

EST-46115: Modelación Bayesiana

Profesor: Alfredo Garbuno Iñigo — Primavera, 2022 — Flujo de trabajo bayesiano.

Objetivo. Que veremos.

Lectura recomendada: Leer Capítulo 9 de [4] y el artículo [3].

1. INTRODUCCIÓN

El desarrollo de algoritmos de muestreo nos permiten **explorar computacionalmente** una distribución de probabilidad de interés $\pi(\cdot)$. En el contexto bayesiano $\pi(\cdot)$ denota la distribución (previa o posterior) para un problema de inferencia, donde queremos reportar resúmenes

$$\pi(f) = \mathbb{E}[f(\theta)] = \int_{\Theta} f(\theta) \pi(\theta|y) d\theta. \quad (1)$$

Para resolver un problema de modelado en el contexto bayesiano debemos de tener en mente los distintos componentes del **modelo** y las **herramientas computacionales**.

- Los datos con los que contamos (o podemos contar), y .
- Nuestra abstracción del modelo generativo, $\pi(y|\theta)$.
- Nuestra matematización de nuestro conocimiento sobre el problema, $\pi(\theta)$.
- Los resúmenes que reportaremos, $f(\theta)$ ó $f(\hat{y})$.
- El mecanismo computacional para resolver integrales, $\int_{\Theta} \cdot \pi(\theta|y) d\theta$.

Exploraremos la idea general de un marco de trabajo bayesiano para después enfocarnos en cada uno de los componentes del proceso. Esto con el objetivo de entender los pasos de este proceso iterativo.

En general cuenta con tres pasos:

1. Inferencia.
2. Exploración y mejora de modelos.
3. Comparación de modelos.

En este contexto no necesariamente queremos escoger el mejor modelo. Lo que buscamos es generar un mejor entendimiento del modelo y esto lo podemos lograr al evaluar las bondades de uno sobre otro. El computo asociado con modelación bayesiana es muy complejo y puede tomar varias iteraciones en lo que estamos seguros de lo que esta realizando. Es decir, nos puede tomar varios pasos estar seguros que podemos confiar en nuestro modelo en relación a los datos que estamos analizando.

Lo importante es considerar la **pregunta objetivo**. Es decir, encontrar la pregunta que esperamos nuestro análisis pueda resolver. Por supuesto en el camino encontraremos preguntas sobre los datos, los modelos, la inferencia. Tener en mente la respuesta que queremos proveer nos permitirá definir muchos de los componentes del flujo que seguirá nuestro trabajo.

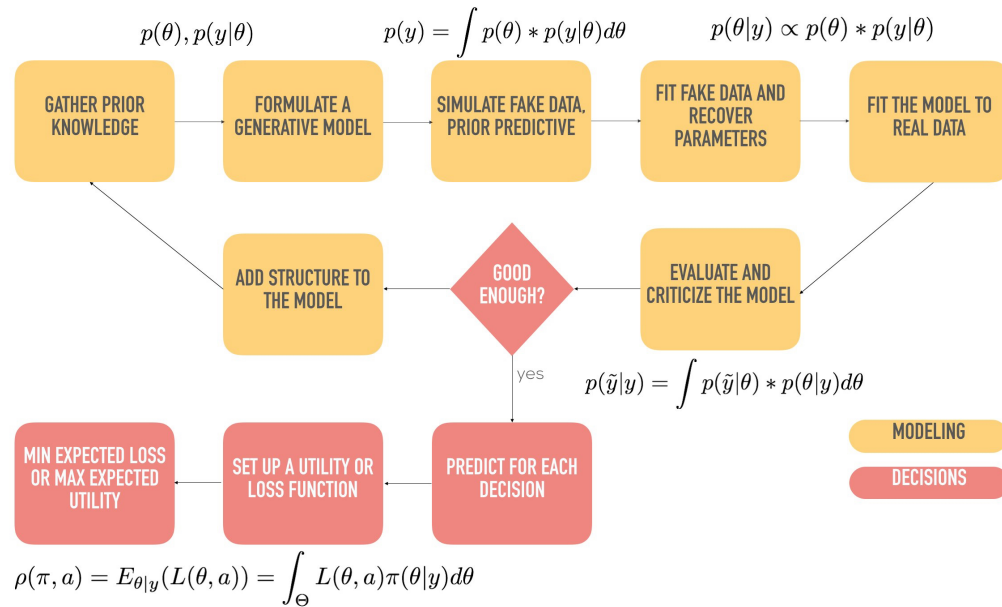


FIGURA 1. Tomada de @BayesDose, Generable.

Es decir, nos permitirá acotar la colección de modelos apropiados, cómo escoger previas, qué esperar de la distribución posterior, cómo seleccionar entre modelos, qué reportar, cómo resumir la información de nuestro modelo, qué conclusiones comunicar.

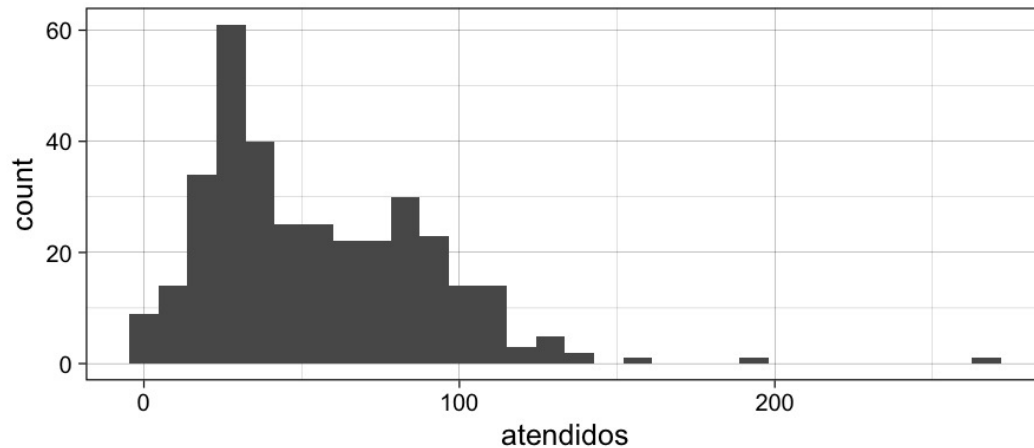
Do not be the Data Scientist and statistician that immediately reaches for Bayesian Methods, Neural Networks, Distributed Computing Clusters, or other complex tools before truly understanding the need —Martin et al. [4].

2. EJEMPLO: DATOS DE CONTEO

```

1 tweets <- read_csv("datos/response_times.csv")
2 tweets <- tweets >
3   mutate(compania = author_id_y,
4          fecha = created_at_x,
5          anio = lubridate::year(fecha),
6          mes = lubridate::month(fecha),
7          dia = lubridate::day(fecha))
8
9 reclamos <- tweets >
10  select(anio, mes, dia, compania, response_time) >
11  filter(anio == 2017, mes %in% c(10, 11),
12         !(compania %in% c("AmericanAir", "Delta", "SouthwestAir"))) >
13  group_by(anio, mes, dia, compania) >
14  summarise(atendidos = n(),
15            respuesta = mean(response_time)) >
16  mutate(compania = factor(compania)) >
17  ungroup()

```



```

1 reclamos >
2   ungroup()>
3   summarise(reclamos = mean(atendidos)) >
4   as.data.frame()

```

2.1. Primer modelo

```

1 data {
2   int N;
3   int y[N];
4 }
5
6 parameters {
7   real<lower=0> lambda;
8 }
9
10 model {
11   lambda ~ gamma(100, 2);
12   y ~ poisson(lambda);
13 }

```

```

1 modelos_files <- "modelos/compilados/reclamaciones"
2 ruta <- file.path("modelos/reclamaciones/modelo-poisson.stan")
3 modelo <- cmdstan_model(ruta, dir = modelos_files)

```

```

1 Running MCMC with 4 sequential chains...
2
3 Chain 1 finished in 0.0 seconds.
4 Chain 2 finished in 0.0 seconds.
5 Chain 3 finished in 0.0 seconds.
6 Chain 4 finished in 0.0 seconds.
7
8 All 4 chains finished successfully.
9 Mean chain execution time: 0.0 seconds.
10 Total execution time: 0.6 seconds.
11 Running MCMC with 4 sequential chains...
12
13 Chain 1 finished in 0.0 seconds.

```

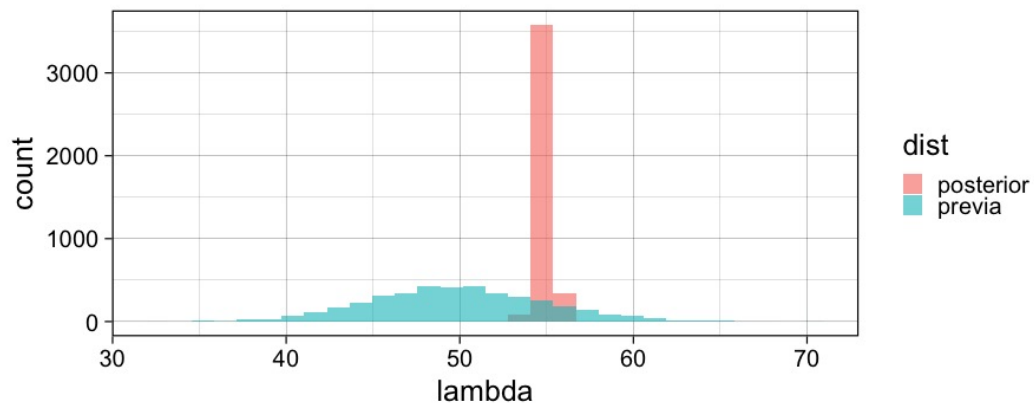
```

14 Chain 2 finished in 0.0 seconds.
15 Chain 3 finished in 0.0 seconds.
16 Chain 4 finished in 0.0 seconds.
17
18 All 4 chains finished successfully.
19 Mean chain execution time: 0.0 seconds.
20 Total execution time: 0.6 seconds.

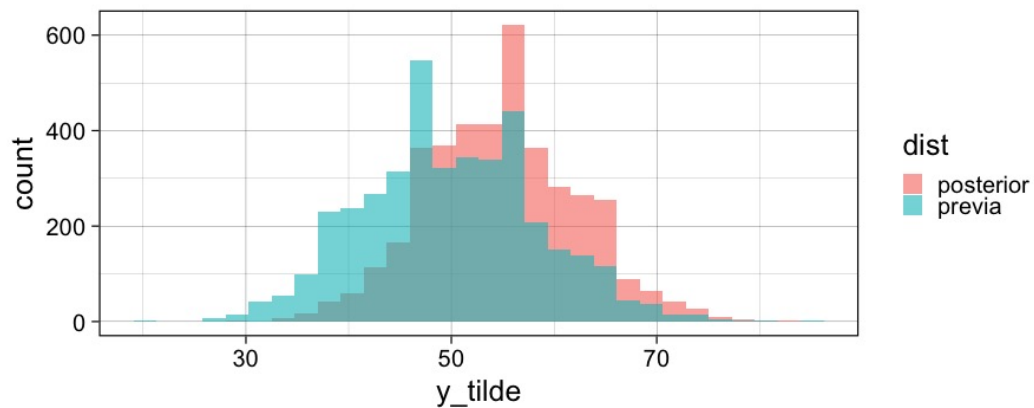
```

