Realizado por: Santiago Robatto, Sofía Terra, Diego Da Rosa y Nahuel Bizoso.

Proyecto final de curso:

Análisis de modelos bayesianos jerárquicos para regresión logística

Basado en el ejercicio 18.10 del libro Bayes Rules!

Nuestro conjunto de datos

- Dataset: **climber_sub**.
- Contiene información de distintas **expediciones al Himalaya.**
- Cuenta con **24 variables** y **2076 observaciones.**
- Algunas variables refieren a los individuos o expediciones y otras son sobre la montaña.

- La variable de interés es el **éxito** o no del individuo (success), es decir que queremos predecir una **variable binaria**.
- Por ello utilizaremos un **modelo logístico.**
- Un éxito se define cuando el escalador llega a la cima de la montaña. Esto es, que la altura a la que llegó el escalador sea igual a la altura de la montaña. (**highpoint_metres = height_metres**)

Nuestras variables (1)

Variable	Tipo_dato	Descripción
expedition_id	numeric	Identificador de la expedición
member_id	numeric	Identificador del miembro
peak_id	character	Identificador del pico
peak_name	character	Nombre del pico
year	numeric	Año de la expedición
season	character	Estación del año en que se realizó la expedición
sex	character	Sexo del escalador
age	numeric	Edad del escalador
citizenship	character	Origen del escalador
expedition_role	character	Rol del escalador en la expedición
hired	numeric	Indica si fue contratado como personal de apoyo
highpoint_metres	numeric	Altura máxima alcanzada (en metros)

Nuestras variables (2)

success	numeric	Exito (Se da cuando la altura alcanzada es igual a la del pico)
solo	numeric	Indica si realizó el ascenso en solitario
oxygen_used	numeric	Indica si usó oxígeno suplementario
died	numeric	Indica si falleció
death_cause	character	Causa de muerte (solo si fallecio)
death_height_metres	numeric	Altura a la que se produjo la muerte (solo si fallecio)
injured	numeric	Indica si resultó herido
injury_type	character	Tipo de herida (solo si resulto herido)
injury_height_metres	numeric	Altura a la que se produjo la herida (solo si resulto herida)
count	numeric	Cantidad de explodores en el grupo
height_metres	numeric	Altura del pico (en metros)
first_ascent_year	numeric	Año del primer ascenso registrado al pico

Resumen de variables numéricas

Variable	Min	Q1	Median	Mean	Q3	Max
year	1978	1996	2006	2002.87331	2011	2019
age	17	29	36	36.95713	43	76
highpoint_metres	4900	6750	7400	7554.86446	8850	8850
death_height_metres	3400	6150	6600	6771.05263	7800	8650
injury_height_metres	3960	5575	6325	6406.42857	6950	8400
count	5	8	11	13.88439	18	44
height_metres	5929	7138	8167	7989.03805	8850	8850
first_ascent_year	1936	1953	1954	1959.22378	1961	2017

Resumen de variables categóricas

Variable	TRUEs	Total	Porcentaje_TRUE
hired	485	2076	23.36
success	807	2076	38.87
solo	2	2076	0.10
oxygen_used	601	2076	28.95
died	19	2076	0.92
injured	24	2076	1.16

Análisis Descriptivo (1)

- Tenemos el **conjunto de montañas del Himalayas (peaks)**
- Las montañas cuentan con **gran variabilidad de altura entre sí (height_metres), lo que genera una variabilidad considerable en la altura máxima que alcanzan los escaladores (highpoint_metres).**
- La **edad** de los participantes también muestra un **rango amplio, pero la mayor parte** de los escaladores se concentra en **edades adultas, entre 29 y 43 años.**
- Otra variable interesante es el **año de la expedición**, que **abarca intentos desde 1978 a 2019, con media en 2002.** Contamos con una **variable que indica el año de la primer expedición (first_ascent_year).**

Análisis Descriptivo (2)

En el caso de las variables binarias, nos enfocamos en la proporción de valores TRUE y FALSE:

- **success** nos muestra que **807 realmente llegaron a la cima, sobre 2076 escaladores. Esto es, el 39% de las observaciones fueron exitosas.**

- **oxygen_used** nos informa que **507 sobre 2076 escaladores usaron oxígeno.** Miraremos mas adelante si hay **diferencias** en el éxito **entre quienes usaron oxígeno y quienes no.**

- **death e injured** tienen **muy pocos valores TRUE, lo que genera que las variables que dependen de ellas tengan valores NA.** Por ende podremos **descartar** estas variables del análisis (altura de muerte, tipo de lesión, etc).

- **solo**, que modela si el individuo fue solo, tiene **únicamente dos valores TRUE**, por lo que la **descartamos.**

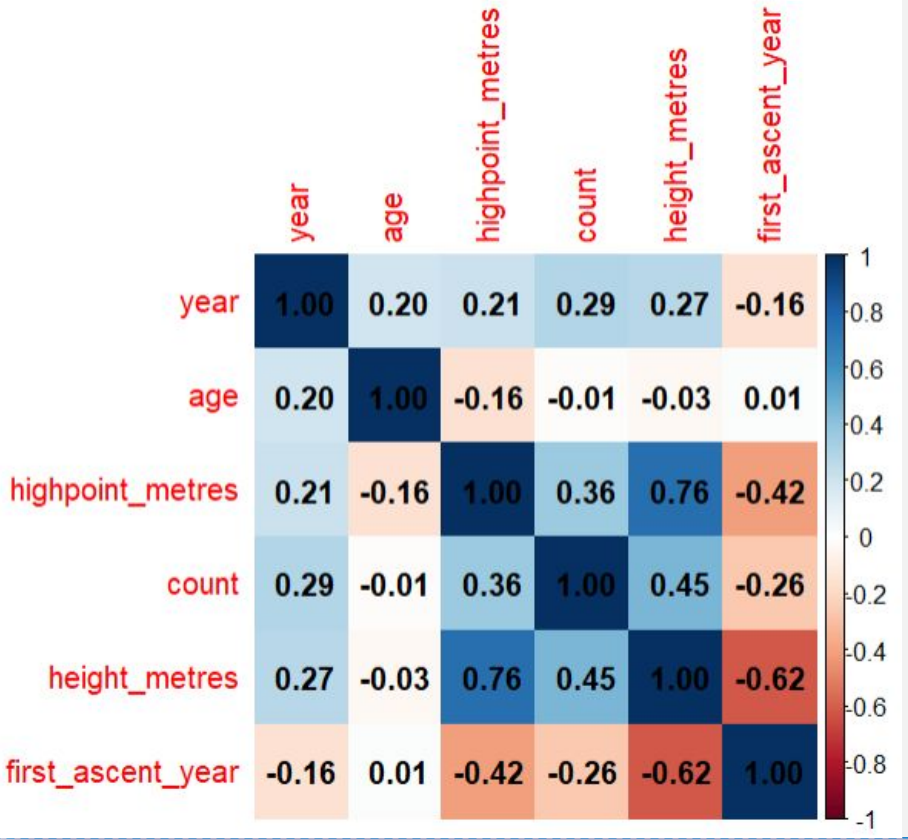
Análisis Descriptivo (3)

Resumen: para tener un dataframe más limpio, ¿cuáles variables descartamos de nuestro análisis?

- **peak_id**(ya que trabajaremos con el nombre de cada pico, estaríamos duplicando información).
- **solo**
- **death_cause**
- **death_height_metres**
- **injury_type**
- **Injury_height_metres**

El nuevo data frame cuenta con 18 variables.

Correlaciones entre variables numéricas:



Matriz de correlación entre los predictores cuantitativos usados para explicar a la variable de interés "success".

