Examen final - Otoño 2020

Alfredo Garbuno Iñigo

Entrega: Enviar la carpeta que el codigo de solución (.Rmd y funciones auxiliares) a mas tardar el 15 de Diciembre antes de las 12:00pm (mediodia), por correo electrónico con el título fundamentos-final, un solo documento por equipo. No se aceptarán entregas extemporáneas. Será mejor entregar un examen resuelto parcialmente, que no entregar nada.

Instrucciones:

- Tus respuestas deben ser claras y debes explicar los resultados, incluye también tus procedimientos/código de manera ordenada, y el código comentado.
- Se evaluará la presentación de resultados (calidad de las gráficas, tablas,...), revisa la sección de visualización en las notas.
- Las sesiones del Martes 8 y Jueves 10 de Diciembre a las 10 am, serán espacios para resolver dudas que puedan surgir del exámen.
- No pueden compartir soluciones entre diferentes equipos, o alumnos del grupo 001 de esta misma materia.
- Al entregar este examen afirmas que el trabajo se realizó sólo con tu compañero de equipo. El material que utilizaste para apoyarte consistió de las notas en clase (pdf en canvas), el codigo fuente de las notas en el repositorio de Github.
- Al entregar estás dando tu consentimiento para que bajo sospecha y suficiente evidencia de copia se anule tu evaluación.

Preparación de ambiente

Asegurate de tener instalado los paquetes que usamos más en las notas del curso. En particular, si usas `renv` como manejador de ambientes puedes instalarlos con las instrucciones de abajo. Sólo necesitarías descomentarlas.

```
# renv::install("tidyverse")
# renv::install("patchwork")
# renv::install("nullabor")
# renv::install("scales")
# renv::install('diegovalle/mxmaps')
# renv::install("nleqslv")
# renv::snapshot()

# Escribe las claves unicas de ambos miembros del equipo, para generar una
# semilla de numeros aleatorios.
claves_unicas <- c(150370, 2)
set.seed(min(claves_unicas))
```

Modelos de conteo

En el curso hemos estudiado las variables aleatorias Gaussianas para modelar eventos aleatorios compuestos de pequeños, pero controlados, efectos. También hemos utilizado variables aleatorias Binomiales para modelar tasas de éxito de algún evento binario de interés. En el contexto Bayesiano, hemos utilizado las distribuciones Beta, Gamma-Inversa, y Normal para realizar análisis conjugado con estos modelos.

En este mini-proyecto, ilustraremos otra familia de distribuciones muy cómunes en la práctica. En particular, veremos la distribución **Poisson** como unmodelo de conteo. Es decir, una variable aleatoria Poisson nos sirve para modelar el número de ocurrencias de un evento en un periodo (tiempo) o área (espacio) base.

Decimos que $x|\theta \sim \mathsf{Poisson}(\theta)$ si los eventos ocurren de manera independiente y a una tasa constante. La función de masa de probabilidad esta dada por

$$p(X = k \,|\, \theta) = \frac{\theta^k \, e^{-\theta}}{k!},$$

donde sabemos que

$$\mathbb{E}[x|\theta] = \theta, \qquad \mathbb{V}[x|\theta] = \theta$$

Al examinar la base de la función de masa de probabilidad notamos que un candidato para un anâlisis conjugado es una distribución Gamma. Es decir, un candidato *natural* para una distribución previa para θ es

$$\theta \sim \mathsf{Gamma}(\alpha, \beta),$$

donde la densidad está dada por

$$p(\theta) \propto \theta^{\alpha-1} \, e^{-\beta \, \theta},$$

y tenemos los siguientes momentos

$$\mathbb{E}[\theta] = \frac{\alpha}{\beta}, \qquad \mathbb{V}[\theta] = \frac{\alpha}{\beta^2}.$$

Pregunta 1) Para una muestra $X_1, \ldots, X_n \overset{iid}{\sim} \mathsf{Poisson}(\theta)$, determina la distribución posterior de θ, y calcula media y varianza de la distribución posterior. ¿Podríamos escribir la media posterior como un promedio ponderado entre datos e información previa? ¿Cómo interpretas los hiper-parámetros (α, β)?

$$\begin{aligned}
\Pi(\theta \mid x_1, \ldots, x_n) &\propto \Pi(x_1, \ldots, x_n \mid \theta) \cdot \Pi(\theta) \\
&\propto \prod_i \Pi(x_i \mid \theta) \cdot \Pi(\theta) \\
&\propto \prod_i \frac{\theta^{x_i} e^{-\theta}}{x_i!} \cdot \theta^{\alpha-1} e^{-\beta\theta} \\
&\propto \frac{\theta^{\sum x_i} e^{-n\theta}}{\prod x_i!} \cdot \theta^{\alpha-1} e^{-\beta\theta} \\
&\propto \theta^{\alpha+\sum x_i - 1} e^{-\theta(\beta+n)}
\end{aligned}$$

