Final-2023

Mariano Villafuerte -156057 Mario Medina - 156940

03 diciembre 2023

Contents

1 Bayesiana y Regularización **1**

2 Apéndice **11**

1 Bayesiana y Regularización

Los datos *pew_research_center_june_elect_wknd_data.dta* tienen información de encuestas realizadas durante la campaña presidencial 2008 de EUA.

NOTA: para cada respuesta se presenta solo un ejemplo del data.frame, al final como anexo se agrega la tabla para los 48 estados considerados.

```
set.seed(156057)
data <- foreign::read.dta("data/pew_research_center_june_elect_wknd_data.dta")
```

- Estima el porcentaje de la población de cada estado (excluyendo Alaska, Hawai, y DC) que se considera *very liberal*, utilizando el estimador de máxima verosimilitud.

Grafica en el eje x el número de encuestas para cada estado y en el eje y la estimación de máxima verosimilitud para very liberal. ¿Qué observas?

R: Excluimos los estados mencionados y utilizamos la variable *weight* como factor de expansión para realizar los calculos del porcentaje de *very liberal*. Además cada registro lo tomamos en cuenta como si fuera 1 encuesta, entonces entendemos por número de encuestas como número de registros agrupados por la variable *state*. El estimador de **máxima verosimilitud** para una proporción, *very liberal* como se ha demostrado varias veces en clase y en la pregunta 2 será $\frac{x_{very.liberal}}{n}$

```
set.seed(156057)
# Consideramos que la población total de mi estado es la suma de "weight"
very_liberal <-
  data %>%
  # Sin alaska, hawaii, washington dc
  dplyr::filter(!(state %in% c("washington dc","hawaii","alaska"))) %>%
  dplyr::group_by(state, ideo) %>%
  dplyr::summarise(poblacion = sum(weight),
                   poblacion_fix = n()) %>%
  tidyr::pivot_wider(id_cols = state, names_from = ideo, values_from = poblacion) %>%
```

```r
  ungroup()

# very_liberal_fixed <-
#   data %>%
#   # Sin alaska, hawaii, washington dc
#   dplyr::filter(!(state %in% c("washington dc","hawaii","alaska"))) %>%
#   dplyr::group_by(state, ideo) %>%
#   dplyr::summarise(poblacion = n()) %>%
#   tidyr::pivot_wider(id_cols = state, names_from = ideo, values_from = poblacion) %>%
#   ungroup()

# Población total = suma de weight; # de encuestas = conteo de renglones
poblacion_encuestas <-
  data %>%
  # Sin alaska, hawaii, washington dc
  dplyr::filter(!(state %in% c("washington dc","hawaii","alaska"))) %>%
  dplyr::group_by(state) %>%
  dplyr::summarise(poblacion = sum(weight),
                   encuestas = n()) %>% ungroup()

# poblacion_encuestas_fixed <-
#   data %>%
#   # Sin alaska, hawaii, washington dc
#   dplyr::filter(!(state %in% c("washington dc","hawaii","alaska"))) %>%
#   dplyr::group_by(state) %>%
#   dplyr::summarise(poblacion = n(),
#                    encuestas = n()) %>% ungroup()

tabla_aux <- dplyr::left_join(very_liberal, poblacion_encuestas, by = "state")
#tabla_aux_fixed <- dplyr::left_join(very_liberal_fixed, poblacion_encuestas_fixed, by = "state")

tabla_aux <- tabla_aux %>%
  dplyr::mutate(perc_very_liberal = round((`very liberal`/poblacion)*100,2))

# tabla_aux_fixed <- tabla_aux_fixed %>%
#   dplyr::mutate(perc_very_liberal = round((`very liberal`/poblacion)*100,2)) %>%
#   dplyr::select(state, perc_very_liberal, encuestas)

kable(tabla_aux %>%
      dplyr::select(state, perc_very_liberal, encuestas) %>% head(),
      caption = "Head por estado con MLE",
      digits = 4,
      format = "latex",
      booktabs = T) %>%
  kable_styling(latex_options = c("striped"),
                bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                full_width = F, fixed_thead = T)
```

```r
#tabla_aux_fixed %>% head()
```

Graficamos... Observamos una **ligera relación positiva: a mayor número de encuestas mayor el porcentaje de población que se identifican con la ideología de very liberal**. Esto es de llamar la atención, no deberían de depender nuestros resultados del número de encuestas podría haber problemas con

