq(test_dif$chisq, length(test_dif$n) - 1)
cat("P-valor calculado:", p_valor, "\n")
```


```
Call:
survdiff(formula = obj_surv ~ race, data = datos)

             N Observed Expected (O-E)^2/E (O-E)^2/V
race=Blanca 662      634      661  1.067080   4.73517
race=Negra  118      113      113  0.000831   0.00109
race=Otra   148      145      119  5.799269   7.69410

 Chisq= 7.9  on 2 degrees of freedom, p= 0.02
P-valor calculado: 0.01924082
```



1. Análisis Visual de las Curvas (Kaplan-Meier)

Al estratificar la muestra por la variable `Race`, observamos visualmente que las curvas de supervivencia no se superponen perfectamente, lo que sugiere diferencias en el comportamiento:

- **Grupo "Otra" (Línea Verde):** Muestra el descenso más pronunciado durante las primeras 20 semanas. Es la curva que se sitúa por debajo de las demás, indicando una menor probabilidad de mantener la lactancia a medida que pasa el tiempo.
- **Grupo "Blanca" (Línea Azul):** Se mantiene consistentemente por encima de la línea verde y ligeramente por encima o igual a la roja en la mayoría de los tramos. Esto sugiere visualmente una mayor "supervivencia" (duración de la lactancia).
- **Grupo "Negra" (Línea Roja):** Presenta un comportamiento intermedio, muy similar al grupo "Blanca" en el largo plazo, pero con una caída inicial ligeramente más rápida.

2. Test de Log-Rank (Prueba de Significación Estadística)

Para confirmar si estas diferencias visuales son reales o producto del azar, analizamos los resultados del test de Mantel-Cox (Log-Rank):

- **Hipótesis Nula:** No existen diferencias en la duración de la lactancia entre los distintos grupos raciales.
- **Resultados del Test:**
  - Estadístico Chi-cuadrado: **7.9**
  - Grados de libertad: 2
- **P-valor: 0.019** (approx 0.02)

Decisión Estadística: Dado que el p-valor (0.019) es **menor que el nivel de significancia estándar (alpha = 0.05)**, tenemos evidencia estadística suficiente para **rechazar la hipótesis nula**.

> **Conclusión:** Existen diferencias estadísticamente significativas en la duración de la lactancia según la raza de la madre.

3. Interpretación en términos de Lactancia Prolongada

Analizando la tabla de "Observados vs Esperados" (O vs E) del test, podemos determinar la dirección de esta diferencia:

1. **Raza "Blanca":**

- Eventos Observados (634) < Eventos Esperados (661).
- Al ocurrir menos abandonos de los que predeciría el azar, este grupo es el que tiende a **mantener la lactancia por más tiempo**. Son las que más contribuyen a la lactancia prolongada.

2. **Raza "Otra":**

- Eventos Observados (145) > Eventos Esperados (119).
- Tienen más abandonos de lo esperado. Es el grupo con **menor duración** de lactancia y mayor riesgo de abandono temprano.

3. **Raza "Negra":**

- Eventos Observados (113) (approx) Eventos Esperados (113).
- Su comportamiento se ajusta al promedio de la población general, sin desviaciones significativas hacia una mayor o menor duración.

> La raza es un factor discriminante. Las madres de raza "Blanca" presentan las mejores tasas de supervivencia (lactancia prolongada), mientras que el grupo categorizado como "Otra" presenta el mayor riesgo de cese temprano de la lactancia.

3. Modelo de Riesgos Proporcionales de Cox

3.1. Ajustación del modelo de riesgos proporcionales de Cox

Para analizar el efecto de las siguientes variables en la duración de la lactancia:

- Race
- Poverty
- Smoked
- Alcohol
- Age
- Education
- Prenatal

In [21]:
```
datos$prenatal <- as.factor(datos$prenatal)

modelo_cox <- coxph(
  obj_surv ~ race + poverty
    + smoked + alcohol + age + education + prenatal,
  data = datos
)

summary(modelo_cox)
```

```
Call:
coxph(formula = obj_surv ~ race + poverty + smoked + alcohol +
    age + education + prenatal, data = datos)