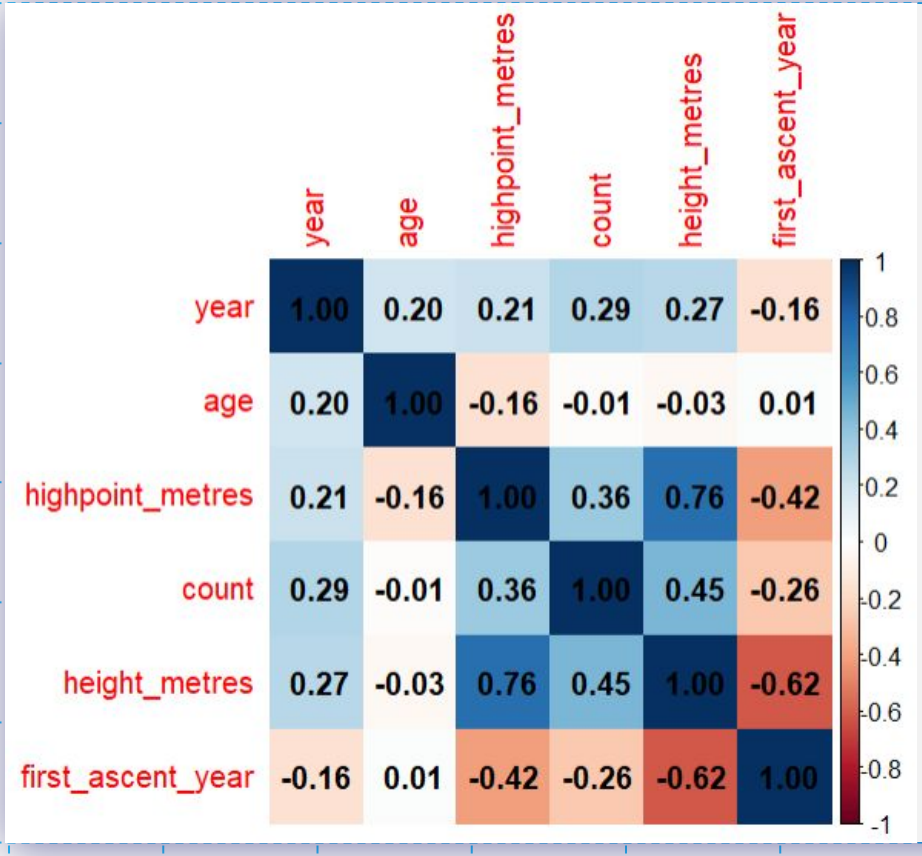
Coeficiente de correlación entre pares de variables cuantitativas. Varia entre 1 y -1

Azul oscuro: correlación positiva elevada. A medida que un predictor aumenta, también lo hace el otro.

Rojo: correlación negativa. A medida que un predictor aumenta, el otro disminuye.

Blanco: no existe correlación.

Análisis de la matriz de correlación (1):

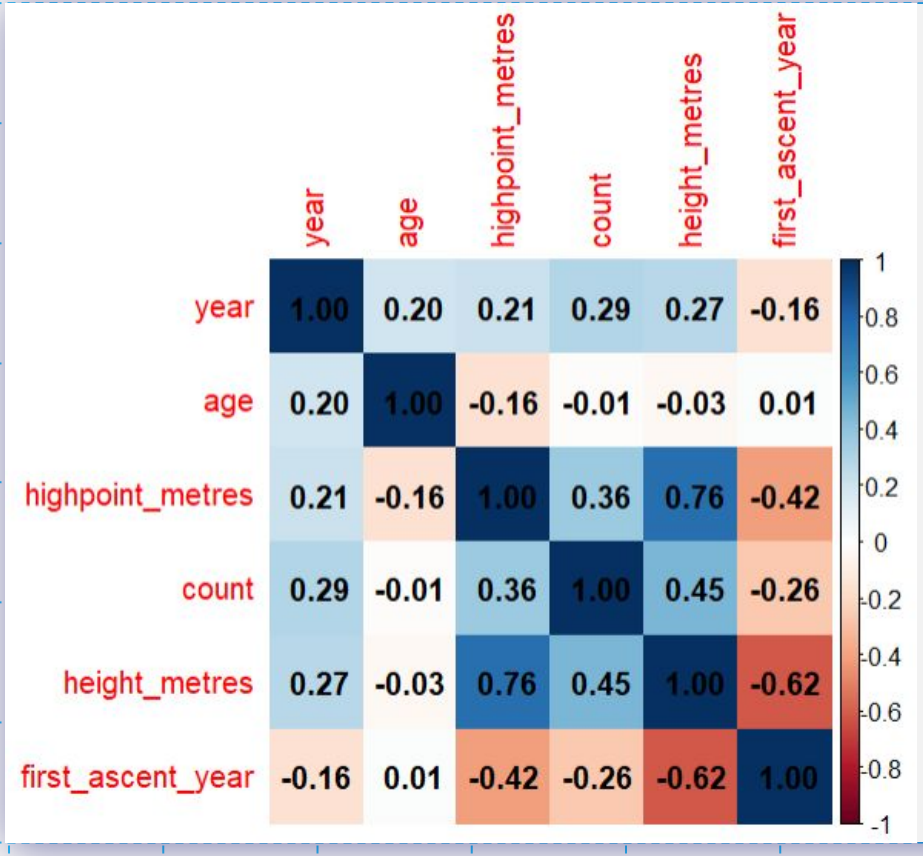


- **height_metres y highpoint_metres:** correlación lineal de 0.76. Es razonable que los escaladores tiendan a alcanzar puntos más altos cuando escalan montañas más altas.

- **height_metres y count** (número de escaladores por expedición), las cuales tienen una correlación lineal de 0.45. Las montañas más altas pueden estar asociadas con expediciones más grandes.

-**highpoint_metres y count** las cuales tienen una correlación lineal de 0.36. Indica que los escaladores que alcanzan puntos más altos tienden a ir en expediciones más grandes.

Análisis de la matriz de correlación (2):



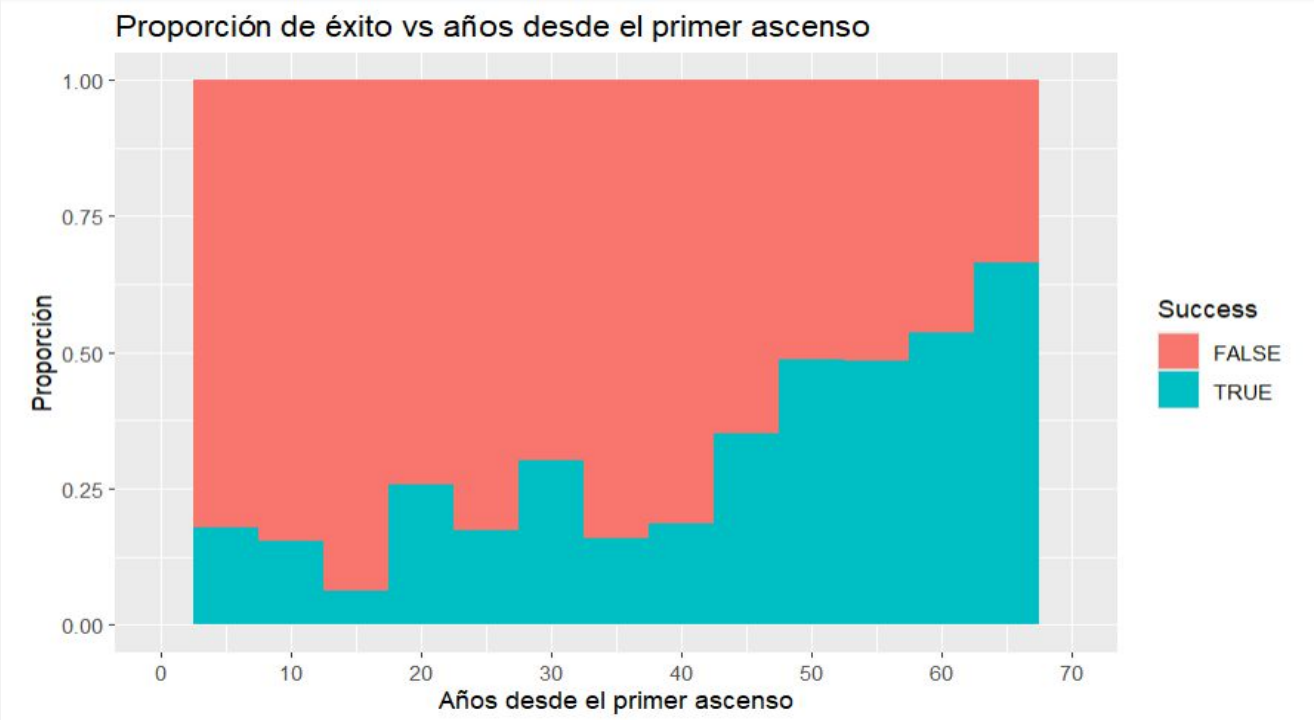
first_year_ascent tiene gran correlación con height_metres y highpoint_metres.