Table 1: Head por estado con MLE

state	perc_very_liberal	encuestas
alabama	7.31	624
arizona	6.42	542
arkansas	1.75	307
california	6.34	2854
colorado	6.20	468
connecticut	2.99	395

el muestro. Sin embargo, al estar hablando de porcentajes "relativamente bajos" el encuestar podría llegarse a ver esta relación.

```
set.seed(156057)
ggplot(tabla_aux, aes(x = encuestas, y = perc_very_liberal)) +
  geom_point() +
  geom_text_repel(aes(label = state), box.padding = 0.5, size = 1.5) +
  #geom_text(aes(label = state), hjust = 1.5, vjust = 0.5, size = 1.5) +
  geom_smooth(method = "lm", se = FALSE, color = "#F8766D", linetype = "dashed") +
  labs(title = "Num. Encuestas vs. Pct Very Liberal") +
  xlab("Num Encuestas") + ylab("Pct Very Liberal (MLE)") +
  ylim(0,12) +
  ggpubr::theme_pubclean(base_size = 12)
```

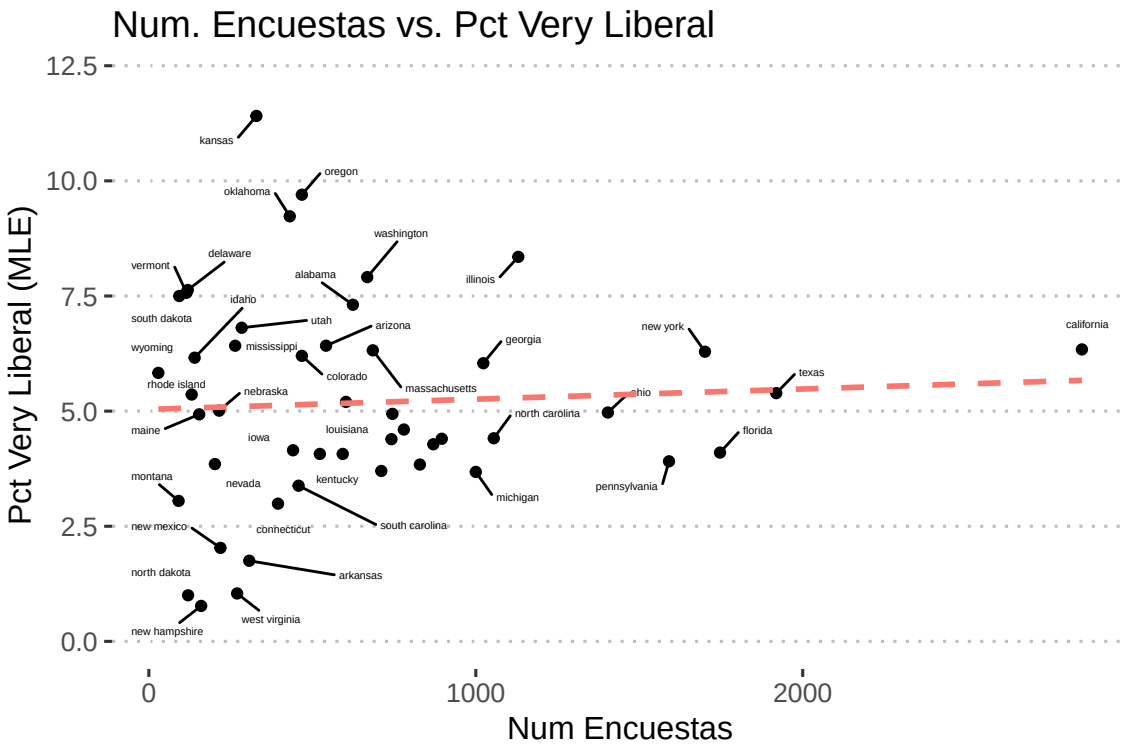

Figure 1: Pct Very Liberal vs. Num. Encuestas (archivo DTA)