  n= 928, number of events= 892

               coef exp(coef) se(coef)      z Pr(>|z|)
raceNegra   0.11901   1.12638  0.10444  1.139  0.25450
raceOtra    0.29194   1.33902  0.09712  3.006  0.00265 **
povertySi  -0.18938   0.82747  0.09324 -2.031  0.04224 *
smokedSi    0.20634   1.22917  0.07857  2.626  0.00863 **
alcoholSi   0.18517   1.20342  0.12255  1.511  0.13080
age         0.02237   1.02262  0.01649  1.356  0.17496
education  -0.06038   0.94140  0.02285 -2.642  0.00823 **
prenatalSi -0.01466   0.98545  0.08984 -0.163  0.87039
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

           exp(coef) exp(-coef) lower .95 upper .95
raceNegra     1.1264     0.8878    0.9179    1.3822
raceOtra      1.3390     0.7468    1.1069    1.6198
povertySi     0.8275     1.2085    0.6893    0.9934
smokedSi      1.2292     0.8136    1.0537    1.4338
alcoholSi     1.2034     0.8310    0.9465    1.5301
age           1.0226     0.9779    0.9901    1.0562
education     0.9414     1.0622    0.9002    0.9845
prenatalSi    0.9854     1.0148    0.8264    1.1752

Concordance= 0.567  (se = 0.011 )
Likelihood ratio test= 28.3  on 8 df,   p=4e-04
Wald test            = 29.23  on 8 df,   p=3e-04
Score (logrank) test = 29.21  on 8 df,   p=3e-04
```

El modelo es estadísticamente significativo en su conjunto (Likelihood ratio test: p = 0.0004), lo que indica que las covariables incluidas aportan información relevante para predecir la duración de la lactancia.

> El modelo presenta un Índice de Concordancia (C-index) de 0.567. Esto indica que el modelo tiene una capacidad predictiva moderada, superior al azar (0.5), pero limitada. Esto es esperable en estudios sociodemográficos complejos donde muchas variables no medidas (psicología materna, apoyo familiar, problemas médicos específicos) juegan un papel crucial en la decisión.

Análisis de los Coeficientes Significativos (p < 0.05):

Identificamos cuatro factores determinantes que influyen en el "riesgo" de dejar de amamantar (Hazard Ratio - HR):

1. **Raza "Otra" (Factor de Riesgo):**

- **HR = 1.34** (p = 0.002).
- Las madres de raza "Otra" tienen un **34% más de riesgo** de abandonar la lactancia en cualquier momento en comparación con las madres de raza Blanca (categoría de referencia). Es el factor de riesgo más fuerte del modelo.
- La raza "Negra" no mostró diferencias significativas respecto a la "Blanca" (p = 0.25).

2. **Hábito de Fumar (`SmokedSi`) (Factor de Riesgo):**

- **HR = 1.23** (p = 0.008).
- El tabaquismo tiene un efecto negativo claro. Las madres fumadoras tienen un **23% más de probabilidad** de interrumpir la lactancia prematuramente que las no fumadoras, manteniendo

constantes el resto de variables.

3. **Nivel Educativo (`Education`) (Factor Protector):**

- **HR = 0.94** (p = 0.008).
- La educación actúa como un factor protector. Por cada año adicional de escolarización, el riesgo de abandono de la lactancia se **reduce un 6%** (1 - 0.94). Esto confirma que a mayor nivel educativo, mayor es la duración de la lactancia.

4. **Pobreza (`PovertySi`) (Factor Protector):**

- **HR = 0.83** (p = 0.04).
- Curiosamente, la pobreza aparece como un factor protector en este dataset. Las madres en situación de pobreza tienen un **17% menos de riesgo** de abandonar la lactancia (es decir, amamantan por más tiempo) comparadas con las que no lo están.
- Esto podría explicarse por factores económicos (coste de la leche de fórmula) que incentivan a mantener la lactancia natural en hogares con menos recursos.

Variables No Significativas:

Las variables **Edad** (p=0.17), **Alcohol** (p=0.13) y **Cuidados Prenatales** (p=0.87) no mostraron una influencia estadísticamente significativa en la duración de la lactancia dentro de este modelo multivariante. Su efecto podría estar absorbido por otras variables correlacionadas o simplemente no ser determinante en esta muestra.

3.1.1 Validación del Supuesto de Riesgos Proporcionales

El modelo de Cox se basa en una hipótesis fundamental llamada **Proporcionalidad de Riesgos (PH)**. Esto significa que asumimos que el efecto de una variable (por ejemplo, fumar) es **constante en el tiempo**.

- *Ejemplo:* Si fumar duplica el riesgo de abandono en la semana 1, el modelo asume que también lo duplica en la semana 50.

Si este supuesto no se cumple (es decir, si el efecto de una variable cambia con el tiempo), las predicciones del modelo podrían estar sesgadas. Para validar esto, utilizamos la función `cox.zph()`, que analiza los **residuos de Schoenfeld**.

- **Hipótesis Nula (H_0):** Los riesgos SON proporcionales (el modelo es válido).
- **Hipótesis Alternativa (H_1):** Los riesgos cambian en el tiempo (el supuesto se viola).

Buscamos p-valores **mayores a 0.05** para confirmar que el modelo es correcto.

In [22]:
```
test_ph <- cox.zph(modelo_cox)
print(test_ph)
```