Parece razonable pensar que desde que se realizaron las primeras expediciones, hubo una mejoría en cuanto a la mayor altura alcanzada, además que parece que se han escalado montañas más altas.

Exploremos esto creando una **nueva variable** que **calcula la diferencia entre year y el primer año registrado de escalada.**

Years_since_first_ascent = Year - first_ascent_yeay

Nueva variable: years_since_first_ascent (1)



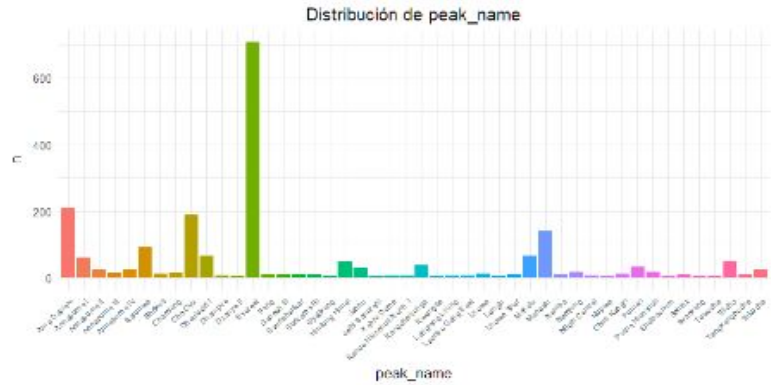
Nueva variable: years_since_first_ascent (2)

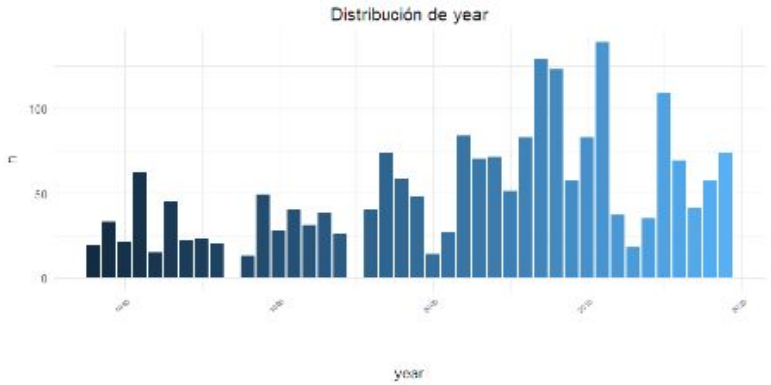
```r
climbers <- climbers %>%
  mutate(year_since_first_ascent = year - first_ascent_year)
```

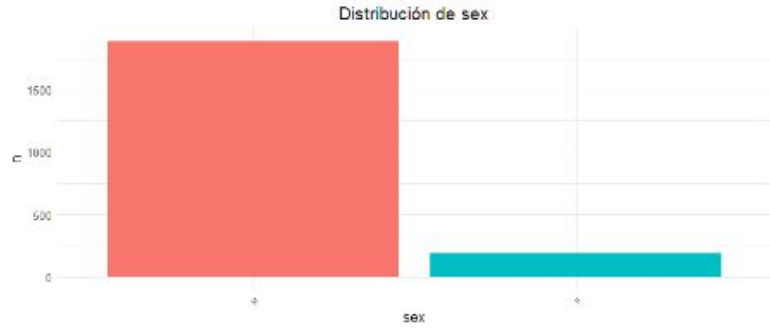
Nuestra nueva variable muestra una **clara tendencia: parece que cuantos más años pasan desde que se escaló por primera vez, mayor es la tasa de éxito.**

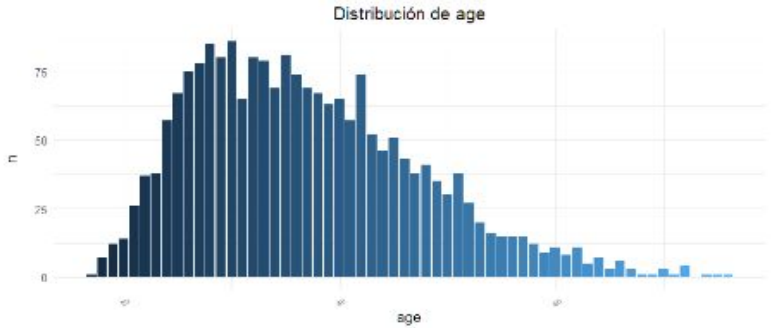
Se realizó el análisis y gráfico agrupando por intervalos (bins) dado que, por tratarse de un ratio, teníamos algunos valores donde el ratio era exactamente 0 o 1, lo que dificultaba el análisis (no se visualizaba bien con scatter plot).

Distribución de variables:









Distribución de variables:

Análisis de la distribución de variables

- Si bien contamos con **muchos picos,** la **gran mayoría de expediciones se realizaron al Everest.**
- La **mayor parte de expediciones** se realizaron de los **2000s en adelante.**
- La **mayor parte** de escaladores son **hombres**.
- La mayor concentración de **edad** se encuentra **entre 20 y 40 años.**
- La **mayor parte** de las expediciones se realizan en **primavera u otoño**. Se llevaron a cabo pocas en invierno o verano.
- El rol más usual con diferencia es el de “Climber”, seguido por “H.A Workers” y “Leaders”.

Distribución agrupando por éxito en boxplots

Análisis de la distribución agrupando por éxito en boxplots

- La **edad** no **representa una diferencia sustancial entre éxito o fracaso**

- La **altura del pico** por sí sola tampoco es suficiente para determinar diferencias, sino que es más relevante la **altura alcanzada**.
  Esto asumimos que es por la gran cantidad de escaladas al Everest. De todos modos no usaremos esta variable como predictora, dado que nos parece “engañoso”.

- **El año de expedición implica una diferencia entre éxito y fracaso.**

Proporción de variables categóricas por éxito

Análisis de la proporción de variables categóricas por éxito

- Claramente el **uso de oxígeno afecta la proporción de éxitos, por lo tanto será utilizado como predictor en el modelo.**

- Si bien se observa una diferencia en la proporcion por **sexo** entre hombres y mujeres, **se descartó la misma dada la diferencia entre cantidad de escaladores.**

- **Se observan ciertas diferencias considerables en la Season. Recordar que para invierno y verano tenemos pocas observaciones. La tasa de éxito en verano es del 100%.**

Definición de los modelos

¿Por qué un modelo logístico?

- **Éxito o fracaso** de la escalar la montaña: **suceso binario**
- Utilizamos variables **Bernoullis**, con probabilidad π de éxito.
- La variable de respuesta **Y** distribuye:

$$Y_i \mid \pi_i \sim \text{Bern}(\pi_i)$$

Con

$$E(Y_i \mid \pi_i) = \pi_i$$

- Utilizaremos las **odds.**
- Son una manera alternativa de expresar la incertidumbre, donde comparamos la probabilidad de éxito contra **su complemento** (la probabilidad de fracaso):

$$\text{odds} = \frac{\pi}{1-\pi}$$

La pregunta que buscamos contestar con odds: **¿Cuántas veces es más probable que el evento ocurra en comparación con que NO ocurra?**

¿Por qué usamos odds?

