Apunte Problemas Resueltos

Marcel Goic

2025-11-11

Contents

Prologo **5**

1 Formulas utiles **7**

2 Modelos de Regresion **11**

2.1 Introducción 11

2.2 Metodología 11

2.3 Problemas Teóricos 12

2.4 Problemas Aplicados 15

3 Modelos Probabilisticos **53**

3.1 Introducción 53

3.2 Metodología 54

3.3 Problemas Teóricos 54

3.4 Problemas Aplicados 57

4 Modelos Estructurales **87**

4.1 Introducción 87

4.2 Metodología 91

4.3 Problemas Teóricos 92

4.4 Problemas Aplicados 94

Disclaimer: Algunos tildes fueron omitidos intencionalmente debido al formato de publicación

Prologo

Dentro del contexto de Marketing, estamos interesados ene estudiar el comportamiento de las personas, de modo de comprenderlo y finalmente, tomar decisiones estratégicas en función de los aprendizajes adquiridos.

En esta línea, se pueden definir dos tipos de enfoques a utilizar, los cuales dependen de los supuestos que tomemos acerca del comportamiento de los agentes (tomadores de decisiones):

- *Modelos Probabilísticos*: En este, se asume que los agentes se comportan en base a decisiones aleatorias. Usualmente se utiliza cuando se tiene información reducida y/o agregada respecto al comportamiento de los agentes de estudio.
- *Enfoque Estructural*: En este, se asume que los agentes se comportan de manera racional, lo cual se traduce a que toman decisiones con el objetivo de maximizar sus utilidades (no necesariamente monetarias)

Chapter 1

Formulas utiles

- Modelo geométrico desplazado:

$$P(T = t|\theta) = \theta(1-\theta)^{t-1}$$

$$P(T > t|\theta) = (1-\theta)^{t}$$

- Modelo beta geométrico desplazado:

$$\mathbb{P}(T = t \mid \alpha, \beta) = \frac{B(\alpha+1, t+\beta-1)}{B(\alpha,\beta)}$$

$$\mathbb{P}(T > t \mid \alpha, \beta) = \frac{B(\alpha, t+\beta)}{B(\alpha,\beta)}$$

- Modelo de Duración Exponencial:

$$P(T \leq t) = 1 - e^{-\lambda t}$$

$$P(T > t) = e^{-\lambda t}$$

- Modelo Gamma Exponencial:

$$\mathbb{P}(T \leq t) = 1 - \left(\frac{\alpha}{\alpha+t}\right)^{r}$$

- Modelo de Duración Weibull:

$$P(T \leq t) = 1 - e^{-\lambda t^c}$$

$$\mathbb{P}(T \leq t \mid \alpha, r, c) = 1 - \left(\frac{\alpha}{\alpha + t^c}\right)^r$$

- Modelo de Conteo:

$$P(N_t = m|\lambda) = \frac{(\lambda t)^m e^{-\lambda t}}{m!}$$

- Modelo Gamma Poisson:

$$\mathbb{P}(N_t = m \mid r, \alpha) = \left(\frac{\alpha}{\alpha + t}\right)^r \left(\frac{t}{\alpha + t}\right)^m \frac{\Gamma(r + m)}{\Gamma(r)m!}$$

- Modelo de elección binomial:

$$P(X_s = x_s|m_s, p_s) = \binom{m_s}{x_s} p_s^{x_s} (1 - p_s)^{m_s - x_s}$$

- Modelo Beta Binomial

$$\mathbb{P}(X_s = x_s \mid \alpha, \beta) = \binom{m_s}{x_s} \frac{B(\alpha + x_s, \beta + m_s - x_s)}{B(\alpha, \beta)}$$

- Esperanzas condicionales:

Modelo	Distribución Incondicional	Distribución Condicional	Esperanza Condicional
Modelo de Tiempo Discreto	$\theta \sim \text{Beta}(\alpha, \beta)$	$\theta \mid t \sim$ $\text{Beta}(\alpha + 1,\ \beta + t)$	$\mathbb{E}[\theta \mid t] = \frac{\alpha + 1}{\alpha + \beta + t}$
Modelo de Tiempo Continuo	$\lambda \sim$ $\text{Gamma}(r, \alpha)$	$\lambda \mid t \sim$ $\text{Gamma}(r + 1, \alpha + t)$	$\mathbb{E}[\lambda \mid t] = \frac{r + 1}{\alpha + t}$
Modelo de Conteo	$\lambda \sim$ $\text{Gamma}(r, \alpha)$	$\lambda \mid x, t \sim$ $\text{Gamma}(r + x,\ \alpha + t)$	$\mathbb{E}[\lambda \mid x, t] = \frac{r + x}{\alpha + t}$

Modelo	Distribución Incondicional	Distribución Condicional	Esperanza Condicional
Modelo de Elección Binaria	$\theta \sim \text{Beta}(\alpha, \beta)$	$\theta \mid x \sim$ $\text{Beta}(\alpha+x,\ \beta+m-x)$	$\mathbb{E}[\theta \mid x] = \dfrac{\alpha + x}{\alpha + \beta + m}$

- Distribución *Beta*:

$$f(x|\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \qquad \mathbb{E}(X) = \frac{\alpha}{\alpha+\beta} \qquad \mathbb{V}ar(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

- Distribución *Gamma*:

$$f(x|r,\alpha) = \frac{\alpha^r x^{r-1} e^{-\alpha x}}{\Gamma(r)} \qquad \mathbb{E}(X) = \frac{r}{\alpha} \qquad \mathbb{V}ar(X) = \frac{r}{\alpha^2}$$

- Función *Gamma*:

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt \qquad \Gamma(z+1) = z\Gamma(z)$$

- Función *Beta*:

$$B(\alpha,\beta) = \int_0^\infty t^{\alpha-1}(1-t)^{\beta-1} dt \qquad B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

- Tasa de Riesgo:

$$h(t) = \frac{f(t)}{1-F(t)} \qquad F(t) = 1 - exp(-\int_0^t h(u)du)$$

- Métricas de ajuste

Métrica de Error	Métrica de Ajuste / Criterio de Información
$MAE = \frac{1}{n}\sum_{i=1}^{n} \lvert y_i - \widehat{y}_i \rvert$	$MAPE = \frac{100}{n}\sum_{i=1}^{n} \left\lvert \frac{y_i - \widehat{y}_i}{y_i} \right\rvert$

Métrica de Error	Métrica de Ajuste / Criterio de Información
$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i}$	$LR = 2(LL_{\mathrm{NR}} - LL_{\mathrm{R}}) \sim \chi^2$
	$AIC = -2\ln(L(\widehat{\theta})) + 2k$
	$BIC = -2\ln(L(\widehat{\theta})) + k\ln(n)$

Chapter 2

Modelos de Regresion

2.1 Introducción

En este capítulo se abarcarán problemas relacionados a modelos de regresión; es decir, aquellos modelos donde la variable dependiente y se escribe como:

$$y = f(\mathbf{X}) = \beta^{'} X + \varepsilon$$

donde $\mathbf{X}$ corresponde a las variables explicativas, $\beta^{'}$ son los parámetros que buscamos para estudiar la influencia de las vairables explicativas en la variable dependiente; o mejor dicho, la influencia de las variaciones en las variables explicativas en las variaciones de la variable depndiente, y ε es el error aleatorio.

Los modelos de regresión tienen la ventaja de ser fáciles de implementar e interpretar, pero se debe tener cuidado al momento de construir, evaluar e interpretar estos modelos. A partir de los datos que uno observa, existen infinitos modelos de regresión para explicar una variable y. Un buen modelo encuentra el balance entre su complejidad y poder explicativo.