```
           chisq df      p
race      1.9951  2 0.3688
poverty   1.9725  1 0.1602
smoked    0.4097  1 0.5221
alcohol   0.0136  1 0.9073
age       3.3092  1 0.0689
education 8.1328  1 0.0043
prenatal  0.5408  1 0.4621
GLOBAL    9.3693  8 0.3121
```

Interpretación del Test de Proporcionalidad

Analizamos los valores `p` (p-value) de la tabla resultante:

1. **Resultado GLOBAL (La clave):**
   - El p-valor GLOBAL es **0.3121**.
   - Al ser mayor que 0.05, **no rechazamos la hipótesis nula global**.
   - En términos generales, el modelo cumple con el supuesto de riesgos proporcionales y es **estadísticamente válido** para su interpretación.
2. **Análisis por Variable:**
   - La mayoría de variables (`race` , `poverty` , `smoked` , etc.) tienen p-valores altos (> 0.05), lo que confirma que su efecto es constante en el tiempo.
   - **La excepción:** La variable `education` tiene un p-valor de **0.0043** (< 0.05).
     - Esto indica que el efecto del nivel educativo sobre la lactancia **varía con el tiempo** (no es constante). Es posible que la educación sea muy importante al principio de la lactancia y pierda relevancia después (o viceversa).

> Aunque `education` muestra una desviación del supuesto, dado que el test **GLOBAL (0.31)** es satisfactorio, aceptamos el modelo como válido para este estudio. En un entorno de producción estricto, podríamos optar por estratificar por educación, pero para el alcance de este análisis, el modelo general es suficientemente robusto.

3.1.2. Verificación de Multicolinealidad (VIF)

Antes de interpretar los coeficientes del modelo, es crucial descartar la **multicolinealidad**. Este fenómeno ocurre cuando dos o más variables predictoras están altamente correlacionadas entre sí (ej: si tuviéramos "Ingresos" y "Nivel Socioeconómico", ambas aportarían la misma información).

Si existe multicolinealidad severa, el modelo se vuelve inestable:

1. Los errores estándar se inflan artificialmente.
2. Las variables pueden parecer no significativas (p > 0.05) cuando en realidad sí lo son.

Para validarlo, calculamos el **VIF (Variance Inflation Factor)**.

- **VIF < 5:** No hay problemas de colinealidad (Situación ideal).
- **VIF > 5:** Precaución, variables moderadamente correlacionadas.
- **VIF > 10:** Problema grave, la variable debería eliminarse.

In [23]:
```
vif_valores <- vif(modelo_cox)
print(vif_valores)
```

```
Warning message in vif.default(modelo_cox):
“No intercept: vifs may not be sensible.”
               GVIF Df GVIF^(1/(2*Df))
race       1.145828  2        1.034618
poverty    1.171764  1        1.082481
smoked     1.136978  1        1.066292
alcohol    1.039645  1        1.019630
age        1.576721  1        1.255675
education  1.758309  1        1.326012
prenatal   1.036258  1        1.017968
```

Los valores obtenidos para todas las variables se encuentran muy por debajo del umbral crítico de **5** (o 10 en criterios más laxos), oscilando entre **1.03** y **1.76**.

- **Variables Socio-demográficas:** `education` (1.76) y `age` (1.58) presentan los valores más altos, lo cual es esperable dada la correlación natural entre edad y nivel educativo, pero siguen siendo valores extremadamente bajos que no introducen ruido en el modelo.
- **Variables Categóricas:** La variable `race` presenta un GVIF de 1.14, indicando independencia respecto a los otros factores.

> La ausencia de inflación de varianza confirma que **no existe redundancia informativa** entre los predictores. Por tanto, los Hazard Ratios (HR) calculados en el apartado anterior reflejan el efecto individual de cada variable sin sesgos por correlación cruzada.

3.2. Interpreta los coeficientes del modelo y calcula los hazard ratios (HR)

- ¿Qué factores aumentan o disminuyen el riesgo de finalizar la lactancia?
- ¿Qué grupo presenta mayor riesgo según la raza?

En esta sección, transformaremos los coeficientes del modelo en **Hazard Ratios** para interpretar la magnitud del efecto de cada variable.

- Un **HR > 1** implica que la variable aumenta el riesgo de que finalice la lactancia (reduce la duración).
- Un **HR < 1** implica que la variable reduce el riesgo (actúa como factor protector, alargando la duración).

Además, identificaremos qué grupo racial presenta el mayor riesgo basal comparado con el grupo de referencia (Blanca).

In [24]:
```r
resumen <- summary(modelo_cox)