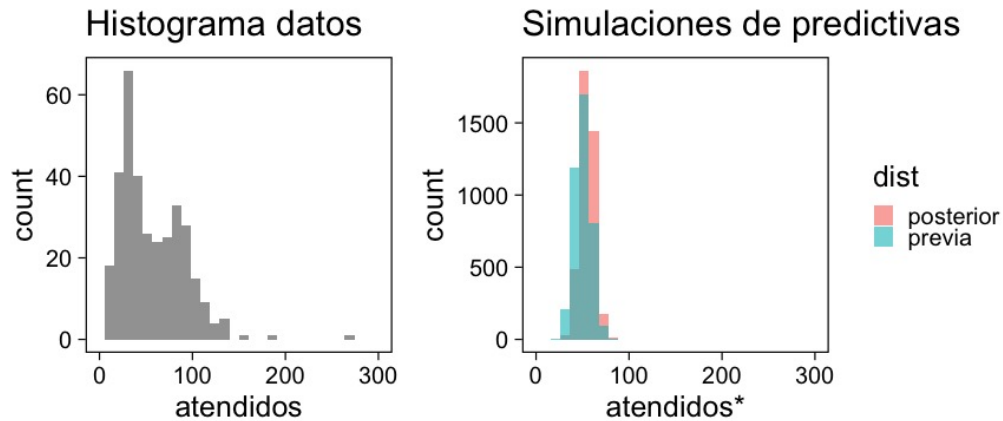
	variable	mean	median	sd	mad	q5	q95	rhat	ess_bulk
1	lp__	57355.57	57355.80	0.6979	0.2965	57354.20	57356.00	1.000	1726
2	lambda	54.85	54.85	0.3898	0.3907	54.23	55.49	1.002	1463
4	ess_tail								
5	1	2236							
6	2	1750							

Simulaciones de parámetro desconocido



Simulaciones de predictivas





2.2. Sobre-dispersión

```
1 promedio  varianza
2 1      54.88      1230
```

Neg-Binom

```
1 data {
2   int N;
3   int y[N];
4 }
5
6 parameters {
7   real<lower=0> lambda;
8   real<lower=0> phi;
9 }
10
11 model {
12   lambda ~ normal(50, 10);
13   phi ~ gamma(1, 1);
14   y ~ neg_binomial_2(lambda, phi);
15 }
16
17 generated quantities {
18   int y_tilde = neg_binomial_2_rng(lambda, phi);
19 }
```

```
1 modelos_files <- "modelos/compilados/reclamaciones"
2 ruta <- file.path("modelos/reclamaciones/modelo-negbinom.stan")
3 modelo <- cmdstan_model(ruta, dir = modelos_files)
```

	variable	mean	median	sd	mad	q5	q95	rhat
1	lp__	-1685.84	-1685.570	0.9435	0.7116	-1687.670	-1684.920	1.0001
2	lambda	54.88	54.818	1.9401	1.9842	51.821	58.050	0.9998
3	phi	2.23	2.228	0.1687	0.1719	1.958	2.511	1.0001
4	y_tilde	55.19	47.000	38.1400	32.6172	10.950	129.000	1.0000
6	ess_bulk	ess_tail						
7	1	1988	2752					
8	2	3504	2764					

```

9 3      3577      2834
10 4      3938      3887

```

2.3. Reparametrizando

```

1  promedio varianza exceso precision
2  1      54.88      1230 0.3901      2.563

```

[ayuda.](#)

2.3.1. Calibración de previa

```

1 functions {
2   // Diferencias para las colas de una Gamma
3   vector tail_delta(vector y, vector theta, real[] x_r, int[] x_i) {
4     vector[2] deltas;
5     deltas[1] = gamma_cdf(theta[1] | exp(y[1]), exp(y[2])) - 0.005;
6     deltas[2] = 1 - gamma_cdf(theta[2] | exp(y[1]), exp(y[2])) - 0.005;
7     return deltas;
8   }
9 }
10
11 transformed data {
12   vector[2] y_guess = [log(9), log(0.5)]; // Valores iniciales
13   vector[2] theta = [1.5, 15]; // Cotas del intervalo
14   vector[2] y;
15   real x_r[0];
16   int x_i[0];
17
18   // Encuentra los parametros de la Gamma para satisfacer que
19   // con 1% de probabilidad estemos en el intervalo [0.5, 20]
20   y = algebra_solver(tail_delta, y_guess, theta, x_r, x_i);
21
22   print("alpha = ", exp(y[1]));
23   print("beta = ", exp(y[2]));
24 }
25
26 generated quantities {
27   real alpha = exp(y[1]);
28   real beta = exp(y[2]);
29 }

```

```

1 modelos_files ← "modelos/compilados/reclamaciones"
2 ruta ← file.path("modelos/reclamaciones/elicit-gamma.stan")
3 modelo ← cmdstan_model(ruta, dir = modelos_files)

```

```

1 solucion ← modelo$sample(iter = 1, iter_warmup = 0, chains = 1, fixed_param =
  TRUE)
2 previa.params ← solucion$draws(format = "df")

```

2.4. Definición del modelo

```

1 data {
2   int N;

```

```

3   int y[N];
4   real<lower=0> gamma_alpha;
5   real<lower=0> gamma_beta;
6 }
7
8 parameters {
9   real log_lambda;
10  real<lower=0> phi;
11 }
12
13 model {
14   log_lambda ~ normal(4, 0.5);
15   phi ~ gamma(gamma_alpha, gamma_beta);
16   y ~ neg_binomial_2_log(log_lambda, phi);
17 }
18
19 generated quantities {
20   int y_tilde = neg_binomial_2_log_rng(log_lambda, phi);
21 }

```

```

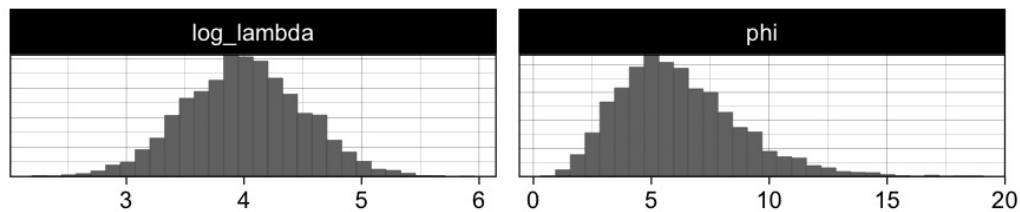
1 modelos_files <- "modelos/compilados/reclamaciones"
2 ruta <- file.path("modelos/reclamaciones/modelo-negbinom-log.stan")
3 modelo <- cmdstan_model(ruta, dir = modelos_files)

```

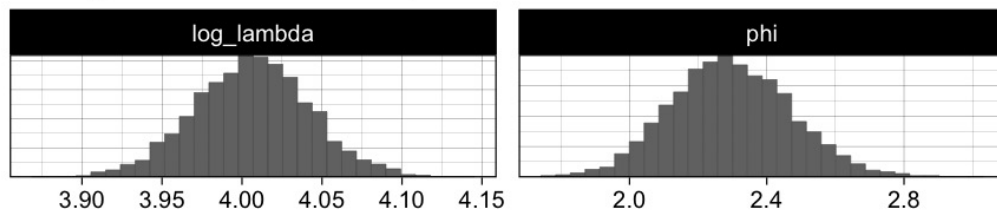
	variable	mean	median	sd	mad	q5	q95	rhat
1	lp__	-1685.764	-1685.430	1.00733	0.69682	-1687.880	-1684.810	1.002
2	log_lambda	4.006	4.007	0.03618	0.03591	3.947	4.066	1.003
3	phi	2.296	2.289	0.17468	0.17565	2.019	2.587	1.001
4	y_tilde	55.131	48.000	36.92204	32.61720	10.000	126.000	1.001

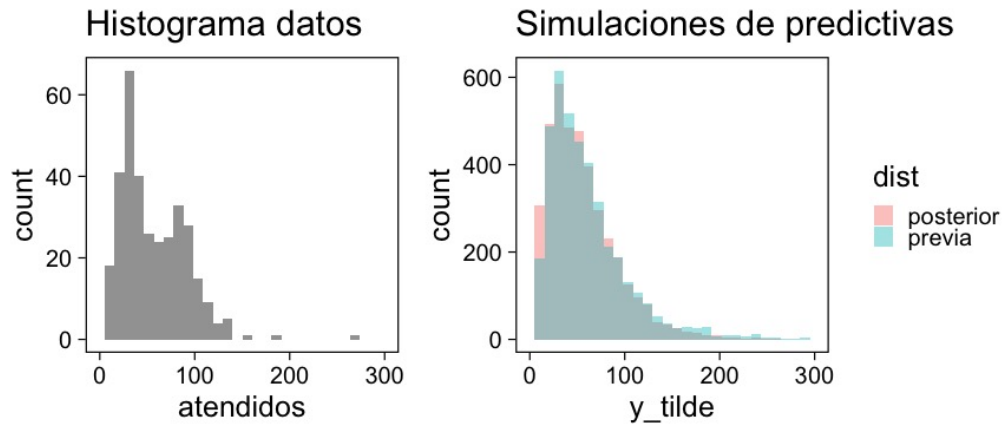
	ess_bulk	ess_tail
1	1858	2521
2	3307	2541
3	3642	2522
4	3397	3806