Table 2: Head por estado con MLE + Obama

state	perc_very_liberal	encuestas	vote_Obama_pct
ALABAMA	7.31	624	38.8
ARIZONA	6.42	542	45.0
ARKANSAS	1.75	307	38.8
CALIFORNIA	6.34	2854	60.9
COLORADO	6.20	468	53.5
CONNECTICUT	2.99	395	60.5

```
# ggplot(tabla_aux_fixed, aes(x = encuestas, y = perc_very_liberal)) +
#   geom_point() +
#   geom_text_repel(aes(label = state), box.padding = 0.5, size = 1.5) +
#   #geom_text(aes(label = state), hjust = 1.5, vjust = 0.5, size = 1.5) +
#   geom_smooth(method = "lm", se = FALSE, color = "#F8766D") +
#   labs(title = "# Encuestas vs. % Very Liberal") +
#   xlab("# Encuestas") + ylab("% Very Liberal") +
#   ggpubr::theme_pubclean(base_size = 12)
```

Grafica en el eje x el porcentaje de votos que obtuvo Obama en la elección para cada estado y en el eje y la estimación de máxima verosimilitud para *very liberal*. ¿Qué observas? (usa los datos *2008ElectionResult.csv*)

R: Realizamos la unión de basos de datos

```
set.seed(156057)
data_xls <- data.table::fread("data/2008ElectionResult.csv") %>%
  dplyr::select(state, vote_Obama_pct)
tabla_aux <- dplyr::mutate(tabla_aux, state = toupper(state))
data_xls <- dplyr::mutate(data_xls, state = toupper(state))

tabla_aux <- left_join(tabla_aux, data_xls, by = "state")

kable(tabla_aux %>%
      dplyr::select(state, perc_very_liberal, encuestas, vote_Obama_pct) %>%
      head(),
      caption = "Head por estado con MLE + Obama",
      digits = 4,
      format = "latex",
      booktabs = T) %>%
  kable_styling(latex_options = c("striped"),
                bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                full_width = F, fixed_thead = T)
```

Graficamos... Observamos una **ligera relación negativa: a mayor número de votantes de Obama menor el porcentaje de población que se identifica con la ideología de very liberal**. Esto hace sentido ya que Obama en general fue percibido como un candidato de *centro-izquierda* más que alguien que representará la ideología muy liberal

```
set.seed(156057)
ggplot(tabla_aux, aes(x = vote_Obama_pct, y = perc_very_liberal)) +
  geom_point() +
```

```
  geom_text_repel(aes(label = state), box.padding = 0.5, size = 1.2) +
  #geom_text(aes(label = state), hjust = 1.5, vjust = 0.5, size = 1.5) +
  geom_smooth(method = "lm", se = FALSE, color = "#F8766D", linetype = "dashed") +
  labs(title = "Pct Obama vs. Pct Very Liberal") +
  xlab("Pct Obama") + ylab("Pct Very Liberal (MLE)") +
  ylim(0,12) +
  ggpubr::theme_pubclean(base_size = 12)
```

