

# Una primera aproximación a la Estadística Bayesiana

Asunción M. Mayoral y Javier Morales. IUI CIO-UMH

Noviembre 2022



# Contents

<b>1</b>	<b>Contexto</b>	<b>5</b>
1.1	Objetivos de aprendizaje . . . . .	5
1.2	Contenidos propuestos . . . . .	6
1.3	Contenidos definitivos . . . . .	6
1.4	INLA . . . . .	6
<b>2</b>	<b>Regresión lineal</b>	<b>9</b>
2.1	Introducción . . . . .	9
2.2	Variables y relaciones . . . . .	10
2.3	Verosimilitud . . . . .	11
2.4	Hiperparámetros . . . . .	12
2.5	Efectos fijos . . . . .	13
2.6	Resultados . . . . .	14
2.7	Distribuciones posteriores . . . . .	18
2.8	Simulación de la posterior . . . . .	21
2.9	Regresión lineal múltiple con INLA . . . . .	24
2.10	Conclusiones . . . . .	32
<b>3</b>	<b>Modelo de ANOVA</b>	<b>35</b>
3.1	Introducción . . . . .	35
3.2	El modelo de ANOVA . . . . .	35
3.3	Anova de una vía . . . . .	36
3.4	Anova de varias vías . . . . .	43

3.5	Análisis de ANCOVA . . . . .	47
3.6	Efectos aleatorios . . . . .	57
3.7	Modelos mixtos . . . . .	65
3.8	Conclusiones . . . . .	80
<b>4</b>	<b>Modelos lineales generalizados</b>	<b>81</b>
4.1	Modelos jerárquicos bayesianos . . . . .	82
4.2	Regresión logística . . . . .	83
4.3	Regresión de Poisson . . . . .	95
	<b>Referencias</b>	<b>105</b>

# Chapter 1

## Contexto

El curso “Una primera aproximación a la Estadística Bayesiana”, con una duración de 10 horas, está ofertado a alumnado de doctorado con una formación científica en el ámbito BIO.

El curso se desarrollará íntegramente online a través de videoconferencia síncrona, durante un total de 10 horas, seccionadas en 4 sesiones de dos horas y media cada una de ellas, los días 8, 10, 15 y 17 de noviembre de 2022, de 16 a 18:30h.

Se presentarán los contenidos a través de ejemplos prácticos programados en R, para que el estudiantado pueda ir generando resultados y comentando las interpretaciones derivadas del análisis.

La evaluación es continua basada en la interacción con el profesorado durante las sesiones de trabajo.

### 1.1 Objetivos de aprendizaje

- Conocer los conceptos básicos en el planteamiento bayesiano de la Estadística.
- Identificar la relevancia de la información previa y de expertos, y la proporcionada por los datos.
- Conocer los procedimientos básicos para conjugar la información disponible.
- Aplicar los conocimientos básicos en problemas sencillos.
- Descubrir las dificultades computacionales en la inferencia bayesiana.

## 1.2 Contenidos propuestos

1. De probabilidad va la historia: la relevancia de las probabilidades condicionadas y el teorema de Bayes.
2. La jerga base: incertidumbre, a priori, a posteriori y verosímil.
3. Manos en la masa 1: ¿con qué probabilidad ocurrió?
4. Manos en la masa 2: ¿con qué abundancia ocurrió?
5. Curioseando para saber más: manuales y software.

## 1.3 Contenidos definitivos

1. SESIÓN 1: Probabilidades, Bayes y las proporciones.
2. SESIÓN 2: INLA y la regresión.
3. SESIÓN 3: INLA y el ANOVA.
4. SESIÓN 4: INLA y los GLM.

## 1.4 INLA

INLA es una librería de R que aproxima la inferencia Bayesiana para modelos gaussianos latentes (LGM). Sus siglas provienen de Integrated Nested Laplace Approximation (INLA), que es un método para aproximar las inferencias bayesianas a través de la aproximación de Laplace.

Aunque la metodología INLA se ha desarrollado sobre modelos que se pueden expresar como campos aleatorios markovianos gaussianos (*Gaussian Markov random fields*, *GMRF*), es viable para una gran familia de modelos habituales en la práctica estadística.

Disponemos de referencias múltiples y documentación de esta librería en la web [r-inla.org](http://r-inla.org), y en particular en el manual de referencia de Gómez-Rubio (2021) titulado *Bayesian inference with INLA*, también publicado por Chapman & Hall-CRC Press.

### 1.4.1 Instalación

Para instalar la librería INLA hemos de ejecutar, desde R, el comando

```
install.packages("INLA", repos=c(getOption("repos"),
                                INLA="https://inla.r-inla-download.org/R/stable"),
                dep=TRUE)
# y a continuación la cargamos con:
library(INLA)
```

Para instalar actualizaciones, basta con ejecutar

```
options(repos = c(getOption("repos"),
                    INLA="https://inla.r-inla-download.org/R/testing"))
update.packages("INLA", dep=TRUE)
```

Las descargas y documentación completa sobre INLA está disponible en R-INLA home.

Ya desde R, para pedir ayuda sobre funciones en INLA, basta usar el comando `inla.doc()`, especificando dentro y entrecomillada, la función/objeto sobre el que se solicita ayuda. Por ejemplo, `inla.doc("ar1")` o `inla.doc("loggamma")`.

También utilizaremos una librería accesoria de INLA, `brinla`, desarrollada por Faraway, Yue y Wan, 2022.

```
install.packages("remotes")
remotes::install_github("julianfaraway/brinla")
```

### 1.4.2 Fundamentos

INLA está basado en la resolución de integrales vía la aproximación de Laplace, que aproxima el integrando a través de una expansión de Taylor de segundo grado que permite calcular la integral analíticamente.

$$\begin{aligned} I_n &= \int_x \exp[nf(x)] dx \\ &\approx \int_x \exp[n(f(x_0) + 1/2(x - x_0)^2 f''(x_0))] dx \\ &= \exp[nf(x_0)] \cdot \sqrt{\frac{2\pi}{-nf''(x_0)}} \end{aligned}$$

Evita así los largos tiempos de simulación de las cadenas de Markov Monte Carlo. Cuando las distribuciones a integrar son Gaussianas, Laplace da órdenes buenos de aproximación. Y este es el principio que usa para modelizar la mayoría de los modelos habituales, que se integran dentro de la amplia clase de los modelos gaussianos latentes, en los que se aplica INLA.

INLA se ha aplicado en mapeo estadístico, modelos de cohorte multidimensionales, modelos de asociación espacial, genética, análisis medioambientales, salud y epidemiología, dinámicas de infecciones, estudios agronómicos, meta-análisis, impacto del cambio climático y muchos más ámbitos (Rue et al, 2017).





## Chapter 2

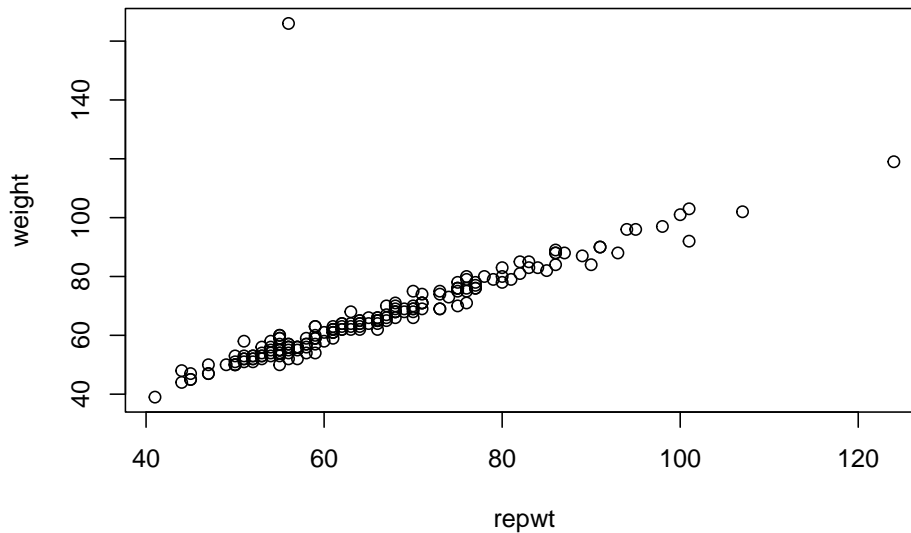
# Regresión lineal

### 2.1 Introducción

Una vez presentados los fundamentos de INLA vamos a utilizarlo para trabajar progresivamente desde los modelos más sencillos a los más sofisticados. Empezamos aquí con el modelo de regresión lineal simple, para continuar generalizando con el de regresión lineal múltiple.

Partimos de la base de datos Davis (en la librería `carData`), que contiene 200 registros de 5 variables relacionadas con un estudio sobre habituación de hombres y mujeres a la realización de ejercicio físico de forma regular. Las variables que se registraron son sexo, peso y altura (reales y reportados). Vamos a indagar la relación entre el peso real (`weight`) y el reportado (`repwt`) a través de un análisis de regresión lineal simple.

```
data(Davis, package="carData")
plot(weight ~ repwt, data=Davis)
```



```
# Excluimos el valor outlier y los NA
davis=Davis %>%
  filter(weight<160) %>%
  slice(which(!is.na(repwt)))
```

## 2.2 Variables y relaciones

Entendemos como variable respuesta  $y=\text{weight}$ , de tipo numérico (continua), y como variable explicativa o covariable,  $x=\text{repwt}$ , también numérica.

La especificación de respuesta  $y$  y predictores  $x_1, x_2, \dots$ , en INLA se registra en una fórmula del tipo:

```
formula= y ~ 1 + x_1 + x_2 +...
```

en la que podemos prescindir del 1 que identifica la interceptación, pues el ajuste, por defecto, siempre se resolverá con su estimación, salvo que en su lugar se escriba un '-1'.

En nuestro problema tendríamos pues:

```
formula = weight ~ repwt
```

A continuación es procedente elegir el modelo sobre la respuesta, o lo que es lo mismo, la verosimilitud.

## 2.3 Verosimilitud

En principio es razonable asumir normalidad en la respuesta, además de independencia entre todas las observaciones. Así, el modelo propuesto para la respuesta es:

$$(y_i|\mu_i, \sigma^2) \sim N(\mu_i, \sigma^2), i = 1, \dots, n$$

donde  $y = \text{weight}$  y  $x = \text{repwt}$ , e  $i = 1, \dots, n$  es un subíndice que identifica a cada uno de los registros disponibles en el banco de datos. La media esperada contiene la relación lineal entre covariable y respuesta,  $\mu_i = \theta + \beta x_i$ , donde  $\theta$  es la interceptación de la recta de regresión y  $\beta$  el coeficiente que explica la relación lineal entre  $x$  e  $y$ . El vector  $(\theta, \beta)$  identifica los **efectos latentes**, en cuya inferencia posterior estamos interesados, y que están involucrados directamente en la media o predictor lineal. El modelo, o lo que es lo mismo, la verosimilitud, depende también de un parámetro de dispersión  $\sigma^2$  sobre el que también queremos inferir.

Veamos cómo especificar este modelo con INLA. La función `names(inla.models())` proporciona un listado de todos los tipos de modelos posibles, tanto para los datos (`likelihood`), para los efectos latentes (`latent`), los parámetros (`prior`), y otras opciones que de momento no nos interesan. El listado completo de todas las distribuciones disponibles para cada uno de los tipos de modelos lo obtenemos con el comando `inla.list.models()`. En particular, si ejecutamos `names(inla.models())$likelihood`, obtenemos todas las distribuciones disponibles para modelizar los datos.

La distribución **gaussian** identifica la distribución normal que hemos planteado en nuestro modelo de regresión lineal. Para obtener información sobre cómo está parametrizada y cuáles son las priors por defecto, basta consultar la documentación *gaussian* con el comando:

```
# documentación (parametrización y priors) del modelo normal
inla.doc("gaussian")
```

Para ajustar un modelo sencillo en INLA hay que echar mano de la función `inla`, en la que introducimos en primer lugar la **formula**, con la relación entre las variables, a continuación el modelo, en el argumento **family**, y después el banco de datos sobre el que trabajamos. Si no especificamos el argumento **family**, la función `inla` interpreta por defecto la opción **gaussian**, esto es, normalidad para los datos, de modo que podríamos excluir dicha especificación cuando modelizamos datos normales.

```
# Asumiendo datos normales
fit=inla(formula,family="gaussian", data)
```

```
# equivalente a
fit=inla(formula, data)
```

Adelantamos pues un paso más en nuestro problema, añadiendo la verosimilitud normal y la base de datos.

```
# ajuste del modelo
formula = weight ~ 1+ repwt
fit=inla(formula,family="gaussian",data=davis)
```

## 2.4 Hiperparámetros

INLA identifica como hiperparámetros todos aquellos parámetros en el modelo que no se corresponden con efectos latentes, esto es, relacionados con el predictor o respuesta esperada. En nuestro modelo, el único hiperparámetro es la varianza  $\sigma^2$ , sobre la que es preciso especificar una distribución a priori. Para la varianza  $\sigma^2$  es habitual asumir una gamma inversa difusa, con media y varianza grandes.

En INLA, en lugar de asignar distribuciones a priori sobre las varianzas, se hace sobre el logaritmo de las precisiones, para facilitar el cálculo del máximo de la log-posterior (obtenida de la log-verosimilitud y la log-prior). Así, asumir una gamma inversa difusa  $GaI(\alpha, \beta)$  para la varianza es equivalente a una Gamma difusa  $Ga(\alpha, \beta)$  para la precisión  $\tau = 1/\sigma^2$ , y una log-gamma difusa  $Log - Gamma(\alpha, \beta)$  para la log-precisión  $\log(\tau)$ .

Por defecto, como ya verificamos en la documentación de la verosimilitud gaussiana, (con `inla.doc("gaussian")`), la distribución a priori por defecto sobre la log-precisión ( $\theta_1$  en la ayuda) es la log-gamma difusa  $LogGa(1, 5 \cdot 10^{-5})$ , lo que da un valor esperado para la precisión  $\tau$  de 20000 y una varianza de  $4 \cdot 10^8$ .

Para modificar la distribución a priori de un parámetro podemos utilizar cualquiera de las distribuciones que ofrece INLA en su listado `names(inla.models())$prior` (siempre buscando coherencia con la información sobre el parámetro en cuestión).

Para definir una prior para los parámetros o hiperparámetros en INLA hay que definir los siguientes argumentos:

- `prior`, el nombre de la distribución a priori (para hiperparámetros, alguna de las opciones en `names(inla.models())$prior`)
- `param`, los valores de los parámetros de la prior
- `initial`, el valor inicial para el hiperparámetro
- `fixed`, variable booleana para decir si el hiperparámetro es fijo o aleatorio.

La modificación la haremos con el argumento `control.family=list(hyper=list(...))` en la función `inla`, al que le proporcionaremos una lista con el nombre de los parámetros (el *short name* con el que los identifica INLA en la documentación), que apunta a una lista con la distribución (*prior*) y los parámetros (*param*) a utilizar.

En nuestro problema, si tenemos información previa sobre la precisión del modelo, reconocida como `prec` (short name), y queremos especificar una a priori  $Ga(1, 0.001)$  para la precisión, habremos de utilizar la siguiente sintaxis:

```
prec.info = list(prior="loggamma", param=c(1,0.001))
fit2=inla(formula,family="gaussian",data=davis,
          control.family = list(hyper = list(prec = prec.info)))
```

Si nos conformamos con la previa por defecto de INLA, el modelo que estamos asumiendo en nuestro problema de regresión lineal simple será:

$$\begin{aligned} (y_i | \mu_i, \sigma^2) &\sim N(\mu_i, \sigma^2), i = 1, \dots, n \\ \tau = 1/\sigma^2 &\sim Ga(1, 0.00005) \end{aligned}$$

## 2.5 Efectos fijos

Una variable explicativa entra en el modelo como **efecto fijo** cuando se piensa que afecta a todas las observaciones del mismo modo, y que su efecto es de interés primario en el estudio.

En un modelo de regresión lineal todos los efectos latentes en el predictor lineal, interceptación y coeficientes de covariables, son efectos fijos. Ante ausencia de información, las priors para los efectos fijos, esto es,  $(\theta, \beta)$  en nuestro caso, se asumen normales centradas en cero y con varianzas grandes. En INLA la interceptación  $\beta$  tiene por defecto una prior gaussiana con media y precisión igual a cero (una distribución plana objetiva, que no integra 1), y los coeficientes  $\beta$  también tienen una prior gaussiana con media cero y precisión igual a 0.001. Estos valores por defecto se pueden consultar con el comando `inla.set.control.fixed.default()`, que da la siguiente información:

- `mean=0` y `mean.intercept=0` son las medias de la distribución normal para los coeficientes  $\beta$  y la interceptación  $\theta$  respectivamente
- `prec=0.001` y `prec.intercept=0` son las precisiones respectivas de las normales para  $\beta$  y  $\theta$ .

Con todo, los parámetros de las priors sobre los efectos fijos se pueden modificar a través del argumento `control.fixed=list(...)` en la función `inla`,

utilizando siempre los nombres que atribuye INLA a los diferentes parámetros e hiperparámetros (*short name*). Por ejemplo, si queremos modificar la precisión de los efectos fijos para hacerla igual a 0.001 (esto es, varianza 1000), utilizamos la siguiente sintaxis:

```
formula = weight ~ 1+ repwt
fit=inla(formula,family="gaussian",data=davis,
        control.fixed=list(prec=0.001,prec.intercept=0.001))
```

Volvemos sobre nuestro ejemplo, y asumiendo las a priori por defecto de INLA tendremos:

$$\begin{aligned} (y_i|\theta, \beta, \sigma^2) &\sim N(\theta + \beta x_i, \sigma^2), i = 1, \dots, n \\ \theta &\sim N(0, \infty) \\ \beta &\sim N(0, 1000) \\ \tau = 1/\sigma^2 &\sim Ga(1, 0.00005) \end{aligned}$$

## 2.6 Resultados

Para mostrar una descriptiva de los resultados del ajuste obtenido con `fit=inla(...)`, utilizamos la sintaxis siguiente:

- `summary(fit)` proporciona una descriptiva del ajuste
- `names(fit$marginals.fixed)` lista los nombres de todos los efectos fijos
- `fit$summary.fixed` resume la inferencia posterior sobre los efectos fijos
- `names(fit$marginals.hyperpar)` lista los nombres de todos los hiperparámetros
- `fit$summary.hyperpar` da un resumen de la inferencia posterior de los parámetros e hiperparámetros
- `fit$summary.fitted.values` resume la inferencia posterior sobre los valores ajustados
- `fit$mlik` da la estimación de la log-verosimilitud marginal, útil para evaluar y comparar modelos.

Veamos los resultados para nuestro problema de regresión.

```
# ajuste del modelo
formula = weight ~ 1+ repwt
fit=inla(formula,family="gaussian",data=davis)
summary(fit)
#>
#> Call:
```

```

#> c("inla.core(formula = formula, family = family,
#> contrasts = contrasts, ", " data = data, quantiles =
#> quantiles, E = E, offset = offset, ", " scale =
#> scale, weights = weights, Ntrials = Ntrials, strata =
#> strata, ", " lp.scale = lp.scale, link.covariates =
#> link.covariates, verbose = verbose, ", " lincomb =
#> lincomb, selection = selection, control.compute =
#> control.compute, ", " control.predictor =
#> control.predictor, control.family = control.family,
#> ", " control.inla = control.inla, control.fixed =
#> control.fixed, ", " control.mode = control.mode,
#> control.expert = control.expert, ", " control.hazard
#> = control.hazard, control.lincomb = control.lincomb,
#> ", " control.update = control.update,
#> control.lp.scale = control.lp.scale, ", "
#> control.pardiso = control.pardiso, only.hyperparam =
#> only.hyperparam, ", " inla.call = inla.call, inla.arg
#> = inla.arg, num.threads = num.threads, ", "
#> blas.num.threads = blas.num.threads, keep = keep,
#> working.directory = working.directory, ", " silent =
#> silent, inla.mode = inla.mode, safe = FALSE, debug =
#> debug, ", " .parent.frame = .parent.frame)")
#> Time used:
#> Pre = 2.56, Running = 0.183, Post = 0.0186, Total = 2.76
#> Fixed effects:
#>          mean      sd 0.025quant 0.5quant 0.975quant mode
#> (Intercept) 2.734 0.814      1.135    2.734      4.332  NA
#> repwt      0.958 0.012      0.935    0.958      0.982  NA
#>          kld
#> (Intercept) 0
#> repwt      0
#>
#> Model hyperparameters:
#>                                mean      sd
#> Precision for the Gaussian observations 0.199 0.021
#>                                0.025quant 0.5quant
#> Precision for the Gaussian observations      0.161    0.198
#>                                0.975quant mode
#> Precision for the Gaussian observations      0.242    NA
#>
#> Marginal log-Likelihood: -426.74
#> is computed
#> Posterior summaries for the linear predictor and the fitted values are computed
#> (Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
fit$mlik

```

```
#>                                     [,1]
#> log marginal-likelihood (integration) -426.7310
#> log marginal-likelihood (Gaussian)    -426.7365
```

Obtenemos pues la salida con un descriptivo de la distribución posterior para los efectos fijos interceptación,  $\theta$ , y el coeficiente del regresor `repwt`,  $\beta$ , con la media, desviación típica y cuantiles con los que podemos evaluar la región creíble al 95%.

También muestra a continuación una tabla con los descriptivos de la distribución posterior para la precisión  $\tau = 1/\sigma^2$  de los datos.

Si queremos obtener los nombres con los que INLA reconoce los efectos fijos (interceptación y coeficiente del regresor) e hiperparámetros (precisión de los datos), llamamos a

```
names(fit$marginals.fixed)
#> [1] "(Intercept)" "repwt"
names(fit$marginals.hyperpar)
#> [1] "Precision for the Gaussian observations"
```

Y podemos pedir descriptivos específicos de las distribuciones de los efectos fijos y de los hiperparámetros.

```
# descriptivos efectos fijos
fit$summary.fixed
#>               mean          sd 0.025quant  0.5quant
#> (Intercept) 2.7338080 0.81425458  1.1351514  2.7338080
#> repwt       0.9583742 0.01213469  0.9345497  0.9583742
#>               0.975quant mode          kld
#> (Intercept)  4.3324649   NA 6.193798e-10
#> repwt       0.9821987   NA 6.194004e-10
# descriptivos varianza
fit$summary.hyperpar
#>                                     mean
#> Precision for the Gaussian observations 0.1991122
#>                                     sd
#> Precision for the Gaussian observations 0.02087927
#>                                     0.025quant
#> Precision for the Gaussian observations 0.1609666
#>                                     0.5quant
#> Precision for the Gaussian observations 0.1983843
#>                                     0.975quant mode
#> Precision for the Gaussian observations 0.2418715   NA
# medias de los efectos fijos
```



```
fit$summary.fixed$mean
#> [1] 2.7338080 0.9583742
# descriptivos primer efecto fijo
fit$summary.fixed[1,]
#>               mean          sd 0.025quant 0.5quant
#> (Intercept) 2.733808 0.8142546   1.135151 2.733808
#>               0.975quant mode          kld
#> (Intercept)  4.332465   NA 6.193798e-10
```

Para describir la marginal posterior sobre cada uno de los datos ajustados:

```
head(fit$summary.fitted.values)
#>               mean          sd 0.025quant 0.5quant
#> fitted.Predictor.001 76.52862 0.2162352   76.10412 76.52862
#> fitted.Predictor.002 51.61089 0.2441104   51.13167 51.61089
#> fitted.Predictor.003 54.48602 0.2189718   54.05614 54.48602
#> fitted.Predictor.004 69.82000 0.1750074   69.47644 69.82000
#> fitted.Predictor.005 59.27789 0.1855715   58.91358 59.27789
#> fitted.Predictor.006 75.57025 0.2087339   75.16047 75.57025
#>               0.975quant mode
#> fitted.Predictor.001  76.95312   NA
#> fitted.Predictor.002  52.09012   NA
#> fitted.Predictor.003  54.91589   NA
#> fitted.Predictor.004  70.16357   NA
#> fitted.Predictor.005  59.64219   NA
#> fitted.Predictor.006  75.98002   NA
```

Si queremos hacer un **análisis de sensibilidad** sobre las distribuciones a priori, reajustamos el modelo con otras priors y comparamos los resultados.

```
# ajuste del modelo
formula = weight ~ 1+ repwt
fit2=inla(formula,family="gaussian",data=davis,
          control.fixed = list(mean.intercept = 100,
                               prec.intercept = 0.001,
                               prec = 0.001),
          control.family = list(hyper = list(
                                prec = list(prior="loggamma", param =c(1,0.001)))))
fit2$summary.fixed
#>               mean          sd 0.025quant 0.5quant
#> (Intercept) 2.7982758 0.81412303  1.2008593 2.7979330
#> repwt      0.9574339 0.01213288  0.9335989 0.9574389
#>               0.975quant mode          kld
#> (Intercept)  4.3976369   NA 6.301531e-10
```

```
#> repwt      0.9812405    NA 6.293895e-10
fit2$summary.hyperpar
#>
#> Precision for the Gaussian observations 0.1990744
#>
#> Precision for the Gaussian observations 0.02084352
#>
#> Precision for the Gaussian observations 0.1609125
#>
#> Precision for the Gaussian observations 0.1983649
#>
#> Precision for the Gaussian observations 0.241728    NA
```

## 2.7 Distribuciones posteriores

Para obtener la distribución marginal de los valores ajustados y predichos necesitamos incorporar a la función `inla` el argumento `control.compute=list(return.marginals.predictor=TRUE)`.

Tras conseguir un ajuste con `inla`, podemos acceder a todas las distribuciones marginales posteriores y predictivas a través de:

- `fit$marginals.fixed` da las distribuciones posteriores marginales de los efectos fijos
- `fit$marginals.fixed$xx` da la distribución del efecto fijo `xx`, y también se puede seleccionar con su ordinal en el conjunto de efectos fijos `fit$marginals.fixed[[i]]`
- `fit.marginals.hyperpar` da las distribuciones posteriores marginales de los parámetros e hiperparámetros
- `fit$marginals.fitted.values` da las distribuciones posteriores marginales para los valores ajustados

Con estas distribuciones, reconocidas como `marginal`, podemos hacer cálculos y gráficos de interés a través de estas funciones que operan sobre las distribuciones y que podemos consultar con `?inla.marginal`:

- `inla.dmarginal(x, marginal, ...)` da la densidad en `x`
- `inla.pmarginal(q, marginal, ...)` da las probabilidades o función de distribución
- `inla.qmarginal(p, marginal, ...)` da los cuantiles
- `inla.rmarginal(n, marginal)` permite obtener  $n$  simulaciones
- `inla.hpdmarginal(p, marginal, ...)` da la región HPD
- `inla.smarginal(marginal, ...)` da un suavizado con splines de la distribución marginal

- `inla.emarginal(fun, marginal, ...)` calcula el valor esperado de una función
- `inla.mmarginal(marginal, ...)` calcula la moda posterior
- `inla.tmarginal(fun, marginal, ...)` transforma la distribución marginal de una función del parámetro
- `inla.zmarginal(marginal, ...)` calcula descriptivos de la marginal.

Veamos algunos ejemplos sobre nuestro problema. Vamos a mostrar a continuación, en un único gráfico, las distribuciones posteriores de los efectos fijos y el parámetro de dispersión de los datos  $\sigma^2$ , con líneas verticales que marquen el valor esperado posterior (en azul) y el HPD95% en rojo.

```
# library(gridExtra)
# library(ggplot2)

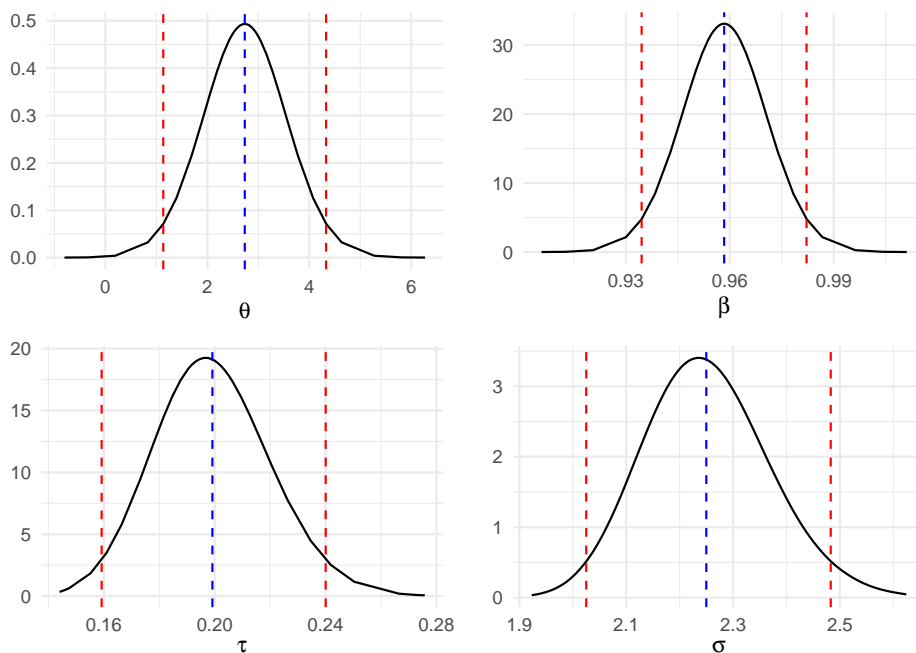
g=list() # lista en que almacenamos los gráficos con d.posteriores
# Efectos fijos
names.fixed=names(fit$marginals.fixed)
n.fixed=length(names.fixed)
names=c(expression(theta),expression(beta))
for (i in 1:n.fixed){
  g[[i]] = ggplot(as.data.frame(fit$marginals.fixed[[i]])) +
    geom_line(aes(x = x, y = y)) +
    labs(x=names[i],y="")+
    geom_vline(xintercept=inla.hpdmarginal(0.95,fit$marginals.fixed[[i]]),
              linetype="dashed",color="red")+
    geom_vline(xintercept=inla.emarginal(function(x) x,fit$marginals.fixed[[i]]),
              linetype="dashed",color="blue")
}

# Parámetros
g[[3]]= ggplot(as.data.frame(fit$marginals.hyperpar[[1]])) +
  geom_line(aes(x = x, y = y)) +
  labs(x=expression(tau),y="")+
  geom_vline(xintercept=inla.hpdmarginal(0.95,fit$marginals.hyperpar[[1]]),
            linetype="dashed",color="red")+
  geom_vline(xintercept=inla.emarginal(function(x) x,fit$marginals.hyperpar[[1]]),
            linetype="dashed",color="blue")

# Transformamos la posterior en tau para obtener la posterior de sigma
sigma.post=inla.tmarginal(function(tau) tau^(-1/2),
  fit$marginals.hyperpar[[1]])
# y la pintamos
g[[4]]= ggplot(as.data.frame(sigma.post)) +
```

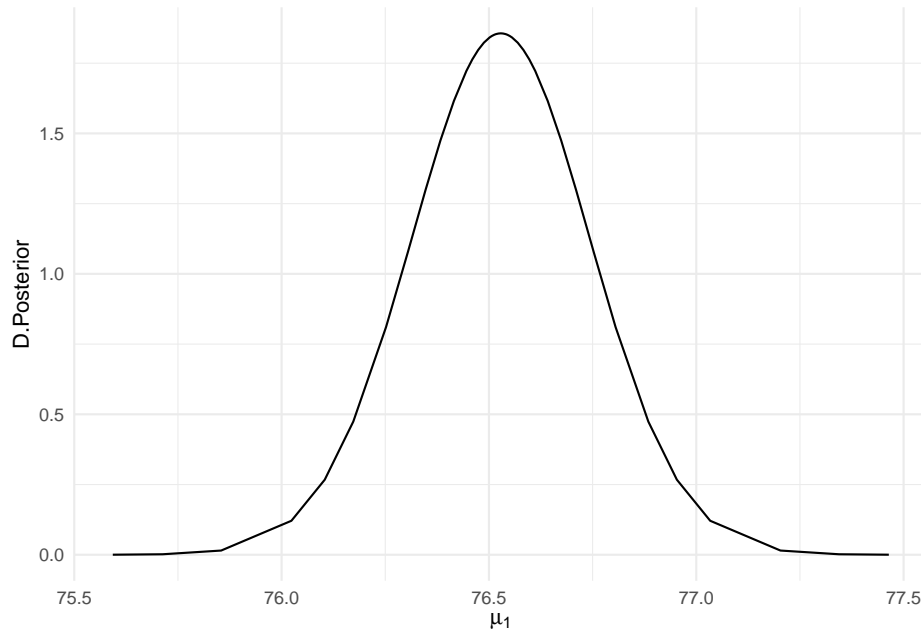
```
geom_line(aes(x = x, y = y)) +
labs(x=expression(sigma),y="")+
geom_vline(xintercept=inla.hpdmarginal(0.95,sigma.post),
           linetype="dashed",color="red")+
geom_vline(xintercept=inla.emarginal(function(x) x,sigma.post),
           linetype="dashed",color="blue")

library(gridExtra)
grid.arrange(g[[1]],g[[2]],g[[3]],g[[4]],ncol=2)
```



Si queremos la distribución posterior de alguna de las medias  $\mu_i = \theta + \beta x_i$ , necesitamos añadir el argumento `control.compute=list(return.marginals.predictor=TRUE)` en `inla`. Así podremos representar, por ejemplo, la distribución posterior sobre el peso esperado para el individuo que aparece en el registro 1 de la base de datos,  $\mu_1 = \theta + \beta x_1$ :

```
fit=inla(formula,family="gaussian",data=davis,
         control.compute=list(return.marginals.predictor=TRUE))
ggplot(as.data.frame(fit$marginals.fitted.values[[1]]),aes(x=x,y=y))+
  geom_line()+
  labs(x=expression(mu[1]),y="D.Posterior")
```



## 2.8 Simulación de la posterior

Cuando queremos inferir sobre funciones de los efectos latentes e hiperparámetros que no proporciona INLA por defecto, podemos recurrir a simular de las distribuciones posteriores de los efectos involucrados, y con ellas evaluar la función que nos interesa, para conseguir una muestra de su distribución posterior.