Histogramas distribución previa



Histogramas distribución posterior





2.5. Modelo jerarquico

	compania	promedio	varianza	exceso	precision
1	AlaskaAir	83.59	1341.1	0.17997	5.557
2	VirginAtlantic	36.44	185.9	0.11255	8.885
3	British_Airways	46.45	734.0	0.31867	3.138
4	JetBlue	85.20	637.3	0.07606	13.148
5	VirginAmerica	32.28	154.2	0.11698	8.548
6	AirAsiaSupport	41.61	1718.4	0.96831	1.033

```

1 data {
2   int N;
3   int y[N];
4   int compania[N];
5   real<lower=0> gamma_alpha;
6   real<lower=0> gamma_beta;
7 }
8
9 parameters {
10  real log_lambda[6];
11  real<lower=0> phi[6];
12 }
13
14 model {
15  log_lambda ~ normal(4, 0.5);
16  phi ~ gamma(gamma_alpha, gamma_beta);
17  y ~ neg_binomial_2_log(log_lambda[compania], phi[compania]);
18 }
19
20 generated quantities {
21  int y_tilde[6];
22  for (ii in 1:6){
23    y_tilde[ii] = neg_binomial_2_log_rng(log_lambda[ii], phi[ii]);
24  }
25 }

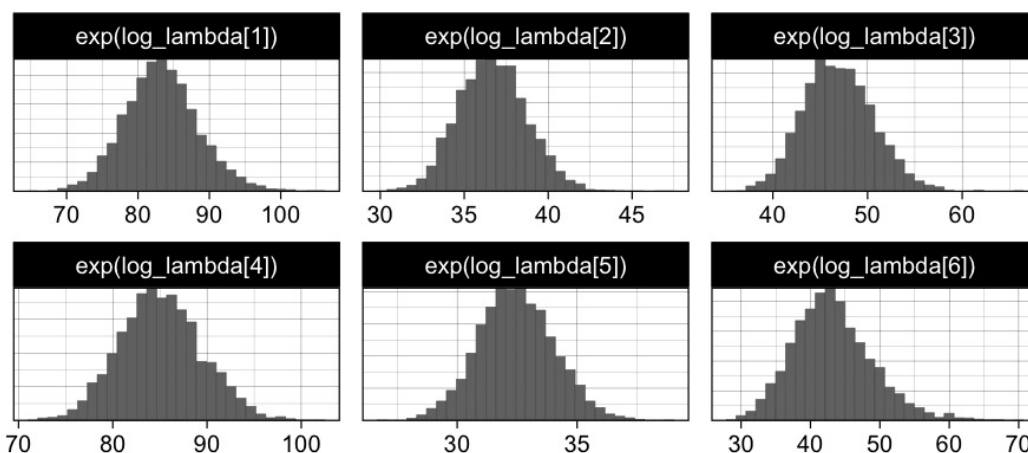
```

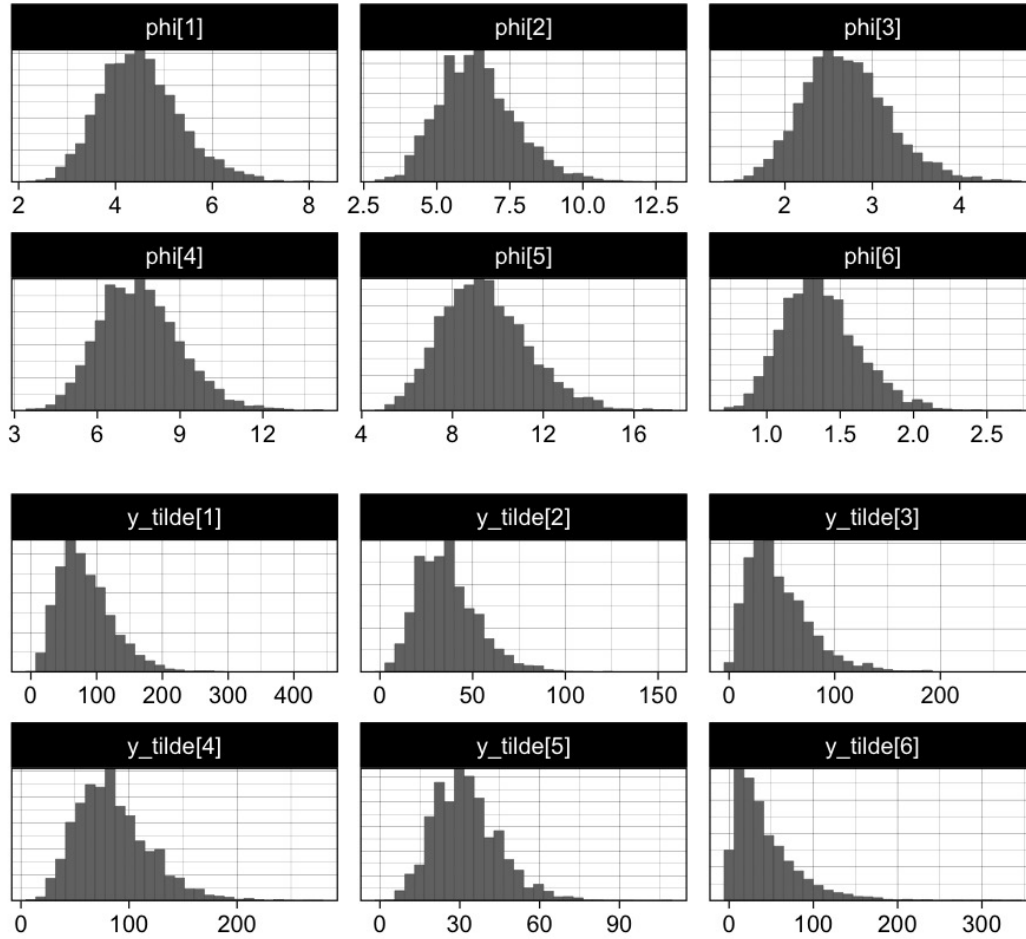
```

1 modelos_files <- "modelos/compilados/reclamaciones"
2 ruta <- file.path("modelos/reclamaciones/modelo-negbinom-jerarquico.stan")
3 modelo <- cmdstan_model(ruta, dir = modelos_files)

```


	variable	mean	median	sd	mad	q5	q95
1							
2	1 lp_	-1561.142	-1560.810	2.49926	2.40181	-1565.7410	-1557.720
3	2 log_lambda[1]	4.421	4.421	0.06188	0.05912	4.3196	4.525
4	3 log_lambda[2]	3.601	3.602	0.05583	0.05598	3.5115	3.692
5	4 log_lambda[3]	3.846	3.845	0.07883	0.07965	3.7189	3.975
6	5 log_lambda[4]	4.443	4.442	0.05118	0.05135	4.3578	4.528
7	6 log_lambda[5]	3.479	3.479	0.04844	0.04876	3.3978	3.558
8	7 log_lambda[6]	3.757	3.754	0.13021	0.12758	3.5439	3.975
9	8 phi[1]	4.519	4.461	0.83398	0.81118	3.2707	6.038
10	9 phi[2]	6.392	6.294	1.32577	1.27740	4.4290	8.763
11	10 phi[3]	2.702	2.664	0.49926	0.47912	1.9402	3.591
12	11 phi[4]	7.546	7.452	1.47012	1.45133	5.3441	10.094
13	12 phi[5]	9.437	9.290	1.91648	1.89374	6.5655	12.851
14	13 phi[6]	1.368	1.348	0.25863	0.25124	0.9832	1.831
15	14 y_tilde[1]	83.304	76.000	42.10947	38.54760	28.0000	162.000
16	15 y_tilde[2]	36.572	34.000	16.48838	14.82600	14.0000	66.000
17	16 y_tilde[3]	46.980	41.000	30.21969	26.68680	11.0000	104.000
18	17 y_tilde[4]	85.118	81.000	33.93328	32.61720	39.0000	147.000
19	18 y_tilde[5]	32.483	31.000	12.37788	11.86080	15.0000	55.000
20	19 y_tilde[6]	43.594	33.000	39.31701	29.65200	4.0000	119.000
21	rhat ess_bulk ess_tail						
22	1 1.0013 1499 2238						
23	2 1.0010 8523 2855						
24	3 1.0004 7864 3024						
25	4 1.0006 6136 2849						
26	5 1.0000 7305 3165						
27	6 1.0012 7819 2843						
28	7 0.9996 7997 2674						
29	8 1.0011 7244 2829						
30	9 1.0002 7781 2356						
31	10 1.0016 7090 3164						
32	11 0.9998 8263 2965						
33	12 1.0031 9427 2987						
34	13 1.0017 9598 3014						
35	14 0.9997 3871 3714						
36	15 0.9999 3861 3929						
37	16 0.9996 3799 3721						
38	17 0.9999 3763 3453						
39	18 0.9995 4171 3791						
40	19 1.0002 4232 3463						



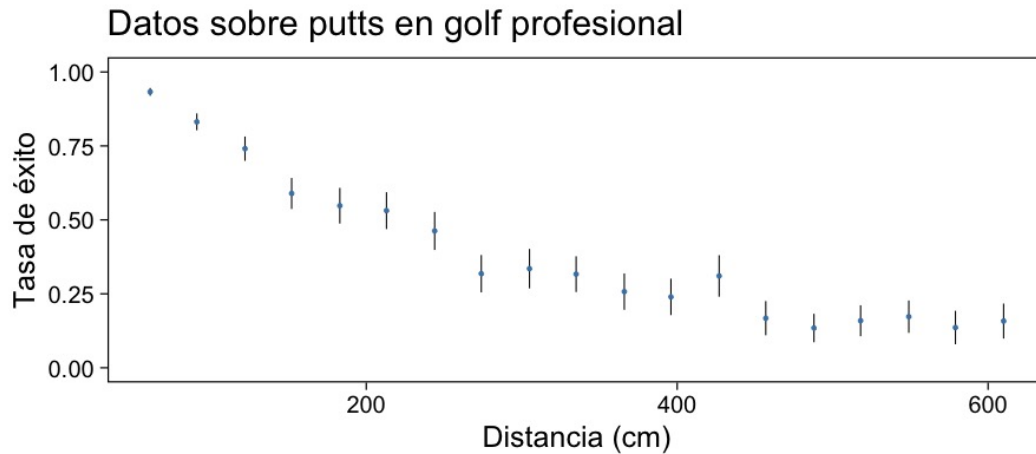


3. EJEMPLO: TIROS DE GOLF

Este ejemplo lo hemos tomado de [1]. El objetivo de este **no** es volvernos expertos en modelar tiros de golf. El objetivo es **conocer de un proceso iterativo para construcción y validación de modelos**.

Queremos entender y modelar la **probabilidad de éxito de /putts** de Golf (*putts*: tiros relativamente cerca del hoyo que buscan que la pelota ruede al hoyo o muy cerca de él). Así como entender la dependencia entre el éxito y la distancia del tiro. Como conclusiones quisiéramos inferir qué tan precisos son los profesionales en sus tiros [2].

Definición (datos): El espacio de observaciones que esperaríamos son del tipo (x, y) donde x es la distancia del *putt* y y indica si se logró o no. Sin embargo, los datos que tenemos son agregados: para cada distancia aproximada x_j tendremos un conteo de intentos n_j y éxitos y_j sobre los tiros de los jugadores profesionales. En total las distancias han sido redondeadas y obtenemos $J = 19$ distancias distintas.



3.1. Modelo logístico

Un primer intento es modelar la probabilidad de éxito a través de una regresión logística.

$$y_j \sim \text{Binomial}(n_j, \text{logit}^{-1}(a + bx_j)),$$

para cada $j = 1, \dots, J$. Este modelo lo escribimos en **Stan** como sigue

```

1 data {
2   int J;
3   int n[J];
4   vector[J] x;
5   int y[J];
6 }
7 parameters {
8   real a;
9   real b;
10 }
11 model {
12   y ~ binomial_logit(n, a*x + b);
13 }
```

LISTING 1. Modelo logístico para tasa de éxito de tiros de golf.