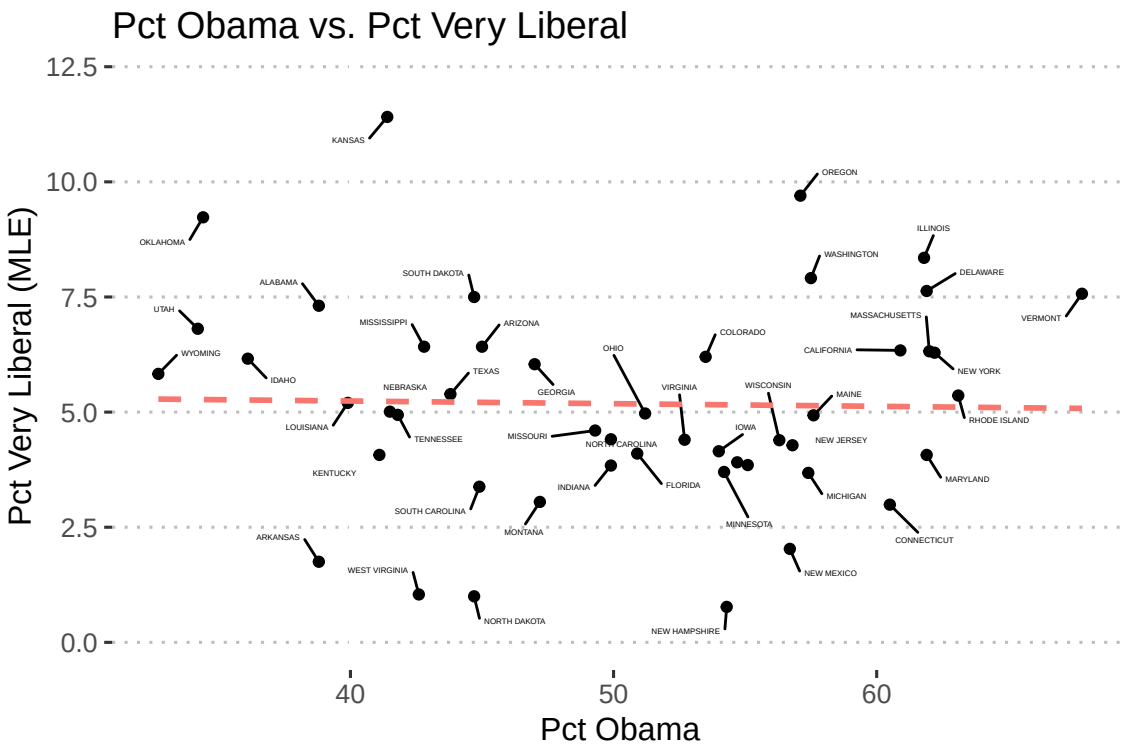

Figure 2: MLE vs. Pct de votantes por Obama

Estima el mismo porcentaje (*very liberal*) usando inferencia bayesiana, en particular la familia conjugada beta-binomial. Deberás estimar la proporción de manera independiente para cada estado, sin embargo, utilizarás la misma inicial a lo largo de todos: Beta(8, 160).

R: Al tener un modelo beta-binomial sabemos que la **posterior tendrá una distribución Beta.**. Previo a establecer una función general para la posterior realizamos simulaciones de la inicial para ver que información previa tenemos.

$$p(\theta) \propto \theta^{8-1}(1-\theta)^{160-1}$$

Como era de esperar tenemos una **media cercana al 5%. No es simétrica ya que considera los posibles escenarios que cubren las colas largas** para porcentajes más altos.

```
set.seed(156057)
sim_inicial <- tibble(theta = stats::rbeta(10000,8,160))
ggplot(sim_inicial) +
  geom_histogram(aes(x = theta, y = ..density..), bins = 15,
                 fill = "#F8766D",
```

```
                alpha=0.5) +
  geom_vline(xintercept = (8/168), color = "red") +
  annotate("text", x = (8/168), y = Inf, label = "Media",
           vjust = 1, hjust = 0.5,colour = "red")+
  labs(title = "Distribución Inicial", subtitle = "Beta(8,160)") +
  xlab("Theta") + ylab("Densidad") +
  ggpubr::theme_pubclean(base_size = 12)
```