Para ello es preciso que al ajustar el modelo con `inla` hayamos incluido el argumento `control.compute=list(config=TRUE)`.

Para obtener simulaciones de las correspondientes distribuciones posteriores, utilizamos las funciones:

- `inla.posterior.sample(n, fit, selection)`, para simular los efectos latentes, donde  $n$  es el número de simulaciones pretendido, `fit` es el ajuste obtenido con `inla` y `selection` es una lista con el nombre de las componentes (efectos latentes) a simular.
- `inla.hyperpar.sample(n, fit, improve.marginals=TRUE)` para simular de parámetros e hiperparámetros.

Para describir las distribuciones de los nuevos parámetros que queremos evaluar con dichas simulaciones, utilizaremos:

- `inla.posterior.eval()` para funciones sobre los efectos fijos
- `inla.hyperpar.eval()` para funciones sobre los hiperparámetros

Imaginemos que queremos obtener la distribución posterior de  $(\theta + \beta \cdot 50)$ , que correspondería con el peso real de un sujeto que ha declarado un peso de 50kg. Hemos de simular pues, de las distribuciones posteriores de  $\theta$  y de  $\beta$ , para luego aplicar la función correspondiente sobre las simulaciones y obtener una muestra posterior de  $\theta + \beta \cdot 50$ .

```
# reajustamos para poder simular de las posteriores
fit <- inla(formula, data = davis,
  control.compute = list(config = TRUE))
# simulamos especificando los parámetros en los que tenemos interés
sims = inla.posterior.sample(100, fit, selection = list("(Intercept)"=1, repwt = 1))
```

Esto nos devuelve una lista de la dimensión del número de simulaciones (cada simulación en un elemento de la lista), y en cada uno de los elementos tenemos los valores simulados de los hiperparámetros (`hyper`), de los efectos latentes (`latent`)

```
length(sims)
#> [1] 100
names(sims[[1]])
#> [1] "hyperpar" "latent" "logdens"
sims[[1]]$hyperpar
#> Precision for the Gaussian observations
#> 0.2539011
sims[[1]]$latent
#> sample:1
#> (Intercept):1 2.6504151
#> repwt:1 0.9604172
```

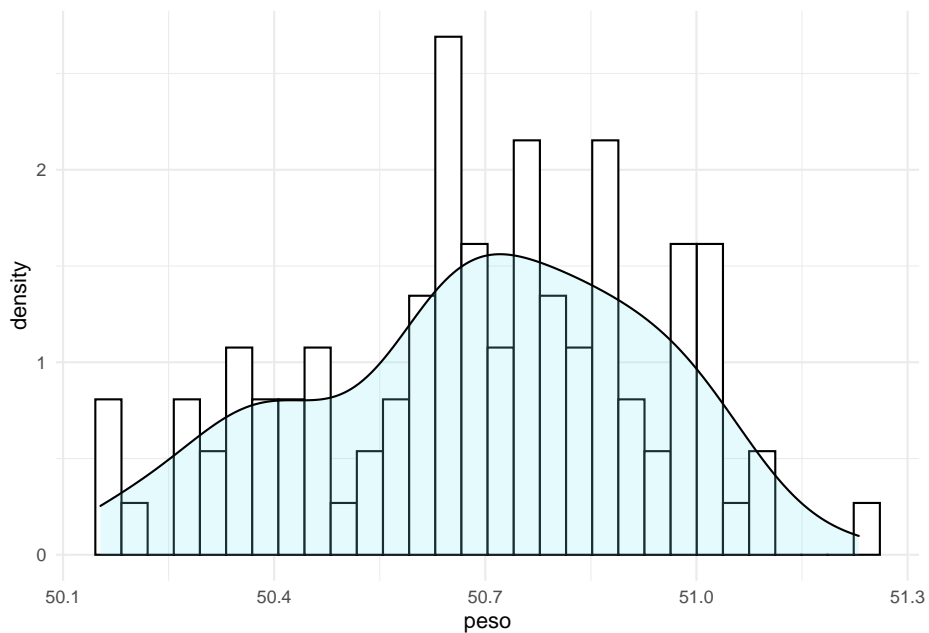
y la log-densidad de la posterior en esos valores (`logdens`)

```
sims[[1]]$logdens
#> $hyperpar
#> [1] -0.1085208
#>
#> $latent
#> [1] 1007.92
#>
#> $joint
#> [1] 1007.812
```

Ahora con la función `inla.posterior.sample.eval`, dado que nuestra función depende de efectos fijos (latentes),  $(\theta, \beta)$ , evaluamos la operación pretendida, y con descriptivos gráficos y numéricos de las simulaciones, podemos aproximar los descriptivos de la distribución posterior.

```
peso_real=inla.posterior.sample.eval(function(...) {(Intercept)+repwt*50},sims)
peso_real
#>      sample:1 sample:2 sample:3 sample:4 sample:5
#> fun[1] 50.67127 50.65628 50.86209 50.8309 50.47244
#>      sample:6 sample:7 sample:8 sample:9 sample:10
#> fun[1] 50.73311 50.41425 50.75967 50.38832 50.93032
#>      sample:11 sample:12 sample:13 sample:14 sample:15
#> fun[1] 50.7573 50.67544 50.88738 50.68311 50.83397
#>      sample:16 sample:17 sample:18 sample:19 sample:20
#> fun[1] 50.98866 50.77479 50.61694 50.91848 50.9645
#>      sample:21 sample:22 sample:23 sample:24 sample:25
#> fun[1] 50.79662 50.76477 50.76447 51.09895 50.80277
#>      sample:26 sample:27 sample:28 sample:29 sample:30
#> fun[1] 50.28143 51.00519 50.88128 50.77139 50.18042
#>      sample:31 sample:32 sample:33 sample:34 sample:35
#> fun[1] 50.63997 50.3894 50.96784 50.35427 50.89216
#>      sample:36 sample:37 sample:38 sample:39 sample:40
#> fun[1] 50.76451 50.65633 50.78713 50.68749 50.71719
#>      sample:41 sample:42 sample:43 sample:44 sample:45
#> fun[1] 50.61936 50.39567 50.33132 50.31954 50.64171
#>      sample:46 sample:47 sample:48 sample:49 sample:50
#> fun[1] 50.8068 51.06663 50.34992 50.69365 50.56718
#>      sample:51 sample:52 sample:53 sample:54 sample:55
#> fun[1] 50.33738 50.72213 50.66545 50.63032 51.01036
#>      sample:56 sample:57 sample:58 sample:59 sample:60
#> fun[1] 50.84434 50.66175 50.17574 50.34754 51.00349
#>      sample:61 sample:62 sample:63 sample:64 sample:65
#> fun[1] 50.88114 50.91519 51.03443 50.64541 50.78457
#>      sample:66 sample:67 sample:68 sample:69 sample:70
#> fun[1] 50.83068 50.65106 50.9818 50.57738 50.28929
#>      sample:71 sample:72 sample:73 sample:74 sample:75
#> fun[1] 50.64571 50.53885 50.96379 50.46517 50.94165
#>      sample:76 sample:77 sample:78 sample:79 sample:80
#> fun[1] 51.08717 51.02017 50.27234 50.76453 50.61683
#>      sample:81 sample:82 sample:83 sample:84 sample:85
#> fun[1] 50.88356 50.67251 50.52058 50.62551 50.56148
#>      sample:86 sample:87 sample:88 sample:89 sample:90
#> fun[1] 50.40734 51.23069 50.87341 50.44473 51.01426
#>      sample:91 sample:92 sample:93 sample:94 sample:95
#> fun[1] 50.99633 50.48536 50.8826 50.87775 50.41319
#>      sample:96 sample:97 sample:98 sample:99 sample:100
```

```
#> fun[1] 50.59677 50.44639 50.72558 50.15308 50.20205
pred=data.frame(peso=as.vector(peso_real))
summary(pred)
#>      peso
#> Min.   :50.15
#> 1st Qu.:50.53
#> Median :50.71
#> Mean   :50.69
#> 3rd Qu.:50.88
#> Max.   :51.23
ggplot(pred,aes(x=peso))+
  geom_histogram(aes(y=..density..), colour="black", fill="white")+
  geom_density(alpha=.2, fill="#80E7F5")
#> `stat_bin()` using `bins = 30`. Pick better value with
#> `binwidth`.
```



## 2.9 Regresión lineal múltiple con INLA

Vemos a continuación cómo se trabaja la regresión múltiple con INLA, simplemente añadiendo más predictores.

El modelo de regresión asume una distribución normal para los datos, datos los efectos latentes (todos los relacionados con el valor esperado o predictor



lineal) y el resto de parámetros o hiperparámetros del modelo (en nuestro caso la varianza de los datos).

Si tenemos  $n$  observaciones

$$(y_i|\mu_i, \sigma^2) \sim N(\mu_i, \sigma^2), \quad i = 1, \dots, n$$

donde  $\mu_i$  representa la media y  $\sigma^2$  la varianza de los datos.

$$E(y_i|\mu_i, \sigma^2) = \mu_i, \quad \text{Var}(y_i|\mu_i, \sigma^2) = \sigma^2$$

Si tenemos varios regresores  $x_1, x_2, \dots, x_J$ , la media  $\mu_i$  coincide con el predictor lineal  $\eta_i$  que se construye a partir de una combinación lineal de los predictores:

$$\mu_i = \eta_i = \theta + \sum_{j=1}^J \beta_j x_{ij}$$

Nos queda a continuación especificar las distribuciones a priori sobre el vector de efectos latentes, en nuestro caso efectos fijos,  $(\theta, \beta_1, \dots, \beta_J)$ , y sobre el parámetro de dispersión o hiperparámetro  $\sigma^2$ .

Cuando no tenemos información previa especificamos distribuciones difusas (vagas) sobre los parámetros. En INLA por defecto tendremos:

$$\begin{aligned} \theta &\sim N(0, \sigma_\theta^2) \\ \beta_j &\sim N(0, \sigma_\beta^2) \\ \log(\tau = 1/\sigma^2) &\sim \text{Log} - \text{Ga}(1, 0.00005) \end{aligned}$$

con  $\sigma_\theta^2 = \infty$  y  $\sigma_\beta^2 = 1000$ .

Ejemplificamos el análisis de regresión lineal múltiple sobre la base de datos **usair**, en la librería **brinla**. Esta BD contiene datos recopilados para investigar los factores determinantes de la polución, utilizando el nivel de SO2 como variable dependiente y las restantes como variables explicativas potenciales. Las relaciones entre las variables que incluye se muestra en la Figura 2.1 a continuación.

```
data(usair, package = "brinla")
library(GGally) # contiene la función de graficado 'ggpairs'
pairs.chart <- ggpairs(usair, lower = list(continuous = "cor"),
                      upper = list(continuous = "points", combo = "dot"))
ggplot2::theme(axis.text = element_text(size = 6))
#> List of 1
#> $ axis.text:List of 11
#> ..$ family      : NULL
#> ..$ face        : NULL
```

```
#> ..$ colour      : NULL
#> ..$ size        : num 6
#> ..$ hjust       : NULL
#> ..$ vjust       : NULL
#> ..$ angle       : NULL
#> ..$ lineheight  : NULL
#> ..$ margin      : NULL
#> ..$ debug       : NULL
#> ..$ inherit.blank: logi FALSE
#> ..- attr(*, "class")= chr [1:2] "element_text" "element"
#> - attr(*, "class")= chr [1:2] "theme" "gg"
#> - attr(*, "complete")= logi FALSE
#> - attr(*, "validate")= logi TRUE
pairs.chart
```

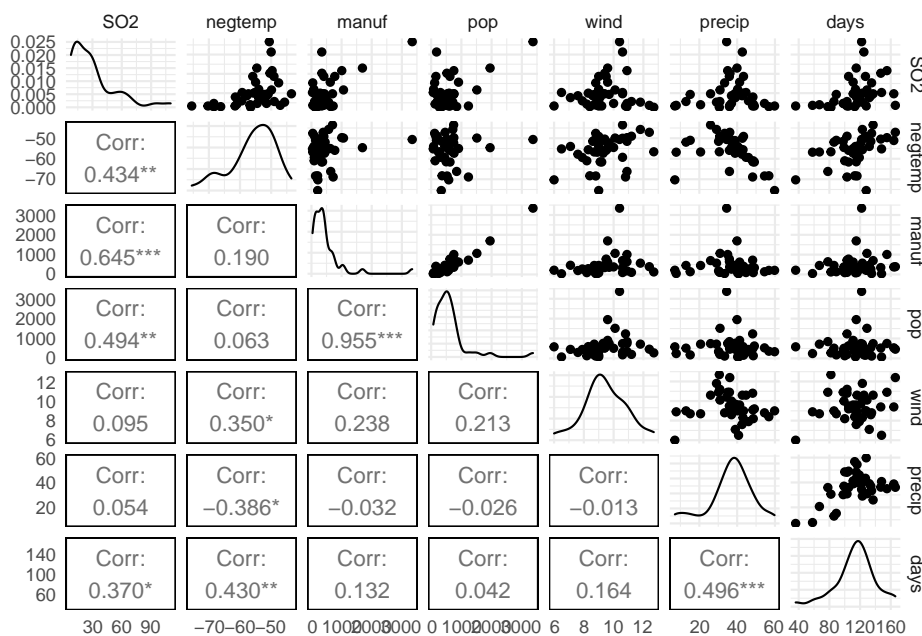


Figure 2.1: Relaciones entre variables en la base de datos usair(brinla).

Apreciamos ya en el gráfico una correlación positiva muy alta entre las variables `pop` y `manuf`, y relevante para `negtemp` y `days` y también para `precip` y `days`, lo que posteriormente condicionará la selección de variables.

### 2.9.1 Selección de variables

Cuando trabajamos con más de una variable predictora en regresión lineal (realmente en cualquier modelo) surge un problema adicional, que es la **selección de variables**, o selección del mejor modelo de predicción. Esto se resuelve en INLA utilizando diversos criterios de selección entre los que destacamos:

- la verosimilitud marginal (valor de la log-verosimilitud): `mlik`; al cambiarle el signo tendremos *-(log-likelihood)*
- el criterio de información de la deviance (DIC): `dic`
- el criterio de información bayesiana ampliado (WAIC): `waic`
- la transformada integral predictiva (PIT): `cpo`

El procedimiento a utilizar para la selección del modelo (de variables) es el siguiente:

1. ajustar todos los modelos resultantes de todas las combinaciones posibles de predictoras,
2. calcular los índices de selección para cada uno de ellos
3. proceder con la selección en base a dichos valores.

Siempre se prefieren modelos con los valores más bajos para estos criterios y se descartan los que proporcionan valores más altos. Cuando no es el mismo modelo el que proporciona el valor mínimo en estos criterios, habremos de optar por alguno de ellos.

Por defecto, al ajustar un modelo con `inla`, nos devuelve la log-verosimilitud marginal (accesible con `fit$mlik` si el ajuste se guardó en el objeto `fit`). Para obtener las otras medidas de selección, hemos de incluir como argumento de la función `inla`, la opción `control.compute = list(dic = TRUE, waic = TRUE)`). Los valores por defecto de `control.compute` los podemos consultar ejecutando el comando `inla.set.control.compute.default()`.

Vayamos pues con el ajuste del modelo de regresión con todas las variables. Recordemos que si no especificamos el argumento `family`, interpreta por defecto la opción `gaussian`, esto es, normalidad para los datos. Así ajustamos el modelo y obtenemos las siguientes inferencias sobre los efectos fijos.

```
formula <- S02 ~ negtemp + manuf + wind + precip + days
fit= inla(formula, data = usair,
          control.compute = list(dic = TRUE, waic = TRUE))
#summary(fit)
round(fit$summary.fixed,3)
#>               mean      sd 0.025quant 0.5quant 0.975quant
#> (Intercept) 135.490 49.837    37.206   135.497   233.733
```

```
#> negtemp      1.769  0.634      0.519   1.769   3.019
#> manuf        0.026  0.005      0.017   0.026   0.035
#> wind        -3.723  1.934     -7.535  -3.723   0.091
#> precip       0.625  0.387     -0.138   0.625   1.388
#> days        -0.057  0.174     -0.400  -0.057   0.287
#>               mode kld
#> (Intercept)    NA    0
#> negtemp        NA    0
#> manuf          NA    0
#> wind           NA    0
#> precip         NA    0
#> days           NA    0
```

La inferencia posterior sobre la desviación típica de los datos,  $\sigma$ , la resolvemos con sus descriptivos.

```
sigma.post= inla.tmarginal(function(tau) tau^(-1/2),
                           fit$marginals.hyperpar[[1]])
inla.zmarginal(sigma.post)
#> Mean      15.6702
#> Stdev      1.85483
#> Quantile  0.025 12.53
#> Quantile  0.25 14.3461
#> Quantile  0.5  15.4915
#> Quantile  0.75 16.7983
#> Quantile  0.975 19.8151
```

Para seleccionar las variables relevantes seguimos el procedimiento descrito anteriormente. Añadimos también el ajuste del modelo de regresión frecuentista, para el que calculamos como criterio de bondad de ajuste el AIC.

```
# variables en la bd
vars=names(usair)[-1] # variables predictoras (excluye v.dpte)
nvars=length(vars) # n° variables predictoras

# Truco para concatenar todos los modelos posibles en una fórmula (de Faraway)
listcombo <- unlist(sapply(1:nvars,
                           function(x) combn(nvars, x, simplify=FALSE)),
                           recursive=FALSE)
predterms <- lapply(listcombo,
                    function(x) paste(vars[x],collapse="+"))
nmodels <- length(listcombo)
coefm <- matrix(NA,length(listcombo),4,
                dimnames=list(predterms,c("AIC","DIC","WAIC","MLIK")))
```

```

# Ajuste de todos los modelos posibles
for(i in 1:nmodels){
  formula <- as.formula(paste("SO2 ~ ",predterms[[i]]))
  # modelo frecuentista
  lmi <- lm(formula, data=usair)
  # modelo bayesiano
  result <- inla(formula, family="gaussian", data=usair, control.compute=list(dic=TRUE, waic=TRUE))
  coefm[i,1] <- AIC(lmi)
  coefm[i,2] <- result$dic$dic
  coefm[i,3] <- result$waic$waic
  coefm[i,4] <- -result$mlik[1]
}

```

Ya solo resta comparar los resultados, respecto de cada uno de los criterios, para seleccionar con qué variables nos quedamos, y por lo tanto con qué modelo de predicción. Basta con encontrar el modelo que proporciona el mínimo valor en cada uno de los criterios.

```

gana.aic = predterms[which.min(coefm[,1])]
gana.dic = predterms[which.min(coefm[,2])]
gana.waic = predterms[which.min(coefm[,3])]
gana.mlik = predterms[which.min(coefm[,4])]
gana.aic;gana.dic
#> [[1]]
#> [1] "negtemp+manuf+pop+wind+precip"
#> [[1]]
#> [1] "negtemp+manuf+pop+wind+precip"
gana.waic;gana.mlik
#> [[1]]
#> [1] "negtemp+manuf+pop+wind+precip"
#> [[1]]
#> [1] "manuf"

```

Concluimos pues que, tanto el criterio AIC en el modelo de regresión frecuentista, como los criterios DIC y WAIC en el modelo de regresión bayesiano, proporcionan el mejor ajuste. Este mejor modelo incluye como predictores las variables ‘negtemp+manuf+pop+wind+precip’. Reajustamos el modelo con estas variables para derivar las inferencias y predicciones.

Las inferencias sobre los efectos fijos y la precisión de los datos se muestran a continuación.

```

formula=SO2 ~ negtemp+manuf+pop+wind+precip
fit=inla(formula, family="gaussian", data=usair, control.compute=list(dic=TRUE, waic=TRUE))
fit$summary.fixed

```

```

#>               mean          sd  0.025quant
#> (Intercept) 100.01589562 30.14775946 40.562018053
#> negtemp      1.12027424  0.41440567  0.303084688
#> manuf         0.06489637  0.01549479  0.034344687
#> pop          -0.03934776  0.01488818 -0.068704986
#> wind         -3.07257825  1.75709018 -6.536341942
#> precip       0.41922271  0.21553521 -0.005784108
#>               0.5quant    0.975quant mode      kld
#> (Intercept) 100.01943874 159.449457726  NA 1.151621e-08
#> negtemp      1.12030060  1.937312486  NA 1.154282e-08
#> manuf         0.06489614  0.095449426  NA 1.155305e-08
#> pop          -0.03934742 -0.009992471  NA 1.155244e-08
#> wind         -3.07283290  0.392644227  NA 1.149181e-08
#> precip       0.41922915  0.844192620  NA 1.155038e-08
fit$summary.hyperpar
#>               mean
#> Precision for the Gaussian observations 0.005066361
#>               sd
#> Precision for the Gaussian observations 0.001179452
#>               0.025quant
#> Precision for the Gaussian observations 0.003034052
#>               0.5quant
#> Precision for the Gaussian observations 0.004975563
#>               0.975quant mode
#> Precision for the Gaussian observations 0.007617574  NA

```

Y en la Figura 2.2 se muestran las distribuciones posteriores de todos los efectos latentes (efectos fijos) en el modelo: interceptación y coeficientes de los regresores.

```

nfixed=length(fit$names.fixed)
g=list()
for(i in 1:nfixed){
  g[[i]]=ggplot(as.data.frame(fit$marginals.fixed[[i]])) +
    geom_line(aes(x = x, y = y)) +
    labs(x=fit$names.fixed[i],y="")+
    geom_vline(xintercept=inla.hpdmarginal(0.95,fit$marginals.fixed[[i]]),
              linetype="dashed",color="red")+
    geom_vline(xintercept=inla.emarginal(function(x) x,fit$marginals.fixed[[i]]),
              linetype="dashed",color="blue")
}
grid.arrange(g[[1]],g[[2]],g[[3]],g[[4]],g[[5]],g[[6]],ncol=2)

```

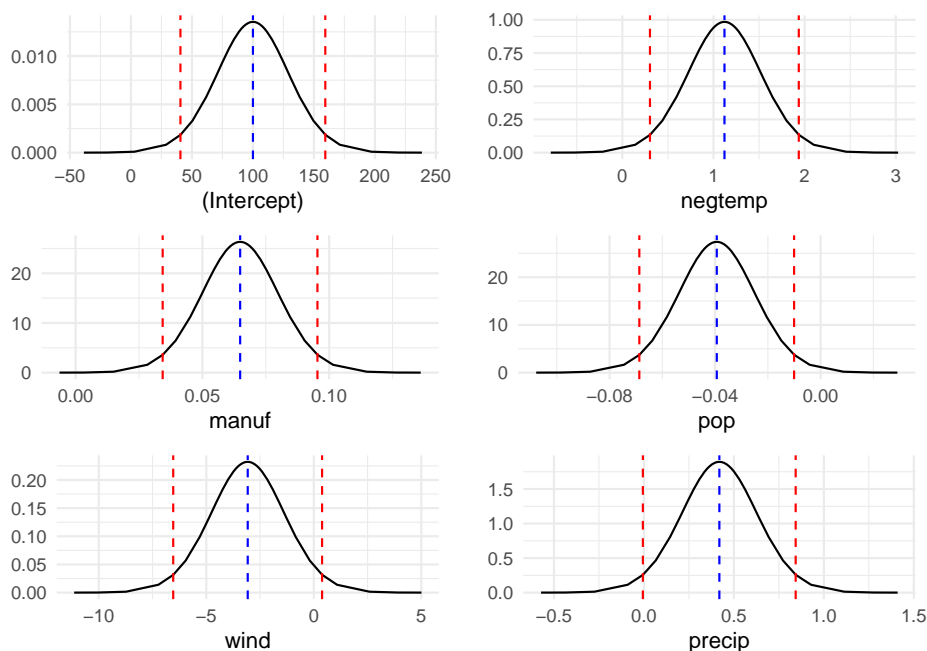


Figure 2.2: Distribuciones posteriores de los efectos latentes.

### 2.9.2 Predicción de medias

Por último, si queremos predecir la respuesta esperada para ciertos valores de las variables explicativas, no necesariamente existentes en la base de datos, utilizamos el argumento `control.predictor` en la función `inla`, especificando en una lista los valores a predecir. Veamos cómo hacerlo con `inla`, ajustando un modelo sobre un `data.frame` combinado, en el que añadimos tantas filas como predicciones queremos conseguir, con los valores deseados para los predictores (y valores faltantes en la respuesta), y especifiquemos los valores a predecir en un vector indicador, a través del argumento `control.predictor(list(link=vector.indicador))`. Las predicciones se obtienen después, con el resumen de los datos ajustados `fitted.values`. Recordemos que para poder mostrar las distribuciones predictivas, hemos de añadir en `inla` el argumento `control.compute=list(return.marginals.predictor=TRUE)`.

```
# Predicción con INLA
## valores de los predictores en los que predecir: 3 escenarios
formula=S02 ~ negtemp+manuf+pop+wind+precip
new.data <- data.frame(negtemp = c(-50, -60, -40),
                      manuf = c(150, 100, 400),
                      pop = c(200, 100, 300),
                      wind = c(6, 7, 8),
```

```

precip = c(10, 30, 20),
days=c(NA, NA,NA))

## añadimos los tres escenarios de predicción a la bd original,
## dejando como faltantes los valores a predecir de la v.dpte
usair.combinado <- rbind(usair, data.frame(SO2=c(NA,NA,NA),new.data))
## creamos un vector con NA's para observaciones y 1's para predicciones
usair.indicador <- c(rep(NA, nrow(usair)), rep(1, nrow(new.data)))
## reajustamos el modelo añadiendo la opción de predicción de datos
fit.pred <- inla(formula, data = usair.combinado,
                 control.compute=list(return.marginals.predictor=TRUE),
                 control.predictor = list(link = usair.indicador))
## y describimos los valores ajustados para los tres escenarios añadidos
fit.pred$summary.fitted.values[(nrow(usair)+1):nrow(usair.combinado),]
#>
#>          mean      sd 0.025quant 0.5quant
#> fitted.Predictor.42 31.62368 8.203273   15.44981 31.62443
#> fitted.Predictor.43 26.42290 5.473326   15.63183 26.42325
#> fitted.Predictor.44 53.16275 7.090693   39.18270 53.16328
#>
#>          0.975quant mode
#> fitted.Predictor.42  47.79310  NA
#> fitted.Predictor.43  37.21180  NA
#> fitted.Predictor.44  67.13953  NA

```

Así graficamos en la Figura 2.3 la distribución predictiva del nivel de SO<sub>2</sub> para una combinación dada de valores de las variables predictivas, específicamente la que aparece en el escenario 1 propuesto, o lo que es lo mismo, en el registro 42 de la base de datos completada con las nuevas predicciones: negtem=-50, manuf=150, pop=200, wind=6 y precip=10.

```

pred=fit.pred$marginals.fitted.values[[42]]
ggplot(as.data.frame(pred)) +
  geom_line(aes(x = x, y = y)) +
  labs(x=expression(eta),y="")+
  geom_vline(xintercept=inla.hpdmarginal(0.95,pred),
            linetype="dashed",color="red")+
  geom_vline(xintercept=inla.emarginal(function(x) x,pred),
            linetype="dashed",color="blue")

```

## 2.10 Conclusiones

En este tema hemos trabajado con el ajuste con INLA de modelos lineales de regresión, esto es, con efectos fijos. En posteriores temas trabajaremos modelos



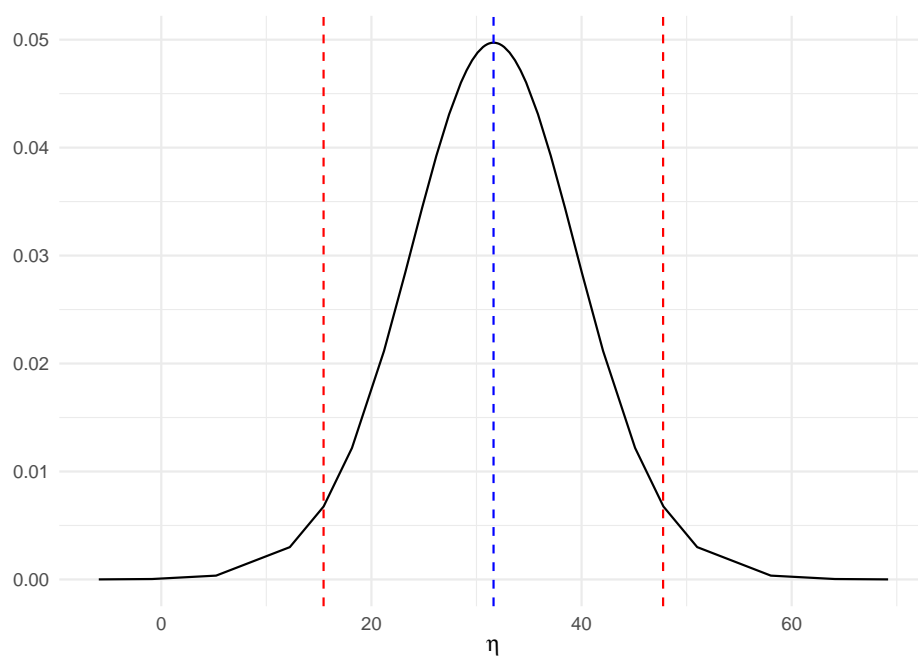


Figure 2.3: Distribución predictiva a posteriori de SO<sub>2</sub> para una configuración dada de los predictores.

más sofisticados en los que incluiremos los efectos aleatorios, generalizaremos el modelo lineal y entenderemos el planteamiento de modelos a través de modelos jerárquicos bayesianos.