Notemos que no hemos especificado una distribución inicial explícita para nuestros parámetros. Por default **Stan** está incorporando una distribución **plana** en todo el espacio $(a, b) \in \mathbb{R}^2$. Podríamos debatir si esto es aceptable y las consecuencias de incluir una distribución inicial de esta naturaleza.

```

1 modelos_files <- "modelos/compilados/golf"
2 ruta <- file.path("modelos/golf/modelo-logistico.stan")
3 modelo <- cmdstan_model(ruta, dir = modelos_files)
```

Utilicemos la siguiente función para evitar *overhead* en el ajuste del modelo.

```

1 Running MCMC with 4 sequential chains...
2
3 Chain 1 finished in 0.1 seconds.
4 Chain 2 Rejecting initial value:
5 Chain 2   Log probability evaluates to log(0), i.e. negative infinity.
6 Chain 2   Stan can't start sampling from this initial value.
```

```

7 Chain_2_finished_in_0.2_seconds.
8 Chain_3_finished_in_0.1_seconds.
9 Chain_4_Rejecting_initial_value:
10 Chain_4_Log_probability_evaluates_to_log(0), i.e. negative infinity.
11 Chain_4_Stan_can't start sampling from this initial value.
12 Chain 4 finished in 0.1 seconds.
13
14 All 4 chains finished successfully.
15 Mean chain execution time: 0.1 seconds.
16 Total execution time: 1.0 seconds.

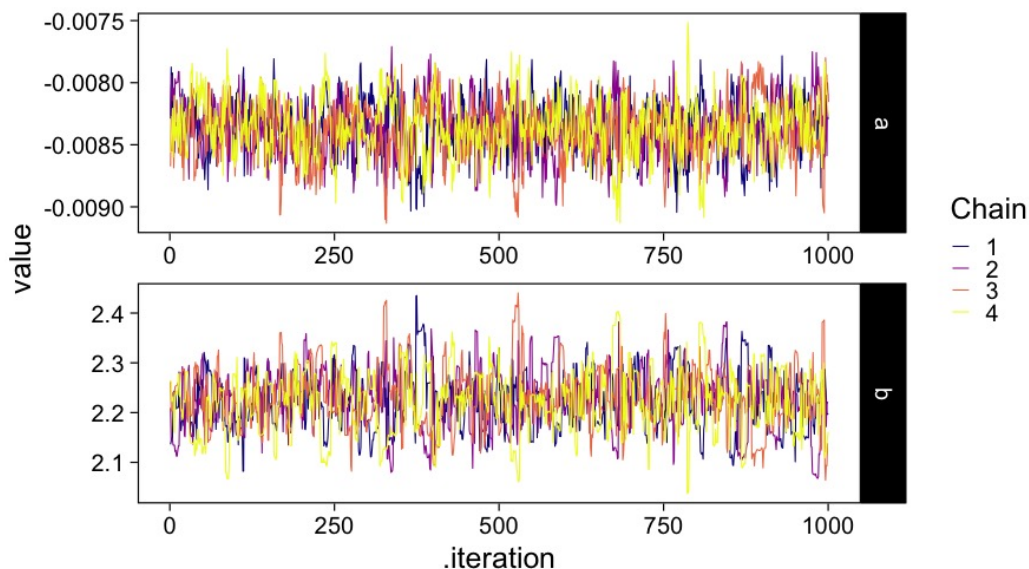
```

A pesar de los problemas en la semillas iniciales parece ser que no hay problema en muestrear del modelo posterior.

	variable	mean	median	sd	mad	q5	q95
1	lp__	-3.022e+03	-3.021e+03	1.0604122	0.7635390	-3.024e+03	-3.021e+03
2	a	-8.372e-03	-8.365e-03	0.0002235	0.0002222	-8.750e-03	-8.008e-03
3	b	2.225e+00	2.225e+00	0.0602230	0.0601194	2.126e+00	2.325e+00

	rhat	ess_bulk	ess_tail
1	1.002	956.7	1213.5
2	1.001	814.1	1205.8
3	1.002	585.3	552.3

Podemos explorar las trayectorias marginales. Todo indica que el ajuste está bien y no hay problemas aparentes con el modelo.



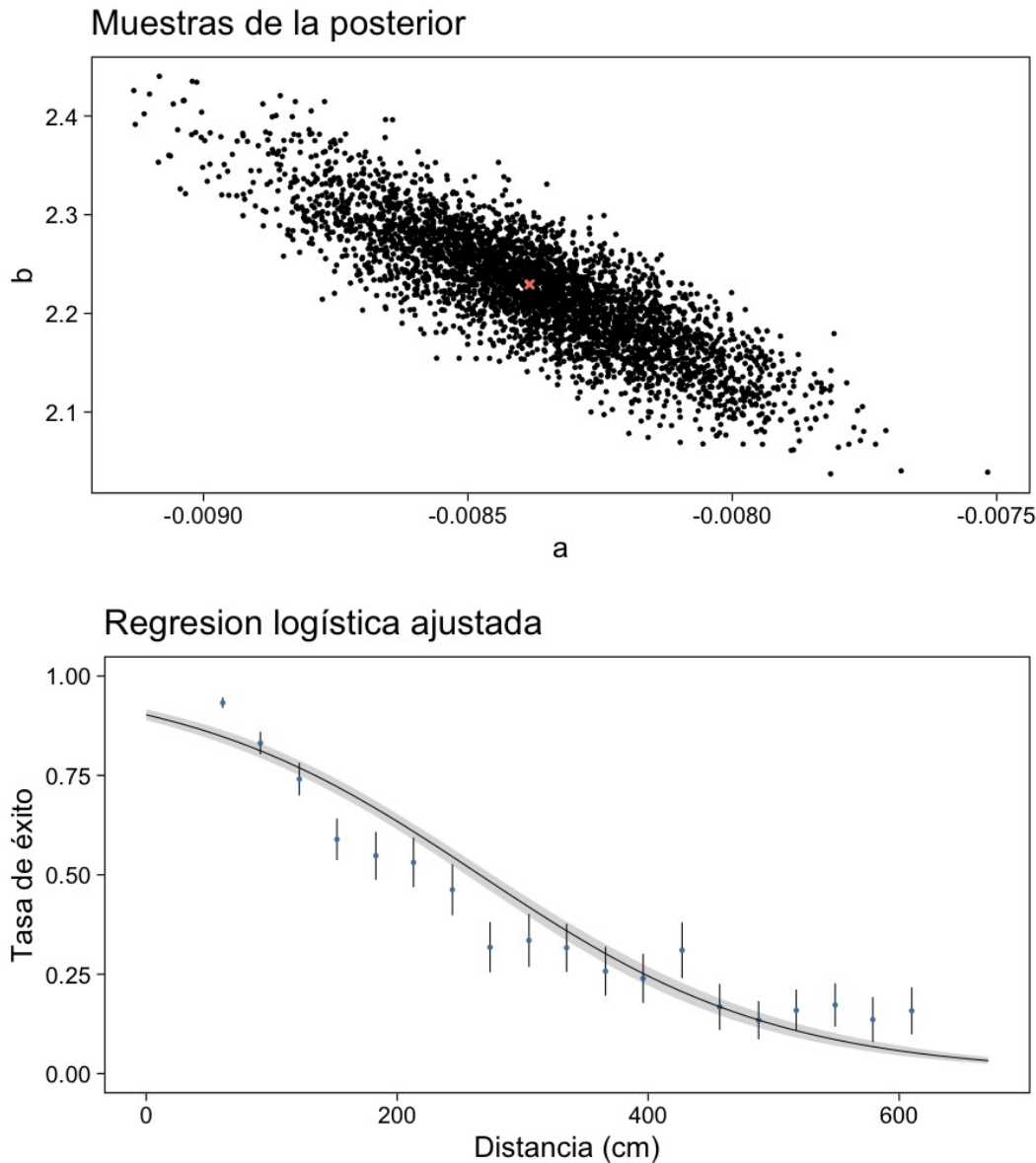
Fun fact: como exploraron en la tarea podemos extraer los puntos que maximizan la distribución posterior, ¿en serio?

```

1 params_map <- modelo$optimize(data = data_list, seed = 108)
2 params_map <- params_map$summary() >
3   pivot_wider(values_from = estimate, names_from = variable)
4 params_map > as.data.frame()

```

Podríamos explorar un gráfico de dispersión para visualizar la correlación posterior de nuestros parámetros y ubicar el valor que maximiza la pseudo-posterior.



La línea sólida representa la mediana de la curva de regresión calculada entre las muestras de la posterior obtenidas. La región sombreada corresponde a la banda del 99 % de credibilidad calculada a partir del mismo conjunto de muestras.

El modelo es razonable, en el sentido de que los parámetros tienen los valores que esperaríamos. La pendiente del modelo de regresión logística es negativa, lo cual interpretamos como la falta de precisión del tirador mientras mas alejado del hoyo. Mientras que para el caso base ($x = 0$) el modelo da una probabilidad de éxito relativamente alta.

En las siguientes secciones ilustraremos el procedimiento para complementar el modelo.

3.2. Análisis conceptual

Podemos pensar en cada intento que hace un golfista como una prueba independiente que puede resultar en éxito o fracaso. El modelo anterior establece la probabilidad de éxito como una función no lineal de la distancia.

El problema es considerablemente complicado conceptualmente ([5]) si consideramos todas las fuentes de variación: ángulo de tiro, potencia de tiro, declive en *greens* y así sucesivamente.

Los supuestos que criticaremos son los siguientes. Seguiremos haciendo la simplificación de superficie plana, pero consideramos dos parámetros para el tiro con distintas condiciones de éxito:

1. El ángulo del tiro.
2. La velocidad con la que la pelota llega (o no llega) al hoyo.

Los radios de una pelota de golf y el hoyo (en centímetros) son de

```

1  pelota hoyo
2  1      2.1  5.4

```

LISTING 2. Radios para pelota y hoyo en una configuración de golf profesional.

Supondremos por el momento que los *greens* de golf (áreas cerca del hoyo) son perfectamente planos (lo cual no es cierto, pero refinaremos después), de modo que el éxito depende de:

1. Tirar la pelota con un ángulo suficientemente cercano a cero con respecto a la línea que va del centro de la pelota al centro del hoyo.

1. Tirar la pelota con una velocidad suficiente para que llegue al hoyo pero no tan alta que vuele por encima del hoyo.

Mejores datos de los tipos de fallo sería útil, pero por el momento no los tenemos disponibles.

3.3. Angulo de tiro

Supongamos que la distancia del centro de la pelota al centro del hoyo es x . Idealmente ésta es la trayectoria que el golfista tendría que ejecutar. Sin embargo, el tiro puede ser inexacto y denotamos por θ el ángulo del tiro realizado. El tiro es exitoso cuando el angulo de tiro satisface

$$|\theta| < \tan^{-1} \left(\frac{R - r}{x} \right). \quad (2)$$

Incorporamos un esquema de esta situación a continuación.