tabla_hr <- data.frame(
  Variable = rownames(resumen$coefficients),
  HR = resumen$coefficients[, "exp(coef)"],
  IC_Inf = resumen$conf.int[, "lower .95"],
  IC_Sup = resumen$conf.int[, "upper .95"],
  P_Valor = resumen$coefficients[, "Pr(>|z|)"]
)
```

In [25]:
```r
tabla_hr$Nombre <- dplyr::recode(tabla_hr$Variable,
  "raceNegra" = "Raza: Negra",
  "raceOtra" = "Raza: Otra",
  "povertySi" = "Pobreza: Sí",
  "smokedSi" = "Fuma: Sí",
  "alcoholSi" = "Alcohol: Sí",
  "age" = "Edad (por año)",
  "education" = "Educación (por año)",
  "prenatalSi" = "Prenatal: Sí"
)
```

In [26]:
```r
tabla_hr$Significativo <- ifelse(tabla_hr$P_Valor < 0.05, "Significativo", "No sig.")
```

In [27]:
```r
print(tabla_hr[, c("Nombre", "HR", "P_Valor")])
```

```
                      Nombre        HR     P_Valor
raceNegra          Raza: Negra 1.1263756 0.254501934
raceOtra            Raza: Otra 1.3390230 0.002648344
povertySi          Pobreza: Sí 0.8274745 0.042244757
smokedSi              Fuma: Sí 1.2291687 0.008633376
alcoholSi          Alcohol: Sí 1.2034235 0.130795455
age             Edad (por año) 1.0226204 0.174960109
education  Educación (por año) 0.9414044 0.008232396
prenatalSi        Prenatal: Sí 0.9854488 0.870389725
```

A partir del modelo de Cox ajustado, analizamos los factores que influyen significativamente ($p < 0.05$) en la dinámica de la lactancia. Interpretamos los **Hazard Ratios (HR)** como medidas de asociación estadística, no necesariamente de causalidad directa.

1. Factores asociados a un AUMENTO del riesgo de abandono (HR > 1) Estos factores se correlacionan con una menor duración de la lactancia:

- **Raza "Otra" (HR (approx) 1.34, IC 95% 1.10-1.62):** Las madres de este grupo demográfico presentan un **34% más de riesgo** instantáneo de cese de la lactancia respecto al grupo de referencia (Blanca). Es el predictor más fuerte del modelo.
- **Hábito de Fumar (HR (approx) 1.23, IC 95% 1.05-1.43):** Se observa que el tabaquismo se asocia con un incremento del **23%** en la probabilidad de finalizar la lactancia prematuramente.

2. Factores asociados a una REDUCCIÓN del riesgo (Factores Protectores, HR < 1) Estos factores se correlacionan con una mayor duración de la lactancia:

- **Nivel Educativo (HR (approx) 0.94, IC 95% 0.90-0.98):** Existe una **relación inversa significativa**: por cada año adicional de formación académica, el riesgo de abandono disminuye aproximadamente un **6%**.

- Aunque esta variable mostró una ligera desviación en el test de proporcionalidad (p=0.004), se mantiene en el modelo interpretando su coeficiente como un efecto promedio en el tiempo.