- La probabilidad se encuentra entre 0 y 1.
- Los odds van de 0 a infinito.
- El logaritmo de los ddds (Log-Odds) abarca todos los reales
- Nos da mayor “libertad matemática”

Definición de un modelo logístico

$$Y_i \mid \boldsymbol{\beta_0}, \beta_1 \dots \beta_n \sim \text{Bern}(\pi_i)\ ,\ \text{con}\ \log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^\top \boldsymbol{\beta}.$$

Componente aleatorio (verosimilitud)

Link function

Donde

Intercepto

$$\beta_0 \sim N(m_0, s_0^2)$$
$$\beta_1 \sim N(m_1, s_1^2)$$
$$\vdots$$
$$\beta_n \sim N(m_n, s_n^2)$$

Prioris

Tenemos:

$$\frac{\pi_i}{1-\pi_i} = e^{\beta_0+\beta_1 X_{i1}} \quad \text{o, despejando} \quad \pi_i = \frac{e^{\beta_0+\beta_1 X_{i1}}}{1+e^{\beta_0+\beta_1 X_{i1}}}$$

"Vuelta a probabilidades". Calculamos la inversa y despejamos.

Modelo Jerárquico logístico

El nuevo modelo es:

$$Y_{ij} \mid \pi_{ij} \sim \text{Bern}(\pi_{ij}) \text{ con } \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_{0j} + \sum_{k=1}^{p} \beta_k X_{ijk}$$

Resultado del individuo i que pertenece al grupo j.

Los datos “vienen” de una Bernouilli.

Efectos fijos de las otras variables (iguales para todos).

$$\beta_0 \sim \text{N}(m_0, s_0^2)$$

Priori intercepto global.

$$\beta_{0j} = \beta_0 + b_{0j}$$

Intercepto específico por grupo. Cada grupo tiene su propio punto de partida.

$$\beta_{0j} | \beta_0, \sigma_0 \overset{ind}{\sim} N(\beta_0, \sigma_0^2)$$

Independencia condicional

$$\sigma_0 \sim \text{Exp}(\lambda)$$

Desvío entre grupos.

Controla qué tan diferentes permitimos que sean los grupos entre sí. Si es baja, los grupos son parecidos, si es alta, los grupos son heterogéneos.heterogéneos..

En resumen:

- **El intercepto pasa a estar formado por dos términos:**
- Por un lado tenemos un **término genérico beta0** que **representa una línea base para los interceptos.** Su interpretación es la habitual, r**epresenta la media cuando los predictores se encuentran en su punto medio (centrados).**
- Por otro lado, vemos los **interceptos específicos de cada grupo beta_0j. Representan la tasa base de éxito (en log-odds) para cada grupo j.** Estos interceptos se modelan con **distribución Normal alrededor del intercepto global beta_0, con desviación estándar sigma_0.**
- Esto permite **diferenciar cada grupo entre sí**, donde por ejemplo algunos grupos tienen niveles de éxito más altos que otros, **pero siempre entorno a una línea base.**
- El resto de interceptos mantienen su interpretación clásica, describiendo el resto de las variables y la relación entre ellas.

Modelo jerárquico para nuestro estudio

Recordamos las diferentes maneras de agrupar los datos (pooling) para justificar nuestra elección de modelado. Le daremos cierta "estructura" a nuestros datos

COMPLETE POOLING: *Ignora cualquier agrupación.* Asume que todas las observaciones son iguales. (Toda la información es compartida).

NO POOLING: *Trata cada montaña/expedición,etc (nuestros grupos) de forma independiente entre sí.* (No comparten ninguna información, se asume independencia entre grupos.)

MODELO JERÁRQUICO (PARTIAL POOLING): *Contempla que cada grupo es único, pero también contempla que todos siguen una distribución común* ("pertenecen a una familia"). Los grupos con pocos datos "toman datos" del promedio global. (captan/aprenden del promedio.)

Prioris Poco Informativas

Se trabajó con prioris poco informativas, de la forma: Normal (0,2.5)

Se decidió así para:

- **Evitar hacer overfitting cuando tenemos pocos datos.**
- En la escala de odds logarítmica, coeficientes mayores a 5 implican certeza prácticamente.
- Nosotros nos **queremos posicionar "en el medio"** y dejar que los datos nos digan qué fue lo que pasó.
- **Nos centramos en cero ya que queremos tener una postura neutral.**
- Asumimos a priori que las variables no tienen efecto sobre el éxito (por eso mu es 0).
- Para el error se utilizó exp(1), la cual es estándar en este tipo de problemas. Contempla que **la varianza no puede ser negativa.**

OBS: podríamos definir prioris informativas. Nos servira de indicio a la hora de modelar las posterioris.

Modelo 1: Complete Pooling (no jerárquico). Definición matemática:

$$Y_i|\beta_0,\beta_1,\beta_2,\beta_3 \sim \text{Bern}(\pi_i) \quad \text{(verosimilitud)}$$

$$\text{donde} \quad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}$$

$$\beta_0 \sim N(0, 2.5^2) \quad \text{(intercepto global)}$$

$$\beta_1,\beta_2,\beta_3 \overset{ind}{\sim} N(0, 2.5^2) \quad \text{(coeficientes fijos)}$$

Variables:

- **Y_i**: Éxito del escalador i.
- **X_i1**: Uso de oxígeno.
- **X_i2**: Estación del año.
- **X_i3**: Años desde el primer ascenso.

Modelo 1: Complete Pooling (no jerárquico)
Código:

```
```{r, echo=TRUE, results='hide'}
modsj <- stan_glm(
  success ~ oxygen_used + season + year_since_first_ascent ,
  ,data = climbers, family = binomial,
  prior_intercept = normal(0, 2.5, autoscale = TRUE),
  prior = normal(0, 2.5, autoscale = TRUE),
  prior_aux = exponential(1, autoscale = TRUE),
  chains = 4, iter = 5000*2, seed = 84735
)
```
```

Modelo 2: Partial pooling por expedicion

$$
\begin{aligned}
Y_{ij} | \beta_{0j}, \beta_1, \beta_2, \beta_3 &\sim \text{Bern}(\pi_{ij}) && \text{(verosimilitud)} \\
\text{with} \quad \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) &= \beta_{0j} + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} && \\
\beta_{0j} | \beta_{0c}, \sigma_{exp} &\overset{ind}{\sim} N(\beta_{0c}, \sigma^2_{exp}) && \text{(variabilidad entre expediciones)} \\
\beta_{0c} &\sim N(0, 2.5^2) && \text{(intercepto global)} \\
\beta_1, \beta_2, \beta_3 &\overset{ind}{\sim} N(0, 2.5^2) && \text{(coeficientes fijos)} \\
\sigma_{exp} &\sim \text{Exp}(1) && \text{(desvío est. entre expediciones)}
\end{aligned}
$$

Variables:

- **Y_i,j:** Éxito del escalador i en la expedición j.
- **X_i,j1:** Uso de oxígeno.
- **X_i,j2:** Estación del año.
- **X_i,j3**: Años desde el primer ascenso.

Jerarquía por expedición j.