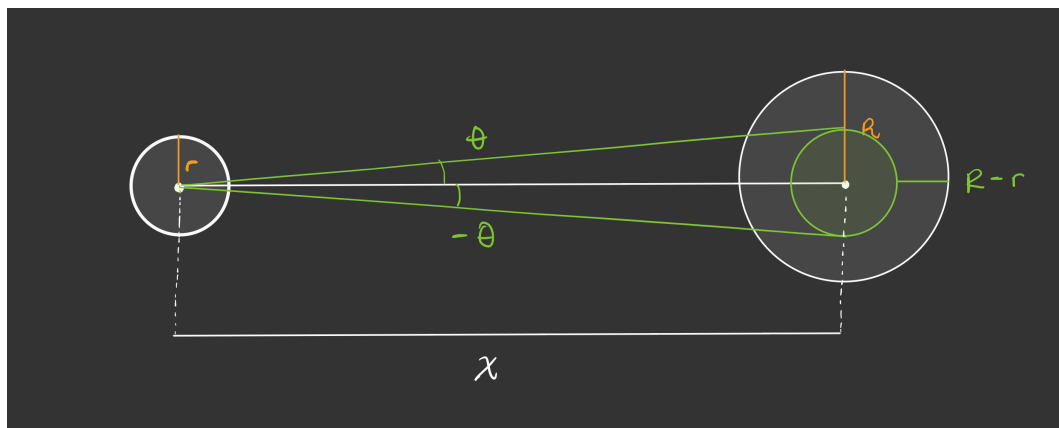


FIGURA 2. Esquema de tiro y condiciones para un tiro exitoso.

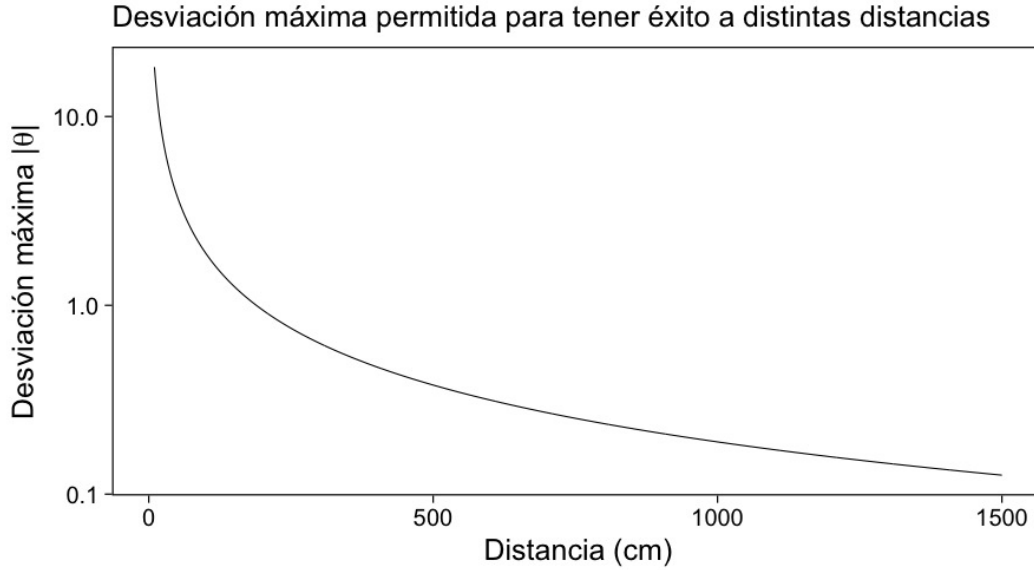
Observación: Aquí hemos hecho un supuesto importante. La **distancia reportada** en los datos, la cual hemos denotado por x , es la distancia entre el centro de la pelota y el centro

del hoyo. ¿Cómo cambiaría nuestra condición de éxito si suponemos que la distancia que viaja la pelota es la registrada?

Para nuestro problema, la condición de éxito es

$$|\theta| < \tan^{-1} \left(\frac{3.3}{x} \right). \quad (3)$$

Mejores golfistas tendrán mejor control sobre θ , y conforme x es más grande, la probabilidad de tener éxito baja:



Observación. Esta curva puede variar dependiendo del jugador, pero vamos a modelar el conjunto de tiros de jugadores profesionales. Suponemos homogeneidad, misma que podríamos checar con datos desagregados por jugador. Estos datos podrían tener sobre-representación de tiradores malos (pues quizá hacen más tiros).

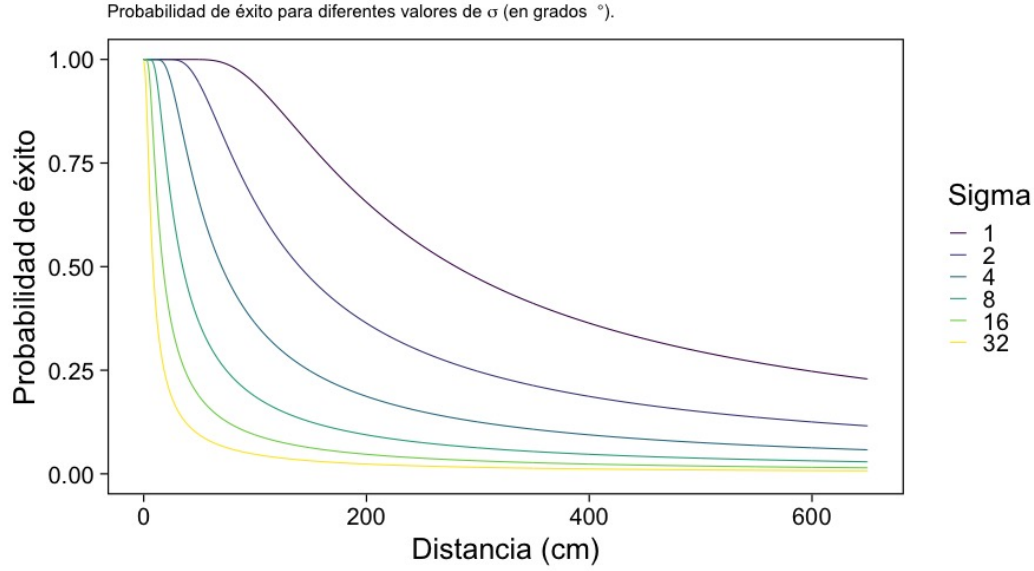
Para modelar θ de manera probabilista asumimos una distribución Gaussiana con media 0 y desviación estándar σ . Este modelo codifica nuestra suposición de que los jugadores en promedio tirarán en la dirección correcta, sin embargo puede haber diversos factores que afectarán este resultado.

Siguiendo esta distribución, la probabilidad de éxito se calcula como

$$\mathbb{P} \left\{ |\theta| < \tan^{-1} \left(\frac{R-r}{x} \right) \right\} = 2 \Phi \left[\frac{\tan^{-1}((R-r)/x)}{\sigma} \right] - 1, \quad (4)$$

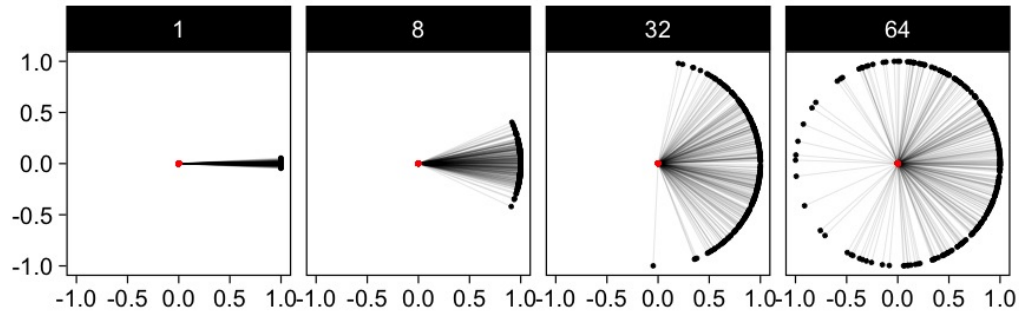
donde Φ es la función de acumulación de una Normal estándar.

El parámetro σ controla la desviación de los tiros en línea recta. Por lo tanto afecta la probabilidad de éxito conforme mas lejos estemos y más grande sea su valor. El gráfico siguiente muestra que si el golfista tiene mejor control sobre su tiro, entonces mayor será su resistencia a encontrarse lejos.



Ahora veamos las distintas realizaciones de tiros a 1 metro de distancia bajo distintos valores de σ . Nota que estamos *traduciendo* el impacto que tiene nuestro modelo previo en términos de observaciones tangibles del modelo.

Posiciones finales de tiro



Notamos que los tiros en general tienen un buen comportamiento. Posiblemente valores de tiros con una desviación de 60° dan lugar a tiros que no tienen sentido. Este punto lo veremos más adelante en caso de que tengamos que refinar. Por el momento, el modelo queda como sigue

$$p_j = 2 \Phi \left(\frac{\tan^{-1}((R-r)/x_j)}{\sigma} \right) - 1, \quad (5)$$

$$y_j \sim \text{Binomial}(n_j, p_j), \quad (6)$$

para $j = 1, \dots, J$.

La gran diferencia del modelo es asumir una relación distinta para la probabilidad de éxito de los experimentos binomiales. Este modelo se ha inferido de primeros principios y un poco de geometría.

3.4. Ajuste modelo

El modelo en **Stan** queda como se muestra. Nota que utilizamos la función de acumulación de una normal estándar [Phi](#).

```

1 data {
2   int J;
3   int n[J];
4   vector[J] x;
5   int y[J];
6   real r;
7   real R;
8 }
9 transformed data {
10   vector[J] threshold_angle = atan((R-r) ./ x);
11 }
12 parameters {
13   real<lower=0> sigma;
14 }
15 model {
16   vector[J] p = 2*Phi(threshold_angle / sigma) - 1;
17   y ~ binomial(n, p);
18 }
19 generated quantities {
20   real sigma_degrees = sigma * 180 / pi();
21 }