2.2 Metodología

Como mencionabámos anteriormente, para poder crear y utilizar los modelos de regresión que nos permite responder de la manera más informada las preguntas de:

1. Definir el nivel de agregación de las variables apropiado para el problema en cuestión, ya sea una agregación temporal (datos diarios, semanales, mensuales, anuales) o lógicos (agregación por producto, categoría, marca, tienda), como también definir los índices de los parámetros del modelo

2. Elegir las variables explicativas a considerar para el modelo, realizar las transformaciones según lo solicitado por el problema de gestión.

3. Decidir si se incluirán interacciones entre las variables explicativas, como también si posee un efecto fijo (como también definir los índices de estos efectos).

4. Evaluar los modelos a partir de distintas métricas, ya sean genéricas (R^2, MAE, $MAPE$, RMS) o de verosimilitud (Likelihood Ratio, AIC, BIC), escogiendo aquél que, a partir de las métricas mencionadas, tengan una buena calibración y una alta capacidad de predicción.

5. Interpretar los coeficientes más significativas de un modelo, para responder la pregunta de gestión original.

2.3 Problemas Teóricos

2.3.1 E1: Signo esperado y significancia

Al evaluar la inclusión de una variable en base a su signo esperado y significancia estadística, ¿cuál de los siguientes casos NO es una buena recomendación para un modelo de pronóstico?

a. Si el parámetro tiene el signo esperado y no es significativo: mantener.
b. Si el parámetro no tiene el signo esperado y no es significativo: eliminar.
c. Si el parámetro no tiene el signo esperado y es significativo: tratar de hacer ajustes al modelo o juntar más datos y variables.
d. El parámetro tiene el signo esperado y es significativo: mantener.
e. Ninguna de las anteriores.

2.3.2 E2: Regresión Lineal vs de Poisson

¿Qué diferencia un modelo de regresión simple de un modelo de regresión de Poisson? (Puede seleccionar más de una opción).

i. La especificación de la distribución del término de error.
ii. Los métodos disponibles para estimar el modelo.
iii. La naturaleza de la variable dependiente.
iv. La posibilidad de realizar t-tests.

2.3.3 E3: Selección automática de variables

Cuál de las siguientes afirmaciones caracterizan adecuadamente a la selección automática de variables para un modelo de regresión lineal (Puede seleccionar más de una opción):

i. Vienen predefinidas en planillas de cálculo.
ii. Son siempre peores que selección manual.
iii. Sólo pueden aplicarse sobre el subconjunto de variables continuas.
iv. No consideran todos los modelos posibles.

2.3.4 E4: Regresión Log-Nivel y Log-Log

Dado el modelo $ln(y) = \beta_0 + \beta_1 x_1 + b_2 ln(x_2)$, se puede asegurar que:

a. Un aumento de una unidad en x_1 causa un aumento de β_0 unidades en y.
b. Un aumento de una unidad en x_2 causa un aumento de $\beta_2\%$ unidades en y.
c. Un aumento de un 1% en x_2 causa un aumento de β_2 unidades en y.
d. Un aumento de un 1% en x_2 causa un aumento de $\beta_2\%$ unidades en y.
e. Ninguna de las anteriores.

2.3.5 E5: Capturar efectos

¿Cuál de las siguientes especificaciones NO permite capturar que las diferencias en el comportamiento y se van incrementando a medida que aumenta la edad de los sujetos?

a. $y = \beta_0 + \sum_i \beta_i Dummy_i, \quad i = \{\text{niños, jóvenes, adultos, ancianos}\}$
b. $y = \beta_0 + \beta_1 ln(edad)$
c. $y = \beta_0 + \beta_1 edad + \beta_2 edad^2$
d. $y = \beta_0 + \beta_1 \exp(edad)$
e. Ninguna de las anteriores.