Modelo 2: Partial pooling por expedicion
Código:

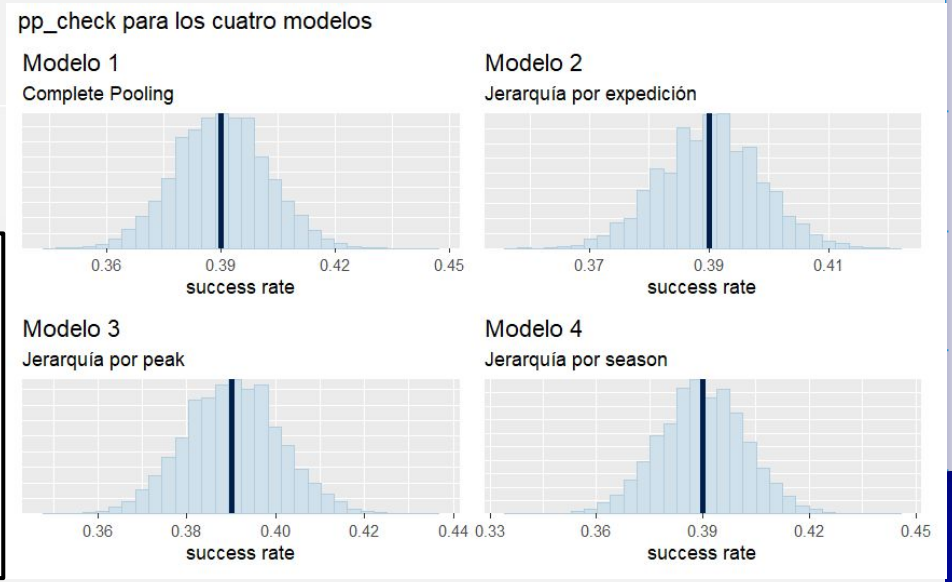
```r
```{r, echo=TRUE, results='hide'}
modje <- stan_glmer(
  success ~ season + year_since_first_ascent + oxygen_used + (1 | expedition_id),
  data = climbers, family = binomial,
  prior_intercept = normal(0, 2.5, autoscale = TRUE),
  prior = normal(0, 2.5, autoscale = TRUE),
  prior_covariance = decov(reg = 1, conc = 1, shape = 1, scale = 1),
  chains = 4, iter = 5000*2, seed = 84735
)
```
```

Modelo 3: Partial pooling por montaña

$$
\begin{aligned}
Y_{ik} | \beta_{0k}, \beta_1, \beta_2, \beta_3 &\sim \text{Bern}(\pi_{ik}) && \text{(verosimilitud)} \\
\text{with} \quad \log\left(\frac{\pi_{ik}}{1-\pi_{ik}}\right) &= \beta_{0k} + \beta_1 X_{ik1} + \beta_2 X_{ik2} + \beta_3 X_{ik3} \\
\beta_{0k} | \beta_{0c}, \sigma_{peak} &\overset{ind}{\sim} N(\beta_{0c}, \sigma^2_{peak}) && \text{(variabilidad entre picos)} \\
\beta_{0c} &\sim N(0, 2.5^2) && \text{(intercepto global)} \\
\beta_1, \beta_2, \beta_3 &\overset{ind}{\sim} N(0, 2.5^2) && \text{(coeficientes fijos)} \\
\sigma_{peak} &\sim \text{Exp}(1) && \text{(desvío est. entre picos)}
\end{aligned}
$$

Variables:

- **Y_i,k:** Éxito del escalador i en el pic k.
- **X_i,k1:** Uso de oxígeno.
- **X_i,k2:** Estación del año.
- **X_i,k3**: Años desde el primer ascenso.

Jerarquía por peak k.

Modelo 3: Partial pooling por montaña
Código:

```
```{r, echo=TRUE, results='hide'}
modjp <- stan_glmer(
  success ~ season + year_since_first_ascent+ + oxygen_used + (1 | peak_name),
  data = climbers, family = binomial,
  prior_intercept = normal(0, 2.5, autoscale = TRUE),
  prior = normal(0, 2.5, autoscale = TRUE),
  prior_covariance = decov(reg = 1, conc = 1, shape = 1, scale = 1),
  chains = 4, iter = 5000*2, seed = 84735
)
```
```

Modelo 4: Partial pooling por estación

$$
\begin{aligned}
Y_{ij} | \beta_{0j}, \beta_1, \beta_2 &\sim \text{Bern}(\pi_{ij}) && \text{(verosimilitud)} \\
\text{donde} \quad \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) &= \beta_{0j} + \beta_1 X_{ij1} + \beta_2 X_{ij2} && \\
\beta_{0j} | \beta_0, \sigma_{season} &\overset{ind}{\sim} N(\beta_0, \sigma^2_{season}) && \text{(intercepto variable por estación)} \\
\beta_0 &\sim N(0, 2.5^2) && \text{(promedio global de estaciones)} \\
\sigma_{season} &\sim \text{Exp}(1) && \text{(variabilidad entre estaciones)} \\
\beta_1, \beta_2 &\overset{ind}{\sim} N(0, 2.5^2) && \text{(coeficientes fijos)}
\end{aligned}
$$

Variables:

- **Y_i,s:** Éxito del escalador i en la season s.
- **X_i,s1:** Uso de oxígeno.
- **X_i,s2:** Estación del año.
- **X_i,s3**: Años desde el primer ascenso.

Jerarquía por season s.

Modelo 4: Partial pooling por estación
Código

```
```{r, echo=TRUE, results='hide'}
modjs <- stan_glmer(
  success ~  year_since_first_ascent + oxygen_used + (1 | season),
  data = climbers, family = binomial,
  prior_intercept = normal(0, 2.5, autoscale = TRUE),
  prior = normal(0, 2.5, autoscale = TRUE),
  prior_covariance = decov(reg = 1, conc = 1, shape = 1, scale = 1),
  chains = 4, iter = 5000*2, seed = 84735
)
```
```

PPCHECK DE LOS MODELOS

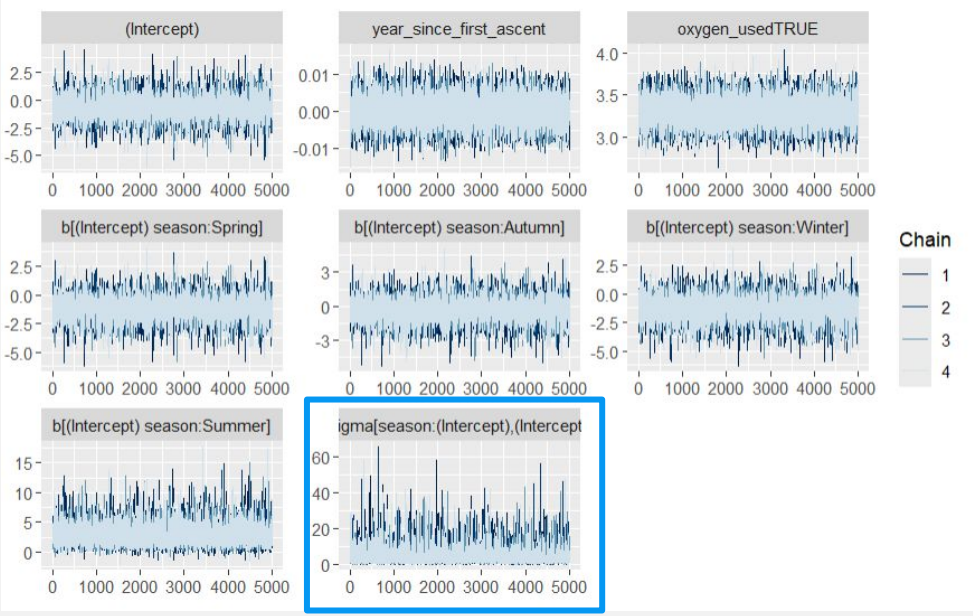
Definimos una función auxiliar para calcular la tasa de éxito de cada modelo y del dataset:

```r
# Define success rate function para los pp check
success_rate <- function(x){mean(x == 1)}
```

La **línea vertical azul** es la tasa de **éxito de los datos reales.** Lo que hace este pp_check es **simular 100 posteriors y agruparlas en histogramas en vez de densidades.** Se observan resultados muy similares para todos los modelos. No obstante, **es levemente más preciso que todos el modelo jerárquico por expedición**.