```

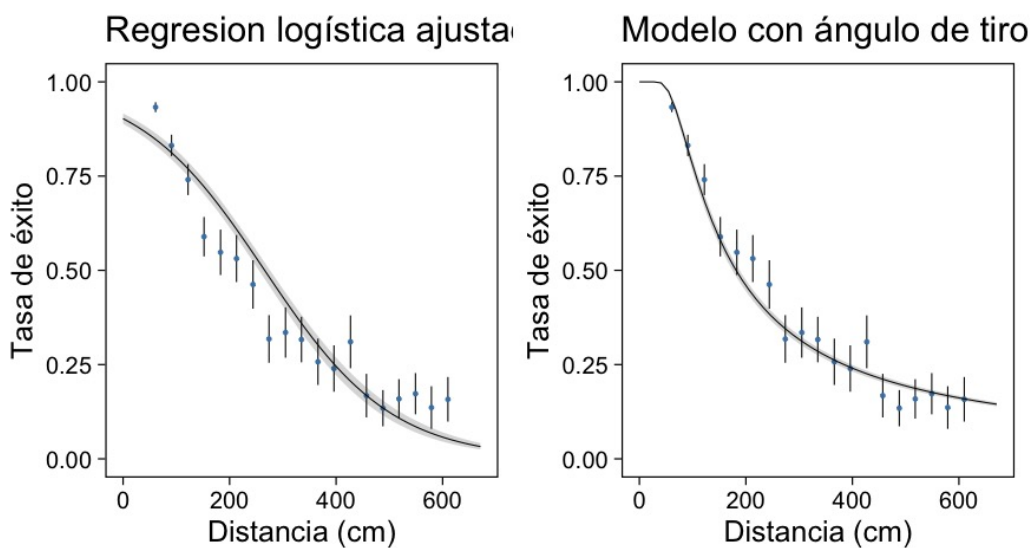
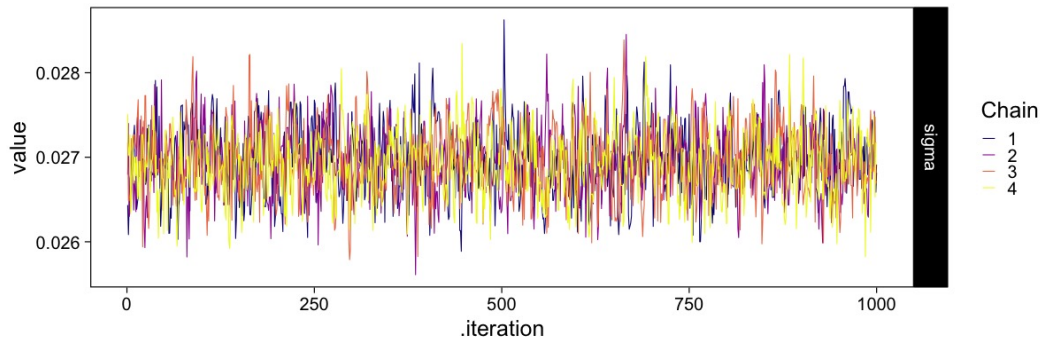
LISTING 3. Modelo con ángulo de tiro y su desviación estándar.

```

1 Model executable is up to date!
2 Running MCMC with 4 sequential chains...
3
4 Chain 1 finished in 0.0 seconds.
5 Chain 2 finished in 0.0 seconds.
6 Chain 3 finished in 0.0 seconds.
7 Chain 4 finished in 0.0 seconds.
8
9 All 4 chains finished successfully.
10 Mean chain execution time: 0.0 seconds.
11 Total execution time: 0.6 seconds.
12
13 variable      mean      median      sd      mad      q5
14 q95
15 1 lp__ -2.926e+03 -2.926e+03 0.6751595 0.2965200 -2.928e+03 -2.926e
16 +03
17 2 sigma 2.696e-02 2.695e-02 0.0003906 0.0003904 2.631e-02 2.761e
18 -02
19 3 sigma_degrees 1.545e+00 1.544e+00 0.0223814 0.0223724 1.508e+00 1.582e
20 +00
21
22 rhat ess_bulk ess_tail
23 1 1.002 1994 NA
24 2 1.001 1530 2002
25 3 1.001 1530 2002

```

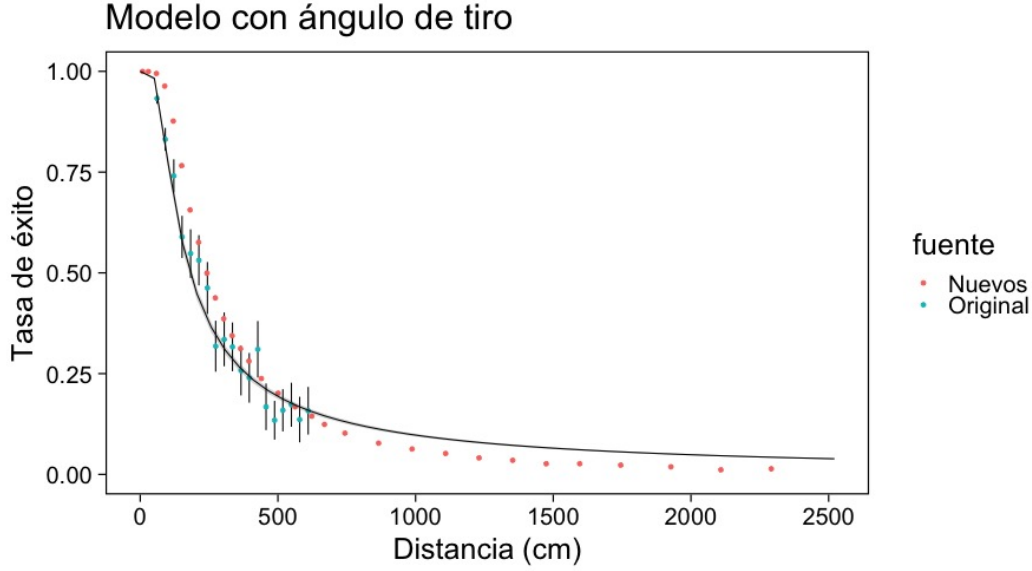
El muestreo del modelo posterior parece no tener problemas. Los diagnósticos se ven bien y las capacidades predictivas dan indicios que se ha podido ajustar un modelo satisfactorio.



3.5. Nuevo conjunto de datos

Después de algunos años se consiguieron mas registros. En particular, el profesor Broadie fue el que brindo dichos datos (comunicación con Andrew Gelman documentada en [1]). La cantidad de datos disponibles es impresionante, basta con observar la dispersión de la probabilidad de éxito bajo el supuesto normal. Los intervalos de confianza son casi imperceptibles para las nuevas observaciones (puntos salmón en el gráfico).

Ajustando el modelo a los datos nuevos vemos que parece no haber un buen ajuste. Subestimamos las tasa de éxito cuando estamos cerca y sobre-estimamos cuando nos encontramos muy lejos.



3.6. Incorporando ángulo de tiro

Para poder hacer un tiro exitoso no sólo es necesario controlar el ángulo de tiro. También es importante tirar con la fuerza suficiente. Siguiendo [5], existe un rango de velocidades iniciales que determinan la condición de éxito.

La condición de éxito en un tiro recto es que la velocidad final v_f (en metros por segundo) de la pelota cumpla con las siguientes condiciones

$$0 < v_f < 1.63.$$

Por otro lado, la aceleración de la pelota al rodar en el *green* satisface

$$a = \left(\frac{10}{7}\right) \rho_r g.$$

donde $\rho_r = \rho/r$, y ρ depende de la superficie donde rueda la pelota, r es el radio de la pelota y g la fuerza de gravedad. Datos experimentales indican que la media en *greens* es de $\rho_r = 0.131$, con un rango de 0.065 a 0.196. De momento, tomaremos $\rho_r = 0.131$.

La velocidad final de la pelota, en términos de la velocidad inicial, utiliza la aceleración en el *green*, lo cual da la siguiente cadena de igualdades

$$v_f^2 = v_0^2 - \left(\frac{10}{7}\right) \rho_r g x_m = v_0^2 - \left(\frac{10}{7}\right) (0.131) (9.81) x_m = v_0^2 - 1.835871 x_m$$

donde x_m es la distancia de la pelota al hoyo en metros. Ahora, podemos despejar para calcular las condiciones de éxito sobre la velocidad inicial v_0

$$c x_m < v_0^2 < (1.63)^2 + c x_m,$$

donde $c = 1.835871$. La condición de éxito se puede escribir en términos de la distancia de la pelota al hoyo. Es decir podemos escribir

$$u \in [x, x + 145],$$

donde $u = v_0^2/c \times 100$ es la distancia en centímetros que la pelota viajaría si no hubiera un hoyo en medio. Esto quiere decir que la pelota debe ser lanzada con fuerza suficiente para alcanzar el hoyo pero no tanta como para sobrepasarse.

Ahora, siguiendo las recomendaciones de Mark Broadie en [1]. Suponemos que los golfistas tienden a tirar con fuerza suficiente para pasarse del hoyo por un pie (30.48 cm), sin embargo la fuerza tiene un error multiplicativo. La intuición es que errores de la misma magnitud afectan en proporción a la distancia de tiro.

La distancia que recorre la pelota esta definida como

$$u = (x + 30.48) \cdot (1 + \varepsilon),$$

donde

$$\varepsilon \sim N(0, \sigma_f^2),$$

y hemos utilizado la notación σ_f^2 para hacer énfasis en el error asociado a la fuerza de tiro. Esto implica que

$$u \sim N(x + 30.48, (x + 30.48)^2 \sigma_f^2),$$

y por lo tanto el éxito debido a la fuerza de tiro —la condición $u \in [x, x + 145]$ — tiene probabilidad de éxito igual a

$$\Phi\left(\frac{114.52}{(x + 30.48)\sigma_f}\right) - \Phi\left(\frac{-30.48}{(x + 30.48)\sigma_f}\right),$$

que es un evento que asumimos independiente del ángulo de tiro.

Para finalizar, utilizamos las condiciones de éxito que definen ambos eventos que asumimos independientes, el ángulo de tiro y la fuerza. Por lo tanto, el modelo lo escribimos como

$$p_j^u = \Phi\left(\frac{114.52}{(x + 30.48)\sigma_f}\right) - \Phi\left(\frac{-30.48}{(x + 30.48)\sigma_f}\right), \quad (7a)$$

$$p_j^\theta = 2\Phi\left(\frac{\tan^{-1}((R - r)/x_j)}{\sigma_\theta}\right) - 1, \quad (7b)$$

$$p_j = p_j^u \cdot p_j^\theta \quad (7c)$$

$$y_j \sim \text{Binomial}(n_j, p_j), \quad (7d)$$

para $j = 1, \dots, J$.