2.3.6 E6: Coeficientes y Parámetros

Se tiene el siguiente modelo lineal de venta de productos

$$Q_{ist} = \theta_s + \mu_{is} P_{ist} + \gamma t + \epsilon_{ist}$$

Donde Q_{ist} es la cantidad vendida del producto i en la tienda s en el año t y P_{ist} el precio. El set de datos contiene registros para 5 productos y 5 tiendas entre 1960 y 2015. A partir de esto, se puede afirmar que (puede elegir más de una opción):

i. Transformar γt en γ_t no influye en el modelo ya que son equivalentes.
ii. Transformar los coeficientes μ_{is} a $\mu_i + \mu_s$ implica calcular menos parámetros.
iii. El modelo no permite capturar comportamiento cíclico de la demanda.
iv. Los parámetros θ_s y μ_{is} no pueden estimarse simultáneamente.

2.3.7 E7: ¿Qué es un modelo de Regresión?

¿Cuál(es) es(son) la utilidad de un modelo de regresión?

i. Permite medir correlación entre variables.
ii. Permite medir la magnitud de una relación causal.
iii. Permite determinar causalidad entre variables.
iv. Dependiendo de la especificación puede determinar causalidad o correlación, pero no ambas simultáneamente.

2.3.8 Solución problemas teóricos

E1: Signo esperado y Significancia
Respuesta correcta: e.
{ Ninguna de las anteriores. Por ejemplo: no tiene signo esperado y no es significativo: Mantener}

E2: Regresión Lineal vs de Poisson
Respuestas correctas: i., ii. y iii.
{ La especificación de la distribución del término de error; Los métodos disponibles para estimar el modelo; La naturaleza de la variable dependiente }

E3: Selección automática de variables
Respuesta correcta: iv.
{ No consideran todos los modelos posibles }

E4: Regresión Log-Nivel y Log-Log
Respuesta correcta: e.
{ Ninguna de las anteriores. Un 1% de aumento en x_2 causa un cambio en $\beta_2/100$ en y.}

E5: Capturar efectos
Respuesta correcta: b.
{ $y = \beta_0 + \beta_1 \ln(edad)$ }

E6: Coeficientes y Parámetros
Respuestas correctas: ii. y iii.
{ Transformar los coeficientes μ_{is} a $\mu_i + \mu_s$ implica calcular menos parámetros; El modelo no permite capturar comportamiento cíclico de la demanda }

E7: ¿Qué es un modelo de Regresión?
Respuestas correctas: i. y ii.
{ Permite medir correlación entre variables; Permite medir la magnitud de una relación causal }

2.4 Problemas Aplicados

2.4.1 Problema 1

Un reconocido grupo empresarial que manufactura productos de aseo y limpieza de diversa índole le pide a usted evaluar el impacto de la pandemia en sus ventas a partir del registro obtenido por distintas tiendas distribuidas a lo largo del país.

El foco del análisis se en la variable q_{ist}, la cual corresponde al número de unidades vendidas del producto i en la tienda s en el día t. Junto a las ventas se cuenta con un vector de características z_t que describe cada día t, entre ellas FDS_t que indica si es un fin de semana y FER_t que indica si el día fue feriado. También, se tiene el vector x_{ist} con características que varían según producto, tienda y día, entre ellas $PRECIO_{ist}$ (precio del producto), $QUAR_{st}$ (si la comunda donde se encuentra la tienda estaba en caurentena o no).

1. Formule un modelo de regresión lineal que capture los siguientes factores:

    a. Existe un efecto fijo por producto y por tienda, es decir, la linea base de las ventas es distinto para cada producto dependiendo de la tienda.
    b. Si un producto no se vende en una tienda en un día dado es probable que aya un quiebre de stock y si un día hay quiebre, es probable que el día siguiente el producto siga quebrado.
    c. El conjunto $\mathcal{PN}$ registra todos los productos de la firma que son de primera necesidad. Las cuarentenass no tienen efecto en os productos de primera necesidad.
    d. El efecto de las cuarentenas puede ser distinto para las ventas de un día de semana con respecto a un fin de semana.
    e. El nivel de progresión de las infecciones ($InfMV_{st}$) afecta el nivel de ventas.
    f. El nivel de progresión de las infecciones afecta el nivel de ventas a través de cuñanto han variado las infecciones con respecto a una semana. Por ejemplo, si hoy hay 100 casos nuevos y hace una semana había solo 800, entonces ese aumento de 200 casos capturarían una fracción relevante de la relación entre la progresión de la pandemia y als ventas de los productos de la firma.