## Chapter 3

# Modelo de ANOVA

### 3.1 Introducción

El modelo de ANOVA se plantea para comparar poblaciones normales, especialmente cuando son más de dos las poblaciones a comparar. Las poblaciones a comparar se identifican a través de una variable clasificadora (de tipo categórico) que actúa como predictora para estimar respuestas medias supuestamente distintas a comparar.

### 3.2 El modelo de ANOVA

Consideremos una variable respuesta  $Y$  que se distribuye normal, y que viene afectada por una variable de clasificación  $A$  con  $a$  niveles de respuesta distintos (uno por cada una de las poblaciones a comparar). Supongamos que tenemos  $n_i$  observaciones de la respuesta para cada uno de los niveles de respuesta de la variable clasificadora,  $i = 1, \dots, a$ . El modelo de ANOVA se plantea asumiendo que en cada nivel o subpoblación, esperamos un valor distinto para la respuesta,

$$(y_{ij}|\mu_i, \sigma^2) \sim N(\mu_i, \sigma^2)$$

de modo que

$$E(y_{ij}|\mu_i, \sigma^2) = \mu_i; \text{Var}(y_{ij}|\mu_i, \sigma^2) = \sigma^2, \quad i = 1, \dots, a; \quad j = 1, \dots, n_i$$

La formulación habitual de este modelo se suele dar en términos de un efecto global y común a todas las observaciones,  $\theta$ , y un efecto diferencial respecto del primer nivel del factor de clasificación  $A$ ,  $\alpha_i$ , con los que se construye la media (identificada generalmente por  $\mu$ ) o predictor lineal (identificada generalmente por  $\eta$ ) y que en el modelo lineal coinciden:

$$\mu_{ij} = \eta_{ij} = \theta + \alpha_i$$

donde  $\alpha_i = \mu_i - \mu_1$  y  $\mu_1 = \theta$ , para  $i \geq 1$ , esto es,  $\alpha_1 = 0$ .

Estamos pues asumiendo que todos los  $n_i$  sujetos en el subgrupo de población  $i$  identificado por la variable clasificadora  $A$ , comparten una media común  $\mu_i$  y cierta variabilidad  $\sigma^2$ . Podríamos asumir varianzas distintas para cada subpoblación, pero por simplicidad consideramos que son iguales.

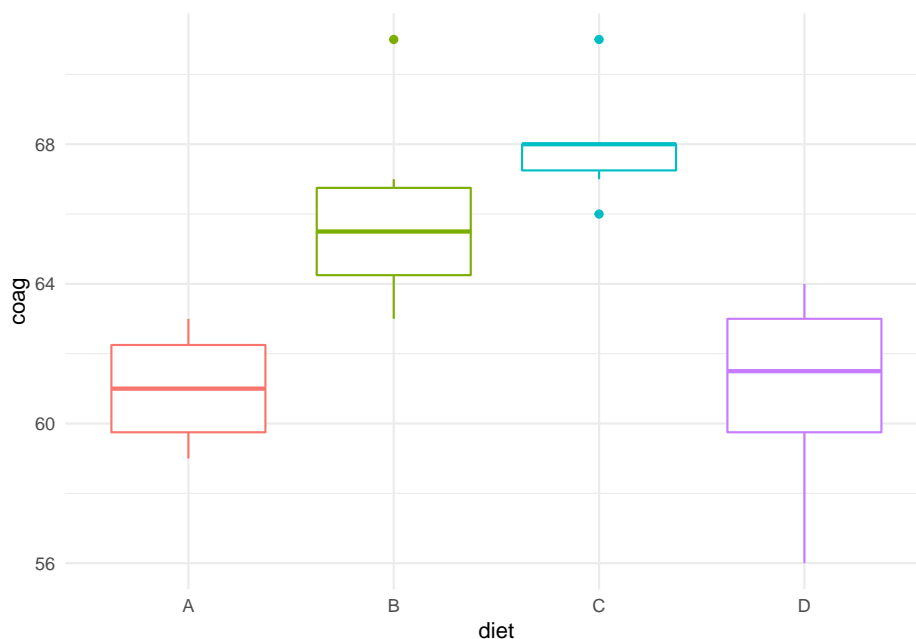
En la modelización bayesiana es preciso añadir distribuciones a priori para cada uno de los parámetros del modelo: los efectos fijos  $\theta, \alpha_i$ , y la varianza  $\sigma^2$  de los datos. Ante ausencia de información, se asumirán las distribuciones difusas habituales en INLA:

$$\begin{aligned} (Y_{ij} | \mu_i, \sigma^2) &\sim N(\mu_i, \sigma^2) \\ \mu_i &= \theta + \alpha_i, i \geq 1 \\ \theta &\sim N(0, \infty) \\ \alpha_i &\sim N(0, 1000), i \geq 1 \\ \tau = 1/\sigma^2 &\sim Ga(1, 0.00005) \end{aligned}$$

### 3.3 Anova de una vía

Vamos a ilustrar el análisis ANOVA en INLA a través de la base de datos `coagulation`, en la librería `faraway`, referidos a un estudio de tiempos de coagulación de la sangre en 24 animales a los que aleatoriamente se les asignó una de entre tres dietas distintas (variable clasificadora `diet`). Posteriormente, para estudiar el efecto de dichas dietas en la coagulación, se tomaron muestras de los tiempos de coagulación (en la variable `coag`, que es la respuesta).

```
data(coagulation, package="faraway")
str(coagulation)
#> 'data.frame': 24 obs. of 2 variables:
#> $ coag: num 62 60 63 59 63 67 71 64 65 66 ...
#> $ diet: Factor w/ 4 levels "A","B","C","D": 1 1 1 1 2 2 2 2 2 2 ...
ggplot(coagulation, aes(x=diet, y=coag))+
  geom_boxplot(aes(color=diet))+
  theme(legend.position="none")
```



Estamos planteando un modelo de Anova como el propuesto en la sección anterior, donde  $\alpha_i$  identifica el efecto diferencial sobre la respuesta con la dieta A, para el resto de las dietas B y C. Los parámetros del modelo son, como en regresión, los efectos fijos  $(\theta, \alpha_i)$  y la varianza  $\sigma^2$ , para los que asumimos las priors difusas que por defecto propone INLA. Ajustamos el modelo y obtenemos las inferencias a posteriori

```
formula=coag ~ diet
fit=inla(formula,family="gaussian",data=coagulation,
         control.compute=list(config=TRUE,return.marginals.predictor=TRUE))
fijos=round(fit$summary.fixed,3);fijos
#>           mean    sd 0.025quant 0.5quant 0.975quant
#> (Intercept) 61.016 1.172    58.700    61.016    63.337
#> dietB       4.979 1.513     1.983     4.980     7.970
#> dietC       6.977 1.513     3.981     6.978     9.968
#> dietD      -0.016 1.435    -2.859    -0.016     2.822
#>           mode kld
#> (Intercept)  NA  0
#> dietB       NA  0
#> dietC       NA  0
#> dietD       NA  0
tau=round(fit$summary.hyperpar,3);tau
#>           mean    sd
#> Precision for the Gaussian observations 0.197 0.059
#>           0.025quant 0.5quant
```

```
#> Precision for the Gaussian observations      0.099      0.191
#>                                           0.975quant mode
#> Precision for the Gaussian observations      0.328      NA
medias=round(fit$summary.linear.predictor,4)
```

Atendiendo a los descriptivos de la distribución posterior para los efectos fijos, concluimos:

- El tiempo esperado de coagulación para los animales que han seguido la dieta A es de 61.016(58.7,63.337).
- Los animales que han seguido la dieta B tienen un tiempo de coagulación esperado superior en 4.979 unidades a los de la dieta A, y dicha diferencia es *significativamente distinta de cero* en el contexto bayesiano, dado que su RC no incluye al cero, (1.983,7.97).
- Una conclusión similar se deriva para la dieta C, que da un tiempo de coagulación esperado superior en 6.977 unidades a los de la dieta A, y una RC (3.981,9.968).
- Las diferencias en los tiempos de coagulación de seguir una dieta D frente a la dieta A no son relevantes. De hecho, la diferencia entre ellos es de -0.016 y el intervalo RC contiene al cero, (-2.859,2.822).

Pintamos a continuación en la Figura 3.1 la distribución posterior de los tiempos esperados de coagulación  $\mu_i$  (o predictores lineales) para cada una de las dietas. En la Figura 3.2 se añaden las medias posteriores y las regiones creíbles.

```
dietas=levels(coagulation$diet)
pred=NULL
for(i in 1:length(dietas)){
  index=which(coagulation$diet==dietas[i])[1]
  # distrib. posterior
  post=as.data.frame(fit$marginals.fitted.values[[index]])
  # media
  e=fit$summary.fitted.values[index,1]
  rc.low=fit$summary.fitted.values[index,3]
  rc.up=fit$summary.fitted.values[index,5]
  pred=rbind(pred,data.frame(dieta=dietas[i],
                             post,e,rc.low,rc.up))
}

ggplot(pred, aes(x = x, y =y)) +
  geom_line(aes(color=dieta))+
  labs(x=expression(paste("Tiempo medio de coagulación:",mu)),
       y="D.Posterior")
```

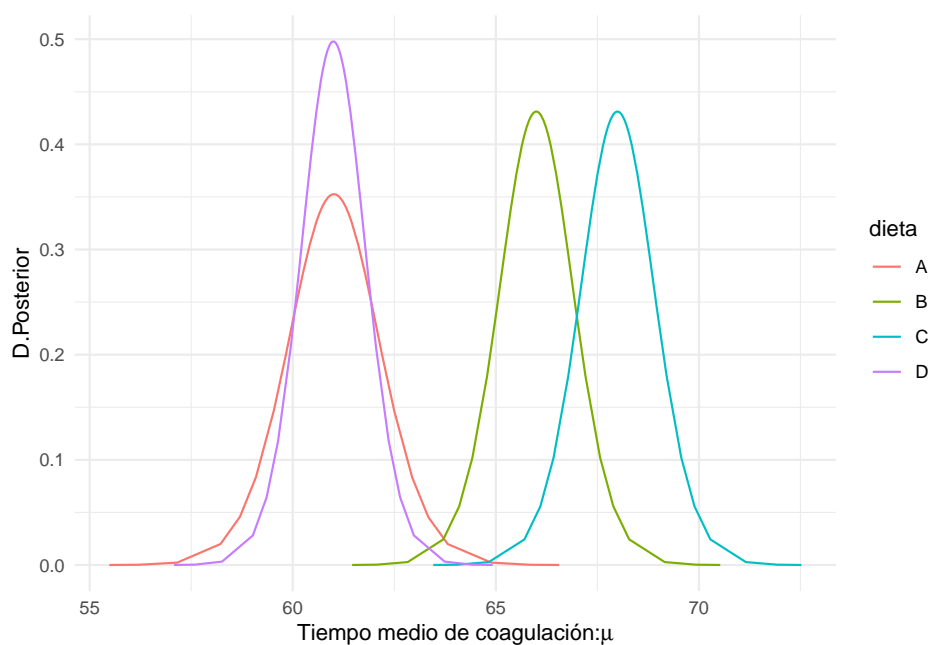


Figure 3.1: Distribución posterior del tiempo medio de coagulación para las 4 dietas.

Podríamos ajustar el modelo prescindiendo del efecto de interceptación y estimar directamente los efectos.

```
formula=coag ~ -1 + diet
fit=inla(formula,family="gaussian",data=coagulation,
         control.compute=list(config=TRUE,return.marginals.predictor=TRUE))
round(fit$summary.fixed,3)
#>      mean    sd 0.025quant 0.5quant 0.975quant mode kld
#> dietA 60.916 1.176    58.577    60.919    63.233   NA  0
#> dietB 65.939 0.961    64.030    65.942    67.832   NA  0
#> dietC 67.937 0.961    66.028    67.940    69.830   NA  0
#> dietD 60.958 0.832    59.306    60.960    62.599   NA  0
round(fit$summary.hyperpar,3)
#>                                     mean    sd
#> Precision for the Gaussian observations 0.197 0.06
#>                                     0.025quant 0.5quant
#> Precision for the Gaussian observations    0.098    0.19
#>                                     0.975quant mode
#> Precision for the Gaussian observations    0.324   NA
```

Con lo cual la representación gráfica se simplifica a través, directamente, de las

distribuciones posteriores de los efectos fijos.

```
pred=NULL
for(i in 1:length(names(fit$marginals.fixed))){
  pred=rbind(pred,data.frame(as.data.frame(fit$marginals.fixed[[i]]),
    dieta=names(fit$marginals.fixed)[i],
    mean=fit$summary.fixed$mean[i],
    rc.low=fit$summary.fixed$'0.025quant'[i],
    rc.up=fit$summary.fixed$'0.975quant'[i]))}

ggplot(pred, aes(x = x, y =y)) +
  geom_line(aes(color=dieta))+
  geom_vline(aes(xintercept=mean,color=dieta),linetype="dashed")+
  geom_vline(aes(xintercept=rc.low,color=dieta),linetype="dotted")+
  geom_vline(aes(xintercept=rc.up,color=dieta),linetype="dotted")+
  facet_wrap(vars(dieta))+
  labs(x=expression(paste("Tiempo medio de coagulación:",mu)),
    y="D.Posterior",title="D.Posterior, medias y RC95%")+
  theme(legend.position="none")
```

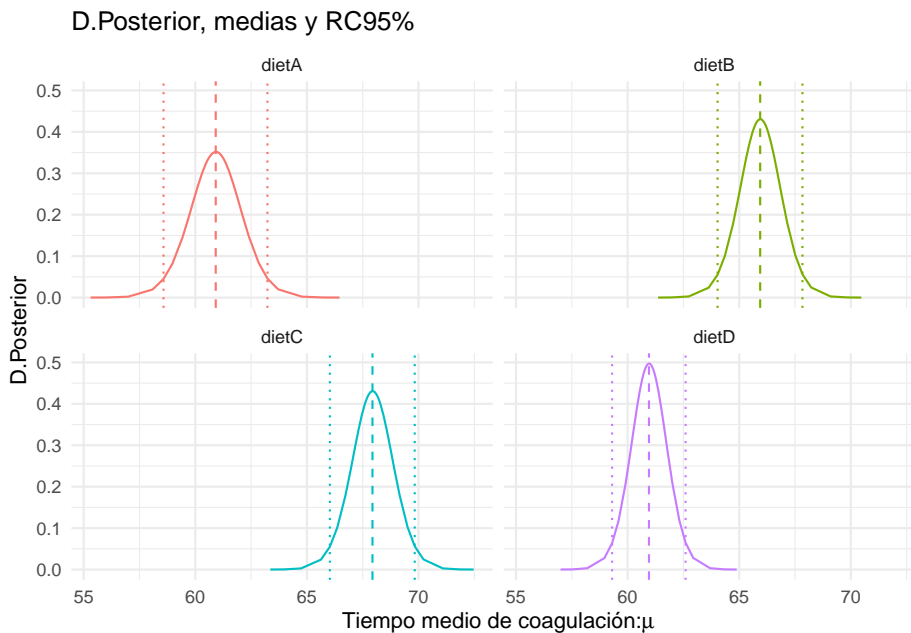


Figure 3.2: Distribuciones posteriores, medias y RC del tiempo esperado de coagulación.

Como ya hacíamos en regresión, podemos inferir sobre la desviación típica de



los datos,  $\sigma$ , transformando la distribución para la precisión  $\tau$ . La distribución posterior junto con su media y RC95% se muestra en la Figura 3.3.

```
sigma.post=inla.tmarginal(function(tau) tau^(-1/2),
  fit$marginals.hyperpar[[1]])
# y la pintamos
ggplot(as.data.frame(sigma.post)) +
  geom_line(aes(x = x, y = y)) +
  labs(x=expression(sigma),y="D.Posterior")+
  geom_vline(xintercept=inla.hpdmarginal(0.95,sigma.post),
    linetype="dotted",color="blue")+
  geom_vline(xintercept=inla.emarginal(function(x) x,sigma.post),
    linetype="dashed",color="blue")
```

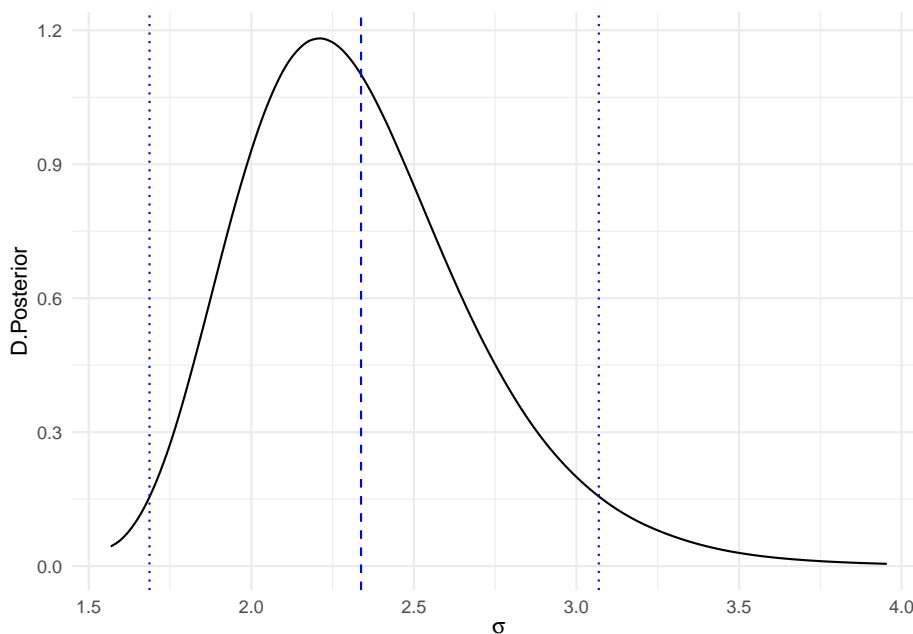


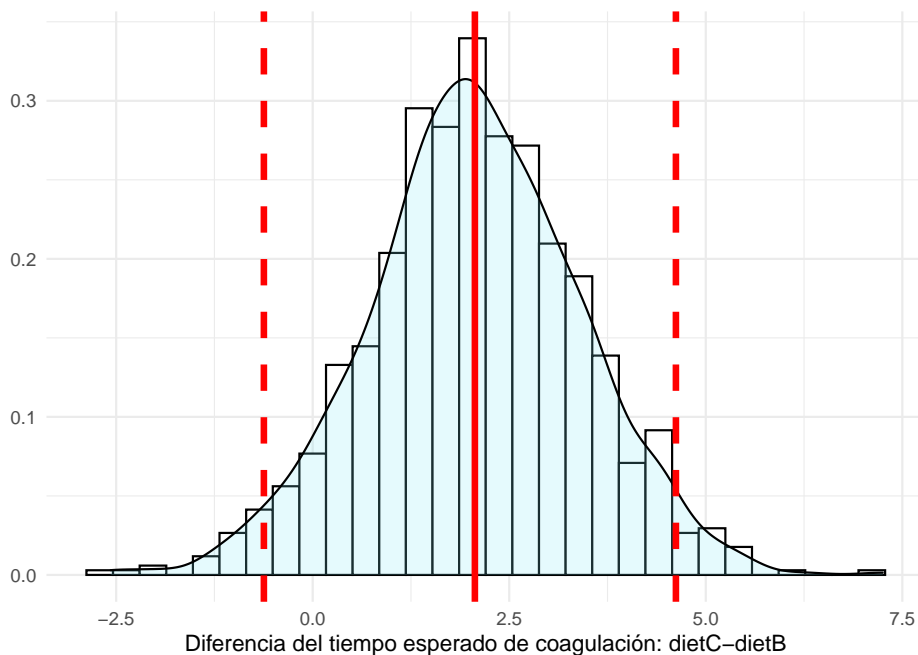
Figure 3.3: Distribución posterior, media y RC, de la desviación típica de los datos (sigma)

```
# Valor esperado
sigma.e=round(inla.emarginal(function(tau) tau^(-1/2),
  fit$marginals.hyperpar[[1]]),4)
# HPD95%
sigma.hpd=round(inla.hpdmarginal(0.95,sigma.post),3)
paste("E(sigma.post)=",sigma.e,"HPD95%=(",sigma.hpd[1],",",sigma.hpd[2],")")
#> [1] "E(sigma.post)= 2.338 HPD95%=( 1.687 , 3.069 )"
```

Si queremos inferir sobre la diferencia entre cualesquiera de los efectos podemos recurrir a simular la distribución posterior de las diferencias. Por ejemplo, supongamos que queremos comparar la dieta B con la dieta C. Simulamos entonces de las distribuciones posteriores de  $\mu_B$  y de  $\mu_C$ , y obtenemos la diferencia  $\mu_B - \mu_C$ .

$$\mu_B^{(i)} \sim \pi(\mu_B|y), \mu_C^{(i)} \sim \pi(\mu_C|y) \Rightarrow \mu_C^{(i)} - \mu_B^{(i)} \sim \pi(\mu_C - \mu_B|y), \quad i = 1, \dots, nsim$$

```
sims=inla.posterior.sample(1000,fit,selection=list(dietB=1,dietC=1))
dif_CB=as.vector(inla.posterior.sample.eval(function(...) dietC-dietB, sims))
pred=data.frame(dif=dif_CB)
ggplot(pred,aes(x=dif))+
  geom_histogram(aes(y=..density..), colour="black", fill="white")+
  geom_density(alpha=.2, fill="#80E7F5")+
  geom_vline(xintercept=mean(dif_CB),color="red",size=1.5)+
  geom_vline(xintercept=quantile(dif_CB,probs=c(0.025,0.975)),color="red",size=1.5,lin
  labs(x="Diferencia del tiempo esperado de coagulación: dietC-dietB",y="")
#> `stat_bin()` using `bins = 30`. Pick better value with
#> `binwidth`.
```



Podríamos así mismo, calcular cualquier probabilidad con ellas, como por ejemplo la probabilidad de que el tiempo de coagulación de un animal que sigue la

dieta C sea 2 unidades superior al de uno que sigue la dieta B.

$$Pr(\mu_C > \mu_B + 2|y) = Pr(\mu_C - \mu_B > 2|y)$$

```
cat(paste("Probabilidad posterior de que muC>muB+2 =",mean(pred$dif>2)))
#> Probabilidad posterior de que muC>muB+2 = 0.515
```

### 3.4 Anova de varias vías

Generalmente, y en especial cuando trabajamos con experimentación, son varios los factores que controlamos para investigar el efecto que producen en una respuesta continua. Hablamos de modelos de Anova de varias vías.

Utilizamos la base de datos `butterfat` en la librería `faraway` para ilustrar el ajuste con INLA de un modelo de Anova de varias vías. Esta base de datos contiene 100 registros del contenido en grasa láctea, `Butterfat`, para muestras aleatorias de 20 vacas (10 de ellas de 2 años y 10 maduras, con más de 4 años -en la variable `Age`) de cada una de cinco razas (en la variable `Breed`).

El objetivo es investigar las diferencias en materia grasa entre razas y edad, con el fin último de identificar cuáles producen más materia grasa y cuáles menos. Veamos los datos en la Figura 3.4.

```
data(butterfat,package="faraway")
str(butterfat)
#> 'data.frame':   100 obs. of  3 variables:
#> $ Butterfat: num  3.74 4.01 3.77 3.78 4.1 4.06 4.27 3.94 4.11 4.25 ...
#> $ Breed    : Factor w/ 5 levels "Ayrshire","Canadian",...: 1 1 1 1 1 1 1 1 1 1 ...
#> $ Age      : Factor w/ 2 levels "2year","Mature": 2 1 2 1 2 1 2 1 2 1 ...
ggplot(butterfat,aes(x=Breed,y=Butterfat))+
  geom_boxplot(aes(color=Age))+
  coord_flip()
```

A la vista del gráfico, apreciamos que por lo general, en la mayoría de las razas, las vacas más jóvenes tienen menor contenido en materia grasa que las más viejas. Sin embargo, tal afirmación no parece tan clara en las razas *Guernsey* y *Holstein-Fresian*, de modo que para modelizar nuestros datos vamos a considerar a priori, la posibilidad de interacciones entre los factores de clasificación `Breed` y `Age`.

Cuando nos enfrentamos a varios factores de clasificación, cabe la posibilidad de que interactúen entre ellos, esto es, que en algunos niveles de un factor actúen de forma diferente a los otros cuando se combinan con los niveles de algún otro factor. El orden de una interacción viene dado por el número de factores de

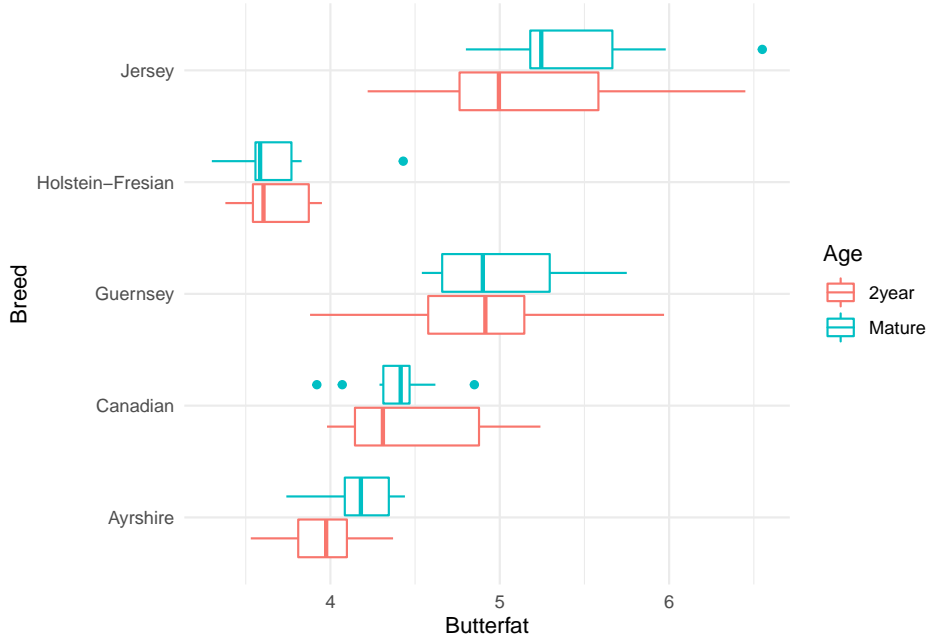


Figure 3.4: Base de datos butterfat, en la librería Faraway

clasificación que involucra, de modo que hablamos de interacciones de orden 2 si consideramos la interacción entre dos factores, de orden 3 si consideramos la interacción entre tres factores, etc. Generalmente trabajamos con interacciones de orden bajo, dada la complejidad de las conclusiones en interacciones de orden alto. Por otro lado, siempre es importante tener en cuenta de cuántos datos disponemos para conocer a priori la posibilidad de estimar con fiabilidad los distintos efectos de interacción: una interacción de dos factores con  $n_1$  y  $n_2$  niveles de clasificación respectivamente, revierte en la estimación de  $(n_1 - 1) \times (n_2 - 1)$  efectos de interacción adicionales.