Nota cómo el cambio que tenemos en nuestro modelo es la composición de dos eventos que esperamos sean independientes: la fuerza y dirección de tiro.

```

1 data {
2   int J;
3   int n[J];
4   vector[J] x;
5   int y[J];
6   real r;
7   real R;
8   real overshoot;
9   real distance_tolerance;
10 }
11 transformed data {
12   vector[J] threshold_angle = atan((R-r) ./ x);
13   vector[J] raw_proportion = to_vector(y) ./ to_vector(n);
14 }
15 parameters {
16   real<lower=0> sigma_angle;
17   real<lower=0> sigma_force;

```

```

18 }
19 transformed parameters {
20   vector[J] p_angle = 2*Phi(threshold_angle / sigma_angle) - 1;
21   vector[J] p_force = Phi((distance_tolerance - overshoot) ./ ((x + overshoot)
    *sigma_force)) -
22     Phi((- overshoot) ./ ((x + overshoot)*sigma_force));
23   vector[J] p = p_angle .* p_force;
24 }
25 model {
26   y ~ binomial(n, p);
27 }
28 generated quantities {
29   real sigma_degrees = sigma_angle * 180 / pi();
30   vector[J] residual = raw_proportion - p;
31 }

```

LISTING 4. Modelo con fuerza y ángulo de tiro.

```

1 data_new <- list(x = datos$x, n = datos$n, y = datos$y, J = nrow(datos),
2                 r = radios$pelota, R = radios$hoyo,
3                 distance_tolerance = 4.5 * 30.48, # 145,
4                 overshoot = 30.48)

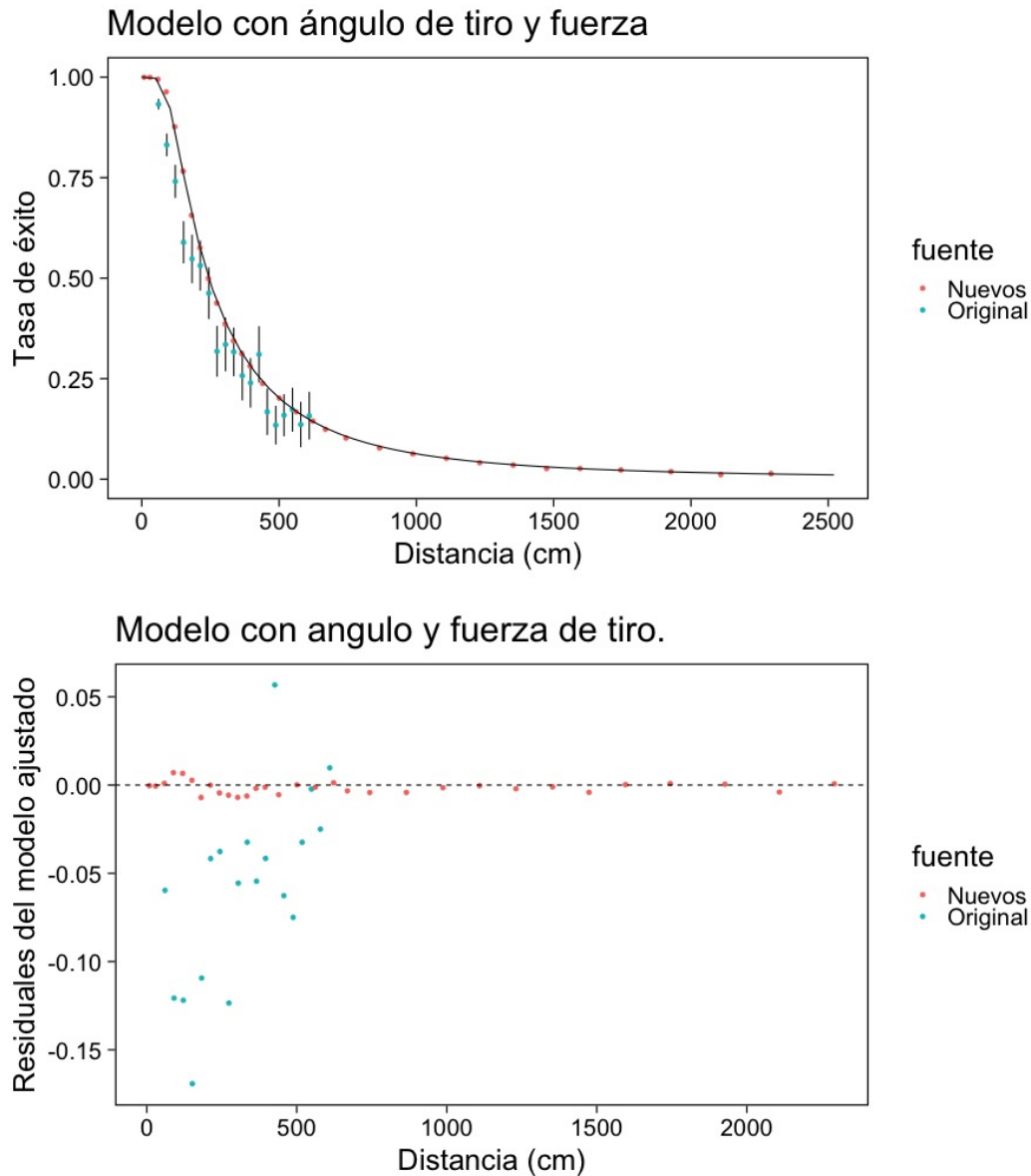
```

```

1 Model executable is up to date!
2 Running MCMC with 4 sequential chains...
3
4 Chain 1 finished in 0.2 seconds.
5 Chain 2 finished in 0.2 seconds.
6 Chain 3 finished in 0.2 seconds.
7 Chain 4 finished in 0.2 seconds.
8
9 All 4 chains finished successfully.
10 Mean chain execution time: 0.2 seconds.
11 Total execution time: 1.4 seconds.
12      variable   mean median      sd      mad      q5      q95  rhat
13 1  sigma_angle 0.0150 0.0150 4.275e-05 4.225e-05 0.01492 0.01506 1.001
14 2  sigma_degrees 0.8592 0.8593 2.449e-03 2.420e-03 0.85491 0.86305 1.001
15 3  sigma_force 0.1362 0.1362 4.882e-04 4.878e-04 0.13548 0.13711 1.002
16  ess_bulk ess_tail
17 1      1321      1536
18 2      1321      1536
19 3      1183      1261

```

Si utilizamos la semilla 2210 (al menos en mi máquina) veríamos que el ajuste del modelo parece indicar ciertos problemas. En particular notemos que podrían ser causados por un punto inicial en una cadena. Después de todo, con 4 cadenas tenemos 25% del esfuerzo computacional en una sola. Además, tenemos alertas en los demás diagnósticos. Con tales resultados nos mostramos un poco escépticos sobre los siguientes resúmenes gráficos.



Al explorar los residuales encontramos que parece haber cierto patrón. Mas aún, el modelo parece estar **muy** seguro de los valores esperados de probabilidad de éxito —lo cual podemos apreciar al incorporar los intervalos de probabilidad de los residuales que se calculan de las muestras. Esto se puede deber a que el número elevado de registros que la nueva base de datos provee.

Alternativamente, podríamos ajustar sólo en los datos nuevos. Pero no tenemos alguna justificación específica para descartar los que ya teníamos. Por lo pronto usaremos ambos conjuntos sin distinción.

3.7. Otro modelo

Una estrategia es incorporar una **aproximación continua** a las proporciones reportadas, misma que podemos utilizar para incorporar un **error de medición latente** (que en este

caso podría ser acertado). El modelo queda especificado como

$$p_j^u = \Phi\left(\frac{114.52}{(x + 30.48)\sigma_f}\right) - \Phi\left(\frac{-30.48}{(x + 30.48)\sigma_f}\right), \quad (8a)$$

$$p_j^\theta = 2\Phi\left(\frac{\tan^{-1}((R-r)/x_j)}{\sigma_\theta}\right) - 1, \quad (8b)$$

$$p_j = p_j^u \cdot p_j^\theta \quad (8c)$$

$$\frac{y_j}{n_j} \sim N\left(p_j, \frac{p_j(1-p_j)}{n_j} + \sigma_{\text{obs}}^2\right), \quad (8d)$$

para $j = 1, \dots, J$.

Por otro lado, el modelo en **Stan** no cambia mucho y se vuelve un poco mas flexible. Lo cual especificamos en el bloque de modelo.

```

1 data {
2   int J;
3   int n[J];
4   vector[J] x;
5   int y[J];
6   real r;
7   real R;
8   real overshoot;
9   real distance_tolerance;
10 }
11 transformed data {
12   vector[J] threshold_angle = atan((R-r) ./ x);
13   vector[J] raw_proportion = to_vector(y) ./ to_vector(n);
14 }
15 parameters {
16   real<lower=0> sigma_angle;
17   real<lower=0> sigma_force;
18   real<lower=0> sigma_obs;
19 }
20 transformed parameters {
21   vector[J] p_angle = 2*Phi(threshold_angle / sigma_angle) - 1;
22   vector[J] p_force = Phi((distance_tolerance - overshoot) ./ ((x + overshoot)
23     *sigma_force)) -
24     Phi((- overshoot) ./ ((x + overshoot)*sigma_force));
25   vector[J] p = p_angle .* p_force;
26 }
27 model {
28   raw_proportion ~ normal(p, sqrt(p .* (1-p) ./ to_vector(n) + sigma_obs^2))
29   ;
30 }
31 generated quantities {
32   real sigma_degrees = sigma_angle * 180 / pi();
33   vector[J] residual = raw_proportion - p;
34 }

```

LISTING 5. Modelo con error de medición.

Podríamos ajustar como lo hemos hecho antes, pero en este caso si tenemos problemas serios en el ajuste.

```

1 Model executable is up to date!
2 Running MCMC with 4 sequential chains...

```

```

3 Chain 1 finished in 0.9 seconds.
4 Chain 2 finished in 0.8 seconds.
5 Chain 3 finished in 0.7 seconds.
6 Chain 4 finished in 0.6 seconds.
7
8
9 All 4 chains finished successfully.
10 Mean chain execution time: 0.7 seconds.
11 Total execution time: 3.3 seconds.
12
13 Warning: 1891 of 4000 (47.0%) transitions ended with a divergence.
14 This may indicate insufficient exploration of the posterior distribution.
15 Possible remedies include:
16 * Increasing adapt_delta closer to 1 (default is 0.8)
17 * Reparameterizing the model (e.g. using a non-centered parameterization)
18 * Using informative or weakly informative prior distributions
19 variable      mean      median      sd      mad      q5      q95      rhat
20 1 sigma_angle 4.261e+307 3.169e+305      Inf 4.698e+305 0.01256 1.585e+308 2.238
21 2 sigma_obs   2.553e-01  2.049e-01 0.2264  2.696e-01 0.02627  5.450e-01 1.760
22 3 sigma_force 4.291e+307 4.381e+305      Inf 6.495e+305 0.07574 1.619e+308 2.153
23 ess_bulk ess_tail
24 1      6.014      171.6
25 2      6.171      136.5
26 3      6.112      97.3

```

Podemos incorporar información **débil** en los parametros de escala, esto es por medio de normales truncadas en la región positiva. El modelo completo sería

$$\sigma^2 \sim \mathcal{N}^+(0, 1) \quad (9a)$$

$$p_j^u = \Phi\left(\frac{114.52}{(x + 30.48)\sigma_f}\right) - \Phi\left(\frac{-30.48}{(x + 30.48)\sigma_f}\right), \quad (9b)$$

$$p_j^\theta = 2\Phi\left(\frac{\tan^{-1}((R-r)/x_j)}{\sigma_\theta}\right) - 1, \quad (9c)$$

$$p_j = p_j^u \cdot p_j^\theta, \quad (9d)$$

$$\frac{y_j}{n_j} \sim \mathcal{N}\left(p_j, \frac{p_j(1-p_j)}{n_j} + \sigma_{\text{obs}}^2\right), \quad (9e)$$

para $j = 1, \dots, J$, donde $\sigma^2 = (\sigma_{\text{obs}}^2, \sigma_\theta^2, \sigma_f^2)$.

```

1 data {
2   int J;
3   int n[J];
4   vector[J] x;
5   int y[J];
6   real r;
7   real R;
8   real overshoot;
9   real distance_tolerance;
10 }
11 transformed data {
12   vector[J] threshold_angle = atan((R-r) ./ x);
13   vector[J] raw_proportion = to_vector(y) ./ to_vector(n);
14 }
15 parameters {
16   real<lower=0> sigma_angle;
17   real<lower=0> sigma_force;

```



```

18   real<lower=0> sigma_obs;
19 }
20 transformed parameters {
21   vector[J] p_angle = 2*Phi(threshold_angle / sigma_angle) - 1;
22   vector[J] p_force = Phi((distance_tolerance - overshoot) ./ ((x + overshoot)
23     *sigma_force)) -
24     Phi((- overshoot) ./ ((x + overshoot)*sigma_force));
25   vector[J] p = p_angle .* p_force;
26 }
27 model {
28   raw_proportion ~ normal(p, sqrt(p .* (1-p) ./ to_vector(n) + sigma_obs^2))
29   ;
30   [sigma_angle, sigma_force, sigma_obs] ~ normal(0, 1);
31 }
32 generated quantities {
33   real sigma_degrees = sigma_angle * 180 / pi();
34   vector[J] residual = raw_proportion - p;
35 }

```

LISTING 6. Modelo completo con información débil.