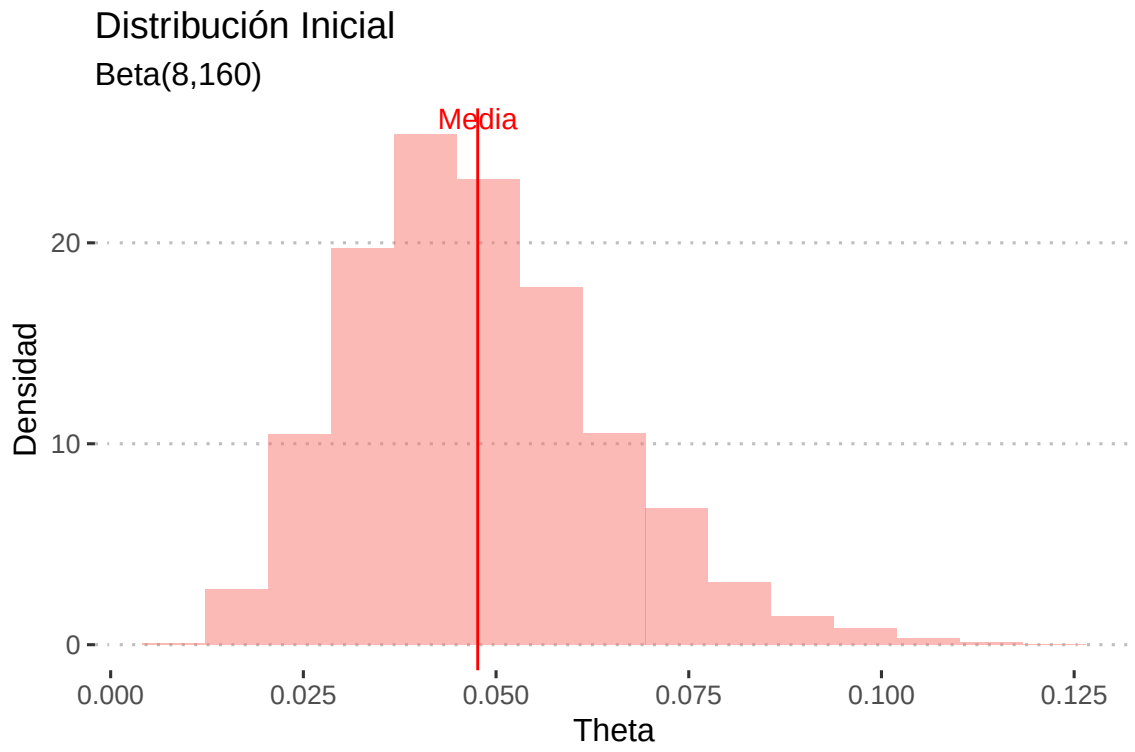

Figure 3: Distribución a priori: Beta (8,160)

La posterior para el *estado i* estará dada por:

$$P(\theta|X) \propto P(X|\theta)P(\theta)$$
$$P(\theta|X) \propto \theta^{k_i+7}(1-\theta)^{n_i+159} \sim Beta(k_i+8, n_i+160)$$

Donde:

- k_i: es el número de individuos identificados con la ideología *very liberal*
- n_i: total de la población

Y nuestro estimador puntual será la media posterior, es decir $\frac{k_i+8}{(k_i+8)+(n_i+160)}$

```
set.seed(156057)
for (i in 1:nrow(tabla_aux)) {
  alpha <- tabla_aux$`very liberal`[i] + 8
  beta <- tabla_aux$poblacion[i] + 160
```

Table 3: Head por estado con estimaciones (+Bayesiano)

state	perc_very_liberal	bayesiano	encuestas
ALABAMA	7.31	6.63	624
ARIZONA	6.42	5.90	542
ARKANSAS	1.75	2.30	307
CALIFORNIA	6.34	5.94	2854
COLORADO	6.20	5.72	468
CONNECTICUT	2.99	3.17	395

```
  tabla_aux$bayesiano[i] <- round((alpha/(alpha+beta)*100),2)
}


kable(tabla_aux %>%
      dplyr::select(state, perc_very_liberal, bayesiano, encuestas) %>%
      head(),
      caption = "Head por estado con estimaciones (+Bayesiano)",
      digits = 4,
      format = "latex",
      booktabs = T) %>%
  kable_styling(latex_options = c("striped"),
                bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                full_width = F, fixed_thead = T)
```

Para dos de los estados: Idaho y Virginia, adicional a calcular la posterior usando las propiedades de la familia conjugada, utiliza Stan para hacer la inferencia, revisa los diagnósticos de convergencia y describe tus observaciones.

R: Realizamos estimaciones vía **Stan*. Dado que es la misma inicial un mismo modelo nos funciona para ambos Estados.

```
set.seed(156057)
archivo_stan <- file.path("stan/modelo_preg3.stan")
# compilar
mod <- cmdstan_model(archivo_stan)
#mod
```

Pasamos datos, muestreamos y revisamos convergencia

IDAHO

```
n <- dplyr::filter(tabla_aux, state == "IDAHO")$poblacion
y <- dplyr::filter(tabla_aux, state == "IDAHO")$`very liberal`
datos_lista <- list(n = n, y = y)
ajuste <- mod$sample(
  data = datos_lista,
  seed = 156057,
  chains = 4,
  iter_warmup = 5000,
  iter_sampling = 20000,
```

```
  parallel_chains = 4,
  show_messages = F)
ajuste$cmdstan_diagnose()
```

```
## Processing csv files: C:/Users/mario/AppData/Local/Temp/Rtmp4CpAII/modelo_preg3-202312032134-1-183a9
##
## Checking sampler transitions treedepth.
## Treedepth satisfactory for all transitions.
##
## Checking sampler transitions for divergences.
## No divergent transitions found.
##
## Checking E-BFMI - sampler transitions HMC potential energy.
## E-BFMI satisfactory.
##
## Effective sample size satisfactory.
##
## Split R-hat values satisfactory all parameters.
##
## Processing complete, no problems detected.
```

```
idaho <- ajuste$summary() %>%
  dplyr::mutate(state = "Idaho") %>%
  dplyr::select(state,variable, mean, sd, rhat, ess_bulk, ess_tail)
```