De lo anterior se puede ver que

$$\Pi(\theta \mid x_1, \ldots, x_n) \sim Gamma(\alpha + \sum x_i, \beta + n)$$

Así,

$$E[\theta \mid x_1, \ldots, x_n] = \frac{\alpha + \sum x_i}{\beta + n} Var[\theta \mid x_1, \ldots, x_n] = \frac{\alpha + \sum x_i}{(\beta + n)^2}$$

Sí se puede ver como un promedio entre los datos y el parámetro α. Entre más datos sean más peso tendrán en la posterior. En cambio si los datos $n \to 0$, la posterior será igual a la distribución a priori.

Otra variable aleatoria de conteo relevante es la **Binomial Negativa.** Esta distribución sirve para modelar el número de éxitos en una secuencia de experimentos Bernoulli antes de encontrar un número específico de fracasos.

Decimos que $X|\alpha, \beta \sim \mathsf{Neg\text{-}Bin}(\alpha, \beta)$, donde X es el número de éxitos que contamos antes de α fracasos, cuando cada fracaso ocurre con probabilidad $\frac{\beta}{\beta+1}$. La función de masa de probabilidad se escribe

$$p(X = k \,|\, \alpha, \beta) = \binom{\alpha + k - 1}{k} \left(\frac{\beta}{\beta + 1}\right)^{\alpha} \left(\frac{1}{\beta + 1}\right)^{k}.$$

Nota que

$$\binom{\alpha + k - 1}{k} = \binom{\alpha + k - 1}{\alpha - 1},$$

es decir, el número de formas que puedes acomodar $\alpha - 1$ fracasos es igual al número de formas que puedes acomodar k éxitos cuando realizaste $\alpha + k - 1$ experimentos y todos los experimentos son independientes. Por otro lado, la definición es

$$\binom{\alpha + k - 1}{k} = \frac{(\alpha + k - 1)!}{k!\,(\alpha - 1)!}.$$

donde $k! = k \times k - 1 \times k - 2 \times \cdots \times 1$, y la función Gamma satisface

$$\Gamma(\alpha) = (\alpha - 1)!.$$

Pregunta 2) Bajo el modelo conjugado que escribiste en la pregunta 1, calcula la **distribución predictiva previa** para una observación Poisson. Es decir, calcula

$$p(y) = \int \mathsf{Poisson}(y \,|\, \theta) \mathsf{Gamma}(\theta \,|\, \alpha, \beta)\, \mathrm{d}\theta.$$

Verifica tu cálculo utilizando las reglas probabilidad condicional. En especifico, utiliza

$$p(y) = \frac{p(y|\theta)p(\theta)}{p(\theta|y)}.$$

¿Qué distribución marginal tiene y bajo el modelo conjugado?

En la práctica, es útil extender el modelo Poisson como sigue

$$x_i|t_i, \theta \sim \mathsf{Poisson}(\lambda_i), \tag{1}$$

$$\lambda_i = t_i\theta, \tag{2}$$

donde la tasa de ocurrencia λ_i ha sido descompuesta en un producto que incorpora la exposición t_i y una tasa de ocurrencia por unidades expuestas θ. En este contexto usualmente tenemos observaciones para x_i y t_i pues conocemos el parámetro de exposición. Por ejemplo, si x_i es el número de personas que se enferman de gripe en la i-ésima ciudad en un año, entonces θ denota la tasa anual por persona de enfermarse de gripe en una población de tamaño t_i.

Pregunta 3) Supongamos que tenemos datos $X_1, \ldots, X_n \sim \text{Poisson}(\lambda_i)$, con $\lambda_i = t_i\theta$ para $i = 1, \ldots, n$. Utilizando el modelo conjugado, ¿cuál es la distribución posterior de θ?

Caso de estudio: Tasas de mortalidad

El INEGI publica para cada año los registros de fallecimiento junto con la causa principal de muerte. En esta sección utilizaremos los modelos descritos anteriormente para inferir tasa de fallecimiento por Neumonía para cada uno de los municipios/delegaciones del país. Contamos con los últimos 5 años de los registros de defunción.

Carga y preparación de datos

Pregunta 4) Empecemos explorando los datos. Carga los datos para un año que elijas. Encontrarás en los archivos en `datos/poblacion/defunciones/<año>` los registros de defunciones por Neumonía para el `<año>` que escojas.

Pregunta 5) De igual forma, carga los datos de población que encontrarás en `datos_examen/poblacion/demograficos`. Por el momento, no necesitamos los grupos de edad (aunque despúes los utilizaremos). Por ahora escribe el codigo necesario para calcular el tamaño de la población en cada uno de los municipios.