Así, en nuestro problema si estamos planteando la posibilidad de que haya interacciones entre los dos factores de clasificación, estamos asumiendo un modelo de tipo siguiente, asumiendo normalidad en la respuesta:

$$(y_{ijk} | \mu_{ij}, \sigma^2) \sim N(\mu_{ij}, \sigma^2)$$

con

$$\mu_{ij} = \theta + \alpha_i + \beta_j + \alpha\beta_{ij}$$

donde en nuestro ejemplo,  $\alpha_i$  es el efecto diferencial (respecto del primer nivel) que aporta el nivel  $i$  de la variable **Breed**,  $\beta_j$  el efecto asociado a la variable **Age**, y  $\alpha\beta$  la correspondiente interacción entre ellas. En  $R$  una interacción de orden 2 entre dos variables  $f_1$  y  $f_2$  se especifica con  $f_1 : f_2$ ; los efectos principales y la

interacción se pueden especificar de varios modos alternativos:

$$f_1 + f_2 + f_1 : f_2 \equiv f_1 * f_2 \equiv (f_1 + f_2)^2$$

Veamos cómo ajustar con INLA este modelo, recabando también los criterios de selección DIC y WAIC.

```
formula=Butterfat ~ Breed * Age
fit=inla(formula,data=butterfat,
         control.compute=list(dic = TRUE, waic = TRUE))
round(fit$summary.fixed,3)
#>
#> (Intercept)          mean      sd 0.025quant
#> BreedCanadian      0.522 0.186      0.157
#> BreedGuernsey       0.933 0.186      0.568
#> BreedHolstein-Fresian -0.303 0.186     -0.668
#> BreedJersey         1.167 0.186      0.802
#> AgeMature           0.188 0.186     -0.177
#> BreedCanadian:AgeMature -0.287 0.263     -0.804
#> BreedGuernsey:AgeMature -0.086 0.263     -0.603
#> BreedHolstein-Fresian:AgeMature -0.175 0.263     -0.692
#> BreedJersey:AgeMature   0.131 0.263     -0.386
#>
#> 0.5quant 0.975quant mode
#> (Intercept)      3.966      4.224  NA
#> BreedCanadian      0.522      0.887  NA
#> BreedGuernsey       0.933      1.298  NA
#> BreedHolstein-Fresian -0.303      0.062  NA
#> BreedJersey         1.167      1.532  NA
#> AgeMature           0.188      0.553  NA
#> BreedCanadian:AgeMature -0.287      0.230  NA
#> BreedGuernsey:AgeMature -0.086      0.431  NA
#> BreedHolstein-Fresian:AgeMature -0.175      0.342  NA
#> BreedJersey:AgeMature   0.131      0.648  NA
#>
#> kld
#> (Intercept)      0
#> BreedCanadian      0
#> BreedGuernsey       0
#> BreedHolstein-Fresian 0
#> BreedJersey         0
#> AgeMature           0
#> BreedCanadian:AgeMature 0
#> BreedGuernsey:AgeMature 0
#> BreedHolstein-Fresian:AgeMature 0
#> BreedJersey:AgeMature 0
fit$dic$dic
#> [1] 120.4861
```

```
fit$waic$waic
#> [1] 121.7376
```

Observamos en la inferencia posterior para los efectos fijos, que todas las RC asociadas a los efectos de interacción contienen al cero, lo que descarta la relevancia de la interacción a la hora de predecir la respuesta. Reajustamos pues el modelo eliminando la interacción, y comprobamos que efectivamente al eliminarla conseguimos reducir los valores del DIC y WAIC que usamos habitualmente para la selección de variables.

```
formula=Butterfat ~ Breed + Age
fit=inla(formula,data=butterfat,
         control.predictor=list(compute=TRUE),
         control.compute=list(return.marginals.predictor=TRUE,
                              dic = TRUE, waic = TRUE))

fit$summary.fixed
#>
#>      mean      sd  0.025quant
#> (Intercept)  4.0077184 0.10124843  3.80873735
#> BreedCanadian  0.3784787 0.13071115  0.12159406
#> BreedGuernsey  0.8899744 0.13071115  0.63308957
#> BreedHolstein-Fresian -0.3905147 0.13071115 -0.64739918
#> BreedJersey  1.2324714 0.13071115  0.97558656
#> AgeMature  0.1045993 0.08266996 -0.05787041
#>
#>      0.5quant 0.975quant mode
#> (Intercept)  4.0077182  4.2067005  NA
#> BreedCanadian  0.3784790  0.6353622  NA
#> BreedGuernsey  0.8899746  1.1468577  NA
#> BreedHolstein-Fresian -0.3905145 -0.1336311  NA
#> BreedJersey  1.2324717  1.4893547  NA
#> AgeMature  0.1045993  0.2670689  NA
#>
#>      kld
#> (Intercept)  2.524760e-09
#> BreedCanadian  2.524640e-09
#> BreedGuernsey  2.524642e-09
#> BreedHolstein-Fresian 2.524639e-09
#> BreedJersey  2.524647e-09
#> AgeMature  2.524872e-09
fit$waic$waic
#> [1] 116.3439
fit$dic$dic
#> [1] 115.3811
```

Observamos ya a partir del modelo ajustado, que el efecto de la edad no es relevante (su RC incluye al cero), pero sin embargo sí que hay diferencias debido a las razas.

Reajustamos de nuevo el modelo, excluyendo la variable `Age`, y verificamos la reducción (ligera) del DIC/WAIC, lo cual justifica usar este modelo para la predicción.

```
formula=Butterfat ~ Breed
fit=inla(formula,data=butterfat,
        control.predictor=list(compute=TRUE),
        control.compute=list(return.marginals.predictor=TRUE,
                             dic = TRUE, waic = TRUE))

fit$summary.fixed
#>
#>      mean      sd 0.025quant
#> (Intercept)  4.0600181 0.09271688  3.8778080
#> BreedCanadian  0.3784786 0.13112176  0.1207926
#> BreedGuernsey   0.8899742 0.13112176  0.6322881
#> BreedHolstein-Fresian -0.3905148 0.13112176 -0.6482006
#> BreedJersey    1.2324713 0.13112176  0.9747851
#>      0.5quant 0.975quant mode
#> (Intercept)  4.0600180 4.2422293  NA
#> BreedCanadian  0.3784788 0.6361634  NA
#> BreedGuernsey   0.8899745 1.1476588  NA
#> BreedHolstein-Fresian -0.3905146 -0.1328298  NA
#> BreedJersey    1.2324715 1.4901558  NA
#>      kld
#> (Intercept)  2.471424e-09
#> BreedCanadian  2.471442e-09
#> BreedGuernsey   2.471445e-09
#> BreedHolstein-Fresian 2.471441e-09
#> BreedJersey    2.471442e-09
fit$waic$waic
#> [1] 115.9279
fit$dic$dic
#> [1] 115.0093
```

Procederíamos igual que en el modelo de Anova de una vía para la representación de las distribuciones posteriores sobre las medias o predictores lineales en cada una de las razas. Igualmente representaremos la distribución posterior del parámetro de dispersión de los datos  $\sigma$ .

## 3.5 Análisis de ANCOVA

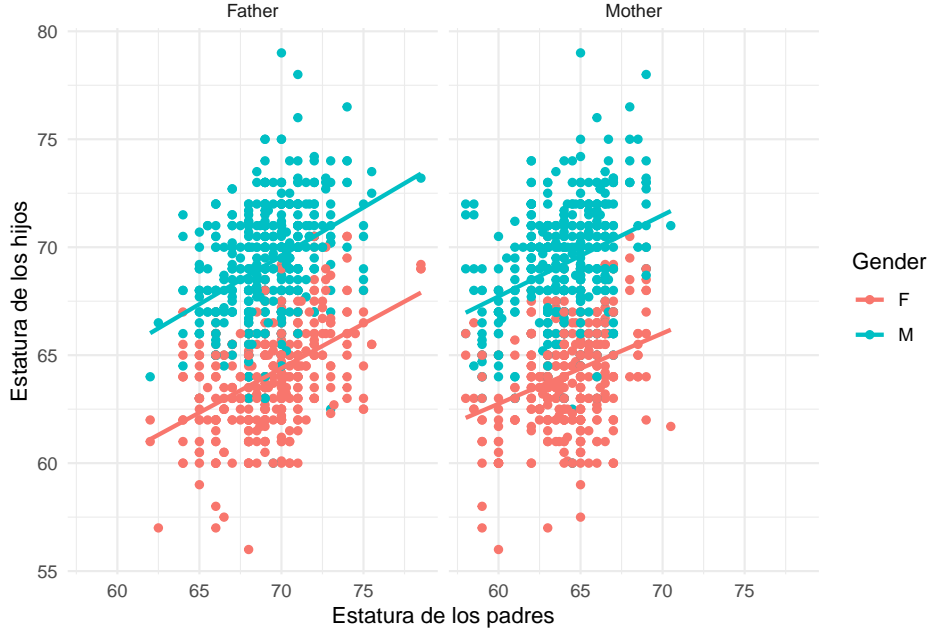
En ocasiones tenemos una variable respuesta de tipo numérico, y como posibles predictores, tanto variables de tipo numérico como variables clasificatorias o factores. Surge entonces la posibilidad de que los predictores numéricos afecten a la respuesta de modo distinto en diferentes niveles de clasificación de los factores;

hablamos entonces de **interacción entre covariables y factores**. Veamos un ejemplo para comprender cómo funcionan estos modelos y cómo se ajustan con INLA.

Consideramos los datos de Galton sobre la regresión de las alturas de los hijos sobre la de los padres (Fte: Galton's Height Data). Tenemos la estatura del padre, de la madre y del hijo/a, identificado/a por su sexo. Vamos a formular un modelo de regresión de la estatura de los hijos en función de la de sus padres y su género.

```
url="https://raw.githubusercontent.com/BayesModel/data/main/Galton.txt"
datos<-read.csv(file=url,header=TRUE,dec=".", sep="")
str(datos)
#> 'data.frame': 898 obs. of 6 variables:
#> $ Family: chr "1" "1" "1" "1" ...
#> $ Father: num 78.5 78.5 78.5 78.5 75.5 75.5 75.5 75.5 75 75 ...
#> $ Mother: num 67 67 67 67 66.5 66.5 66.5 66.5 64 64 ...
#> $ Gender: chr "M" "F" "F" "F" ...
#> $ Height: num 73.2 69.2 69 69 73.5 72.5 65.5 65.5 71 68 ...
#> $ Kids : int 4 4 4 4 4 4 4 4 2 2 ...
datos %>%
  pivot_longer(cols=c("Father","Mother"),
               names_to = "Parents",values_to="Height_parents") %>%
  ggplot(aes(x=Height_parents,y=Height))+
  geom_point(aes(color=Gender))+
  geom_smooth(method="lm",aes(color=Gender),se=FALSE)+
  facet_wrap(vars(Parents))+
  labs(x="Estatura de los padres",y="Estatura de los hijos")
#> `geom_smooth()` using formula 'y ~ x'
```





Asumimos pues como respuesta la variable  $y = \text{Height}$ , como regresores las variables  $x_1 = \text{Father}$  y  $x_2 = \text{Mother}$  con las estaturas del padre y la madre respectivamente, y con factor de clasificación la variable  $G = \text{Gender}$ , con niveles M/F. En principio cabrían posibles interacciones entre los regresores (estaturas del padre y de la madre) y los factores de clasificación (sexo del sujeto). Esto implicaría que las pendientes de relación ‘estatura padres’ versus ‘estatura hijos’ no serían paralelas para los sujetos hombres y mujeres. Planteamos pues, para predecir la respuesta del sujeto  $j$  en el grupo  $i$  del factor de clasificación ( $\text{Gender}$ ),  $y_{ij}$ , el modelo:

$$(y_{ij} | \mu_{ij}, \sigma^2) \sim N(\mu_{ij}, \sigma^2)$$

con el predictor lineal

$$\eta_{ij} = \mu_{ij} = \beta_0 + \alpha_M + (\beta_1 + \beta_1^M)x_{1j} + (\beta_2 + \beta_2^M)x_{2j}; \quad j = 1, \dots, n_i; i = M, F$$

donde  $\alpha_M$  es el efecto diferencial global de los hombres frente a las mujeres al predecir la estatura, y  $\beta_1^M, \beta_2^M$  son los efectos diferenciales que afectan a los regresores para los sujetos varones, y por lo tanto que provocan pendientes distintas al predecir la estatura del sujeto con las de los padres, en función de si este es hombre o mujer.

Asumimos una distribución vaga sobre todos los efectos fijos y  $\tau = 1/\sigma^2$ , y ajustamos el modelo Gausiano en INLA:

```

formula = Height ~ 1+(Father+Mother)*Gender
fit = inla(formula,family = "gaussian",data=datos)
round(fit$summary.fixed,3)
#>               mean      sd 0.025quant 0.5quant 0.975quant
#> (Intercept)   16.652 3.887      9.028   16.652    24.276
#> Father         0.400 0.039      0.324    0.400     0.477
#> Mother         0.307 0.045      0.218    0.307     0.396
#> GenderM        2.707 5.428     -7.940    2.707    13.354
#> Father:GenderM  0.012 0.058     -0.103    0.012     0.126
#> Mother:GenderM  0.027 0.062     -0.096    0.027     0.149
#>               mode kld
#> (Intercept)    NA    0
#> Father          NA    0
#> Mother          NA    0
#> GenderM         NA    0
#> Father:GenderM  NA    0
#> Mother:GenderM  NA    0

```

Observamos que ninguna de las interacciones tienen un efecto a considerar (su RC posterior incluye al cero), de modo que las descartamos y reajustamos el modelo sin ellas.

$$\eta_{ij} = \mu_{ij} = \beta_0 + \alpha_M + \beta_1 x_{1j} + \beta_2 x_{2j}; \quad j = 1, \dots, n; i = M, F.$$

```

formula = Height ~ Father+Mother+Gender
fit = inla(formula,family = "gaussian",data=datos,
           control.predictor=list(compute=TRUE),
           control.compute=list(return.marginals.predictor=TRUE,
                                dic = TRUE, waic = TRUE))
round(fit$summary.fixed,3)
#>               mean      sd 0.025quant 0.5quant 0.975quant
#> (Intercept)  15.345 2.747      9.957   15.345    20.733
#> Father        0.406 0.029      0.349    0.406     0.463
#> Mother        0.321 0.031      0.260    0.321     0.383
#> GenderM       5.226 0.144      4.943    5.226     5.508
#>               mode kld
#> (Intercept)    NA    0
#> Father          NA    0
#> Mother          NA    0
#> GenderM         NA    0

```

Ahora todos los efectos fijos son relevantes para predecir la estatura de los hijos. Utilizamos este modelo para derivar las inferencias.

Representamos a continuación en la Figura 3.5 las distribuciones posteriores de los efectos fijos:

```
fixed=names(fit$marginals.fixed)
g=list()
for(i in 1:length(fixed)){
  g[[i]]=ggplot(as.data.frame(fit$marginals.fixed[[i]]),aes(x=x,y=y))+
    geom_line()+
    geom_vline(xintercept=fit$summary.fixed$mean[i],linetype="dashed")+
    geom_vline(xintercept=fit$summary.fixed[i,3],linetype="dotted")+
    geom_vline(xintercept=fit$summary.fixed[i,5],linetype="dotted")+
    labs(x=fixed[i],y="D.posterior")
}
grid.arrange(g[[1]],g[[2]],g[[3]],g[[4]],ncol=2)
```

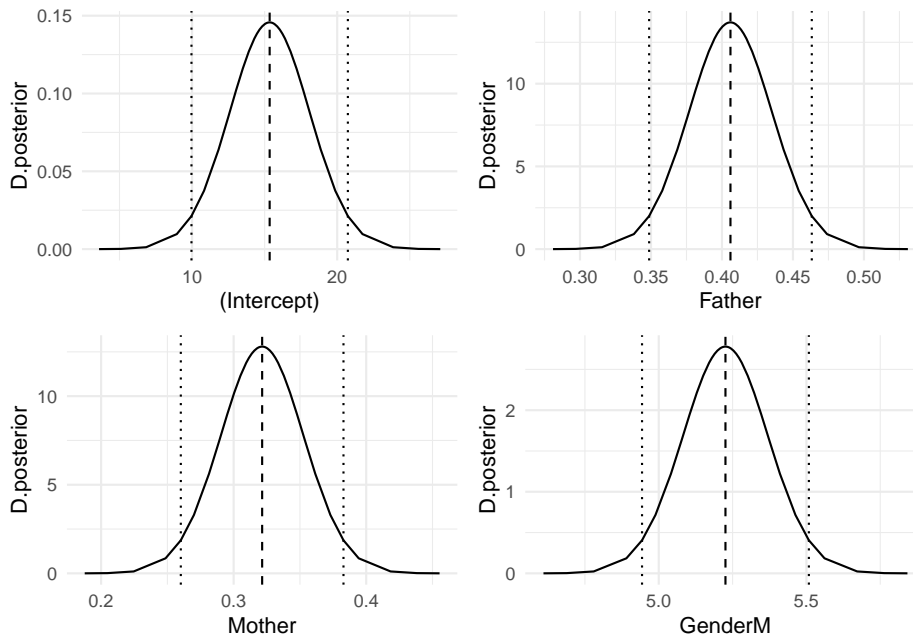


Figure 3.5: Distribución posterior de los efectos fijos

En media observamos que la estatura de los hombres es 5.23 unidades superior a la de las mujeres.

Con estas distribuciones podemos calcular cualquier probabilidad, como por ejemplo, la probabilidad de que la estatura de un hombre supere en 5 unidades a la de una mujer, independientemente de cómo sean sus padres, esto es,

$$Pr(\alpha_M > 5 | \text{datos})$$

```
1-inla.pmarginal(5,fit$marginals.fixed$"GenderM")
#> [1] 0.9414322
```

Podemos acceder a las distribuciones posteriores de la estatura de cualquiera de los sujetos en la muestra y posicionar las estaturas de sus padres, que se muestran en la Figura 3.6

```
# la predicción del predictor lineal para cada sujeto es:
pred<-fit$marginals.linear.predictor
# que en este caso coincide con los valores ajustados
fitted<-fit$marginals.fitted.values
ggplot(as.data.frame(pred$Predictor.1),aes(x=x,y=y))+
  geom_line()+
  labs(x="Estatura media del sujeto 1",y="D.posterior")+
  geom_vline(xintercept=fit$summary.fitted.values$mean[1],linetype="dashed")+
  geom_vline(xintercept=datos$Father[1],linetype="dashed",color="blue")+
  geom_vline(xintercept=datos$Mother[1],linetype="dashed",color="pink")+
  annotate("text",x=datos$Mother[1]+1,y=1,label="Madre")+
  annotate("text",x=datos$Father[1]-1,y=1,label="Padre")
```

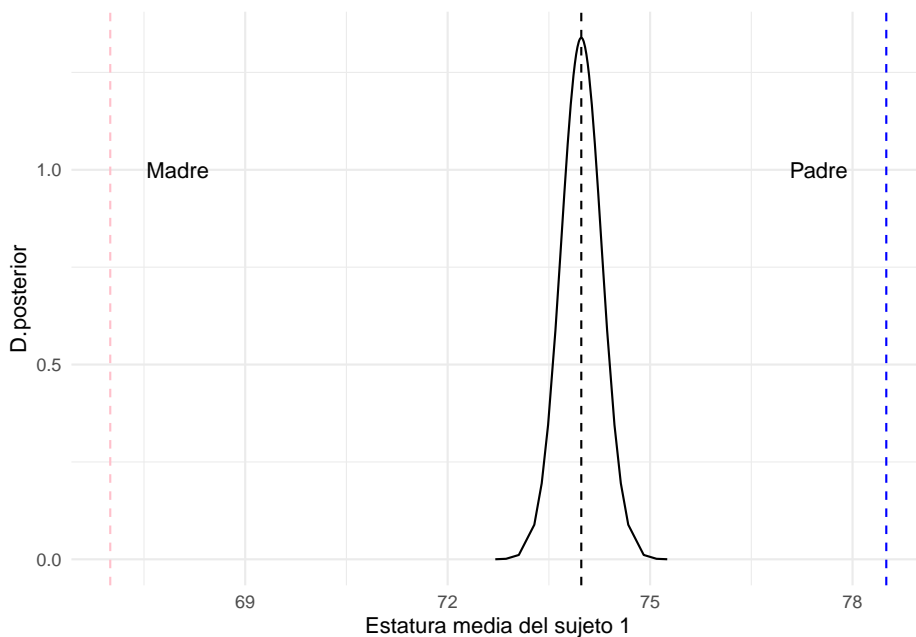


Figure 3.6: Predicción de la estatura del primer sujeto en la muestra

Podemos ir más allá, infiriendo sobre la estatura (esperada) de un sujeto, sea

hombre o mujer, cuando su padre mide 1.75m (68.9 pulgadas) y su madre 1.70m (66.9 pulgadas). Expresamos los resultados en centímetros.

```
new.data=data.frame(Father=c(68.9,68.9),
                    Mother=c(66.9,66.9),
                    Gender=c("M","F"),
                    Height=c(NA,NA))
datos.combinado <- rbind(datos, data.frame(Family=c(NA,NA),new.data,Kids=c(NA,NA)))

## creamos un vector con NA's para observaciones y 1's para predicciones
datos.indicador <- c(rep(NA, nrow(datos)), rep(1, nrow(new.data)))
## reajustamos el modelo añadiendo la opción de predicción de datos
fit.pred <- inla(formula, data = datos.combinado,
                control.compute=list(return.marginals.predictor=TRUE),
                control.predictor = list(link = datos.indicador))
## y describimos los valores ajustados para los escenarios añadidos
round(fit.pred$summary.fitted.values[(nrow(datos)+1):nrow(datos.combinado),]*2.54,1)
#>               mean sd 0.025quant 0.5quant
#> fitted.Predictor.899 177.9 0.3      177.3    177.9
#> fitted.Predictor.900 164.7 0.3      164.0    164.7
#>               0.975quant mode
#> fitted.Predictor.899   178.6   NA
#> fitted.Predictor.900   165.3   NA
```

También graficar las distribuciones posteriores y calcular las probabilidades, por ejemplo, de que dicho sujeto supere el 1.65m si es mujer, o el 1.78m si es hombre (Figura 3.7).

```
# Distribuciones predictivas
pred.M=as.data.frame(fit.pred$marginals.fitted.values[[nrow(datos)+1]])*2.54
pred.F=as.data.frame(fit.pred$marginals.fitted.values[[nrow(datos)+2]])*2.54
d.pred=rbind(pred.M,pred.F)

# atributo Gender
d.pred$Gender=rep(c("M","F"),c(nrow(pred.M),nrow(pred.F)))
# objetivo de estatura
d.pred$obj=rep(c(178,165),c(nrow(pred.M),nrow(pred.F)))

# cálculo de probabilidades
p165F=round(1-inla.pmarginal(165,pred.F),2)
cat(paste("Pr(estatura>165|mujer,padre=175,madre=170)=",p165F))
#> Pr(estatura>165|mujer,padre=175,madre=170)= 0.16
cat("\n")
p178M=round(1-inla.pmarginal(178,pred.M),2)
cat(paste("Pr(estatura>178|hombre,padre=175,madre=170)=",p178M))
```

```
#> Pr(estatura>178/hombre,padre=175,madre=170)= 0.42

d.pred$prob=rep(c(p178M,p165F),c(nrow(pred.M),nrow(pred.F)))

ggplot(d.pred,aes(x=x,y=y))+
  geom_line(aes(color=Gender))+
  geom_vline(aes(xintercept=obj),linetype="dashed")+
  facet_wrap(vars(Gender),scales="free")+
  labs(x="Estatura",y="D.posterior")+
  theme(legend.position = "none")
```

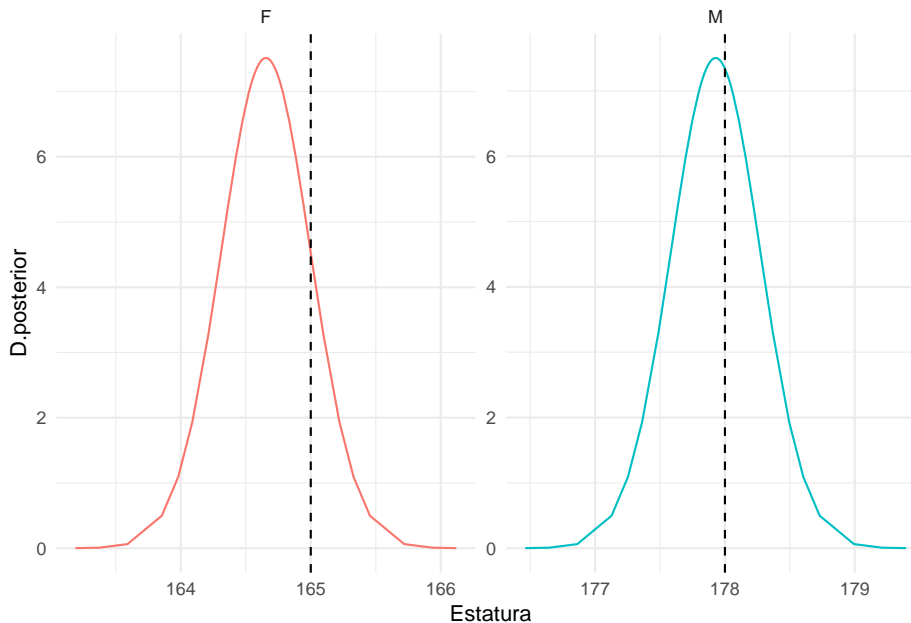


Figure 3.7: Distribución posterior de la estatura de un sujeto cuyo padre mide 1,75cm y madre 1,07cm.

Podríamos también, modificar las especificaciones a priori sobre los parámetros  $\beta_0$  y  $\beta_1$  mediante el comando `control.fixed`. Por ejemplo, queremos asumir a priori  $\beta_0 \sim N(0, 10^4)$  y  $\beta_1 \sim N(0, 100)$  y ver cómo afecta a las inferencias.

```
fit<-inla(formula,family="gaussian",data=datos,
          control.fixed=list(mean=0,prec=0.01,
                             mean.intercept=0, prec.intercept=0.0001))
round(fit$summary.fixed,3)
#>          mean      sd 0.025quant 0.5quant 0.975quant
```

```
#> (Intercept) 15.335 2.745      9.950 15.335 20.721
#> Father      0.406 0.029      0.349 0.406 0.463
#> Mother      0.322 0.031      0.260 0.322 0.383
#> GenderM     5.225 0.144      4.942 5.225 5.507
#>
#> mode kld
#> (Intercept) NA 0
#> Father      NA 0
#> Mother      NA 0
#> GenderM     NA 0
```

Si queremos especificar medias a priori diferentes para los coeficientes de los distintos regresores, hemos de especificarlos con listas.

```
fit = inla(formula,family = "gaussian",data=datos,
           control.fixed=list(mean=list(Father=0.2,Mother=0.1)))
round(fit$summary.fixed,3)
#>
#> mean sd 0.025quant 0.5quant 0.975quant
#> (Intercept) 15.345 2.746      9.958 15.345 20.733
#> Father      0.406 0.029      0.349 0.406 0.463
#> Mother      0.321 0.031      0.260 0.321 0.383
#> GenderM     5.226 0.144      4.943 5.226 5.508
#>
#> mode kld
#> (Intercept) NA 0
#> Father      NA 0
#> Mother      NA 0
#> GenderM     NA 0
```

Si queremos modificar la especificación de la prior en  $\sigma^2$ , o lo que es equivalente, en la precisión  $\tau$ , con la distribución  $\log(\tau) \sim N(0,1)$  en lugar de  $\tau \sim Ga(1,10^{-5})$ , vemos cómo afecta a la inferencia posterior sobre la precisión.

```
fit_n = inla(formula,family="gaussian", data=datos,
             control.family=list(hyper=list(
               prec=list(prior="gaussian",param=c(0,1)))))
# con el modelo log-gamma para precisión
round(fit$summary.hyperpar,3)
#>
#> mean sd
#> Precision for the Gaussian observations 0.216 0.01
#>
#> 0.025quant 0.5quant
#> Precision for the Gaussian observations 0.197 0.216
#>
#> 0.975quant mode
#> Precision for the Gaussian observations 0.236 NA
# con el modelo normal para precisión
round(fit_n$summary.hyperpar,3)
#>
#> mean sd
```

```
#> Precision for the Gaussian observations 0.216 0.01
#>                                     0.025quant 0.5quant
#> Precision for the Gaussian observations 0.197 0.216
#>                                     0.975quant mode
#> Precision for the Gaussian observations 0.237 NA
```

Cuando tenemos información previa disponible sobre la variación de los datos, será generalmente más intuitivo expresarla en términos de la desviación estándar  $\sigma$ . Bastará con conseguir la equivalencia en la escala de  $\log(\tau)$  para incluirla en el modelo. Por ejemplo, si sabemos que la desviación típica está entre 2 y 14,  $\sigma \sim \text{Unif}(2, 14)$ , podemos calcular una prior para la log-precisión del siguiente modo:

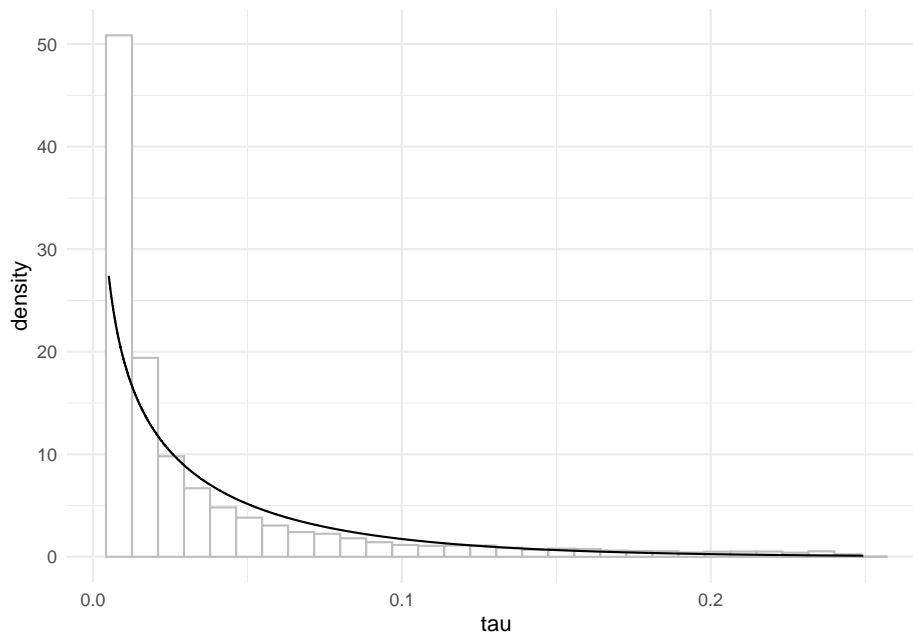
1. simular una muestra de  $\sigma \sim \text{Unif}(2, 14)$
2. transformar a precisiones
3. calcular los parámetros de la Gamma para la precisión, a partir de su media y varianza

Hacemos los cálculos y graficamos la prior en la Figura 3.8.

```
# parámetros para sigma~Un(a1,b1)
a1<-2
b1<-14
# simulamos sigma de una distribución Unif(a1,b1)
sigma<-runif(n=10000,min=a1,max=b1)
# obtenemos la precisión
tau<-1/sigma^2
# Calculamos los parámetros alpha,beta de una distrib. Gamma para la precisión
# mean=alpha/beta; var=alpha/beta^2
beta= mean(tau)/var(tau)
alpha<-mean(tau)*beta
# dibujamos los valores de la precisión
tau.seq=sort(tau)
# seq(min(tau),max(tau),length=1000)
prior=data.frame(tau=tau.seq,dprior=dgamma(tau.seq,alpha,beta))
ggplot(prior, aes(x=tau))+
  geom_histogram(aes(y=..density..),color="grey",fill="white")+
  geom_line(aes(y=dprior))
#> `stat_bin()` using `bins = 30`. Pick better value with
#> `binwidth`.
```

Utilicemos pues esos parámetros para especificar la prior sobre  $\tau$  en INLA:



Figure 3.8: Distribución a prior para tau con  $\sigma \sim \text{Uniforme}(2,14)$ 

```
fit = inla(formula,family="gaussian",data=datos,
           control.family=list(hyper=list(
             prec=list(prior="loggamma",param=c(alpha,beta))))))
round(fit$summary.hyperpar,3)
#>
#> Precision for the Gaussian observations 0.214 0.01
#>
#> 0.025quant 0.5quant
#> Precision for the Gaussian observations 0.195 0.214
#>
#> 0.975quant mode
#> Precision for the Gaussian observations 0.234 NA
```

### 3.6 Efectos aleatorios

Desde una perspectiva frecuentista un modelo básico de Anova podría ser un modelo de efectos fijos, pero también de efectos aleatorios. Así por ejemplo el ‘tratamiento’ dado en un ensayo clínico es un efecto relevante para comparar y diferenciar cómo afecta a la respuesta; ‘tratamiento’ sería entonces un efecto fijo, por ser un efecto de interés primario. En otro ejemplo, se han aplicado varios fertilizantes a cultivos en fincas distintas; el interés primario será comparar los fertilizantes, pero la diversidad de fincas solo se ha incluido para introducir

variabilidad e incrementar, por supuesto, el número de datos en el estudio; así pues, ‘fertilizante’ será un efecto fijo, pero no es un objetivo comparar las fincas, por lo que se considerará como un efecto aleatorio.