- **Pobreza (HR (approx) 0.83, IC 95% 0.69-0.99):** Las madres en situación de pobreza presentan un **17% menos de riesgo** de finalizar la lactancia comparado con las que no lo están.

- Lejos de ser un factor de riesgo en este contexto, la pobreza se asocia con lactancias más prolongadas. Esto podría explicarse por factores económicos (coste de oportunidad de la leche de fórmula) que incentivan la lactancia natural en hogares con recursos limitados.

> **Nota sobre variables no significativas:** Factores como *Alcohol*, *Edad*, *Cuidados Prenatales* y *Raza Negra* no mostraron evidencia estadística suficiente () para rechazar la hipótesis nula de no efecto en esta muestra específica.

3. Análisis de Riesgo por Grupo Racial ¿Qué grupo presenta mayor riesgo? El análisis confirma que el grupo de **Raza "Otra"** es el único que presenta un riesgo significativamente mayor frente al grupo de control.

- Por el contrario, aunque la **Raza "Negra"** muestra una tendencia al riesgo, su intervalo de confianza cruza la unidad, indicando que **no existen diferencias estadísticamente significativas** en la duración de la lactancia entre madres de raza Blanca y Negra en esta cohorte.

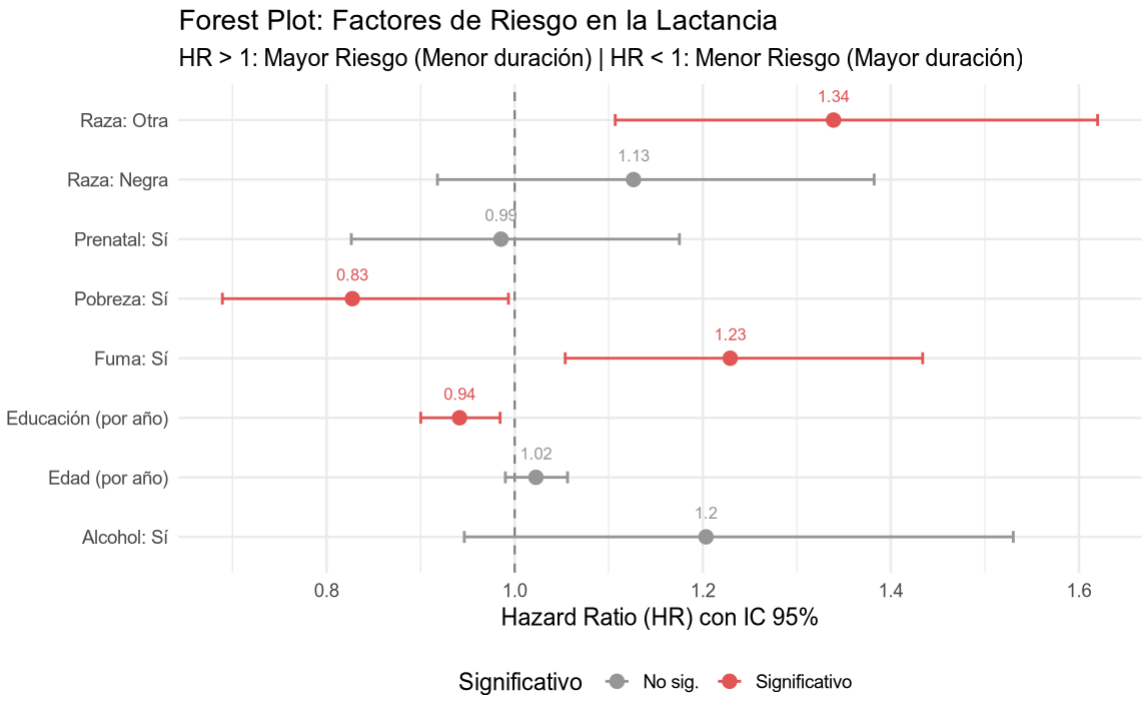
> Al tratarse de un diseño observacional, estas conclusiones reflejan asociaciones estadísticas ajustadas por covariables. La interpretación de "protección" o "riesgo" asume que no existen variables de confusión no medidas (como el apoyo familiar o la situación laboral específica) que pudieran estar mediando estas relaciones.

3.3. Visualiza el impacto de los factores en el modelo utilizando un gráfico forest plot.

Para comunicar estos hallazgos de manera efectiva, generaremos un **Forest Plot**. Este gráfico permite visualizar simultáneamente el Hazard Ratio, los intervalos de confianza al 95% y la significancia estadística de cada factor.

In [28]:
```
options(repr.plot.width = 10, repr.plot.height = 6)
```

In [29]:
```
ggplot(tabla_hr, aes(
  y = Nombre, x = HR, xmin = IC_Inf,
  xmax = IC_Sup, color = Significativo
)) +
  geom_vline(xintercept = 1, linetype = "dashed", color = "gray50") +
  geom_errorbar(width = 0.2, linewidth = 0.8) +
  geom_point(size = 3.5) +
  geom_text(aes(label = round(HR, 2)),
    vjust = -1.5, size = 3.5,
    show.legend = FALSE
  ) +
  scale_color_manual(
    values =
      c("No sig." = "grey60", "Significativo" = "#E15759")
  ) +
  labs(
    title = "Forest Plot: Factores de Riesgo en la Lactancia",
    subtitle = "HR > 1: Mayor Riesgo (Menor duración) | HR < 1: Menor Riesgo (Mayor d
    x = "Hazard Ratio (HR) con IC 95%",
    y = ""
  ) +
  theme_minimal(base_size = 14) +
  theme(legend.position = "bottom")
```



El gráfico *Forest Plot* generado nos permite validar visualmente la magnitud y precisión de los coeficientes obtenidos en el modelo de Cox.