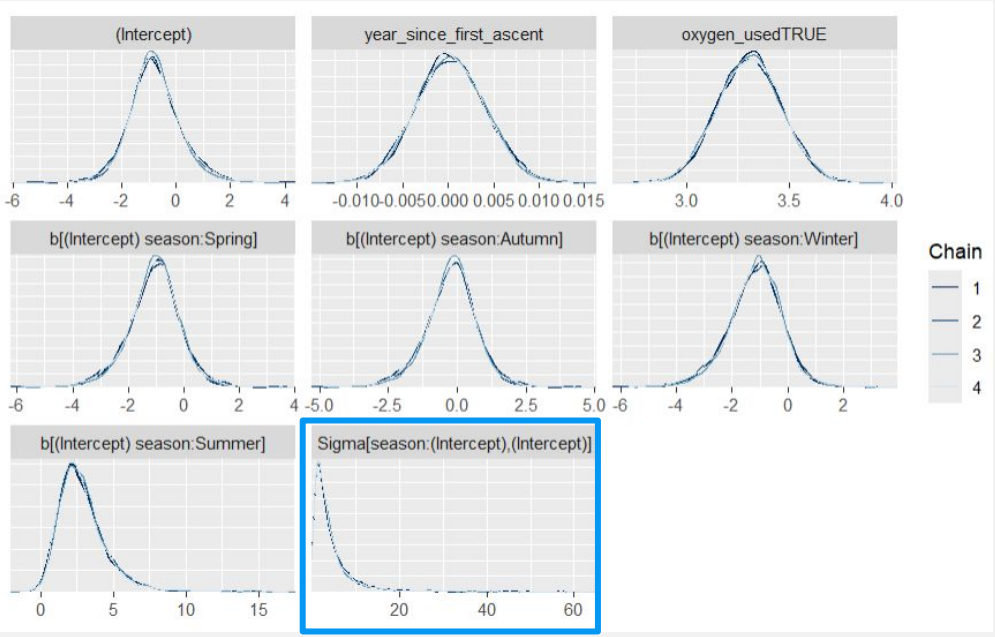
MCMC diagnostics del modelo 4: Traza



Se revisó la **convergencia** de todos los **3 modelos** anteriores a este, la cual fue la **adecuada.**

En el caso de jerarquía por season, los resultados son levemente inestables. Se detectan ciertos problemas de convergencia en alguna de las cadenas. Esto ocurre dada la poca cantidad de datos para summer.

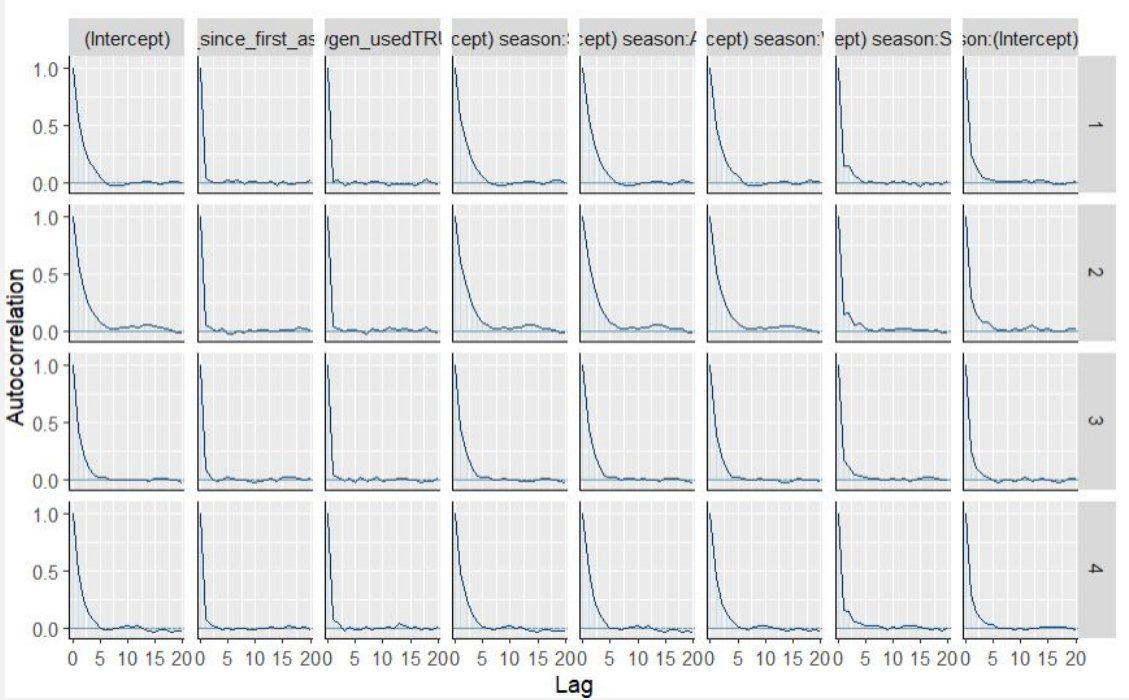
MCMC diagnostics del modelo 4: Densidades



Nuevamente, la convergencia de los otros **tres modelos** fue correcta, generando **densidades adecuadas**.

En **jerárquico por season**, los resultados de las **densidades de los parámetros por cadena son consistentes entre las distintas cadenas.**

MCMC diagnostics del modelo 4: Correlación



Nuevamente se encuentra **todo correcto para los otros tres modelos.**

En **jerárquico por season**, los resultados de las correlaciones de los parámetros por cadena muestran que cierto nivel (bajo) de correlación.

Comparación entre Modelos: Método LOO.

	elpd_diff	se_diff
modje	0.0	0.0
modjp	-267.9	22.7
modsj	-451.7	26.1
modjs	-453.2	26.1

El método **LOO** sirve para **comparar modelos según su capacidad predictiva usando Leave-One-Out Cross-Validation (LOO).**
El resultado muestra, para cada modelo, **cuánto peor es respecto al mejor (el que tiene 0 en elpd_diff)**. Cuanto más negativo es, peor predice. Si la diferencia entre modelos supera aproximadamente 2 x se_diff se considera relevante.

El modelo modje (**jerárquico por expedición**) es claramente el **mejor modelo**, con una diferencia considerable respecto al resto.
Esto tiene sentido: **el éxito personal de un individuo depende en gran parte de la expedición a la que pertenece.**
En cambio, **no depende de la montaña que sube o la época del año en la que lo hace.**

Análisis de los posteriors del modelo 2

Mediante la función tidy () se realizó el análisis de los posteriors.
Para los IC se toma 95% de confianza:

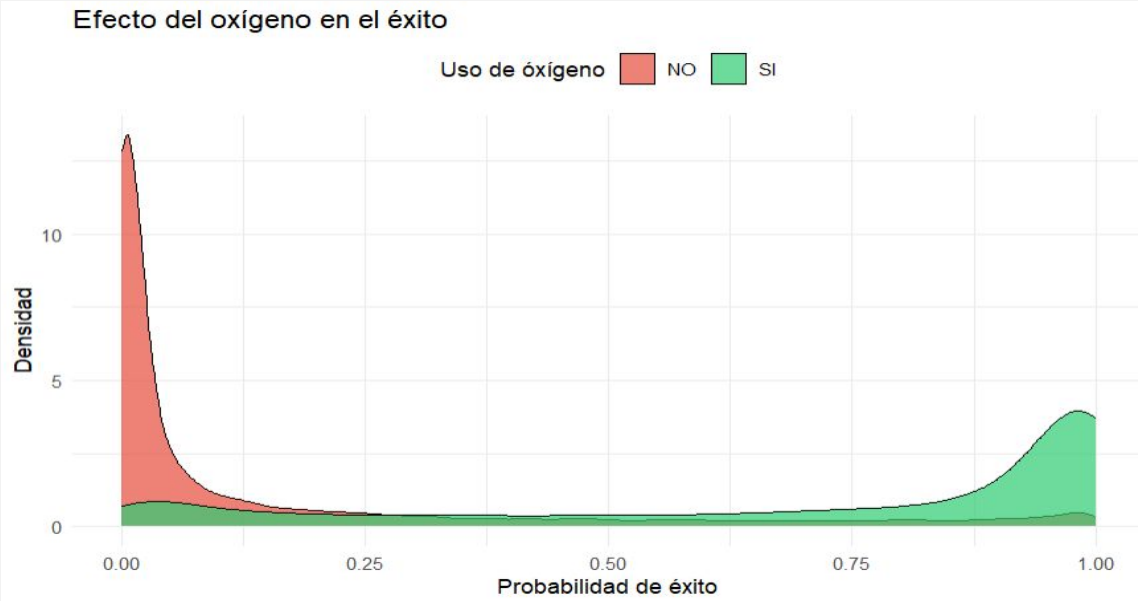
Estimacion de efectos fijos

term	estimate	std.error	conf.low	conf.high
(Intercept)	-3.987	0.918	-5.858	-2.253
seasonAutumn	1.726	0.619	0.533	2.940
seasonWinter	-0.106	1.739	-3.706	3.246
seasonSummer	35.814	26.553	5.827	110.084
year_since_first_ascent	-0.001	0.018	-0.035	0.035
oxygen_usedTRUE	6.146	0.522	5.204	7.246

La variable **years_since_first_accent no tiene relevancia predictiva dado que está prácticamente centrada en cero.**
El resto de variables tienden a ser relevantes.
El error estándar se summer se debe a los pocos datos.