Una variable predictiva, numérica o categórica, entra en el modelo como **efecto fijo** cuando se piensa que afecta a todas las observaciones del mismo modo, y que su efecto es de interés primario en el estudio. En un contexto bayesiano un efecto fijo tendrá un coeficiente asociado al que se le asigna a menudo una distribución a priori vaga (mínimo informativa), como una gaussiana con media cero y varianza (conocida) grande. En cualquier caso, la distribución a priori que se asume para los efectos fijos es siempre una distribución conocida.

Un **efecto aleatorio** identifica a variables de tipo categórico que no son de interés primario en la investigación, pero que se considera que añaden incertidumbre y por lo tanto variabilidad a la respuesta. La modelización habitual de los efectos aleatorios es una prior gaussiana con media cero y una precisión desconocida, para la que será preciso asignar, así mismo, una distribución a priori. La distribución a priori de los efectos aleatorios tiene parámetros desconocidos, llamados **hiperparámetros**, a los que habrá que asignar también distribuciones a priori.

Puesto que no salimos del modelo lineal, seguiremos asumiendo una respuesta normal, *gaussian*, con media igual a un predictor lineal  $\mu = \eta = \theta + Zu$ , donde  $Z$  es la correspondiente matriz de diseño para los efectos aleatorios  $z_1, z_2, \dots$ . Se asume además una varianza desconocida  $\sigma^2$ .

En INLA la fórmula de predicción de una respuesta  $y$  a partir de un conjunto de efectos aleatorios  $z_1, z_2, \dots$  se especifica como:

```
formula = y ~ 1 + f(z1, model="") + f(z2,model="")
```

donde la función  $f()$  especifica la relación entre el predictor lineal de la respuesta y los efectos aleatorios  $z$ . La función  $f()$  tiene muchos argumentos, que se pueden consultar con el comando `?f`. El tipo de relación asumida se incluye en el argumento `model` o modelo latente, que tiene como posibilidades `names(inla.models())$latent`. En el modelo lineal, la opción habitual es `model="iid"`, que asume efectos aleatorios independientes e idénticamente distribuidos.

```
names(inla.models())$latent)
#> [1] "linear"      "iid"         "mec"
#> [4] "meb"        "rgeneric"    "cgeneric"
#> [7] "rw1"        "rw2"        "crw2"
#> [10] "seasonal"    "besag"       "besag2"
#> [13] "bym"        "bym2"       "besagproper"
#> [16] "besagproper2" "fgn"        "fgn2"
#> [19] "ar1"        "ar1c"       "ar"
```

```
#> [22] "ou"           "intslope"      "generic"
#> [25] "generic0"      "generic1"      "generic2"
#> [28] "generic3"      "spde"          "spde2"
#> [31] "spde3"         "iid1d"         "iid2d"
#> [34] "iid3d"         "iid4d"         "iid5d"
#> [37] "iidkd"         "2diid"         "z"
#> [40] "rw2d"         "rw2diid"       "slm"
#> [43] "matern2d"      "dmatern"       "copy"
#> [46] "clinear"       "sigm"          "reusigm"
#> [49] "log1exp"       "logdist"
```

Veamos cómo ajustar un modelo de efectos aleatorios a partir de un ejemplo sencillo. Comenzamos con la base de datos **broccoli** en la librería **faraway**. Varios cultivadores suministran brócoli a una planta de procesamiento de alimentos. La planta da instrucciones a los cultivadores para que empaquen el brócoli en cajas de tamaño estándar. Debe haber 18 racimos de brócoli por caja y cada racimo debe pesar entre 1,33 y 1,5 libras. Debido a que los productores utilizan diferentes variedades, métodos de cultivo, etc., hay cierta variación en el peso de los racimos. El responsable de la planta seleccionó 3 cultivadores al azar y luego 4 cajas al azar suministradas por estos cultivadores. Se seleccionaron 3 racimos de brocoli de cada caja (a modo de repeticiones).

La variable de interés es el peso del racimo de brócoli, en la variable **wt**. Sin embargo, dado cómo se ha seleccionado la muestra, el objetivo no es ni la comparación entre cultivadores (**grower**), ni entre cajas (**box**), asumiendo que tenemos varios racimos (**cluster**) en cada una de las combinaciones de los anteriores factores. Sin embargo, de manera lógica intuimos que habrá variabilidad entre cajas (efecto aleatorio **box**) y también entre cultivadores (efecto aleatorio **grower**), lo que nos conduce a un modelo en el que todos los predictores, **box** y **grower** intervienen como efectos aleatorios; la variable **cluster** la aprovechamos a modo de repeticiones de medidas en una misma caja de un mismo cultivador.

La base de datos cuenta con 36 registros (3 observaciones en cada combinación **grower-box**).

$$(y_{ijk} | \mu_{ij}, \sigma^2) \sim N(\mu_{ij}, \sigma^2)$$

con

$$\eta_{ij} = \mu_{ij} = \theta + \alpha_i^G + \beta_j^B; \quad i = 2, 3; j = 2, 3, 4$$

el peso medio que comparten todos los racimos en cada combinación de agricultor-caja:  $k = 1, 2, 3$ , y donde  $\alpha^G$  representa el efecto aleatorio asociado al cultivador (**grower**) y  $\beta^B$  a la caja (**box**).

Así el vector de efectos latentes está compuesto por el efecto fijo de interceptación  $\theta$  y los efectos aleatorios  $u = (\alpha_2^G, \alpha_3^G, \beta_2^B, \beta_3^B, \beta_4^B)$ .

El siguiente paso es especificar una distribución a priori sobre los parámetros. INLA por defecto asigna una prior difusa sobre la interceptación  $\theta$  y también sobre la precisión de los datos  $\tau = 1/\sigma^2$ . Dado que los  $\alpha_i^G$  representan el efecto diferencial asociado al cultivador, es razonable asumir independencia entre todos estos parámetros y una distribución idéntica, centrada en el cero (ante ausencia de información) y con una varianza desconocida. Con esto estamos diciendo que en principio no tenemos información sobre que efectivamente el efecto cultivador sea relevante (media cero), pero sí que añade cierta variabilidad  $\sigma_\alpha^2$  a la respuesta. Del mismo modo, se asume que los  $\beta_j^B$  son a priori independientes e idénticamente distribuidos (iid) con una normal centrada en el cero (ante ausencia de información) y con varianza desconocida  $\sigma_\beta^2$ .

$$\begin{aligned}\theta &\sim N(0, \sigma_\theta^2), \sigma_\theta^2 = \infty \\ \log(\tau) &\sim \text{Log} - \text{Ga}(1, 5 \cdot 10^{-5}) \\ \alpha_i^G &\sim_{iid} N(0, \sigma_\alpha^2), i = 2, 3 \\ \beta_j^B &\sim_{iid} N(0, \sigma_\beta^2), j = 2, 3, 4\end{aligned}$$

Surgen pues, dos nuevos parámetros en las a priori, o hiperparámetros,  $\sigma_\alpha^2$  y  $\sigma_\beta^2$ , a los que también habrá que asignar una distribución a priori. Dado que se trata de varianzas, por defecto INLA asume gammas inversas difusas, o lo que es lo mismo, log-gammas difusas para las precisiones

$$\begin{aligned}\tau_\alpha = 1/\sigma_\alpha^2 &\sim \text{Ga}(1, 5 \cdot 10^{-5}) \\ \tau_\beta = 1/\sigma_\beta^2 &\sim \text{Ga}(1, 5 \cdot 10^{-5})\end{aligned}$$

Surgen pues, tres niveles de especificación del modelo: datos, parámetros e hiperparámetros, que generan un **modelo jerárquico de tres niveles**, y sobre el que hablaremos más adelante.

```
data(broccoli, package="faraway")
str(broccoli)
#> 'data.frame': 36 obs. of 4 variables:
#> $ wt      : num 352 369 383 339 367 328 376 359 388 365 ...
#> $ grower  : Factor w/ 3 levels "1","2","3": 1 1 1 2 2 2 3 3 3 1 ...
#> $ box     : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 2 ...
#> $ cluster: Factor w/ 3 levels "1","2","3": 1 2 3 1 2 3 1 2 3 1 ...
formula = wt ~ f(grower,model="iid")+ f(box,model="iid")
fit = inla(formula, family="gaussian",data=broccoli,
           control.compute = list(dic=TRUE,waic=TRUE))
```

Cuando queremos mostrar los resultados a posteriori sobre los efectos aleatorios a partir de un ajuste `fit` con `inla`, tenemos las siguientes opciones:

- `fit$summary.random` resume la inferencia posterior sobre los efectos aleatorios
- `names(fit$marginals.random)` lista los nombre de todos los efectos aleatorios
- `fit$marginals.random` da las distribuciones posteriores marginales de los efectos aleatorios

```
fit$summary.random
#> $grower
#>   ID          mean          sd 0.025quant    0.5quant
#> 1  1  1.733783e-06 0.01192254 -0.02570770  1.024464e-06
#> 2  2 -1.213642e-05 0.01192255 -0.02574427 -7.171168e-06
#> 3  3  1.040269e-05 0.01192254 -0.02568487  6.146754e-06
#>   0.975quant mode          kld
#> 1 0.02571684   NA 0.001147299
#> 2 0.02568031   NA 0.001147313
#> 3 0.02573970   NA 0.001147309
#>
#> $box
#>   ID          mean          sd 0.025quant    0.5quant
#> 1  1  1.712530e-05 0.01061304 -0.02318818  9.345921e-06
#> 2  2 -9.818468e-06 0.01061302 -0.02326660 -5.358246e-06
#> 3  3 -4.338381e-06 0.01061301 -0.02325063 -2.367567e-06
#> 4  4 -2.968336e-06 0.01061301 -0.02324664 -1.619892e-06
#>   0.975quant mode          kld
#> 1 0.02328791   NA 0.003488241
#> 2 0.02320942   NA 0.003488125
#> 3 0.02322537   NA 0.003488079
#> 4 0.02322935   NA 0.003488073
```

Sin embargo, lo relevante en un modelo de efectos aleatorios es la inferencia sobre las varianzas asociadas a los datos, pero también la variabilidad extra que añaden los efectos aleatorios:

```
fit$summary.hyperpar
#>
#> Precision for the Gaussian observations 3.796674e-03
#> Precision for grower 1.007448e+04
#> Precision for box 1.822326e+04
#>
#> sd
#> Precision for the Gaussian observations 9.517889e-04
#> Precision for grower 1.134337e+04
#> Precision for box 2.091938e+04
#>
#> 0.025quant
#> Precision for the Gaussian observations 2.068520e-03
```

```
#> Precision for grower 1.035933e+03
#> Precision for box 8.755046e+02
#> 0.5quant
#> Precision for the Gaussian observations 3.764493e-03
#> Precision for grower 6.683418e+03
#> Precision for box 1.163807e+04
#> 0.975quant mode
#> Precision for the Gaussian observations 5.739575e-03 NA
#> Precision for grower 3.980374e+04 NA
#> Precision for box 7.386217e+04 NA
```

Vemos que tanto la precisión asociada al efecto aleatorio caja (box), como al efecto cultivador, **grower**, son muy grandes, lo que implica varianzas muy pequeñas que posiblemente nos permitiría prescindir de dichos efectos aleatorios para ajustar un mejor modelo. Cuando transformamos a escala de desviaciones estándar, tenemos la distribución posterior para los tres tipos de error, representados en la Figura 3.9.

```
nombres=c("sigma","grower","box")
sigma.post=as.data.frame(
  inla.tmarginal(function(tau) tau^(-1/2),
    fit$marginals.hyperpar[[1]]))
sigma.grower.post =as.data.frame(
  inla.tmarginal(function(tau) tau^(-1/2),
    fit$marginals.hyperpar[[2]]))
sigma.box.post = as.data.frame(
  inla.tmarginal(function(tau) tau^(-1/2),
    fit$marginals.hyperpar[[3]]))

sigma=rbind(sigma.post,sigma.grower.post,sigma.box.post)
sigma$efecto=rep(c("sigma","grower","box"),
  c(nrow(sigma.post),
    nrow(sigma.grower.post),
    nrow(sigma.box.post)))

ggplot(sigma,aes(x=x,y=y)) +
  geom_line(aes(color=efecto)) +
  labs(x=expression(sigma),y="D.Posterior")+
  facet_wrap(vars(efecto),scales = "free")+
  theme(axis.text.x = element_text(angle = 45),
    legend.position="none")
```

No obstante, antes de tomar una decisión sobre la exclusión de los efectos aleatorios, vamos a hacer una aproximación del porcentaje de varianza explicada por

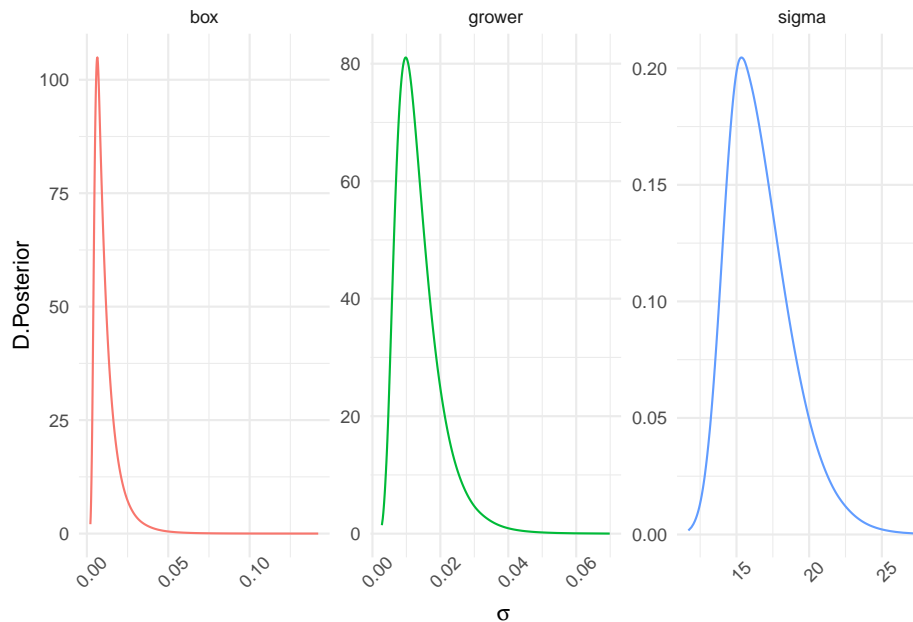


Figure 3.9: Distribución posterior de la desviación típica para las tres fuentes de error: datos, caja y cultivador

cada una de estas fuentes de variación. Utilizando simulaciones de las distribuciones posteriores de  $\sigma^2$ ,  $\sigma_\alpha^2$  y  $\sigma_\beta^2$  vamos a calcular la contribución a la varianza del efecto cultivador,  $\sigma_\alpha^2/(\sigma^2 + \sigma_\alpha^2 + \sigma_\beta^2)$  y la contribución a la varianza del efecto caja,  $\sigma_\beta^2/(\sigma^2 + \sigma_\alpha^2 + \sigma_\beta^2)$ . Esto es, vamos a simular de las distribuciones posteriores de las contribuciones a la varianza, y calcular con dichas distribuciones, la probabilidad de que sea suficientemente grande, por ejemplo de que dicha contribución sea mayor a un 1%. Si dicha probabilidad es considerable, estaremos diciendo que el efecto cultivador (o caja) está provocando demasiada variabilidad, indicador de que no se están cumpliendo los estándares de calidad.

$$Pr\left(\frac{\sigma_\alpha^2}{\sigma^2 + \sigma_\alpha^2 + \sigma_\beta^2} | y\right) \geq 0.01; \quad Pr\left(\frac{\sigma_\beta^2}{\sigma^2 + \sigma_\alpha^2 + \sigma_\beta^2} | y\right) \geq 0.01$$

```
n=1000
tau=as.data.frame(
  inla.hyperpar.sample(n,fit,improve.marginals=TRUE))
sigma2=apply(tau,2,function(x) 1/x)
colnames(sigma2)=c("sigma2d","sigma2G","sigma2B")
sigma2=as.data.frame(sigma2)
sigma2=sigma2 %>%
  mutate(contrib.G=sigma2G/(sigma2d+sigma2G+sigma2B),
```

```

contrib.B=sigma2B/(sigma2d+sigma2G+sigma2B))

# contribución a la varianza de grower: Pr(contrib.G>1/100)
cat(paste("Prob.contribución de grower a la varianza > 1%=",
          mean(sigma2$contrib.G > 0.01),"\n"))
#> Prob.contribución de grower a la varianza > 1%= 0
# contribución a la varianza de box: Pr(contrib.B>1/100)
cat(paste("Prob.contribución de box a la varianza > 1%=",
          mean(sigma2$contrib.B > 0.01)))
#> Prob.contribución de box a la varianza > 1%= 0

```

Ante estos resultados, con probabilidad 0 de que dichos efectos aporten a la varianza más de un 1%, se justifica la opción de prescindir de los efectos `grower` y `box` como efectos aleatorios y ajustar el modelo con un único efecto fijo global.

```

formula = wt ~ 1
fit = inla(formula, family="gaussian",data=broccoli,
           control.compute = list(dic=TRUE,waic=TRUE))
fit$summary.fixed
#>              mean          sd 0.025quant 0.5quant
#> (Intercept) 358.1666 2.776124   352.6935 358.1666
#>              0.975quant mode          kld
#> (Intercept)  363.6397  NA 1.142468e-08
fit$summary.hyperpar
#>
#>              mean
#> Precision for the Gaussian observations 0.003776544
#>
#>              sd
#> Precision for the Gaussian observations 0.0008759687
#>
#>              0.025quant
#> Precision for the Gaussian observations 0.002268571
#>
#>              0.5quant
#> Precision for the Gaussian observations 0.003707896
#>
#>              0.975quant mode
#> Precision for the Gaussian observations 0.005693093  NA

```

Vemos que la variación en los indicadores DIC (308.0208063) y WAIC (307.7907593) es despreciable para este nuevo modelo.

Inferimos a continuación con las distribuciones posteriores para la media global y la varianza de los datos.

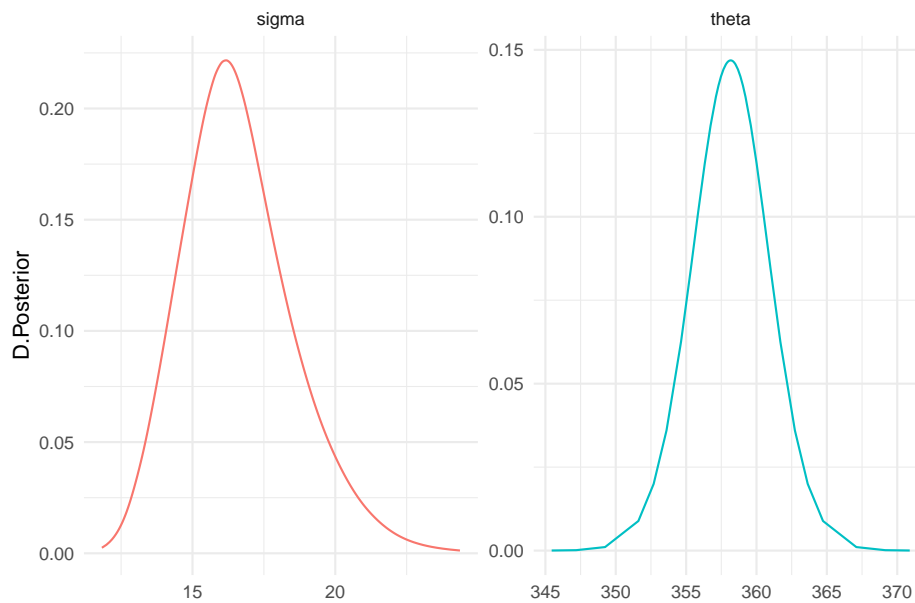
```

theta.post = as.data.frame(fit$marginals.fixed[[1]])
sigma.post=as.data.frame(inla.tmarginal(function(tau) tau^(-1/2),
  fit$marginals.hyperpar[[1]]))

```



```
posterior=rbind(theta.post,sigma.post)
posterior$efecto=rep(c("theta","sigma"),
                    c(nrow(theta.post),nrow(sigma.post)))
ggplot(posterior,aes(x=x,y=y)) +
  geom_line(aes(color=efecto)) +
  labs(x="",y="D.Posterior")+
  facet_wrap(vars(efecto),scales = "free")+
  theme(legend.position = "none")
```



Hemos concluido con este análisis, que todos los cultivadores han respetado los protocolos de calidad establecidos para el empaquetado en cajas.

### 3.7 Modelos mixtos

En ocasiones cuando ajustamos un modelo lineal tendremos algunos factores de clasificación que operan como efectos fijos y otros que operan como efectos aleatorios. Estaremos ante **modelos lineales mixtos**. Siendo estrictos, realmente el modelo con solo efectos aleatorios ya es un modelo mixto, puesto que incluye como efecto fijo una interceptación global.

En un modelo lineal mixto seguimos asumiendo una respuesta normal, *gaussian*, con media igual a un predictor lineal  $\eta = X\beta + Zu$ , donde  $X$  es una matriz de diseño con los efectos fijos  $x_1, x_2, \dots$ , y  $Z$  la correspondiente para los efectos aleatorios  $z_1, z_2, \dots$ . Se asume además una varianza desconocida que puede ser distinta para distintos niveles de los predictores, e incluso contener correlaciones

entre niveles distintos, y que en general se suele expresar a través de una matriz de covarianzas  $\Sigma$ ,  $(y|\eta, \Sigma) \sim N(\eta, \Sigma)$ .

En INLA la fórmula de predicción de una respuesta  $y$  a partir de un conjunto de efectos fijos  $x_1, x_2, \dots$ , y un conjunto de efectos aleatorios  $z_1, z_2, \dots$  se especifica como:

```
formula = y ~ 1 + x1 + x2 + f(z1, model="") + f(z2, model="")
```

De nuevo mencionar que la opción más habitual para los efectos aleatorios en un modelo lineal es `model="iid"`.

### 3.7.1 Datos longitudinales con pendientes iguales

Veamos cómo resolver las inferencias a través de un ejemplo disponible en R-bloggers, proporcionado por Patrick Curran y descargable desde Github. Se refieren estos datos, a un estudio con 405 niños en los dos primeros años de la escuela infantil, medidos a lo largo de cuatro instantes equidistantes (no medidos todos en todos los sujetos) para registrar su progreso en lectura y en comportamiento antisocial. Nos centramos aquí exclusivamente en intentar predecir los progresos en lectura (variable `read`) para cada sujeto (identificado como `id`) a lo largo de los 4 instantes de medición (`occasion`).

Cargamos los datos, prescindimos de los que tienen valores faltantes, y los inspeccionamos en la Figura 3.10.

```
url="https://raw.githubusercontent.com/BayesModel/data/main/curran_dat.csv"
curran_dat=read.csv(url) %>%
  select(id, occasion, read) %>%
  filter(complete.cases(.))
# el identificador de cada sujeto lo convertimos a factor
curran_dat$id=as.factor(curran_dat$id)
curran_dat$occasion=as.double(curran_dat$occasion)
# Relaciones
g1=ggplot(curran_dat, aes(x=as.factor(occasion), y=read))+
  geom_boxplot()
g2=ggplot(curran_dat, aes(x=occasion, y=read))+
  geom_line(aes(group=id), color="grey", size=0.4)
grid.arrange(g1, g2, ncol=2)
```

Como base vamos a asumir normalidad en la respuesta de un sujeto  $i$  en un instante  $j$ , y plantear un modelo lineal para obtener nuestras conclusiones.

$$(y_{ij}|\mu_{ij}, \sigma^2) \sim N(\mu_{ij}, \sigma^2); i = 1, \dots, 450; j = 1, 2, 3, 4$$

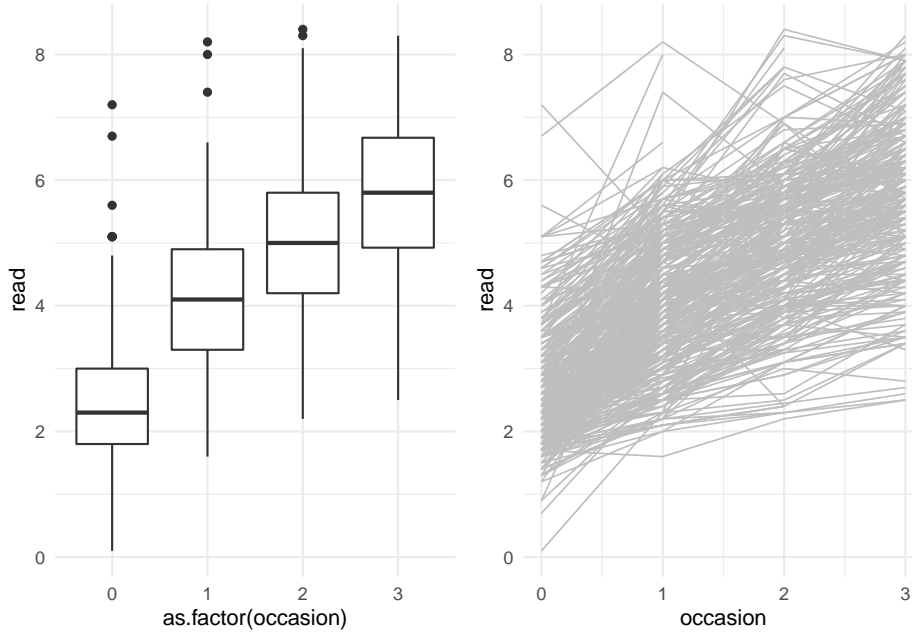


Figure 3.10: Descripción de la BD CurranLong sobre desarrollo de las habilidades lectoras en niños.

A la vista de la Figura 3.10 podríamos considerar el tiempo afecta de modo positivo y lineal sobre las habilidades lectoras (a más tiempo, mejores habilidades), lo que convierte a la variable `occasion` en una covariable numérica (efecto fijo): nos interesará cuantificar cómo afecta el tiempo a la capacidad lectora.