VIRGINIA

```
n <- dplyr::filter(tabla_aux, state == "VIRGINIA")$poblacion
y <- dplyr::filter(tabla_aux, state == "VIRGINIA")$`very liberal`
datos_lista <- list(n = n, y = y)
ajuste <- mod$sample(
  data = datos_lista,
  seed = 156057,
  chains = 4,
  iter_warmup = 5000,
  iter_sampling = 20000,
  parallel_chains = 4,
  show_messages = F)
ajuste$cmdstan_diagnose()
```

```
## Processing csv files: C:/Users/mario/AppData/Local/Temp/Rtmp4CpAII/modelo_preg3-202312032134-1-835f9
##
## Checking sampler transitions treedepth.
## Treedepth satisfactory for all transitions.
##
## Checking sampler transitions for divergences.
## No divergent transitions found.
##
## Checking E-BFMI - sampler transitions HMC potential energy.
## E-BFMI satisfactory.
##
## Effective sample size satisfactory.
```

Table 4: Estimaciones via Stan para Idaho y Virginia

state	variable	mean	sd	rhat	ess_bulk	ess_tail
Idaho	lp___	-126.9621	0.7103	1.0000	35612.33	41426.59
Idaho	theta	0.0572	0.0096	1.0000	30507.51	34989.49
Idaho	theta_inicial	0.0477	0.0165	1.0000	79460.34	79388.52
Virginia	lp___	-451.7294	0.7154	1.0001	35564.62	43272.79
Virginia	theta	0.0440	0.0041	1.0001	29322.89	37589.43
Virginia	theta_inicial	0.0477	0.0165	1.0000	80329.04	79109.27

```
##
## Split R-hat values satisfactory all parameters.
##
## Processing complete, no problems detected.
```

```
virginia <- ajuste$summary() %>%
  dplyr::mutate(state = "Virginia") %>%
  dplyr::select(state,variable, mean, sd, rhat, ess_bulk, ess_tail)
```

Observamos que el valor de la posterior tiene efectos diferentes entre ambos estados. Para Idaho la media posterior aumenta, pasando de 0.047 a 0.057 y para Virginia disminuye ligeramente de 0.047 a 0.043

```
stan_resumen <- rbind(idaho, virginia)
kable(stan_resumen,
      caption = "Estimaciones via Stan para Idaho y Virginia",
      digits = 4,
      format = "latex",
      booktabs = T) %>%
  kable_styling(latex_options = c("striped"),
                bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                full_width = F, fixed_thead = T)
```

Realizamos las mismas gráficas; sin embargo, ahora utilizamos la media posterior como estimadores puntuales.

La relación encontrada se mantiene respecto a la relación entre número de encuestas y el porcentaje estimado de la población con ideología *very liberal*

```
set.seed(156057)
tabla_aux_plot <-
  tabla_aux %>%
  dplyr::select(state, perc_very_liberal, vote_Obama_pct, bayesiano, encuestas) %>%
  dplyr::rename(MLE=perc_very_liberal) %>%
  tidyr::pivot_longer(
    cols = c(MLE, bayesiano),
    names_to = "Metodologia",
    values_to = "value"
  )
ggplot(tabla_aux_plot, aes(x = encuestas, y = value, color = Metodologia)) +
  geom_point() +
  geom_text_repel(aes(label = state), box.padding = 0.5, size = 1.5) +
```

```
  #geom_text(aes(label = state), hjust = 1.5, vjust = 0.5, size = 1.5) +
  #geom_smooth(method = "lm", se = FALSE, color = Metodologia) +
  geom_smooth(method = "lm", se = FALSE,
              aes(group = Metodologia, color = Metodologia),
              linetype = "dashed") +
  labs(title = "Num. Encuestas vs. Pct Very Liberal", subtitle = "MLE vs. Bayesiano") +
  xlab("Num. Encuestas") + ylab("Pct Very Liberal") +
  ylim(0,12) +
  ggpubr::theme_pubclean(base_size = 12)
```