Efecto del oxígeno en el posterior

Se graficó la probabilidad de éxito en función de si el individuo usa o no oxígeno.



Cuando se usa oxígeno, la distribución se mueve hacia valores claramente más altos, mostrando que aumenta la probabilidad de éxito.
Cuando no se usa oxígeno, la probabilidad es muy baja (la densidad está concentrada cerca de 0). Se observa una cola de probabilidad inferior.

Caterpillar Plot basado en Log-Odds

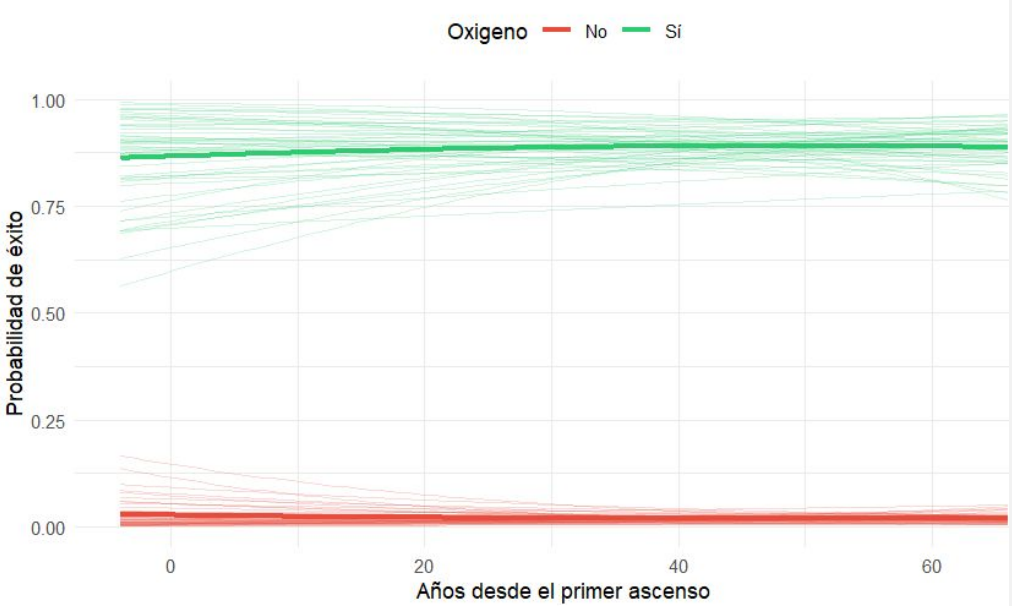
Se toman 30 expediciones al azar. Para cada una se calculan sus log-odds y se comparan con 0. Se toma como base 0 dado que 0 es el logaritmo del odd neutro (1).

Los valores a la **derecha** del 0 son aquellos en los que es más **probable** que hayan sido **éxitos.**
Los valores a la **izquierda** son los que son más **probable** que haya sido **fracaso**.

Tabla de conversión entre log-odds, odds y probabilidad

Log-Odds	Odds (Veces más probable)	% Probabilidad
4.6	100 a 1	99%
2.2	9 a 1	90%
1.1	3 a 1	75%
0.7	2 a 1	66%
0	1 a 1	50%
-0.7	1 a 2	33%
-1.1	1 a 3	25%
-2.2	1 a 9	10%
-4.6	1 a 100	1%

¿Cómo influye el oxígeno junto a year_since_first_ascent?



Claramente los años entre la expedición y la primer expedición no influyen en la probabilidad de éxito ni en el uso o no de oxígeno.
Esta nueva evidencia, sumada a la obtenida en el análisis de los posteriors, nos da a pensar que **no es una variable relevante del modelo.**

¿Qué pasa si eliminamos la variable oxígeno?

Según el gráfico anterior, vimos que la variable oxígeno es sumamente relevante a la hora de modelar la probabilidad de éxito.
Vamos a comprobar la importancia de incluir el oxígeno corriendo el modelo sin dicha variable.

```
modje_sin_oxi <- stan_glmer(
  success ~ season + year_since_first_ascent + (1 | expedition_id),
```

Mediante el método LOO comprobamos que este modelo es considerablemente peor que el modelo con oxígeno como variable, por lo que comprobamos que es necesario incluirla en el modelo.

	elpd_diff	se_diff
modje	0.0	0.0
modje_sin_oxi	-167.3	19.1

¿Qué pasa si eliminamos la variable de años?

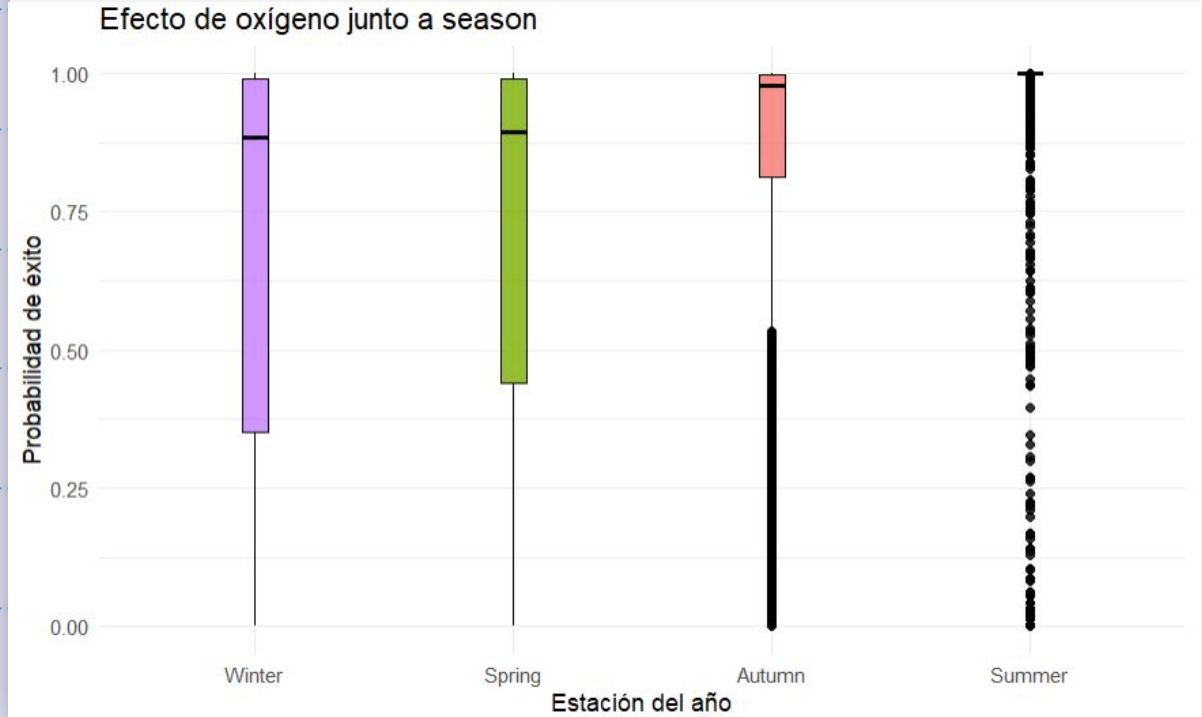
Vimos que la variable de años desde la expedición no era relevante a la hora de modelar la probabilidad de éxito.
Vamos a comprobar si realmente vale la pena o no incluirla corriendo el modelo sin la misma.

```
modje_sin_anios <- stan_glmer(
  success ~ season + oxygen_used + (1 | expedition_id),
```

Mediante el método LOO comprobamos que este modelo es igual de preciso que el anterior, por lo que podríamos eliminar esta variable del análisis sin perder precisión.

	elpd_diff	se_diff
modje_original	0.0	0.0
modje_sin_anios	-0.8	0.4

¿Cómo influye el uso de oxígeno junto a season?