Sin embargo, también en el gráfico apreciamos que cada sujeto arranca de un inicio diferente, o lo que es lo mismo, su recta de predicción tiene una interceptación distinta. Puesto que no nos interesa comparar los individuos, planteamos incorporar un efecto aleatorio del sujeto (variable `id`). Estamos pues, hablando de predecir la habilidad lectora de un sujeto  $i$  en un instante  $t_{ij} = j$  con:

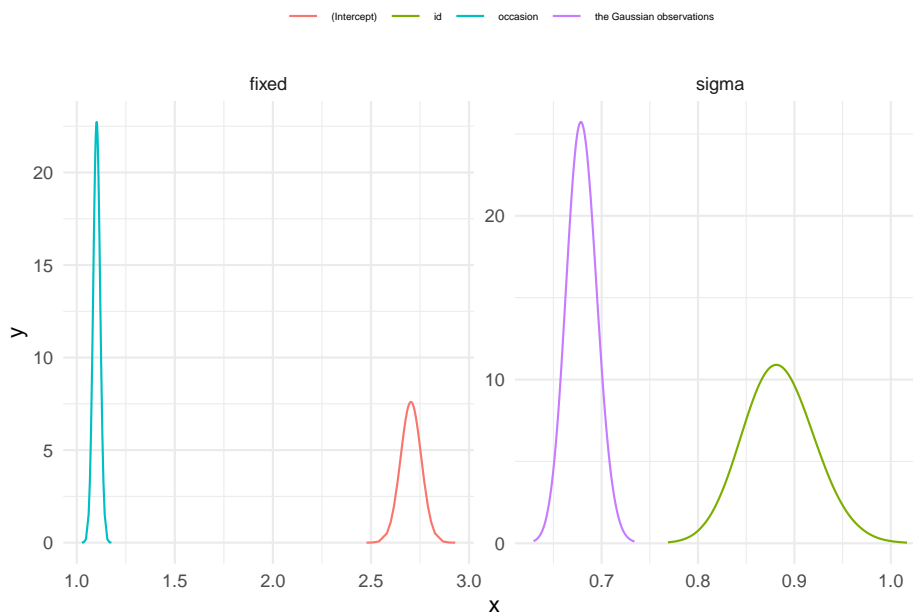
$$\mu_{ij} = \theta + \alpha_i + \beta \cdot t_{ij}$$



```

                                tipo="fixed"))
}
for(j in 1:nhyp){
  res=rbind(res,data.frame(
    inla.tmarginal(function(tau) tau^(-1/2),fit$marginals.hyperpar[[j]]),
    id=str_sub(names(fit$marginals.hyperpar)[j], start =15),
    tipo="sigma"))
}
ggplot(res,aes(x=x,y=y))+
  geom_line(aes(color=id))+
  facet_wrap(vars(tipo),scales="free")+
  theme(legend.position="top",
        legend.title=element_blank(),
        legend.text = element_text(size=5))

```



### Datos longitudinales con pendientes distintas

Belenky et al. (2003) describen un estudio de los tiempos de reacción en pacientes a los que se ha privado de sueño durante 10 días; cada día se ha ido registrando la respuesta para cada uno de los 18 sujetos en el estudio. Los datos están disponibles como `sleepstudy` en la librería `lme4` y tienen como variables el tiempo medio de reacción en microsegundos (`Reaction`), el número de días con privación de sueño (`Days`) y un id para cada sujeto (`Subject`). Los tiempos de reacción se transforman a segundos para tener mayor estabilidad. Aun así, en la Figura 3.11 se aprecia que el número de días de falta de sueño afecta de modo distinto a cada sujeto.

```
data(sleepstudy, package="lme4")
sleepstudy$Reaction <- sleepstudy$Reaction / 1000
ggplot(sleepstudy, aes(x=Days, y=Reaction)) +
  geom_point(size=0.5) +
  geom_smooth(method="lm", color="blue", size=0.5) +
  facet_wrap(vars(Subject), ncol=6) +
  theme(axis.text.x = element_text(size=5),
        axis.text.y = element_text(size=5))
#> `geom_smooth()` using formula 'y ~ x'
```

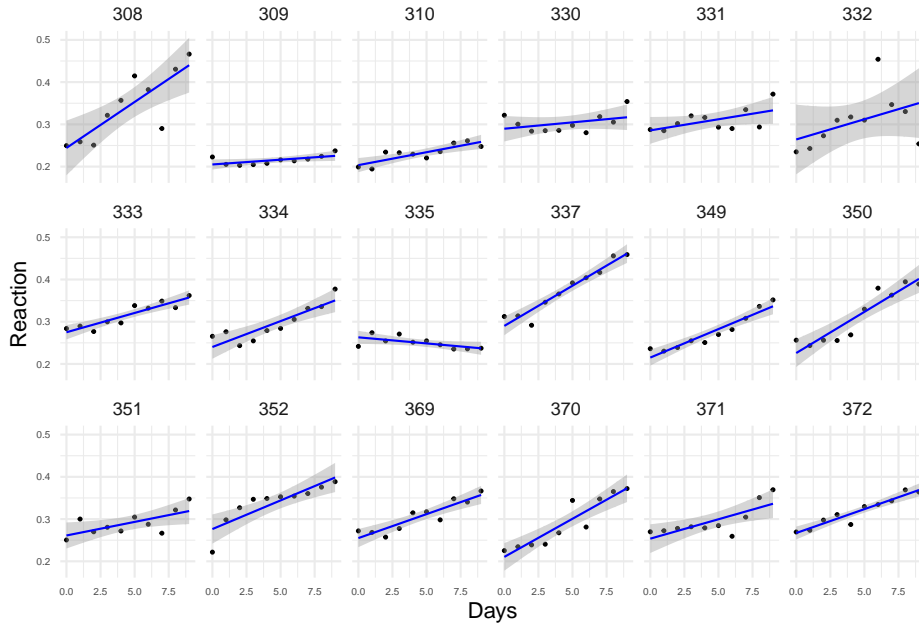


Figure 3.11: Tiempos de reacción en función del número de días con falta de sueño (sleepstudy) para los 18 sujetos en el estudio

Un modelo razonable para estos datos es un modelo lineal que relacione los tiempos de reacción con los días, pero que tenga interceptaciones y pendientes diferentes para cada sujeto. El efecto sujeto entraría en el modelo como un efecto aleatorio para relacionar todos los datos del mismo sujeto sin perder la asunción de independencia entre las observaciones de sujetos distintos. Si llamamos  $y = \text{Reaction}$  a la respuesta, estaríamos planteando el siguiente modelo:

$$y_{ij} | \mu_{ij}, \sigma^2 \sim N(\mu, \sigma^2), i = 1, \dots, 18; j = 1, \dots, 10$$

con

$$\mu_{ij} = \theta + \alpha_i + \beta \cdot x_{ij} + \gamma_{ij}$$

donde el predictor  $x$  es la variable **Days**,  $(\theta, \beta)$  se tratarían como efectos fijos con a priori difusas ante falta de información, y  $(\alpha_i, \gamma_{ij})$  como efectos aleatorios, con normales centradas en cero y una varianza desconocida a la que habría que asignar así mismo, una distribución a priori. El modelo jerárquico que surge es pues:

$$\begin{aligned} &\text{Nivel I} \\ y_{ij} | \mu_{ij}, \sigma^2 &\sim N(\mu, \sigma^2), i = 1, \dots, 18; j = 1, \dots, 10 \\ &\text{Nivel II} \\ \theta &\sim N(0, 1000) \\ \beta &\sim N(0, 1000) \\ \alpha_i &\sim N(0, \sigma_\alpha^2) \\ \gamma_{ij} &\sim N(0, \sigma_\gamma^2) \\ \tau = 1/\sigma^2 &\sim Ga(0.001, 0.001) \\ &\text{Nivel III} \\ \tau_\alpha = 1/\sigma_\alpha^2 &\sim Ga(0.001, 0.001) \\ \tau_\gamma = 1/\sigma_\gamma^2 &\sim Ga(0.001, 0.001) \end{aligned}$$

En INLA modelizamos esta propuesta utilizando como predictores;

- la covariable para generar una interceptación ‘media’ (efecto fijo),
- el efecto aleatorio de cada sujeto para generar interceptaciones distintas,
- la interacción entre la covariable y el efecto aleatorio, a través de la matriz de diseño que hemos de construir específicamente, y definir en paralelo un índice de la misma dimensión de los datos, para aplicarla. En la interacción la primera variable define el número de grupos y la segunda el valor de la covariable.

```
prec.prior=list(prec=list(param=c(0.001,0.001)))
# matriz de diseño para la interacción
Z <- as(model.matrix( ~ 0 + Subject:Days, data = sleepstudy), "Matrix")
# índice para aplicar la matriz de diseño
DayR=1:nrow(sleepstudy)
formula= Reaction ~ Days + f(Subject,model="iid",hyper=prec.prior)+
  f(DayR,model="z",Z=Z,hyper = prec.prior)
fit=inla(formula,family="gaussian",data=sleepstudy,
  control.compute=list(config=TRUE))
#> Warning in inla.model.properties.generic(inla.trim.family(model), mm[names(mm) == : Model 'z'
```

```
#> Use this model with extra care!!! Further warnings are disabled.
#> as(<dgCMatrix>, "dgTMatrix") is deprecated since Matrix 1.5-0; do as(., "TsparseMat
```

Obtenemos en consecuencia, efectos fijos e hiperparámetros, cuyas inferencias posteriores se resumen con:

```
round(fit$summary.fixed,3)
#>               mean      sd 0.025quant 0.5quant 0.975quant mode
#> (Intercept) 0.251 0.008      0.236      0.251      0.267  NA
#> Days        0.010 0.003      0.004      0.010      0.017  NA
#>               kld
#> (Intercept) 0
#> Days        0
round(fit$summary.hyperpar,3)
#>               mean      sd
#> Precision for the Gaussian observations 1566.714 182.872
#> Precision for Subject                  1314.751 571.884
#> Precision for DayR                    6067.452 2128.981
#>               0.025quant 0.5quant
#> Precision for the Gaussian observations 1231.362 1558.540
#> Precision for Subject                  520.882 1209.119
#> Precision for DayR                    2819.867 5762.942
#>               0.975quant mode
#> Precision for the Gaussian observations 1950.856  NA
#> Precision for Subject                  2728.308  NA
#> Precision for DayR                    11105.451  NA
```

Y los efectos aleatorios:

```
names(fit$marginals.random)
#> [1] "Subject" "DayR"
head(fit$summary.random$Subject)
#>   ID      mean      sd 0.025quant 0.5quant
#> 1 308 -0.003768856 0.01446451 -0.0322859126 -0.003748474
#> 2 309 -0.037614872 0.01485942 -0.0674455257 -0.037392697
#> 3 310 -0.038204772 0.01487068 -0.0680610254 -0.037981071
#> 4 330 0.028705919 0.01469937 0.0002843497 0.028533663
#> 5 331 0.025993579 0.01465096 -0.0023753045 0.025835592
#> 6 332 0.009899219 0.01447471 -0.0184009868 0.009836377
#>               0.975quant mode      kld
#> 1 0.024632666  NA 5.824669e-09
#> 2 -0.009029197  NA 2.027899e-08
#> 3 -0.009601924  NA 2.046429e-08
#> 4 0.058095377  NA 1.427017e-08
```

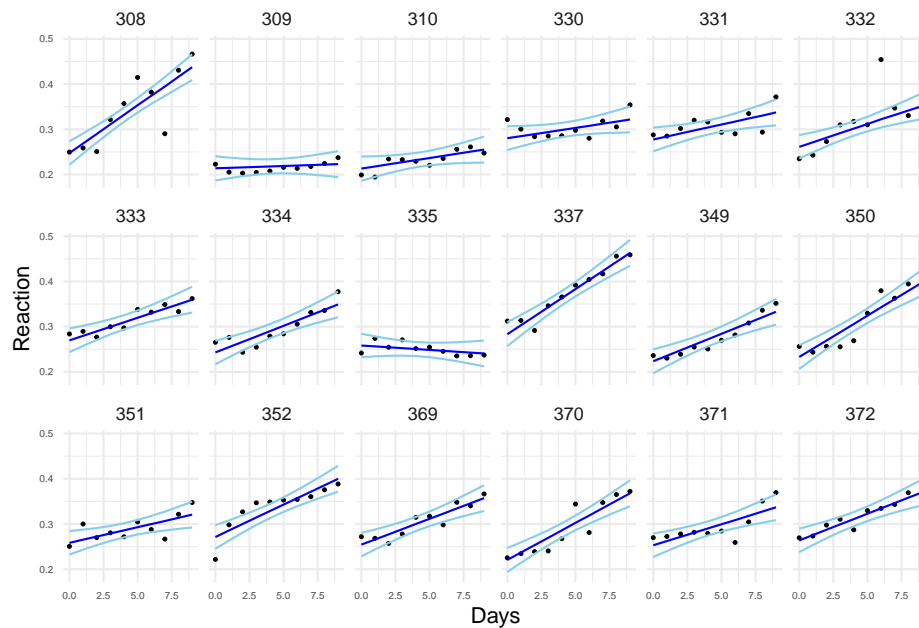


```
#> 5  0.055251050  NA 1.290464e-08
#> 6  0.038554228  NA 6.836835e-09
head(fit$summary.random$DayR)
#>   ID      mean      sd  0.025quant  0.5quant
#> 1  1 6.924185e-06 0.001501481 -0.002937713 6.924389e-06
#> 2  2 1.056644e-02 0.004274275  0.002188710 1.055373e-02
#> 3  3 2.106571e-02 0.008121900  0.005144076 2.103764e-02
#> 4  4 3.184263e-02 0.012056265  0.008205548 3.180029e-02
#> 5  5 4.249501e-02 0.016013228  0.011099082 4.243830e-02
#> 6  6 5.322667e-02 0.019979610  0.014052986 5.315582e-02
#>   0.975quant mode      kld
#> 1 0.00295156  NA 5.527427e-11
#> 2 0.01901743  NA 7.952084e-09
#> 3 0.03714923  NA 1.020978e-08
#> 4 0.05572393  NA 1.066177e-08
#> 5 0.07421816  NA 1.083077e-08
#> 6 0.09280918  NA 1.090781e-08
```

En la Figura ?? mostramos los datos y también los valores ajustados para las rectas, en términos de las interceptaciones y pendientes medias de las correspondientes distribuciones posteriores, además de la banda de estimación que construimos con los correspondientes percentiles de las posteriores.

```
sleepstudy.pred = sleepstudy %>%
  mutate(fitted=fit$summary.fitted.values$mean,
         rc.low=fit$summary.fitted.values$"0.025quant",
         rc.up=fit$summary.fitted.values$"0.975quant")

ggplot(sleepstudy.pred, aes(x=Days, y=Reaction)) +
  geom_point(size=0.5) +
  geom_line(aes(y=fitted), color="blue") +
  geom_line(aes(y= rc.low), color="skyblue") +
  geom_line(aes(y=rc.up), color="skyblue") +
  facet_wrap(vars(Subject), ncol=6) +
  theme(axis.text.x = element_text(size=5),
        axis.text.y = element_text(size=5))
```



En la Figura 3.12 mostramos la distribución posterior de los errores de datos y aleatorios.

```
nhyp=length(names(fit$marginals.hyperpar))
res=NULL
for(j in 1:nhyp){
  res=rbind(res,data.frame(
    inla.tmarginal(function(tau) tau^(-1/2),fit$marginals.hyperpar[[j]]),
    id=str_sub(names(fit$marginals.hyperpar)[j], start =15)))
}
ggplot(res,aes(x=x,y=y))+
  geom_line(aes(color=id))+
  labs(x=expression(sigma),y="")+
  theme(legend.position="top",
        legend.title=element_blank(),
        legend.text = element_text(size=5))
```

### 3.7.2 Efectos anidados

Hablamos de efectos anidados cuando cada miembro de un grupo está contenido completamente dentro de una única unidad de otro grupo. Que un factor A esté anidado en otro B, implica que cada nivel de B contiene niveles distintos de A, esto es, cada nivel de A está vinculado solo a algún nivel de B.

La base de datos `eggs` en la librería `faraway` nos resulta útil para describir este

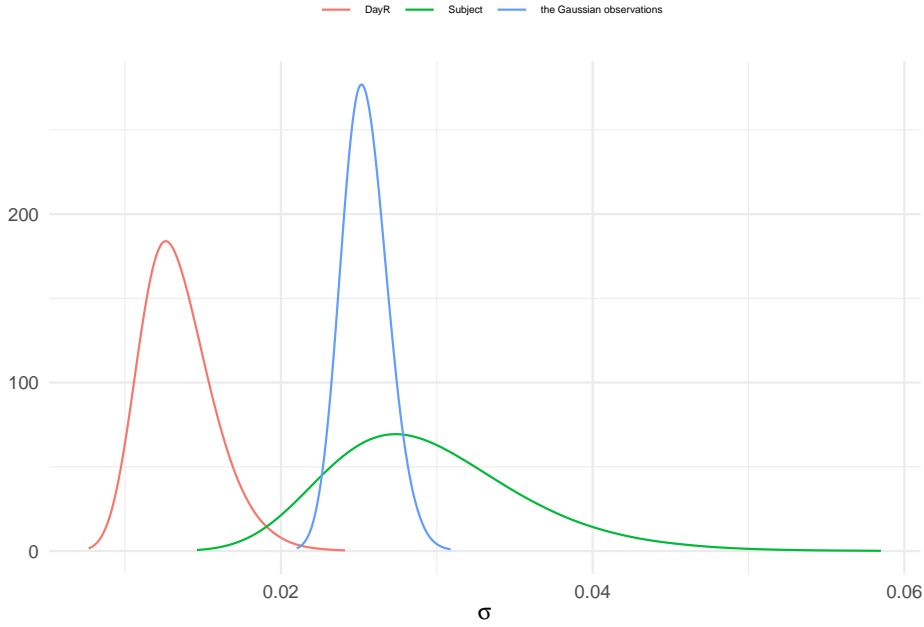


Figure 3.12: Distribución posterior del error de los datos y el error aleatorio

tipo de modelos con efectos anidados. Estos datos son los resultantes de un experimento para testar la consistencia en los tests de laboratorio que realizaban laboratorios distintos, técnicos distintos. Para ello se dividió en varias muestras un tarro de polvo de huevo seco homogeneizado (con idéntico contenido graso). Se enviaron 4 muestras a cada uno de los 6 laboratorios. De esas 4 muestras, 2 se etiquetaron como G y 2 como H (aun siendo idénticas). Se dieron instrucciones a los laboratorios de dar dos muestras a dos técnicos distintos. Los técnicos recibieron instrucciones de dividir sus muestras en dos partes y medir el contenido graso de cada una. Así, cada laboratorio reportó 8 mediciones del contenido graso (**Fat**), cada técnico 4 mediciones, con 2 réplicas en cada una de las dos muestras.

Realmente, el laboratorio, el técnico y la muestra solo deberían generar variabilidad en la respuesta, pero en ningún caso generar mediciones distintas. Estamos pues interesados en investigar la magnitud del error debido al laboratorio, al técnico y a la identificación de muestras. Es por ello que tiene sentido considerarlos efectos aleatorios.

Tenemos así en este ejemplo, a los técnicos (**Technician**) anidados en los laboratorios (**Lab**). En la Figura 3.13 se muestra claramente la variación entre laboratorios, entre técnicos, y debida al efecto irreal de tener dos muestras distintas G y H (**Sample**).

```
data(eggs, package="faraway")
ggplot(eggs, aes(x=Lab, y=Fat)) +
  geom_boxplot(aes(color=Technician)) +
  facet_wrap(vars(Sample))
```

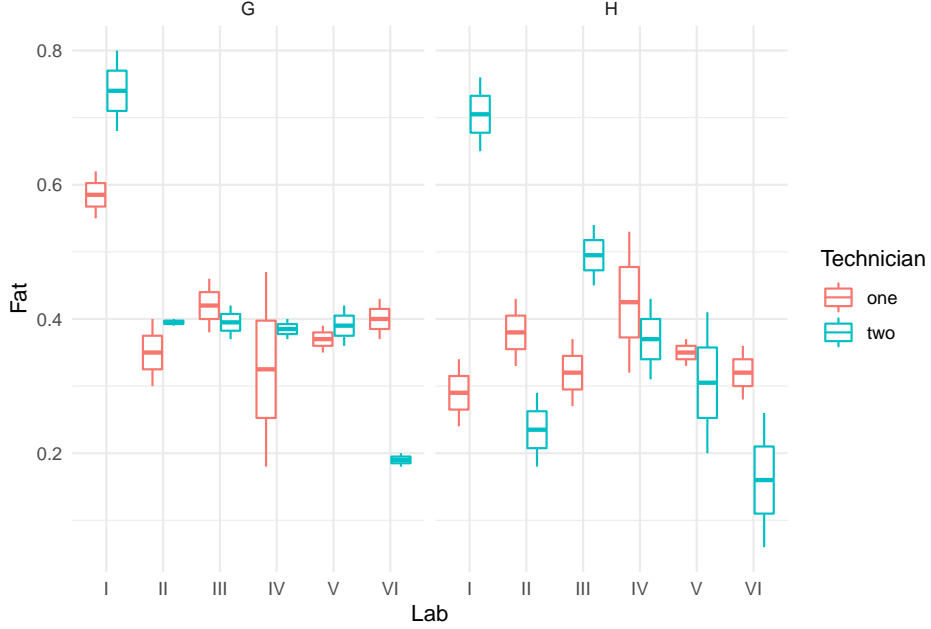


Figure 3.13: Descripción de eggs: variación entre laboratorios y técnicos.

El modelo que planteamos para estimar la respuesta  $y_{ijk}$ , contenido graso de la muestra  $k$  ( $k = 1, 2$ ) del laboratorio  $i$  ( $i = 1, \dots, 6$ ), por el técnico  $j$  ( $j = 1, 2$ ) está basado como siempre, en el modelo normal,  $(y_{ijk} | \mu_{ijk}, \sigma^2) \sim N(\mu_{ijk}, \sigma^2)$ , con una media o predictor lineal representado por:

$$\mu_{ijk} = \theta + \alpha_i^{lab} + \beta_{j:i}^{tec} + \gamma_{k:(j:i)}^{sam}$$

y asumiendo en un segundo nivel del modelo las distribuciones a priori:

$$\begin{aligned} \theta &\sim N(0, 1000) \\ \tau = 1/\sigma^2 &\sim Ga(0.001, 0.001) \\ \alpha_i^{lab} &\sim N(0, \sigma_{lab}^2); i = 1, \dots, 6 \\ \beta_{j:i}^{tec} &\sim N(0, \sigma_{tec}^2); j : i = 1, \dots, 12 \\ \gamma_{k:(j:i)}^{sam} &\sim N(0, \sigma_{sam}^2); k : (j : i) = 1, \dots, 24 \end{aligned}$$

El tercer nivel recibiría las distribuciones a priori para los hiperparámetros

$\sigma_{lab}^2, \sigma_{tec}^2, \sigma_{sam}^2$ , sobre los que interesa inferir. A priori, con mínima información asumiremos  $GaI(0.001, 0.001)$ .

Para especificar en INLA los efectos anidados hemos de recurrir a la matriz del modelo, `model.matrix()`, para crear las matrices de los efectos aleatorios anidados.

A continuación hemos de crear los correspondientes índices, de longitud similar a la del número de registros en la base de datos, para aplicarles las correspondientes matrices de efectos anidados, y ya proceder con el ajuste como habitualmente hacemos.

```
# matrices de efectos aleatorios anidados
Zlt <- as(model.matrix( ~ 0 + Lab:Technician, data = eggs), "Matrix")
Zlts <- as(model.matrix( ~ 0 + Lab:Technician:Sample, data = eggs), "Matrix")

# índices para aplicar los efectos aleatorios
eggs$IDt = eggs$IDts = 1:nrow(eggs)

# Ajuste
prec.prior=list(prec=list(param=c(0.001,0.001)))
formula = Fat ~ 1 + f(Lab,model="iid",hyper=prec.prior) +
  f(IDt,model="z",Z=Zlt,hyper=prec.prior)+
  f(IDts,model="z",Z=Zlts,hyper=prec.prior)

fit <- inla(formula,data = eggs,
            control.predictor = list(compute = TRUE),
            control.family=list(hyper=prec.prior),
            control.fixed=list(prec.intercept=0.001))

# inferencias de interés
round(fit$summary.hyperpar,4)
#>
#> Precision for the Gaussian observations 142.0228 39.6759
#> Precision for Lab 349.9044 651.1324
#> Precision for IDt 206.0281 210.4672
#> Precision for IDts 366.9348 335.2773
#> 0.025quant 0.5quant
#> Precision for the Gaussian observations 77.8362 137.4892
#> Precision for Lab 22.4894 170.2020
#> Precision for IDt 26.6558 144.1364
#> Precision for IDts 59.8942 270.7215
#> 0.975quant mode
#> Precision for the Gaussian observations 232.9082 NA
#> Precision for Lab 1808.5998 NA
#> Precision for IDt 762.1211 NA
#> Precision for IDts 1256.5370 NA
```

```

nhyp=length(names(fit$marginals.hyperpar))
res=NULL
for(j in 1:nhyp){
  res=rbind(res,data.frame(
    inla.tmarginal(function(tau) tau^(-1/2),fit$marginals.hyperpar[[j]]),
    id=str_sub(names(fit$marginals.hyperpar)[j], start =15)))
}
ggplot(res,aes(x=x,y=y))+
  geom_line(aes(color=id))+
  labs(x=expression(sigma),y="")+
  theme(legend.position="top",
        legend.title=element_blank(),
        legend.text = element_text(size=5))

```

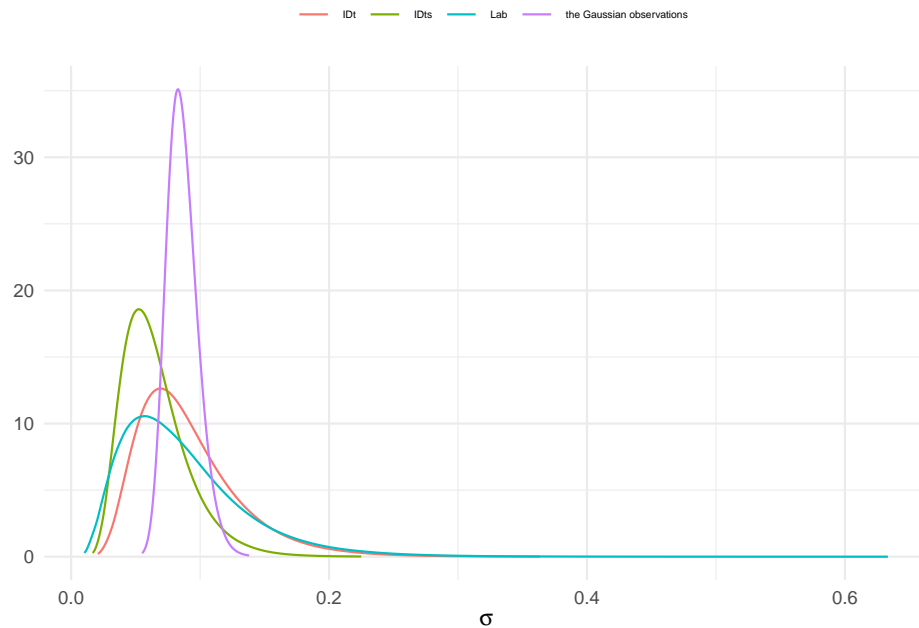


Figure 3.14: Distribución posterior del error de los datos y el error aleatorio

Alternativamente podríamos crear una variable índice a partir de las matrices de efectos aleatorios, para utilizarlas con `model="iid"` para describir los efectos aleatorios:

```

eggs$labtech <- as.factor(apply(Zlt, 1, function(x){names(x)[x == 1]}))
eggs$labtechsamp <- as.factor(apply(Zlts, 1, function(x){names(x)[x == 1]}))

formula=Fat ~ 1 + f(Lab, model = "iid", hyper = prec.prior) +

```

```

    f(labtech, model = "iid", hyper = prec.prior) +
    f(labtechsamp, model = "iid", hyper = prec.prior)
fit=inla(formula, data = eggs,
        control.predictor = list(compute = TRUE),
        control.family=list(hyper=prec.prior),
        control.fixed=list(prec.intercept=0.001))
round(fit$summary.fixed,4)
#>               mean      sd 0.025quant 0.5quant 0.975quant
#> (Intercept) 0.3875 0.0554      0.2756   0.3875     0.4994
#>               mode kld
#> (Intercept)   NA    0
round(fit$summary.hyperpar,4)
#>
#>               mean      sd
#> Precision for the Gaussian observations 141.9191 39.6192
#> Precision for Lab                      349.8807 651.1461
#> Precision for labtech                  206.0172 210.4586
#> Precision for labtechsamp              366.9533 335.2914
#>
#>               0.025quant 0.5quant
#> Precision for the Gaussian observations  77.8097 137.3969
#> Precision for Lab                      22.4870 170.1808
#> Precision for labtech                  26.6559 144.1279
#> Precision for labtechsamp              59.8971 270.7363
#>
#>               0.975quant mode
#> Precision for the Gaussian observations 232.6614  NA
#> Precision for Lab                      1808.5416  NA
#> Precision for labtech                  762.0849  NA
#> Precision for labtechsamp             1256.5921  NA

```

En la Figura 3.15 se muestra la distribución posterior del error de los datos y de los errores aleatorios.

```

nhyp=length(names(fit$marginals.hyperpar))
res=NULL
for(j in 1:nhyp){
  res=rbind(res,data.frame(
    inla.tmarginal(function(tau) tau^(-1/2),fit$marginals.hyperpar[[j]]),
    id=str_sub(names(fit$marginals.hyperpar)[j], start =15)))
}
ggplot(res,aes(x=x,y=y))+
  geom_line(aes(color=id))+
  labs(x=expression(sigma),y="")+
  theme(legend.position="top",
        legend.title=element_blank(),
        legend.text = element_text(size=5))

```

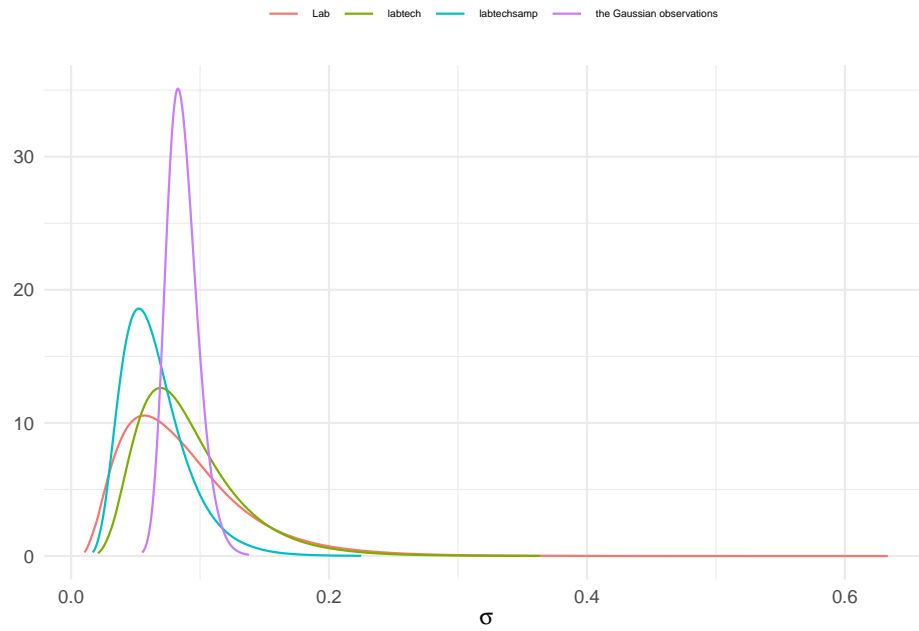


Figure 3.15: Distribución posterior del error de los datos y el error aleatorio

### 3.8 Conclusiones

Hasta aquí desarrollamos los modelos lineales basados en Anova, esto es, en la integración de factores de clasificación como variables que van a explicar diferencias en la respuesta, como los efectos fijos, o variabilidad extra en los datos, como los efectos aleatorios, a veces incluso con otros predictores de tipo numérico, e incluso interaccionando con ellos.



## Chapter 4

# Modelos lineales generalizados

Los modelos lineales generalizados, conocidos en inglés como *Generalized Linear Models* o GLM, son una clase de modelos introducidos por Nelder y Wedderburn (1972) y McCullagh y Nelder (1989), con el objetivo de extender la regresión lineal al caso en el que la variable dependiente no se distribuye necesariamente según una normal, pero su distribución todavía pertenece a la familia exponencial. Estamos hablando de distribuciones como la Binomial, Poisson, Gamma, o Gausiana inversa, básicamente. Trabajamos a continuación con dos de los GLM más comunes en epidemiología y ciencias sociales: la regresión logística y la de Poisson, mostrando cómo resolver este tipo de problemas con INLA.

Un modelo lineal generalizado está basado en asumir, además de una distribución de los datos dentro de la familia exponencial, una relación lineal entre cierta transformación del valor esperado de la respuesta y los predictores disponibles, sean covariables, efectos fijos, efectos aleatorios, o incluso alguna función de estos.

Si  $y$  representa una respuesta observada,  $x_1, x_2, \dots$  una serie de covariables o efectos fijos, y  $z_1, z_2, \dots$  efectos aleatorios, y el valor esperado de la respuesta lo denotamos como  $\mu$ ,

$$E(y|x, z, \theta) = \mu$$

la relación entre esta media  $\mu$  y un predictor lineal  $\eta$  que construimos a partir de una combinación lineal de los predictores disponibles, viene dado por una *función link*  $g$  tal que:

$$g(\mu) = \eta = X\beta + Zu$$

Todos los parámetros involucrados en el predictor lineal  $\eta$  son los efectos latentes del modelo (fijos o aleatorios). Estos, junto con el resto de parámetros definidos

en este primer nivel de la modelización (nivel de datos), han de modelizarse a continuación, en un segundo nivel del modelo, con sus correspondientes distribuciones a priori.

Los modelos lineales que hemos visto antes (regresión, anova, ancova, modelos mixtos) se engloban dentro del modelo lineal generalizado.

El argumento `control.predictor=list(compute=TRUE)` en la función `inla` permite obtener las distribuciones predictivas para el predictor lineal, que en estos modelos será distinto al valor medio ajustado `fitted`. Además para obtener la distribución marginal de los valores ajustados y predichos necesitamos incorporar a la función `inla` el argumento `control.compute=list(return.marginals.predictor=TRUE)`. Ya con todo ello podremos resumir las inferencias posteriores sobre los predictores lineales con el comando `fit$summary.linear.predictor` y obtener para graficar sus distribuciones posteriores con `fit$marginals.linear.predictor`.

## 4.1 Modelos jerárquicos bayesianos

A lo largo del curso ya hemos ido comentando algo sobre la especificación de un modelo en varios niveles. Presentamos ya de lleno estos modelos lineales generalizados como modelos multi-nivel o modelos jerárquicos, denominados así porque se va especificando por niveles (o jerarquías) la información disponible sobre todo aquello que es desconocido, distribución de los datos y parámetros.