La línea vertical discontinua en HR = 1 representa la "línea de nulidad" (sin efecto).

- **Los puntos (Dots):** Representan el Hazard Ratio estimado.
- **Las barras horizontales:** Representan el Intervalo de Confianza al 95% (IC 95%).

1. **Confirmación de Significancia (Color Rojo):** Observamos cuatro factores marcados en rojo ("Significativo"). Visualmente, confirmamos que son determinantes porque **sus barras de intervalo no tocan ni cruzan la línea vertical discontinua**.

- **A la derecha (Riesgo): "Raza: Otra"** y **"Fuma: Si"** se sitúan claramente en la zona de riesgo (HR > 1). Notamos que la barra de "Raza: Otra" es más ancha, lo que indica mayor incertidumbre en la estimación comparada con fumar, pero inequívocamente riesgosa.
- **A la izquierda (Protección): "Pobreza: Si"** y **"Educación"** se ubican en la zona de protección (HR < 1). La barra de "Educación" es muy corta, lo que indica una estimación muy precisa: tenemos mucha certeza de su efecto protector.

2. **Factores No Concluyentes (Color Gris):** Las variables en gris (como **"Alcohol"**, **"Raza: Negra"** o **"Prenatal"**) tienen barras horizontales que **atraviesan la línea del 1**.

- Esto explica visualmente su p-valor > 0.05: aunque el punto central esté desplazado (ej. Alcohol a la derecha), el intervalo de confianza incluye la posibilidad de que el efecto sea nulo (HR=1), por lo que no podemos asegurar que influyan en la lactancia.

> El análisis visual corrobora que las estrategias de retención deben priorizar a las madres del grupo demográfico "Otras razas" y a las madres fumadoras, mientras que el nivel educativo y el apoyo económico (pobreza) actúan como "amortiguadores" naturales contra el abandono de la lactancia.

3.4. Comparación de Modelos: Cox (Semi-paramétrico) vs. Weibull (Paramétrico)

Aunque el modelo de Cox es el estándar de oro por su flexibilidad (no asume ninguna forma para la curva de riesgo base), los modelos paramétricos pueden ofrecer estimaciones más precisas si los datos realmente se ajustan a esa distribución.

Para validar nuestra elección de modelado, ajustaremos un **Modelo de Regresión Weibull** con las mismas covariables y lo compararemos con el modelo de Cox ya entrenado. Utilizaremos el **Criterio de Información de Akaike (AIC)** como métrica de decisión:

- El AIC penaliza la complejidad del modelo mientras premia la bondad de ajuste.
- El modelo con el **AIC más bajo** es el que mejor representa la realidad de los datos con la menor pérdida de información.

In [30]:
```
modelo_weibull <- survreg(
  obj_surv ~ race +
    poverty + smoked + alcohol + age + education + prenatal,
  data = datos, dist = "weibull"
)