Pregunta 6) Ahora necesitamos *cruzar* las tablas de defunciones y población para crear una tabla con ambos registros. Para esto necesitarás la función `dplyr::full_join`.

Pregunta 7) Con esto tendrás conocimiento de cómo cargar la información relevante (número de defunciones y población total en cada municipio). Sin embargo, tenemos información para las defunciones de los últimos 5 años. Carga la información que encontrarás en `/defunciones/` y agrupa de tal forma que tengas una tabla como la anterior. **Importante:** Para fines de este proyecto no necesitamos los conteos por año, sólo el agrupado. Es decir, el número de defunciones totales de los 5 años por municipio.

Cáclulo de estadístico de interés

Lo que nos interesa en particular son las tasas de mortalidad anual en los municipios del país. Para esto utilizaremos el modelo Poisson que vimos en la primera parte. Si denotamos por y_i el número total de defunciones por neumonia en el i-ésimo municipio; θ_i, la tasa de mortalidad por individuo por año, entonces

$$y_i \,|\, n_i, \theta_i \sim \text{Poisson}(\lambda_i),$$

donde n_i denota la población total del municipio i-ésimo y λ_i la tasa con la que ocurren las muertes por neumonía en el periodo observado para la población del municipio i-ésimo.

Pregunta 8) ¿Cómo escribirías λ_i en función de θ_i?

Ahora, utilizaremos un mapa para ver si podemos observar algún patrón en las tasas de mortalidad por individuo por año θ_i. Por ejemplo, podríamos esperar que algunas zonas del país concentren las tasas mas altas. Por ejemplo, podemos crear mapas con los municipios con las tasas mas bajas y altas. Digamos que sólo queremos ver el 25% mas bajo y alto. Los mapas los obtenemos con las funciones `mxmaps::mxmunicipio_choropleth`.

La estructura que necesita esta función es una tabla con una columna que se llame `region` donde venga el codigo identificador del municipio. Por ejemplo, para el municipio `001` en el estado `24` el codigo de region será `24001`. Otra columna necesaria es el valor con el que "coloreará" el municipio en el mapa y se tiene que llamar `value` y puede ser una variable `Boolean` o `double`.

Pista. Para este punto, podrías necesitar la función `dplyr::row_number`. De igual forma podrías ocupar una indicadora para decir cuáles son los municipios con las tasas mas altas y cuáles son los que tienen las mas bajas. Al final, podrías presentar esto como dos mapas separados.

¿Qué observas? No hay patrón tan claro. Especialmente si observamos lo que sucede en Chihuahua, Durango y Coahuila, donde tenemos municipios de ambas categorías. ¿Cómo puede ser que un mismo estado tenga las tasas mas altas y bajas al mismo tiempo?

¡El problema es el tamaño de muestra! Considera un municipio de 1,000 habitantes. Muy probablemente en 5 años no veamos una muerte por neumonía, lo cual convertiría la tasa observada en 0. Sin embargo, si ocurriera una muerte entonces la tasa sería de 1/5,000 por año, lo cual sería muy elevado con respecto a otros municipios con poblaciones grandes y mayor número de casos.

Inferencia Bayesiana para tasa de mortalidad

Utilizaremos inferencia Bayesiana para regularizar el problema. Seguiremos suponiendo que

$$y_i \,|\, n_i, \theta_i \sim \mathsf{Poisson}(\lambda_i),$$

pero ahora necesitamos una distribución previa para θ_i. Sabemos, por lo anterior, que el modelo Poisson-Gamma es conjugado. Por lo tanto requerimos una distribución Gamma. Sólo falta elicitar los hiperparámetros.

No todos somos expertos en salud ni tenemos conocimiento previo. Sin embargo, podemos visitar esta página para darnos una idea de las tasas de mortalidad por neumonía en el resto del mundo.

A continuación se muestra las tasas de mortalidad para los últimos años para algunos paises y la region de America Latina y el Caribe.

Considera los siguientes puntos:

- Las tasas anteriores han sido calculadas con un método que incorpora la estructura demográfica de cada país y la estandariza con respecto a la pirámide poblacional mundial. En nuestro ejemplo, nuestras tasas no serán ajustada de tal forma (este método se conoce en inglés como *age-standarized mortality rates*).
- Las tasas reportadas tienen una base distinta, pues son reportadas con respecto a una población de 100,000 habitantes. Es decir, son tasas de mortalidad anuales para poblaciones de 100K habitantes. Por ejemplo, un valor de 5 significa que en promedio 5 habitantes por cada 100K mueren de neumonia al año.