Un modelo bayesiano se modeliza a través de un modelo jerárquico o multinivel en el que en el nivel I se define la distribución asumida sobre la variable respuesta y que determina la verosimilitud. Esta variable depende de unos parámetros que definen los efectos fijos y aleatorios, y para los que hay que proporcionar la información previa disponible a través de una distribución a priori en el segundo nivel del modelo jerárquico. La distribución a priori para los efectos fijos generalmente será común a todos ellos, mientras que la distribución a priori para los efectos aleatorios estará vinculada a otros hiperparámetros para los que también será preciso especificar una distribución a priori en un tercer nivel de la modelización, y así sucesivamente. Podríamos esquematizar un modelo jerárquico, a grosso modo, a través de la siguiente jerarquía.

$$\begin{array}{ll}
\text{Nivel I} & \\
(y|\beta, u, \Sigma) & \sim f(y|\beta, u, \Sigma) \\
& E(y|\beta, u, \Sigma) = \mu; \text{Var}(y|\beta, u, \Sigma) = \Sigma \\
& g(\mu) = \eta = X\beta + Zu \\
\text{Nivel II} & \\
\beta & \sim N(0, \sigma_\beta^2), \sigma_\beta^2 \text{ fijo} \\
(u|\sigma_u^2) & \sim_{iid} N(0, \sigma_u^2) \\
\Sigma|\sigma & \sim F_{\Sigma|\sigma} \\
\text{Nivel III} & \\
\sigma_u^2 & \sim F_{\sigma_u} \\
\sigma & \sim F_\sigma
\end{array}$$

## 4.2 Regresión logística

La regresión logística es el modelo estándar para respuestas binarias (éxitos/fracasos). Tiene dos variaciones, en función de si la respuesta viene dada en función de las observaciones individuales (0/1), o de los conteos (número de éxitos en  $n$  pruebas).

Si las observaciones de éxitos-fracasos vienen individualizadas, tenemos que la respuesta es directamente modelizable con una distribución *Bernoulli*:

$$Y|\pi \sim \text{Ber}(\pi) \rightarrow E(Y) = \pi$$

En el caso de que los datos vengan agrupados a modo de conteos de éxitos en diversos grupos poblacionales, modelizaremos con distribuciones Binomiales de tamaño  $n$ , que podría variar en cada grupo:

$$(Z = \sum_{i=1}^n Y_i)|\pi \sim \text{Bin}(n, \pi) \rightarrow E(Z) = n\pi$$

En cualquiera de los casos, el objetivo es inferir sobre la probabilidad de éxito  $\pi$ . El modelo logit utiliza la función logit para relacionar el predictor lineal  $\eta$  y la probabilidad  $\pi$ :

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \eta = X\beta + Zu$$

de forma que

$$\pi = \text{logit}^{-1}(X\beta + Zu) = \frac{\exp(X\beta + Zu)}{1 + \exp(X\beta + Zu)}$$

Una vez especificado el modelo, si no hay información previa disponible sobre los efectos latentes  $(\beta, u)$ , se asumen las distribuciones a priori habituales, generalmente normales independientes con media cero y varianza grande para efectos fijos y varianza desconocida para efectos aleatorios. En un tercer nivel se habrá de modelizar la información a priori de las varianzas para los efectos aleatorios.

### 4.2.1 Interpretación de los coeficientes en la regresión logit

Por simplificar la interpretación de los coeficientes, vamos a asumir que la respuesta es Bernoulli y todos los efectos latentes son fijos, de modo que el predictor lineal se puede escribir como:

$$\eta = X\beta = \beta_0 + \sum_{j=1}^M \beta_j x_j$$

A partir de esta expresión se entiende claramente que la interceptación del predictor lineal  $\beta_0$  se interpreta como el valor del predictor cuando las variables predictivas toman el valor cero si son numéricas, o cuando se refieren al nivel de referencia si son categóricas, es decir,

$$\beta_0 = \eta_{(X=0)} = \text{logit}(\pi_{(X=0)})$$

En consecuencia, el logit inverso de  $\beta_0$  se interpreta como “la probabilidad de éxito cuando los predictores están en su nivel de referencia o son cero”.

$$\text{logit}^{-1}(\beta_0) = \pi_{(X=0)} = \Pr(Y = 1|X = 0)$$

En cuanto a la interpretación de cualquier otro coeficiente de regresión en el predictor lineal, como  $\beta_1$ , echamos mano del concepto de *odds* y *odds ratio*.

Los *odds* de un evento  $E$  se definen a través de las posibilidades de que se dé dicho evento, y se calcula con el cociente de su probabilidad y la probabilidad de que no se dé:

$$\text{odds}(E) = \frac{\Pr(E)}{1 - \Pr(E)}$$

Los *odds ratio*, *OR* de un evento relativo a dos condiciones A y B, comparan, a través de un cociente, las posibilidades a favor del evento  $E$  bajo condiciones A y sus posibilidades bajo condiciones B. Nos sirve para evaluar cuánto afectan a las posibilidades del evento el hecho de variar las condiciones de B a A.

$$\text{OR}(A, B) = \frac{\Pr(E|A)/(1 - \Pr(E|A))}{\Pr(E|B)/(1 - \Pr(E|B))}.$$

En el modelo logístico, nos interesa saber el efecto que tiene sobre la respuesta (realmente sobre la probabilidad de éxito) el incremento de una unidad en la

variable predictora  $X$ , y para ello consideramos los odds bajo  $X = x$  y los odds bajo  $X = x + 1$ , esto es, las posibilidades de éxito bajo  $x$  y bajo  $x + 1$ . Si además consideramos estos odds en escala logarítmica, esto es, los *log-odds*, tenemos:

$$\begin{aligned} \log.odds(x + 1) &= \log \left( \frac{P(y = 1|x + 1)}{P(y = 0|x + 1)} \right) \\ &= \text{logit}(\pi_{(x+1)}) = \beta_0 + \beta_1(x + 1) \\ \log.odds(x) &= \log \left( \frac{P(y = 1|x)}{P(y = 0|x)} \right) \\ &= \text{logit}(\pi_{(x)}) = \beta_0 + \beta_1 x \end{aligned}$$

Así tendremos que el logaritmo del *odds-ratio* coincide con el coeficiente que relaciona  $X$  con el predictor lineal, esto es,  $\beta_1$ :

$$\begin{aligned} \log(OR(x + 1, x)) &= \log \left( \frac{odds(x + 1)}{odds(x)} \right) \\ &= \log.odds(x + 1) - \log.odds(x) = \beta_1 \end{aligned}$$

Es decir, la exponencial del coeficiente  $\beta_1$  nos da el odds ratio asociado a dicha covariable, esto es, el cambio en los odds a favor de un éxito cuando se incrementa en una unidad la variable predictora  $X$ . Esta interpretación es muy común en Epidemiología.

$$\exp(\beta_1) = \frac{odds(x + 1)}{odds(x)} = OR(x + 1, x).$$

Cuando estamos ante un predictor categórico, la exponencial del coeficiente estimado para un nivel o categoría  $i$  representa la variación en los odds que se produce al pasar del nivel base del factor al nivel  $i$ . Si dicha variación es superior a 1, entonces hablamos de que es más probable el evento, y si es inferior a 1, decimos que es menos probable cuando estamos en el nivel  $i$  que en el nivel 1.

### 4.2.2 Intención de voto feb2022

Tenemos acceso a los datos completos obtenidos en la encuesta encargada por El País y la Cadena Ser a la empresa “40dB”, en febrero de 2022, sobre la intención de voto nacional en España (fuente).

Queremos predecir la probabilidad de votar al partido que gobierna mayoritariamente en la actualidad, PSOE, registrado en la variable `psoe`. Vamos a utilizar como predictores dos factores que nos dicen si el sujeto tiene simpatía por ese partido, `psoe_sim`, y si votó PSOE en las últimas elecciones `psoe_past`; también utilizaremos la comunidad autónoma `ccaa` como un efecto aleatorio, para

contabilizar posible variación extra (igualmente podríamos considerarla como un efecto fijo).

```
url="https://raw.githubusercontent.com/BayesModel/data/main/barometro_feb22.csv"
barometro_feb22=read.csv(url)
datos=barometro_feb22 %>%
  select(id,p2,p3,p5,ccaa) %>%
  mutate(psoe=1*(p2=="PSOE (Partido Socialista Obrero Español)"),
         psoe_simp=1*(p3=="PSOE (Partido Socialista Obrero Español)"),
         psoe_past=1*(p5=="PSOE (Partido Socialista Obrero Español)"))
#summary(datos)
```

Especificamos pues un modelo que asume para los efectos fijos las distribuciones difusas por defecto en INLA, y para la varianza de los efectos aleatorios una  $GaI(0.001, 0.001)$ .

$$\begin{aligned}
 y_{ij}|\pi_i &\sim Br(\pi_i) \\
 \text{logit}(\pi_i) &= \eta_i = \theta + \alpha_{simp} + \alpha_{past} + \gamma_{ca} \\
 \theta &\sim N(0, \infty) \\
 \alpha_{simp} &\sim N(0, 1000) \\
 \alpha_{past} &\sim N(0, 1000) \\
 \gamma_{ca} &\sim N(0, \sigma_{ca}^2) \\
 1/\sigma_{ca}^2 &\sim Ga(0.001, 0.001)
 \end{aligned}$$

Las inferencias posteriores se resumen en los siguientes descriptivos.

```
prec.prior=list(prec=list(param=c(0.001,0.001)))
formula = psoe ~ psoe_simp + psoe_past+ f(ccaa,model="iid",hyper=prec.prior)
fit=inla(formula,family="binomial",data=datos)
fit$summary.fixed
#>               mean          sd 0.025quant  0.5quant
#> (Intercept) -3.590590 0.1897365 -3.998981 -3.578298
#> psoe_simp    3.085809 0.1865735  2.723938  3.084374
#> psoe_past    2.352356 0.1856828  1.989733  2.351810
#>               0.975quant mode          kld
#> (Intercept) -3.251374    NA 5.805021e-07
#> psoe_simp    3.455828    NA 3.434752e-07
#> psoe_past    2.718075    NA 9.427354e-07
fit$summary.hyperpar
#>               mean          sd 0.025quant  0.5quant
#> Precision for ccaa 70.85958 251.0441  1.978298 10.68807
#>               0.975quant mode
```

```
#> Precision for ccaa 589.4174 NA
head(fit$summary.fitted.values)
#>
#>      mean      sd 0.025quant
#> fitted.Predictor.0001 0.02613268 0.006597318 0.01417253
#> fitted.Predictor.0002 0.03313666 0.006805633 0.02191503
#> fitted.Predictor.0003 0.02920946 0.005829130 0.01886787
#> fitted.Predictor.0004 0.02920946 0.005829130 0.01886787
#> fitted.Predictor.0005 0.03313666 0.006805633 0.02191503
#> fitted.Predictor.0006 0.88231903 0.023396647 0.83526830
#>
#>      0.5quant 0.975quant mode
#> fitted.Predictor.0001 0.02592325 0.03990485 NA
#> fitted.Predictor.0002 0.03236610 0.04864296 NA
#> fitted.Predictor.0003 0.02882857 0.04186095 NA
#> fitted.Predictor.0004 0.02882857 0.04186095 NA
#> fitted.Predictor.0005 0.03236610 0.04864296 NA
#> fitted.Predictor.0006 0.88269663 0.92602566 NA
head(fit$summary.linear.predictor)
#>
#>      mean      sd 0.025quant 0.5quant
#> Predictor.0001 -3.651129 0.2694963 -4.242176 -3.626350
#> Predictor.0002 -3.393915 0.2091688 -3.798424 -3.397742
#> Predictor.0003 -3.523517 0.2069641 -3.951246 -3.517136
#> Predictor.0004 -3.523517 0.2069641 -3.951246 -3.517136
#> Predictor.0005 -3.393915 0.2091688 -3.798424 -3.397742
#> Predictor.0006 2.034601 0.2317236 1.623435 2.018218
#>
#>      0.975quant mode      kld
#> Predictor.0001 -3.180534 NA 1.900590e-07
#> Predictor.0002 -2.973382 NA 6.892912e-08
#> Predictor.0003 -3.130640 NA 1.991678e-07
#> Predictor.0004 -3.130640 NA 1.991678e-07
#> Predictor.0005 -2.973382 NA 6.892912e-08
#> Predictor.0006 2.527184 NA 7.424809e-08
```

La variabilidad para los efectos aleatorios es relevante, como se manifiesta a través de la distribución posterior de su varianza, que se muestra en la Figura 4.1 en términos de  $\sigma_{ca}$ .

```
sigma.post=inla.tmarginal(function(tau) tau^(-1/2),
  fit$marginals.hyperpar[[1]])
# y la pintamos
ggplot(as.data.frame(sigma.post)) +
  geom_line(aes(x = x, y = y)) +
  labs(x=expression(sigma),y="D.Posterior")+
  geom_vline(xintercept=inla.hpdmarginal(0.95,sigma.post),
    linetype="dotted",color="blue")+
  geom_vline(xintercept=inla.emarginal(function(x) x,sigma.post),
```

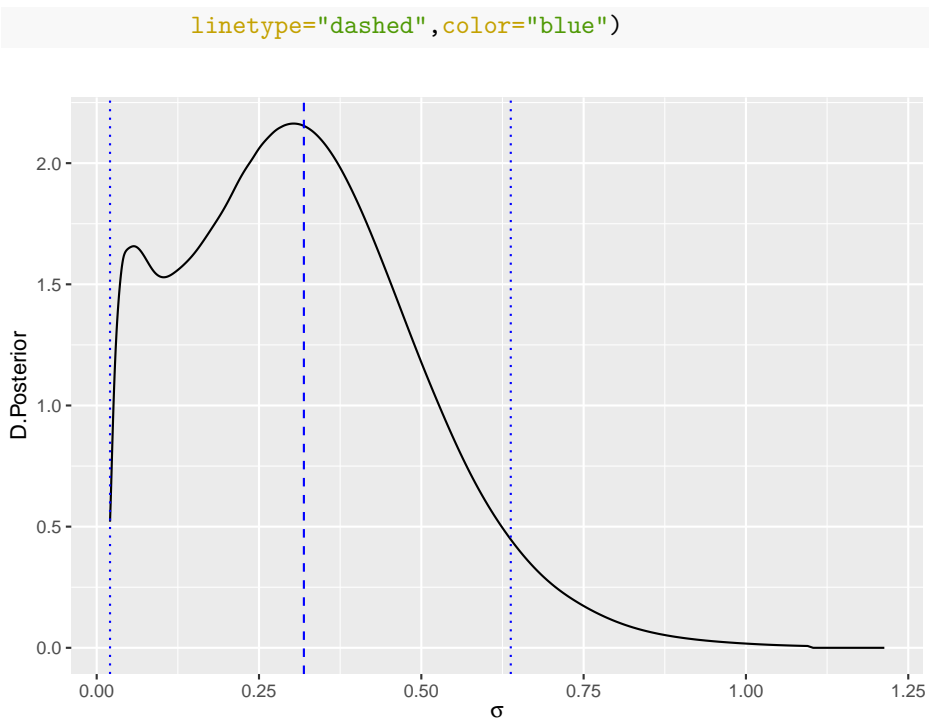


Figure 4.1: Distribución posterior, media y RC, de la desviación típica de los efectos aleatorios

```
# Valor esperado
sigma.e=round(inla.emarginal(function(tau) tau^(-1/2),
  fit$marginals.hyperpar[[1]]),4)
# HPD95%
sigma.hpd=round(inla.hpdmarginal(0.95,sigma.post),3)
paste("E(sigma.post)=",sigma.e,
  "HPD95%=(",sigma.hpd[1],",",sigma.hpd[2],")")
#> [1] "E(sigma.post)= 0.3189 HPD95%=( 0.021 , 0.638 )"
```

Las distribuciones posteriores de los exponenciales de los efectos aleatorios (que representan los *odds-ratios*) se muestran en la Figura 4.2, en términos de medias y regiones creíbles. Están identificadas en verde las comunidades con efectos positivos hacia el voto PSOE ( $\text{odds} > 1$ ), y en rojo las de efectos negativos ( $\text{odds} < 1$ ). Estas distribuciones de efectos aleatorios nos permiten diferenciar qué comunidades autónomas son más (verde) y menos (rojo) favorables a votar por el PSOE, en términos de valor esperado (a posterior), así como descubrir la incertidumbre y variabilidad existente en dichas afirmaciones, dada por la región creíble.



```

random = as.data.frame(fit$summary.random)
random$pro=1*exp(random$ccaa.mean)>1
ggplot(random,aes(x=exp(ccaa.mean),y=ccaa.ID)) +
  geom_point(aes(color=pro))+
  geom_errorbarh(aes(xmin=exp(ccaa.0.025quant),
                      xmax=exp(ccaa.0.975quant),color=pro))+
  geom_vline(xintercept=1,linetype="dotted")+
  labs(x="Medias y RC posteriores para los odds-ratio",
       y="Comunidad autónoma")+
  theme(legend.position="none")

```

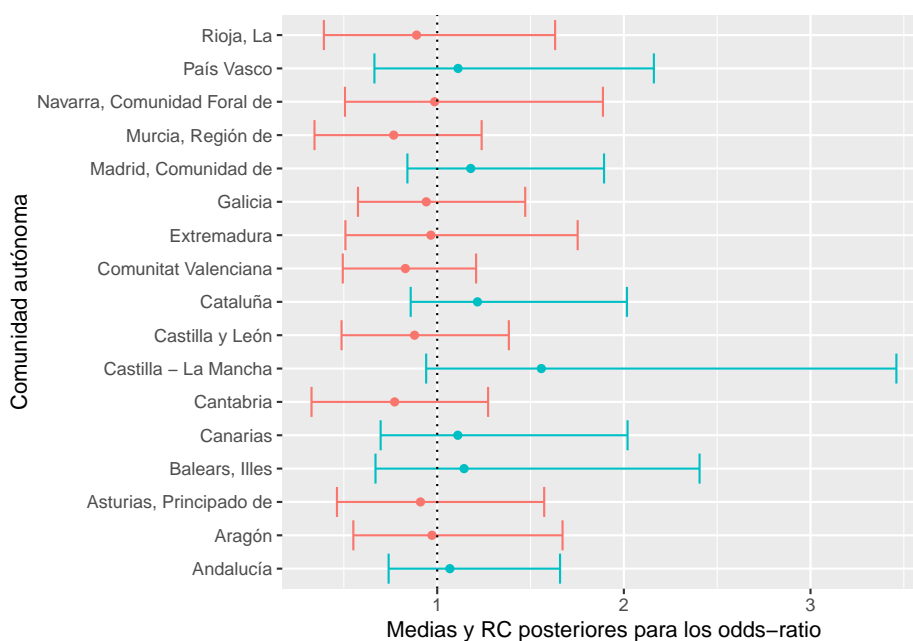


Figure 4.2: Medias y RC posteriores para los odds-ratios de los efectos aleatorios

La distribución posterior de la probabilidad de voto para el PSOE la conseguimos a través de los valores ajustados, `fitted`. En la Figura 4.3 está representada la inferencia posterior sobre la probabilidad de voto para cada uno de los cuatro colectivos que identificamos en función de su simpatía por el PSOE y su voto en el pasado, en la comunidad autónoma con más variabilidad en el efecto aleatorio, esto es, Castilla-La Mancha.

```

datos_pred = datos %>%
  mutate(f.post=round(fit$summary.fitted.values$mean,3),
         f.rc.low=round(fit$summary.fitted.values$"0.025quant",3),
         f.rc.up=round(fit$summary.fitted.values$"0.975quant",3)) %>%

```

```

distinct(f.post, .keep_all = TRUE) %>%
filter(ccaa=="Castilla - La Mancha") %>%
mutate(simpast=str_c(psoe_simp,psoe_past))

ggplot(datos_pred,aes(x=f.post,y=simpast))+
  geom_point(aes(color=simpast))+
  geom_errorbarh(
    aes(xmin=f.rc.low,xmax=f.rc.up,color=simpast),
    height=0.2)+
  labs(x="Medias y RC posteriores para la probabilidad de voto PSOE",
       title="Castilla - La Mancha")+
  scale_y_discrete(name="",
    labels=c("No simpatía/No votó PSOE","No simpatía/Votó PSOE",
             "Simpatía/No votó PSOE","Simpatía/Votó PSOE"))+
  theme(legend.position="none")

```

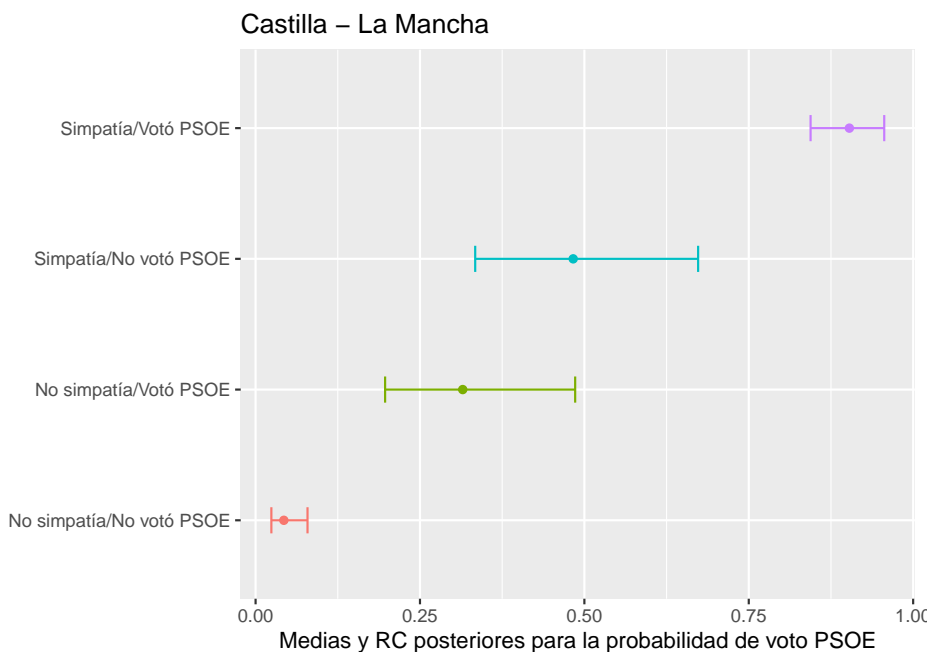


Figure 4.3: Medias y RC posteriores para la probabilidad de voto PSOE

### 4.2.3 Mortalidad por infarto en Sheffield

Utilizamos los datos *stroke*, disponibles en datasets in SSTM-RINLA relativos al periodo 1994-1999. El objetivo es evaluar la asociación entre los niveles de

NOx y el infarto en Sheffield (UK). Utilizaremos la siguiente información:

- número de infartos y en cada distrito,
- una agrupación de distritos en base a cierto índice relativo al nivel de desventajas y privación en dicho distrito, **Townsend** (en escala 1-5),
- la concentración anual media de NOx, categorizada en cinco niveles (escala 1-5)
- el tamaño de la población en cada distrito, **pop**
- el riesgo base ajustado por sexo y edad para el número de infartos, calculado con estandarización indirecta con ratios de referencia internos basados en 18 estratos (9 para edad y 2 para género), en la variable **stroke\_exp** (Maheswaran et al.2006).

La respuesta  $y_i$  relativa al número de infartos en el distrito  $i$ , con una población  $n_i$ , se puede modelizar con:

$$y_i | \pi_i \sim \text{Bin}(n_i, \pi_i)$$

Planteamos las asociaciones que intuimos a través del predictor lineal, definido en función del nivel de NOx y el nivel de privación **Townsend**, ambos como efectos fijos, así como de *offset* que representa el riesgo base corregido por el tamaño del distrito,  $\tilde{p}_i = \text{stroke\_exp}/\text{pop}$ , en escala logit.

$$\eta_i = \text{logit}(\pi_i) = \theta + \alpha_i^{\text{NOx}} + \alpha_i^{\text{Town}} + \text{logit}(\tilde{p}_i)$$

Ajustamos pues el modelo con los efectos fijos NOx y **Townsend**, y el *offset* que calculamos y llamamos **logit.adjusted.prob**, que proporciona una estandarización del riesgo en base al tamaño de población en cada distrito, y para el que no se estima coeficiente.

```
url="https://raw.githubusercontent.com/BayesModel/data/main/Stroke.csv"
Stroke <- read.csv(url,sep="," ,dec=".",header=TRUE)
# conversión a factores y cálculo del riesgo base
stroke=Stroke %>%
  mutate(NOx=as.factor(NOx),Townsend=as.factor(Townsend),
         adjusted.prob=stroke_exp/pop,
         logit.adjusted.prob=log(adjusted.prob/(1-adjusted.prob)))

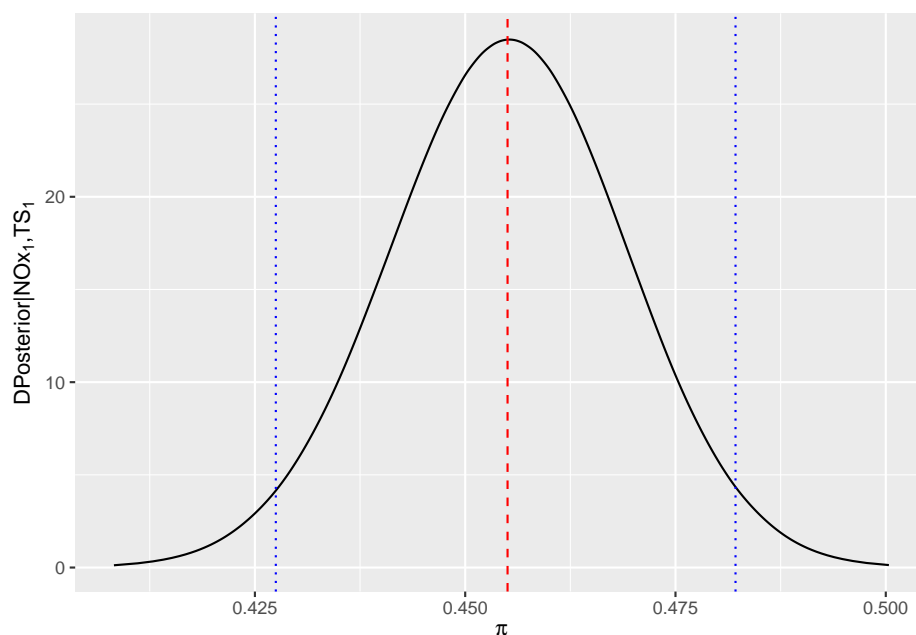
# ajuste del modelo
formula.inla <- y ~ 1 + NOx + Townsend + offset(logit.adjusted.prob)
fit <- inla(formula.inla, family="binomial", Ntrials=pop, data=stroke)
round(fit$summary.fixed[,1:5],3)
#>               mean      sd 0.025quant 0.5quant 0.975quant
#> (Intercept) -0.181 0.057      -0.293   -0.180    -0.071
#> NOx2         0.132 0.059       0.016    0.132     0.248
```

```
#> NOx3      0.105 0.061      -0.014      0.105      0.225
#> NOx4      0.261 0.059      0.144      0.260      0.377
#> NOx5      0.425 0.062      0.302      0.425      0.547
#> Townsend2 0.077 0.061     -0.043      0.077      0.198
#> Townsend3 0.137 0.060      0.020      0.137      0.255
#> Townsend4 -0.132 0.063     -0.255     -0.132     -0.009
#> Townsend5 -0.118 0.067     -0.250     -0.118      0.014
```

Inferimos sobre la probabilidad esperada de infarto en un distrito con niveles base en privación Townsend y NOx, esto es, Townsend=1 y NOx=1, con la distribución posterior del logit inverso de la interceptación  $\theta$ . Esta distribución posterior la obtendremos entonces a partir de simulaciones de la distribución posterior de  $\theta$  y utilizando la función logit-inversa

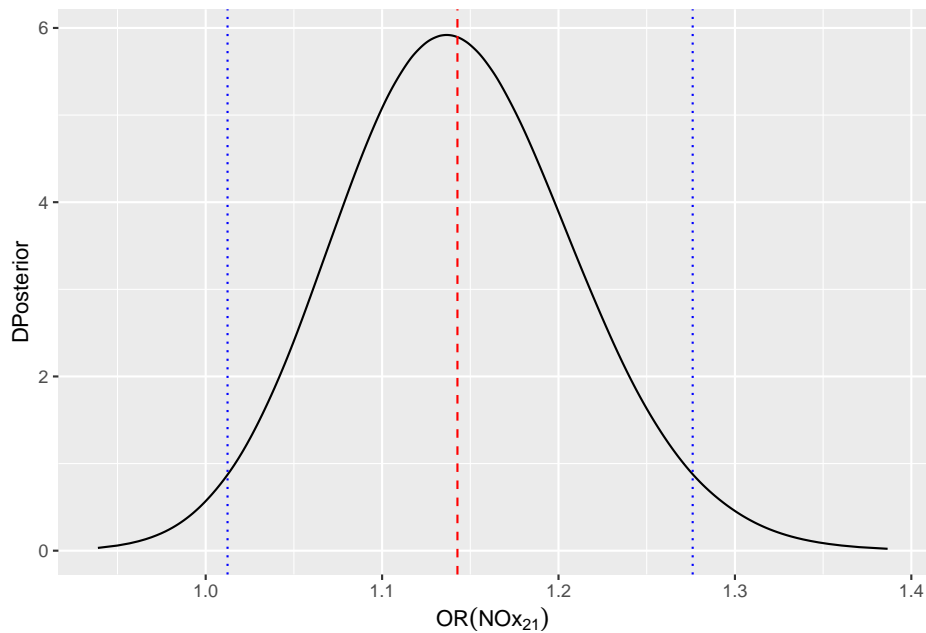
$$\text{logit}^{-1}(\theta) = \frac{\exp(\theta)}{1 + \exp(\theta)}$$

```
prob.stroke <- inla.tmarginal(function(x) exp(x)/(1+exp(x)), fit$marginals.fixed[[1]])
des=inla.zmarginal(prob.stroke)
#> Mean      0.455034
#> Stdev     0.0139168
#> Quantile 0.025 0.427486
#> Quantile 0.25  0.445615
#> Quantile 0.5   0.455082
#> Quantile 0.75  0.464485
#> Quantile 0.975 0.482137
ggplot(data.frame(prob.stroke), aes(x=x, y=y)) +
  geom_line() +
  geom_vline(xintercept=des$mean, linetype="dashed", color="red") +
  geom_vline(xintercept=c(des$quant0.025, des$quant0.975),
             linetype="dotted", color="blue") +
  labs(x=expression(pi),
       y= expression(paste("DPosterior|", NOx[1], ",", "TS[1]")))
```



El efecto  $\alpha_2^{NOx}$  representa el efecto diferencial en los log-odds del infarto al estar en el nivel  $NOx = 2$  frente al de estar en el nivel de referencia  $NOx = 1$ . Si queremos evaluar el odds-ratio, simplemente calculamos la distribución posterior de  $\exp(\alpha_2^{NOx})$  con `inla.tmarginal`. Si sólo estamos interesados en sus descriptivos posteriores bastaría utilizar `inla.zmarginal`.

```
odds.nox21 <- inla.tmarginal(exp,
                             fit$marginals.fixed$NOx2)
m=inla.emarginal(exp,fit$marginals.fixed$NOx2)
hpd=inla.hpdmarginal(0.95,odds.nox21)
ggplot(data.frame(odds.nox21),aes(x=x,y=y))+
  geom_line()+
  labs(x=expression(OR(NOx[21])),
       y= "DPosterior")+
  geom_vline(xintercept=m,linetype="dashed",color="red")+
  geom_vline(xintercept=hpd,
             linetype="dotted",color="blue")
```



Concluimos pues, que la probabilidad de muerte por infarto se incrementa en un -54.5 cuando la exposición de NOx cambia del primer al segundo nivel (la región creíble da un margen de crecimiento entre -57.25 y -51.79).