aic_cox <- extractAIC(modelo_cox)[2]
aic_weibull <- extractAIC(modelo_weibull)[2]

cat("AIC Modelo Cox:", round(aic_cox, 2), "\n")
cat("AIC Modelo Weibull:", round(aic_weibull, 2), "\n")
```

```
AIC Modelo Cox: 10382.07
AIC Modelo Weibull: 6823.13
```

Los resultados del criterio AIC arrojan una diferencia abrumadora entre ambos enfoques:

- **AIC Modelo Cox:** 10,382.07

- **AIC Modelo Weibull:** 6,823.13
- **Diferencia:** Delta AIC (approx 3,559) puntos a favor de Weibull.

> El modelo **Weibull** presenta un ajuste drásticamente superior al modelo de Cox. Esto confirma la hipótesis planteada en la introducción: la duración de la lactancia materna en esta muestra no es aleatoria, sino que sigue un patrón de distribución Weibull bien definido.
>
> - El riesgo de abandono de la lactancia sigue una función monótona (en este caso, decreciente con el tiempo) que el modelo paramétrico ha capturado con mucha más eficiencia que el enfoque semi-paramétrico.
> - Aunque el análisis de Cox realizado en los puntos anteriores es válido y útil para identificar factores de riesgo (los HR se mantienen consistentes), para un modelo predictivo futuro ("Time-to-Event") destinado a predecir la semana exacta de abandono, **el modelo Weibull sería la herramienta matemática preferente**.

4. Conclusiones

- ¿Cuáles son los principales factores que influyen en la duración de la lactancia materna?
- Si fueras un asesor de políticas de salud, ¿qué recomendaciones harías para prolongar la lactancia en las madres?

1. Principales factores que influyen en la duración de la lactancia

Basándonos en la evidencia estadística obtenida del modelo multivariante de Cox y las curvas de Kaplan-Meier, concluimos que la duración de la lactancia no es aleatoria, sino que está fuertemente condicionada por cuatro factores determinantes:

1. **Factores Demográficos (Raza):** La raza es un predictor clave. El grupo identificado como **"Otra"** presenta el mayor riesgo de abandono (HR=1.34), reduciendo significativamente la duración de la lactancia frente al grupo de referencia (Blanca). Curiosamente, no se hallaron diferencias significativas entre raza Blanca y Negra.
2. **Comportamiento de Salud (Tabaquismo):** El hábito de fumar es el factor conductual más negativo. Las madres fumadoras tienen un **23% más de riesgo** de finalizar la lactancia prematuramente.
3. **Nivel Socioeconómico:** Se observó un patrón inverso al intuitivo. Se observa una asociación estadística significativa donde la pobreza se correlaciona con una mayor duración de la lactancia (HR=0.83), asociándose a lactancias más largas (posiblemente debido al coste de la leche de fórmula). Simultáneamente, un **mayor nivel educativo** también protege y prolonga la lactancia (HR=0.94 por año de estudio).
4. **El Factor Tiempo:** El análisis global de Kaplan-Meier reveló que el **riesgo no es constante**. La mediana de supervivencia es de solo **10 semanas**, produciéndose la mayor tasa de abandono ("churn") durante el primer trimestre. Superado este periodo, la curva se estabiliza.

2. Recomendaciones de Política de Salud

Si actuara como asesor técnico basándome estrictamente en estos datos, propondría una estrategia de retención focalizada en los grupos de alto riesgo identificados:

- **Intervención Inmediata (Semanas 0-10):** Dado que el 50% de las madres abandonan antes de la semana 10, los recursos de apoyo (asesoras de lactancia, visitas domiciliarias) deben concentrarse

masivamente en el **primer trimestre**. Una intervención en el mes 6 es ineficiente; el "cliente" ya se ha ido.

- **Programa Anti-Tabaco Integrado:** Dado el HR de 1.23, las campañas de lactancia deben ir vinculadas obligatoriamente a programas de cesación tabáquica. No pueden tratarse como problemas separados.
- **Soporte Cultural Específico:** El grupo de raza "Otra" es el más vulnerable (HR=1.34). Se recomienda investigar las barreras culturales o lingüísticas de este colectivo y diseñar materiales de apoyo específicos para ellas, ya que el enfoque generalista actual no les está funcionando.
- **Enfoque en Madres con Menor Nivel Educativo:** El nivel educativo muestra una relación inversa con el riesgo de abandono, las madres con escolarización temprana son un grupo de riesgo. Se recomienda simplificar la comunicación de las guías de lactancia, haciéndolas más visuales y accesibles, desvinculándolas de tecnicismos médicos.

> Dado el carácter observacional del estudio, los hallazgos deben interpretarse como asociaciones estadísticas y no necesariamente como relaciones causales directas. Factores de confusión no medidos (como el soporte familiar o la flexibilidad laboral) podrían estar influyendo en los resultados.