Pregunta 9) Con esto en mente, escribe los límites necesarios para encontrar una distribución Gamma adecuada. Encuentra la solución al sistema de ecuaciones no lineales por medio de la función `nleqslv::nleqslv`. Escribe tu razonamiento para seleccionar dichos valores.

```
limits <- # escribe los intervalos adecuados aqui

gamma.limits <- function(x){
  # reparametrizamos para que el problema sea mas "fácil" en términos numéricos.
  log_alpha <- x[1]
  log_beta  <- x[2]

  # definimos las cotas de probabilidad
  p_cota <- # define un valor adecuada
  c(    pgamma(limits[1], exp(log_alpha), rate = exp(log_beta)) - p_cota,
    1 - pgamma(limits[2], exp(log_alpha), rate = exp(log_beta)) - p_cota)
}

initial_guess <- c(log(1), log(1))
```

```
results <- nleqslv(initial_guess, gamma.limits)

params.prior <- exp(results$x)
```

Pregunta 10) Grafica los histogramas de una variable aleatoria Gamma con los valores iniciales para el problema de optimización y con los finales de dicho algorimo. Esto te servirá de verificación que el método funciona adecuadamente.

Pregunta 11) ¿Cómo se compara la distribución a priori con las tasas observadas en los municipios? Puedes utilizar histogramas para estas comparaciones. Por otro lado, no te preocupes si no se ven identicas. El punto es ver que nuestras creencias iniciales se ven coherentes.

Pregunta 12) Utiliza un gráfico de dispersión para comparar las tasas observadas contra la población del municipio. ¿Qué observas? Utiliza los ejes en escala logaritmica. Para esto checa la función: `ggplot2::scale_x_log10` y `ggplot2::scale_y_log10`

Pregunta 13) Ahora usaremos la distribución predictiva **previa** para explorar los posibles valores que tendrían los casos de muerte bajo nuestro modelo para municipios de distintos tamaños. Para este punto considera que la predictiva es una mezcla de Poisson con Gamma, como se expresa en

$$p(y) = \int \mathsf{Poisson}(y|n,\theta)\,\mathsf{Gamma}(\theta|\alpha,\beta)\,\mathrm{d}\theta,$$

o bien, la forma en especifico de la predictiva preevia. ¡Esto ya lo has resuelto en la primera parte del examen!

Usa histogramas para ver los números de muertes en municipios hipotéticos de tamaño $n = 10^3, 10^4, 10^5$.

Pregunta 14) Calcula los valores posteriores de las tasas de mortalidad bajo nuestro modelo bayesiano y compara con los estimadores de máxima verosimilitud. Para esto puedes utilizar un gráfico de dispersión cómo el visto en clase o los anteriores. ¿Observas regularización en nuestras estimaciones? ¿Qué observas si haces un grafico como el de la pregunta 12?

Pregunta 15) Utiliza la distribución predictiva *posterior* para verificar el ajuste del modelo. Para esto, escoge tres municipios al azar de distintos tamaños (chico, mediano, grande) y haz un *lineup* para cada uno para observar si las predicciones posteriores son consistentes con los datos.

Incorporando Grupos de Edad

Se sabe que las muertes por neumonia no son uniformes y las tasas de mortalidad son mas altas en niños y personas mayores. Ahora realizaremos el mismo análisis considerando grupos de edad. Para esto ampliaremos nuestro modelo

$$y_{k,i} \,|\, n_{k,i}, \theta_{k,i} \sim \mathsf{Poisson}(\lambda_{k,i}),$$

donde utilizamos el sub-indice k, i para denotar el k-ésimo grupo de edad en el i-ésimo municipio.

Pregunta 16) Genera histogramas para cada grupo de edad y discute si el supuesto anterior esta soportado por los datos. Para esto calcula las tasas de mortalidad adecuadas. Auxiliate de `ggplot2::facet_wrap`.

Por motivos de simplicidad usaremos la misma distribución previa para cada tasa de mortalidad asociada a grupos de edad y municipio que en los puntos anteriores.

Pregunta 17) Utiliza graficos de dispersión para determinar si hay efectos de regularización. Las ideas las encuentras arriba en la pregunta 12 y 14.

Pregunta 18) Para uno de los tres municipios que escogiste anteriormente utiliza la distribución predictiva posterior para verificar el ajuste del modelo para los grupos de edad: $[0, 3), [18, 25)$ y $[64, \infty)$. Puedes hacer un *lineup*.

Conclusiones:

¡El último modelo (edad-municipio) incorpora alrededor de 17K parámetros distintos! Sin duda, no es parsimonioso. De hecho, este modelo representa el extremo en complejidad para esta situación. Podemos incorporar una estructura jerárquica donde podemos interpretar una estructura multi-nivel en cuanto al conocimiento que podemos generar. Esto es por que en la estuctura de dependencia dejamos la misma distribucion previa sin importar municipio o grupo de edad. En cursos posteriores exploraremos estas opciones. Pero ahora, ¡a descansar!