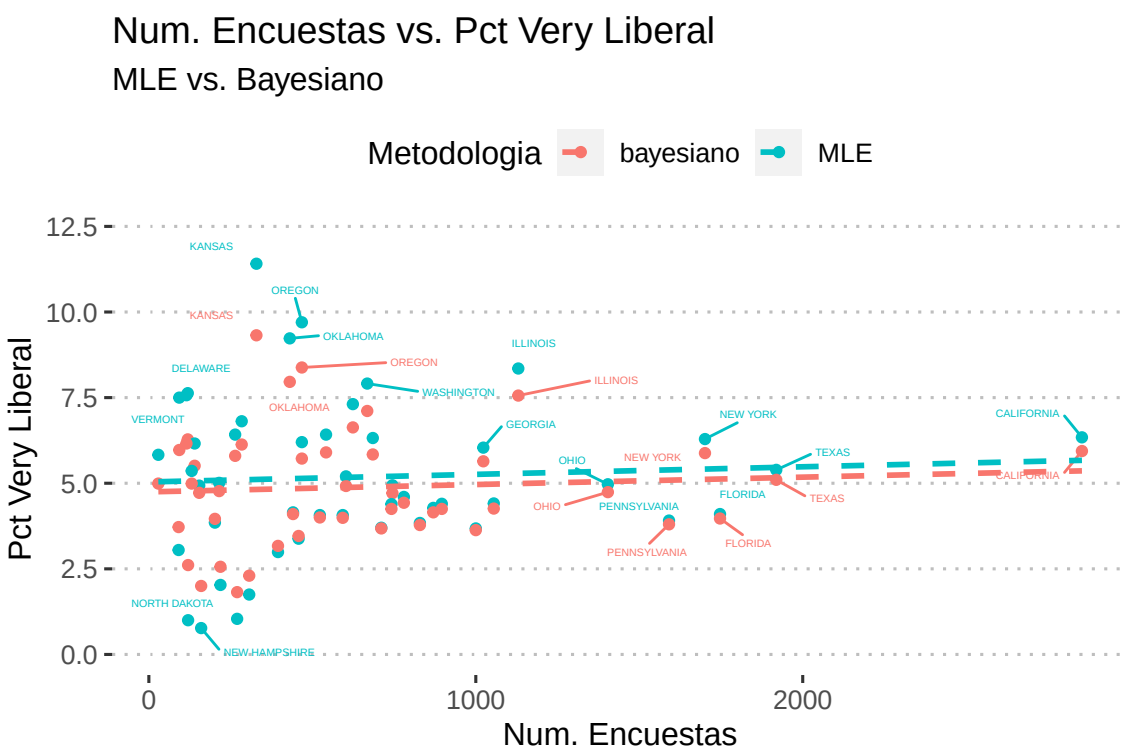

Figure 4: Bayes-MLE y Num. Encuestas

La relación encontrada se mantiene respecto a la relación entre porcentaje de población votante por Obama y el porcentaje estimado de la población con ideología *very liberal*. Un factor a considerar es que en esta gráfica es mucho más notorio que la estimación bayesiana en probabilidades bajas obtenidas por MLE son un poco mayores y probabilidades relativamente más altas por MLE ahora son un poco menores.

```
ggplot(tabla_aux_plot, aes(x = vote_Obama_pct, y = value, color = Metodologia)) +
  geom_point() +
  geom_text_repel(aes(label = state), box.padding = 0.5, size = 1.5) +
  #geom_text(aes(label = state), hjust = 1.5, vjust = 0.5, size = 1.5) +
  #geom_smooth(method = "lm", se = FALSE, color = Metodologia) +
  geom_smooth(method = "lm", se = FALSE,
              aes(group = Metodologia, color = Metodologia),
              linetype = "dashed") +
  labs(title = "Num Encuestas vs. Pct Very Liberal", subtitle = "MLE vs. Bayesiano") +
  xlab("Pct Obama") + ylab("Pct Very Liberal") +
```

```
  ylim(0,12) +
  ggpubr::theme_pubclean(base_size = 12)
```