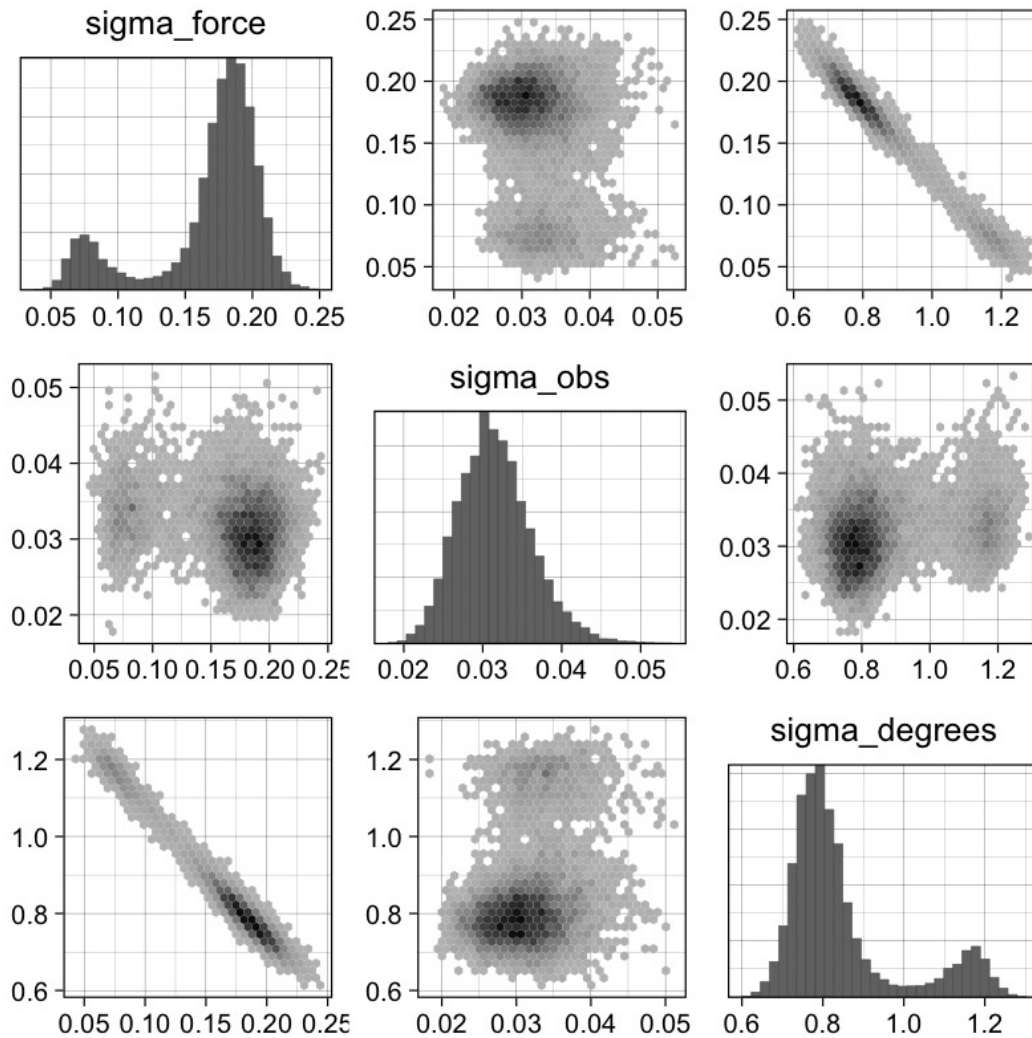
```

1 Model executable is up to date!
2 Running MCMC with 4 sequential chains...
3
4 Chain 1 finished in 2.6 seconds.
5 Chain 2 finished in 2.4 seconds.
6 Chain 3 finished in 2.4 seconds.
7 Chain 4 finished in 2.3 seconds.
8
9 All 4 chains finished successfully.
10 Mean chain execution time: 2.4 seconds.
11 Total execution time: 10.0 seconds.
12   variable      mean  median      sd      mad      q5      q95  rhat
13 1  sigma_angle  0.01481 0.01400 0.002499 0.001330 0.01229 0.02059 1.003
14 2  sigma_degrees 0.84874 0.80225 0.143164 0.076183 0.70402 1.17956 1.003
15 3  sigma_force  0.16709 0.18002 0.041690 0.023237 0.07159 0.21112 1.003
16 4  sigma_obs    0.03155 0.03123 0.004613 0.004471 0.02457 0.03965 1.001
17  ess_bulk ess_tail
18 1    921.8    1389
19 2    921.8    1389
20 3    926.6    1266
21 4   2450.5    4272

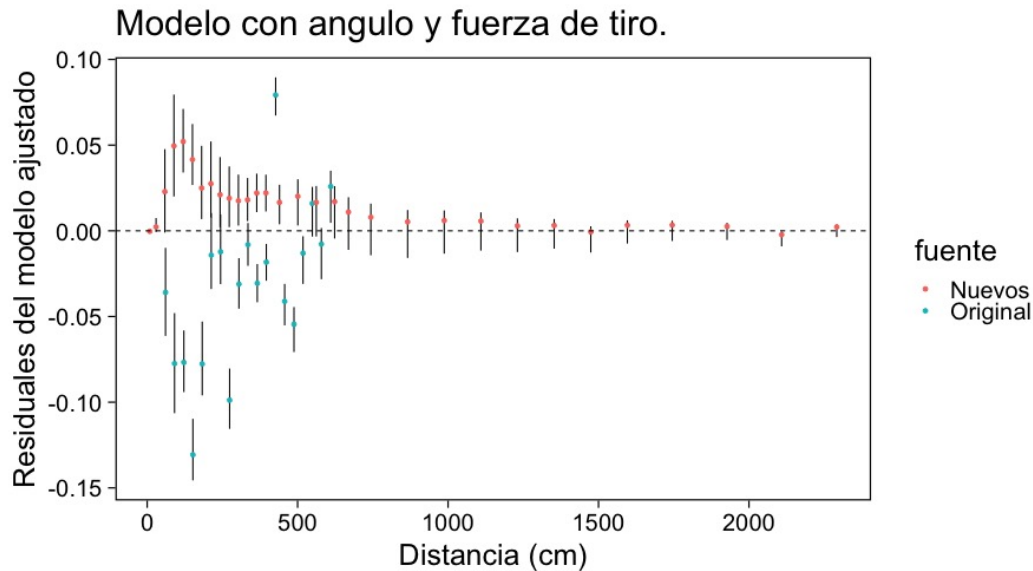
```

Los parámetros estimados los interpretamos como sigue:

- σ_θ tiene un valor cercano a 0.015 que corresponde a $\sigma_{\text{grados}} = 0.8$. De acuerdo a los datos obtenidos los jugadores de golf cometen errores de ángulo de **casi** un 1° . Si comparamos este valor con el de modelos anteriores podemos notar que al incluir errores de precisión en la fuerza de tiro ésta desviación disminuye. Ya no es necesario corregir con ángulos lo que se puede explicar de otra forma, esta correlación la podemos ver gráficamente por medio de un diagrama de dispersión como abajo.
- σ_f tiene un valor esperado de 0.17, lo cual implica un error del 17 % debido a la errores en distancia producto de la fuerza de tiro.
- σ_{obs} tiene un valor de 0.03 lo cual incide en errores atribuibles a medición del 3 puntos porcentuales.



La aparente bimodalidad de los gráficos de dispersión se podría explicar a través del efecto de tener mediciones de dos tipos. Un tipo son los datos originales en los que parece haber un número limitado de registrados, y las nuevas observaciones de Broadie que tienen un número muy grande observaciones a distintas distancias.



3.7.1. *Tarea:* Exploraremos algunas rutas de mejora del modelo.

1. Por un lado exploraremos eliminar uno de los componentes redundantes. Para esto elimina el supuesto de la fuerza de tiro y reajusta el modelo con la aproximación continua.
2. Incorpora un modelo jerárquico para ajustar el modelo que incorpore errores observacionales para las dos poblaciones de datos. Es decir, un modelo que tenga una $\sigma_{\text{obs},1}$ para los datos del primer conjunto de observaciones y $\sigma_{\text{obs},2}$ para los datos del segundo.
3. ¿Qué conclusiones obtienes?

4. MENSAJE

- Es fácil escribir modelos Bayesianos y hacer inferencia.
- Difícil mantener en producción: limitar el alcance del modelo.
- Reparametrización, previas informativas.
- El muestreo podría no escalar.

REFERENCIAS

- [1] A. Gelman. Model building and expansion for golf putting. <https://mc-stan.org/users/documentation/case-studies/golf.html>, 2019. [10](#), [18](#), [20](#)
- [2] A. Gelman and D. Nolan. A Probability Model for Golf Putting. *Teaching Statistics*, 24(3):93–95, 2002. ISSN 1467-9639. . [10](#)
- [3] A. Gelman, A. Vehtari, D. Simpson, C. C. Margossian, B. Carpenter, Y. Yao, L. Kennedy, J. Gabry, P.-C. Bürkner, and M. Modrák. Bayesian workflow. *arXiv preprint arXiv:2011.01808*, 2020. [1](#)
- [4] O. A. Martin, R. Kumar, and J. Lao. *Bayesian Modeling and Computation in Python*. Chapman and Hall/CRC, Boca Raton, first edition, nov 2021. ISBN 978-1-00-301916-9. . [1](#), [2](#)
- [5] A. R. Penner. The physics of putting. *Canadian Journal of Physics*, 80(2):83–96, feb 2002. ISSN 0008-4204, 1208-6045. . [13](#), [19](#)