Los odds-ratio para el resto de los niveles resultan igualmente “significativos” desde el punto de vista bayesiano, teniendo en cuenta sus estimaciones (media de la distribución posterior) e intervalos creíbles:

```
OR31=inla.zmarginal(inla.tmarginal(function(x) exp(x),
                                   fit$marginals.fixed$NOx3))

#> Mean          1.1131
#> Stdev          0.0673074
#> Quantile 0.025 0.986696
#> Quantile 0.25 1.06638
#> Quantile 0.5 1.11095
#> Quantile 0.75 1.1574
#> Quantile 0.975 1.25097
OR41=inla.zmarginal(inla.tmarginal(function(x) exp(x),
                                   fit$marginals.fixed$NOx4))

#> Mean          1.29988
#> Stdev          0.0766197
#> Quantile 0.025 1.15596
#> Quantile 0.25 1.24669
#> Quantile 0.5 1.29744
#> Quantile 0.75 1.35031
#> Quantile 0.975 1.4568
```

```

OR51=inla.zmarginal(inla.tmarginal(function(x) exp(x),
                                   fit$marginals.fixed$N0x5))

#> Mean          1.53223
#> Stdev         0.0950874
#> Quantile 0.025 1.35392
#> Quantile 0.25  1.46617
#> Quantile 0.5   1.52909
#> Quantile 0.75  1.59475
#> Quantile 0.975 1.72726

rbind(OR31,OR41,OR51)[,c(1,2,3,7)]
#>      mean      sd      quant0.025 quant0.975
#> OR31 1.113102 0.06730743 0.9866958 1.250967
#> OR41 1.29988  0.07661972 1.155961  1.456801
#> OR51 1.532227 0.09508741 1.353922  1.727261

```

### 4.3 Regresión de Poisson

La regresión de Poisson es útil cuando la variable respuesta representa conteos y estos toman valores discretos entre 0 y  $+\infty$ , sin una cota superior de referencia. El parámetro de interés es el número promedio de eventos  $\lambda = E(y)$  y el link natural es el logaritmo, de modo que el predictor lineal está ligado con las covariables y factores según:

$$y \sim Po(\lambda)$$

$$\eta = \log(\lambda) = X\beta + Zu, \quad \lambda = \exp(X\beta + Zu)$$

Para completar el modelo bayesiano, se especifican distribuciones a priori para los coeficientes de los efectos fijos y aleatorios,  $(\beta, u)$ , típicamente como normales con media cero y una varianza grande para efectos fijos  $\beta$  cuando no hay información previa disponible, y varianzas desconocidas para los efectos aleatorios  $u$ . En el siguiente nivel se especificarán gammas dispersas para la precisión de los efectos aleatorios.

Los coeficientes se interpretan a través de la función exponencial. Supongamos por simplicidad, un predictor lineal de la forma  $\eta = \theta + \sum_{i=1}^M \beta_i x_i$ :

- $\exp(\theta)$  representa el valor esperado de la respuesta cuando todas las covariables son cero, o si son categóricas para el primer nivel de las categorías posibles.
- $\exp(\beta_i)$  es el cambio que se produce en la respuesta promedio  $y$  cuando  $x_i$  se incrementa en una unidad si dicha variable es numérica, o el cambio que se produce al pasar del primer nivel o categoría al nivel  $i$ .

La mayoría de las veces que se utiliza la regresión de Poisson, el interés recae en las ratios o riesgos relativos, más que en el número esperado de casos  $\lambda$ . Para cambiar la escala en términos de riesgo  $\rho$ , generalmente expresado con  $\lambda = \rho \cdot E$ , ha de utilizarse un offset  $E$  como factor de corrección en la especificación del modelo. Este offset representa el denominador común del riesgo, o riesgo base, y entra en la regresión en una escala logarítmica (dada la función link), asumiendo que tiene un coeficiente de regresión fijado a 1, es decir, no se estima coeficiente para él.

$$\eta = \log(\lambda) = X\beta + Zu + \log(E)$$

Así, el riesgo relativo de que se produzca un evento se calculará dividiendo por el riesgo base, que en escala logarítmica será

$$\log\left(\frac{\lambda}{Offset}\right)$$

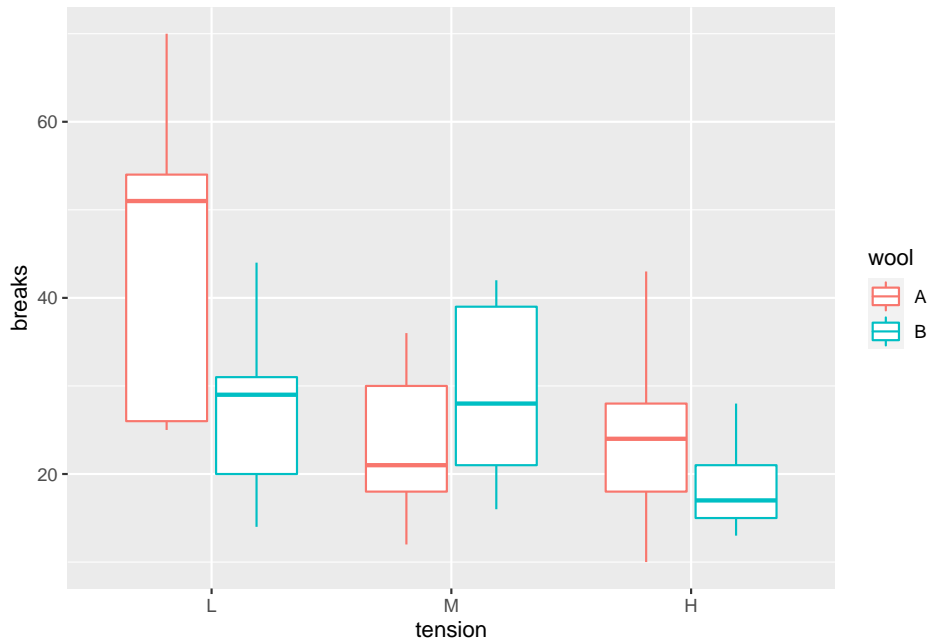
y al exponenciar los coeficientes  $\exp(\beta)$ , podremos interpretarlos como el cambio que se produce en el riesgo relativo al incrementar en una unidad el predictor correspondiente, si es continuo, o al pasar del nivel base al nivel que representa dicho coeficiente.

### 4.3.1 Roturas de hilo

La base de datos **warpbreaks** en la librería **datasets** proporciona el número de roturas **breaks** en el hilado en cada telar, para una serie de telares de longitud común. Contiene también información sobre el tipo de lana (A y B), **wool** y el nivel de tensión del telar (L,M,H), **tension**. En la Figura ?? se representan los datos.

```
data(warpbreaks, package="datasets")
ggplot(warpbreaks, aes(x=tension, y=breaks))+
  geom_boxplot(aes(color=wool))
```





Queremos pues predecir el número de roturas en función del tipo de lana y la tensión del telar, que consideramos como efectos fijos.

$$y_i | \lambda_i \sim Po(\lambda_i) \log(\lambda_i) = \eta_i = \theta + \alpha_i^w + \alpha_i^t$$

Ajustamos el modelo anterior con las priors por defecto y obtenemos:

```
formula= breaks ~ wool + tension
fit=inla(formula,family="poisson",data=warpbreaks,
         control.compute=list(return.marginals.predictor=TRUE))
round(fit$summary.fixed[,1:5],3)
#>           mean      sd 0.025quant 0.5quant 0.975quant
#> (Intercept)  3.692 0.045      3.602   3.692   3.780
#> woolB       -0.206 0.052     -0.307  -0.206  -0.105
#> tensionM    -0.321 0.060     -0.440  -0.321  -0.203
#> tensionH    -0.518 0.064     -0.645  -0.518  -0.393
```

Los efectos fijos expresados con sus distribuciones posteriores completas los representamos en la Figura 4.4.

```
fixed=names(fit$marginals.fixed)
g=list()
for(i in 1:length(fixed)){
  g[[i]]=ggplot(as.data.frame(fit$marginals.fixed[[i]]),aes(x=x,y=y))+
```

```

geom_line()+
geom_vline(xintercept=fit$summary.fixed$mean[i],linetype="dashed")+
geom_vline(xintercept=fit$summary.fixed[i,3],linetype="dotted")+
geom_vline(xintercept=fit$summary.fixed[i,5],linetype="dotted")+
labs(x=fixed[i],y="D.posterior")
}
grid.arrange(g[[1]],g[[2]],g[[3]],g[[4]],ncol=2)

```

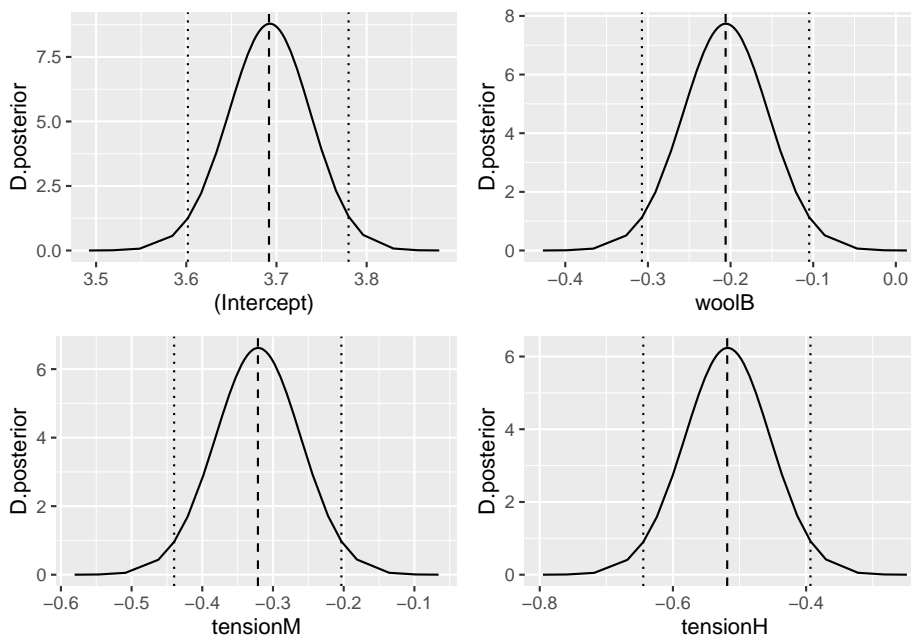


Figure 4.4: Distribución posterior de los efectos fijos

Y si queremos intentar una interpretación en términos de exponenciales de los efectos, tenemos sus distribuciones posteriores representadas en la Figura 4.5.

```

fixed=names(fit$marginals.fixed)
g=list()
for(i in 1:length(fixed)){
  exp.eff=inla.tmarginal(function(x) exp(x),fit$marginals.fixed[[i]])
  g[[i]]=ggplot(as.data.frame(exp.eff),aes(x=x,y=y))+
    geom_line()+
    geom_vline(
      xintercept=inla.emarginal(function(x) exp(x),
        fit$marginals.fixed[[i]]),
      linetype="dashed")+
    geom_vline(xintercept=inla.hpdmarginal(0.95,exp.eff),

```

```

linetype="dotted")+
  labs(x=fixed[i],y="D.posterior")
}
grid.arrange(g[[1]],g[[2]],g[[3]],g[[4]],ncol=2)

```

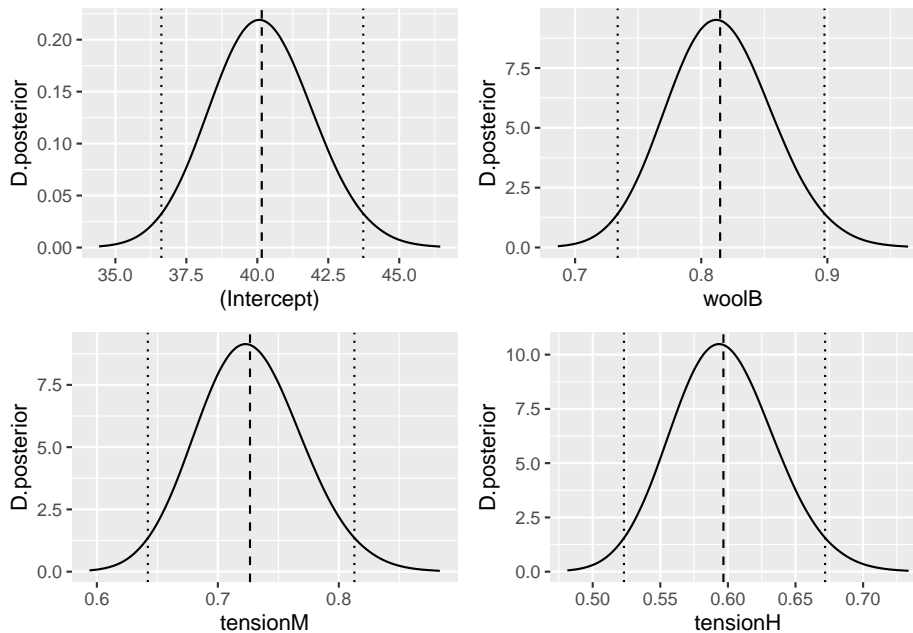


Figure 4.5: Distribución posterior de los exponenciales de los efectos fijos

Si quisiéramos reajustar el modelo utilizando como nivel de referencia en la variable `tension` el nivel H y en `wool` el nivel B, lo hacemos con la función `relevel`.

```

warp=warpbreaks %>%
  mutate(wool=relevel(wool,"B"),
          tension=relevel(tension,"H"))
formula= breaks ~ wool + tension
fit=inla(formula,family="poisson",data=warp,
          control.compute=list(return.marginals.predictor=TRUE))

```

Y obtenemos entonces las distribuciones posteriores para los exponenciales de los efectos fijos, en la Figura 4.6.

```

fixed=names(fit$marginals.fixed)
g=list()

```

```

for(i in 1:length(fixed)){
  exp.eff=inla.tmarginal(function(x) exp(x),fit$marginals.fixed[[i]])
  g[[i]]=ggplot(as.data.frame(exp.eff),aes(x=x,y=y))+
    geom_line()+
    geom_vline(
      xintercept=inla.emarginal(function(x) exp(x),
        fit$marginals.fixed[[i]]),
      linetype="dashed")+
    geom_vline(xintercept=inla.hpdmarginal(0.95,exp.eff),
      linetype="dotted")+
    labs(x=fixed[i],y="D.posterior")
}
grid.arrange(g[[1]],g[[2]],g[[3]],g[[4]],ncol=2)

```

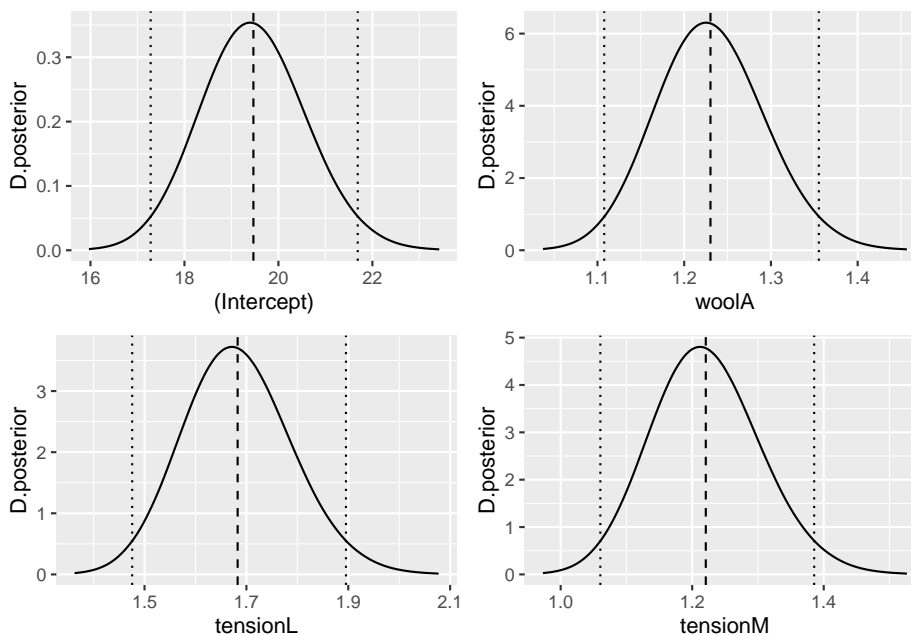


Figure 4.6: Distribución posterior de los exponentes de los efectos fijos, con niveles de referencia wool=B y tension=H

### 4.3.2 Incidentes en barcos

Utilizamos los datos `ShipsIncidents` en `datasets` for `SSTM-RINLA` para estimar el riesgo mensual de incidentes en barcos. Los factores potenciales del riesgo son el periodo de construcción (*built*), el periodo de operación (*oper*) y el tipo de barco (*type*).

El modelo se formula considerando que el número de incidentes será proporcional al número de meses que ha navegado el barco, de modo que utilizaremos como offset el  $\log(months)$ , donde `months` son los meses que ha navegado el bargo y que ponderan en consecuencia el riesgo de incidentes  $\rho_i$ .

El modelo con el offset será entonces

$$y_i \sim \text{Poisson}(E_i \rho_i) \eta_i = \log(\rho_i) = \theta + \alpha_i^b + \alpha_i^o + \alpha_i^t + \text{Offset}(\log(month))$$

Para los efectos fijos especificaremos las distribuciones a priori por defecto que proporciona INLA.

```
url="https://raw.githubusercontent.com/BayesModel/data/main/Ships.csv"
ShipsIncidents <- read.csv(url,sep=",")

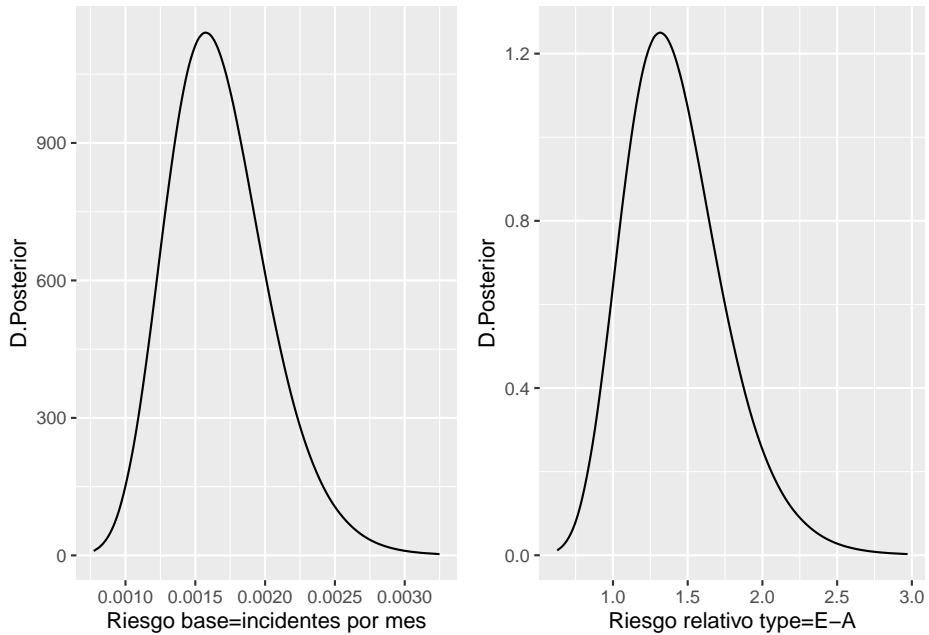
formula.inla <- y ~ 1 + built + oper + type
fit <- inla(formula.inla,family="poisson", data=ShipsIncidents, offset=log(months))
round(fit$summary.fixed[,1:5],3)
#>               mean      sd 0.025quant 0.5quant 0.975quant
#> (Intercept) -6.416 0.217      -6.852   -6.413    -5.998
#> built65-69   0.696 0.150       0.406    0.695     0.994
#> built70-74   0.819 0.170       0.487    0.818     1.153
#> built75-79   0.453 0.233      -0.012    0.455     0.903
#> oper75-79    0.384 0.118       0.153    0.384     0.617
#> typeB        -0.543 0.178      -0.882   -0.546    -0.185
#> typeC        -0.688 0.329      -1.366   -0.678    -0.075
#> typeD        -0.075 0.291      -0.664   -0.069     0.476
#> typeE         0.326 0.236      -0.141    0.327     0.785
```

Vamos pues a interpretar algunos de los efectos fijos. El parámetro  $\exp(\theta)$  representa el ratio medio de incidentes por mes navegado para los barcos que fueron construidos entre el 60 y el 64, han operado entre el 60 y el 74 y son de tipo A (las categorías de referencia). El parámetro  $\exp(\alpha_E^t)$  proporciona el incremento respecto de la categoría base `type=A` que se produce en el ratio mensual de incidentes en los barcos de `type=E`. Las distribuciones posteriores de dichos efectos se muestra en la Figura ??.

```
#names(fit$marginals.fixed)
# ratio medio de incidentes por mes en las categorías base
rbase=inla.tmarginal(function(x) exp(x),fit$marginals.fixed[[1]])
# riesgo relativo de barcos tipo E
riesgoE=inla.tmarginal(function(x) exp(x),fit$marginals.fixed$typeE)

g1=ggplot(as.data.frame(rbase),aes(x=x,y=y))+
  geom_line()+labs(x="Riesgo base=incidentes por mes",y="D.Posterior")
g2=ggplot(as.data.frame(riesgoE),aes(x=x,y=y))+
```

```
geom_line()+labs(x="Riesgo relativo type=E-A",y="D.Posterior")
grid.arrange(g1,g2,ncol=2)
```



Para el riesgo base tenemos:

```
# descriptivos
des.rbase=inla.zmarginal(rbase)
#> Mean          0.00167376
#> Stdev          0.000362092
#> Quantile 0.025 0.00105947
#> Quantile 0.25  0.00141477
#> Quantile 0.5   0.00163997
#> Quantile 0.75  0.00189552
#> Quantile 0.975 0.0024769
des.rbase
#> $mean
#> [1] 0.001673763
#>
#> $sd
#> [1] 0.0003620923
#>
#> $quant0.025
#> [1] 0.001059467
#>
#> $quant0.25
```

```
#> [1] 0.001414768
#>
#> $quant0.5
#> [1] 0.001639975
#>
#> $quant0.75
#> [1] 0.001895523
#>
#> $quant0.975
#> [1] 0.002476898
```

esto es, el ratio de incidentes por cada 1000 meses navegados para los niveles base será de 1.67 incidentes.

Para el incremento del riesgo de los barcos de tipo E respecto de los de tipo A, tenemos

```
des.riesgoE=inla.zmarginal(riesgoE)
#> Mean          1.42397
#> Stdev         0.336616
#> Quantile  0.025 0.870054
#> Quantile  0.25  1.18227
#> Quantile  0.5   1.38686
#> Quantile  0.75  1.62474
#> Quantile  0.975 2.18636
des.riesgoE
#> $mean
#> [1] 1.423969
#>
#> $sd
#> [1] 0.3366165
#>
#> $quant0.025
#> [1] 0.8700537
#>
#> $quant0.25
#> [1] 1.182271
#>
#> $quant0.5
#> [1] 1.386862
#>
#> $quant0.75
#> [1] 1.624742
#>
#> $quant0.975
```

```
#> [1] 2.186358
```

es decir, el riesgo de incidente por usar un barco de tipo E respecto a usar uno de tipo A se incrementa en un 42.4%.



# Referencias

- Berger, J. O. (2010). Statistical Decision Theory and Bayesian Analysis. Springer New York.
- Bolstad, W. M. (2004). Introduction to Bayesian Statistics. Wiley.
- Buck, C. E., Rakow, T., Garthwaite, P. H., O'Hagan, A., Oakley, J. E., Daneshkhah, A., Eiser, J. R., & Jenkinson, D. J. (2006). Uncertain Judgements: Eliciting Experts' Probabilities. Wiley.
- Buck, C. E., Rakow, T., Garthwaite, P. H., O'Hagan, A., Oakley, J. E., Daneshkhah, A., Eiser, J. R., & Jenkinson, D. J. (2006). Uncertain Judgements: Eliciting Experts' Probabilities. Wiley.
- Cameletti, M., & Blangiardo, M. (2015). Spatial and Spatio-temporal Bayesian Models with R - INLA. Wiley. Code and data in SSTBM-RINLA <https://sites.google.com/a/r-inla.org/stbook/datasets>
- Carlin, J. B., Gelman, A., Vehtari, A., Rubin, D. B., Dunson, D. B., & Stern, H. S. (2013). Bayesian Data Analysis. CRC Press. Available at [stat.columbia.edu](http://stat.columbia.edu). Course by Aki Vehtari [https://github.com/avehtari/BDA\\_course\\_Aalto](https://github.com/avehtari/BDA_course_Aalto)
- Christensen, R., Hanson, T. E., Branscum, A., & Johnson, W. (2011). Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians. Taylor & Francis.
- Congdon, P., & Congdon, P. P. (2003). Applied Bayesian Modelling (Wiley Series in Probability and Statistics). Wiley.
- CRAN Task View: Bayesian Inference. (n.d.). The Comprehensive R Archive Network. Retrieved November 17, 2022, from <https://cran.r-project.org/web/views/Bayesian.html>
- Donovan, T. M., & Mickey, R. M. (2019). Bayesian Statistics for Beginners: A Step-by-step Approach. Oxford University Press.

- Dorazio, R. M., & Royle, J. A. (2008). Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities. Elsevier Science.
- Faraway, J. (n.d.). Examples using INLA for Bayesian Regression Models. Julian Faraway. Retrieved October 27, 2022, from <http://julianfaraway.github.io/brinla/examples/>
- Faraway, J. (2022, August 23). Package faraway: Functions and Datasets for Books by Julian Faraway. The Comprehensive R Archive Network. Retrieved October 27, 2022, from <https://cran.r-project.org/web/packages/faraway/faraway.pdf>
- Faraway, J. J., Wang, X., & Yue, Y. (2018). Bayesian Regression Modeling with INLA. CRC Press.
- Gabry, J., & Mahr, T. (n.d.). Plotting for Bayesian Models • bayesplot. Stan. Retrieved November 17, 2022, from <http://mc-stan.org/bayesplot/>
- Gabry, J., Simpson, D., Vehtari, A., Betancourt, M. and Gelman, A. (2019), Visualization in Bayesian workflow. J. R. Stat. Soc. A, 182: 389-402. <https://doi.org/10.1111/rssa.12378>
- Gelman, A., Fagan, J., & Kiss, A. (2007). An Analysis of the New York City Police Department's "Stop-and-Frisk" Policy in the Context of Claims of Racial Bias. Journal of the American Statistical Association, 102(479), 813-823. 10.1198/016214506000001040
- Gómez-Rubio, V. (2020). Bayesian Inference with INLA. CRC Press. Available at <https://becarioprecario.bitbucket.io/inla-gitbook/index.html>
- Hoff, P. D. (2009). A First Course in Bayesian Statistical Methods. Springer.
- Kurt, W. (2019). Bayesian Statistics the Fun Way: Understanding Statistics and Probability with Star Wars, LEGO, and Rubber Ducks. No Starch Press.
- Link, W. A., Link, W. A., & Barker, R. J. (2010). Bayesian Inference: With Ecological Applications (R. J. Barker & W. A. Link, Eds.). Elsevier Science.
- Malouche, D. (n.d.). R tools for Bayesian Statistics. Kaggle. Retrieved November 17, 2022, from <https://www.kaggle.com/code/dhafer/r-tools-for-bayesian-statistics>
- McGrayne, S. B. (2011). The Theory that Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, & Emerged Triumphant from Two Centuries of Controversy. Yale University Press.

- McGrayne, S. B. (2011). *The Theory that Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, & Emerged Triumphant from Two Centuries of Controversy*. Yale University Press.
- plot: Default INLA plotting in andrewzm/INLA: Functions which allow to perform full Bayesian analysis of latent Gaussian models using Integrated Nested Laplace Approximation. (n.d.). RDRR.io. Retrieved November 17, 2022, from <https://rdr.io/github/andrewzm/INLA/man/plot.html>
- R-INLA project. (2022). R-INLA Project. Retrieved November 17, 2022, from <https://www.r-inla.org/>
- Robert, C. P., & Marin, J.-M. (2013). *Bayesian Essentials with R*. Springer New York.
- Ross, K. (2022). *An Introduction to Bayesian Reasoning and Methods*. Online publication available at [https://bookdown.org/kevin\\_davisross/bayesian-reasoning-and-methods/poisson.html](https://bookdown.org/kevin_davisross/bayesian-reasoning-and-methods/poisson.html). Consultado en octubre 2022.
- Rue, H. et al (2017). Bayesian Computing with INLA: A Review. *Annu.Rev.Stat.Appl.* 4: 395:421. DOI: 10.1146/annurev-statistics-060116-054045.
- Samanta, T., Ghosh, J. K., & Delampady, M. (2006). *An Introduction to Bayesian Analysis: Theory and Methods* (T. Samanta, M. Delampady, & J. K. Ghosh, Eds.). Springer.
- Statisticat, LLC. (2015). *LaplacesDemon Examples*. Retrieved October 23, 2022, from <https://cran.r-project.org/web/packages/LaplacesDemon/vignettes/Examples.pdf>
- Statisticat, LLC. (2015). *Bayesian Inference*. The Comprehensive R Archive Network. Retrieved October 23, 2022, from <https://cran.r-project.org/web/packages/LaplacesDemon/vignettes/BayesianInference.pdf>
- Stigler, S. (2016, April 12). *LaplacesDemon: A Complete Environment for Bayesian Inference within R*. The Comprehensive R Archive Network. Retrieved October 23, 2022, from <https://cran.r-project.org/web/packages/LaplacesDemon/vignettes/LaplacesDemonTutorial.pdf>
- Wang, X., Yue, R., & Faraway, J. (n.d.). *Bayesian Regression with INLA | brinla*. Julian Faraway. Retrieved October 27, 2022, from <http://julianfaraway.github.io/brinla/>