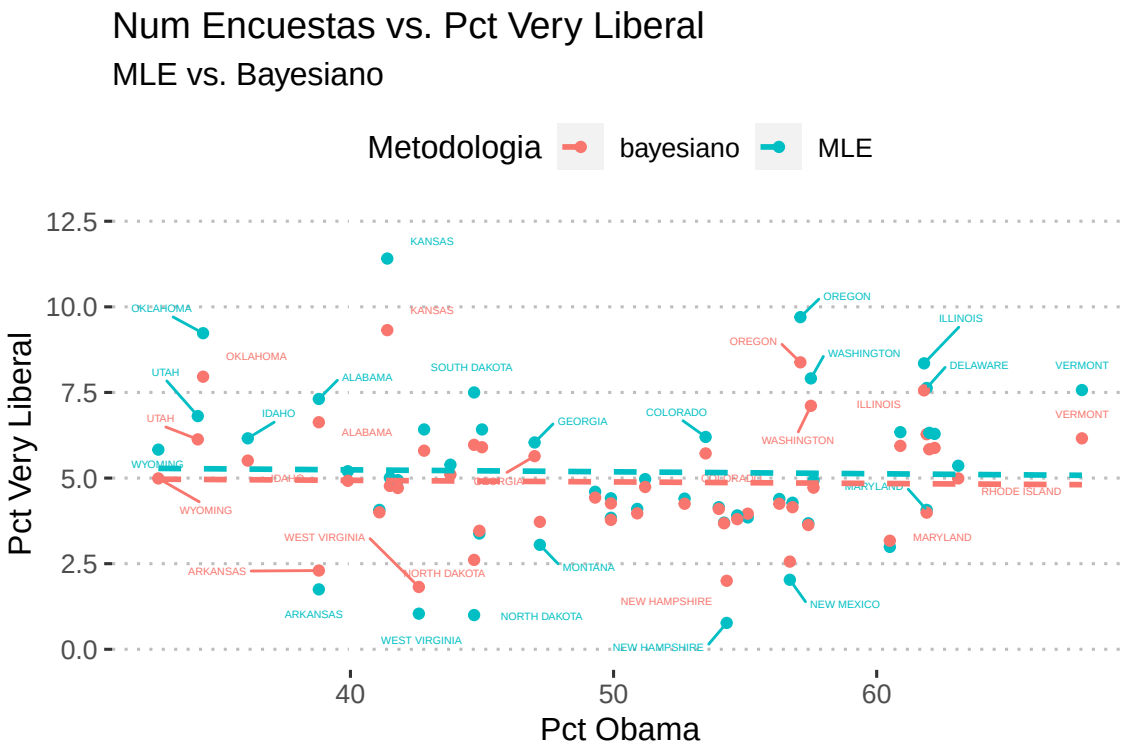

Figure 5: Bayes vs. MLE y Pct Obama

2 Apéndice

Tabla con todas las estimaciones para todos los estados.

```
kable(tabla_aux %>%
        dplyr::select(state, perc_very_liberal, bayesiano) %>%
        dplyr::rename(MLE = perc_very_liberal, Bayesiano = bayesiano),
      caption = "Estimaciones por estado",
      digits = 4,
      format = "latex",
      booktabs = T) %>%
  kable_styling(latex_options = c("striped"),
                bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                full_width = F, fixed_thead = T)
```

Table 5: Estimaciones por estado

state	MLE	Bayesiano
ALABAMA	7.31	6.63
ARIZONA	6.42	5.90
ARKANSAS	1.75	2.30
CALIFORNIA	6.34	5.94
COLORADO	6.20	5.72
CONNECTICUT	2.99	3.17
DELAWARE	7.63	6.28
FLORIDA	4.10	3.97
GEORGIA	6.04	5.64
IDAHO	6.16	5.51
ILLINOIS	8.35	7.56
INDIANA	3.84	3.78
IOWA	4.15	4.10
KANSAS	11.41	9.32
KENTUCKY	4.07	4.00
LOUISIANA	5.20	4.92
MAINE	4.93	4.72
MARYLAND	4.07	3.99
MASSACHUSETTS	6.32	5.84
MICHIGAN	3.68	3.63
MINNESOTA	3.70	3.68
MISSISSIPPI	6.42	5.80
MISSOURI	4.60	4.43
MONTANA	3.05	3.72
NEBRASKA	5.01	4.77
NEVADA	3.85	3.96
NEW HAMPSHIRE	0.77	2.00
NEW JERSEY	4.28	4.15
NEW MEXICO	2.03	2.56
NEW YORK	6.29	5.88
NORTH CAROLINA	4.41	4.26
NORTH DAKOTA	1.00	2.61
OHIO	4.97	4.74
OKLAHOMA	9.23	7.96
OREGON	9.70	8.38
PENNSYLVANIA	3.91	3.80
RHODE ISLAND	5.36	4.99
SOUTH CAROLINA	3.38	3.46
SOUTH DAKOTA	7.50	5.97
TENNESSEE	4.94	4.71
TEXAS	5.39	5.10
UTAH	6.81	6.13
VERMONT	7.57	6.16
VIRGINIA	4.40	4.25
WASHINGTON	7.91	7.11
WEST VIRGINIA	1.04	1.82
WISCONSIN	4.39	4.25
WYOMING	5.83	4.99