Otoño: eficacia y consistencia (tenemos mínima incertidumbre).

Invierno y Primavera: Alta mediana, pero con gran variabilidad.

Verano: tenemos muy pocas observaciones, lo que hace que la simulación genere muchos datos atípicos y ninguna caja.

Nuestra cantidad de observaciones:

Spring	Autumn	Winter	Summer
1129	883	58	6

Distribución Predictiva Posteriori (1)

Creación de nuevas expediciones

Queremos ver la capacidad de generalización del modelo. Ya sabemos lo que ocurrió (caterpillar plot), pero nos interesa ver qué puede pasar a futuro.

```r
#definir datos para las nuevas expediciones
#vamos a ver como afecta el oxígeno en el exito, teniendo en cuenta season
new_expedition_sin_anios <- data.frame(
  oxygen_used = c(FALSE, TRUE, FALSE, TRUE, FALSE, TRUE, FALSE, TRUE), #se usa oxígeno o no
  season = c("Spring", "Spring", "Winter", "Winter", "Autumn", "Autumn", "Summer", "Summer"), #estaciones
  expedition_id = c("Exp_Futura", "Exp_Futura", "Exp_Futura", "Exp_Futura", "Exp_Futura", "Exp_Futura",
"Exp_Futura", "Exp_Futura") #id nuevo
)

#predicciones posterioris
set.seed(84735)
# allow.new.levels = TRUE es necesario para predecir una id nueva
pred_binaria <- posterior_predict(modje_sin_anos, newdata = new_expedition_sin_anos, allow.new.levels = TRUE)

#cálculo de probabilidad promedio de éxito
#promediamos los 0s y 1s de las 4000 simulaciones
prob_exito <- colMeans(pred_binaria)
```

Distribución Predictiva Posteriori (2)

Predicción para una nueva expedición

Oxigeno	Estacion	Probabilidad_Exito
SI	Spring	0.1506
NO	Spring	0.7051
SI	Winter	0.1537
NO	Winter	0.6807
SI	Autumn	0.2688
NO	Autumn	0.8450
SI	Summer	0.9782
NO	Summer	0.9977

Si la nueva exploración utiliza oxígeno tiene mayor probabilidad de tener éxito

Evaluación Predictiva (1)

¿Qué es la matriz de clasificación y confusión?

Es una tabla que **compara lo que el modelo predice contra lo que realmente ocurrió.**

Resume **en cuántos casos** el modelo:

1. **Predice bien** (verdaderos positivos y verdaderos negativos)
2. **Es erróneo por exceso** (falsos positivos)
3. **Es erróneo por defecto** (falsos negativos)

Es una forma de evaluar qué tan bien acierta el modelo cuando estamos trabajando con una variable de decisión binaria, en nuestro caso el éxito de escalada.

¿Por qué decimos que clasificamos? Aquí es donde se introduce el concepto de **CUTOFF:**
Es el **umbral el cual utilizamos para clasificar un éxito de escalada o no.**

Si la probabilidad de éxito es mayor al cutoff definido entonces clasificamos a ese climber como exitoso.

Evaluación Predictiva (2)

Classification summary y confusion matrix con un cutoff de 0.5

```
$confusion_matrix
      y    0   1
 FALSE 1158 111
  TRUE   74 733

$accuracy_rates
```

sensitivity	0.9083024
specificity	0.9125296
overall_accuracy	0.9108863

- 0 y 1 representan la predicción del modelo
- TRUE y FALSE representan la realidad

Sensitivity: capacidad para detectar los éxitos. El modelo logró identificar correctamente al **90.8%**

Specificity: capacidad para prever los fracasos. El modelo clasificó correctamente al **91.3%**.

- **Muy buen desempeño global (teniendo en cuenta nuestro conjunto de datos)**
- **Muy balanceado.**
- **Acierta en más de 9 de cada 10 casos.**

Evaluación Predictiva (3)

¿Qué pasa si cambiamos el cutoff?

cutoff = 0.8

```
$confusion_matrix
      y    0   1
 FALSE 1246  23
  TRUE  267 540

$accuracy_rates
```

sensitivity	0.6691450
specificity	0.9818755
overall_accuracy	0.8603083

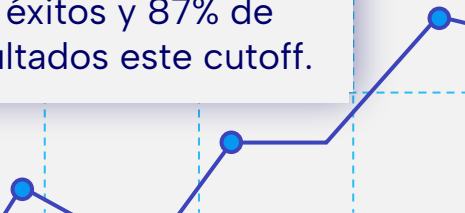
Es muy exigente para detectar los éxitos. Casi no tiene falsos positivos y su capacidad para prever el fracaso es muy buena. Sin embargo, se le “escapan” éxitos. Predice que 267 serán fracasos pero en realidad fueron éxitos.

cutoff = 0.3

```
$confusion_matrix
      y    0   1
 FALSE 1109 160
  TRUE   38 769

$accuracy_rates
```

sensitivity	0.9529120
specificity	0.8739165
overall_accuracy	0.9046243

No es exigente para detectar los éxitos. Detecta correctamente el 95% de los éxitos y 87% de fracasos. Nos da buenos resultados este cutoff.



Evaluación Predictiva (4)

¿Cómo se compara nuestro modelo con el del libro con un cutoff de 0.5?

Nuestro modelo

- Jerarquía por expedición.
- Variable de respuesta: success
- **Predictoras: season y oxygen_used**

```
$confusion_matrix
      y    0   1
 FALSE 1158 111
  TRUE   74 733

$accuracy_rates
```

sensitivity	0.9083024
specificity	0.9125296
overall_accuracy	0.9108863

Modelo del libro

```
$confusion_matrix
      y    0   1
 FALSE 1174  95
  TRUE   77 730

$accuracy_rates
```

sensitivity	0.9046
specificity	0.9251
overall_accuracy	0.9171

- Jerarquía por expedición.
- Variable de respuesta: success
- **Predictoras: age y oxygen_used**

Ambos dan resultados muy buenos

¿Cómo se diferencia nuestro modelo?

Nuestro modelo

- Jerarquía por expedición.
- Variable de respuesta: success
- **Predictoras: season y oxygen_used**

Categórica | Discreto

Season refiere a lo ambiental y logístico.

Modelo del libro

- Jerarquía por expedición.
- Variable de respuesta: success
- **Predictoras: age y oxygen_used**

Numérica | Continuo

Age refiere a lo biológico

Mientras que la edad es una característica intrínseca del individuo (no se puede cambiar), la estación nos ayuda a definir cuándo deberíamos ir, teniendo en cuenta también el uso de oxígeno.

Resumen de nuestro análisis de modelos

Propusimos **cuatro candidatos:**

1. Complete pooling
2. Partial pooling con jerarquía por expedición
3. Partial pooling con jerarquía por montaña
4. Partial pooling con jerarquía por estación

Con las **predictoras**: oxygen_used, season y year_since_first_ascent

Utilizamos validación cruzada mediante **LOO** para medir la capacidad predictiva fuera de la muestra y penalizar el overfitting. **El modelo jerárquico por expedición fue el mejor.**

Se demostró que la predictora **year_since_first_ascent no aportaba nada estadísticamente significativo, por lo que se eliminó del modelo**

Nuestro **modelo resultante fue muy bueno** y lo comparamos con el del libro.

Gracias :)

Santiago Robatto
Sofía Terra
Nahuel Bizoso
Diego Da Rosa

Inferencia II – 2025
Docentes: Marco Scavino y Fabricio Camacho

FCEA