2. Considere los siguientes dos modelos, los cuales contienen variables mencionadas anteriormente, junto con otras que no vale la pena nombrar:

	M1	M2
...	...	...
Quar	-0.06 (0.02)	-0.07 (0.01)
FER	0.11 (0.03)	0.12 (0.02)
FDS	0.14 (0.02)	0.13 (0.02)
Quar x FER	0.07 (0.02)	0.06 (0.03)
Quar x FDS	0.06 (0.03)	0.08 (0.01)
Npar	23	27
R^2	0.63	0.69
$R^2_{adj.}$	0.62	0.61
RMSE train	0.34	0.33
RMSE test	0.34	0.38

Tabla: Fragmento de parámetros y métricas de ajuste para dos modelos alternativos.Para los par

Además de incluir el valor de los coeficientes, se entrega para cada modelo 4 métricas de ajuste. ¿Cuál de los dos modelos recomendaría usar para ver el efecto de las cuarentenas en las ventas?

3. Inspeccionando los resultados del modelo de regresión lineal que consideró en la parte anterior, ¿cómo describiría el rol que han jugado las cuarentenas en las ventas de la firma?

Solución

1. Agregamos progresivamente un término adicional a la regresión:

$$q_{ist} = \alpha_{is} + \beta_1 \cdot \mathbf{1}_{[q_{ist-1}=0]} + \beta_2 \cdot QUAR_{st} \times \mathbf{1}_{[i \notin \mathcal{PN}]} + \beta_3 \cdot QUAR_{st} \times FDS_t + \beta_4 \cdot InfMV_{st} + \beta_5 \cdot (InfMV_{st} -$$

Observación: El modelo anterior puede permitir algunas variantes. Por ejemplo, aunque la condición (a) indique que existe un factor fijo por tienda y producto, una variante (no tan correcta) es considerar por separador los factores. Esto es, en vez de tomar α_{is} se elige $(\alpha_i + \alpha_s)$.

También, al considerar la condición (c) al momento de abordar el factor en (d), se puede decir que el efecto de las cuarentenas en los fines de semana no afecta a los productos de primera necesidad (como lo dice la condición (c)). De este modo, este factor se puede considerar como $\beta_3 \cdot QUAR_{st} \times FDS_t \times \mathbf{1}_{[i \notin \mathcal{PN}]}$.

2. El modelo 1 es mejor que el modelo 2 en casi todas las métricas relevantes, excepto en la métrica de R^2. Esto se explica porque tiene más parámetros y queda corregido al mirar el coeficiente $R^2_{adj.}$.

3. Para el modelo preferido (1) todos los coeficientes asociados a las cuarentenas y sus interacciones son significativos y por lo tanto mirar los estimadores nos dan una razonable primera aproximación. Acá vemos que las cuarentenas tienen un efecto negativo siginificativo en reducir las ventas. Sin embargo, este efecto solo se da en los días de semana. Tanto para feriados como fines de semana el efecto se va a cero los fines de semana y (marginalmente) positivo para feriados.

2.4.2 Problema 2

Un cine pequeño quiere impulsar la venta de entradas para este verano. Para atraer de manera eficiente a los clientes deciden ser selectivos en las películas que tendrán en cartelera. Para ello, desean determinar las características de las películas que provocan un mayor ingreso. El equipo de ventas dispone de un panel de datos con el historial de las 3,800 funciones del año de las 1,250 películas presentadas con los siguientes datos:

Variable	Descripción	Media
I dFunción	Identificador de la función	-
Fecha	Fecha de la función	-
Hora	Hora de la función	-
Película	Nombre de la película a proyectar	-
Duración	Duración de la película presentada [en Minutos]	103.32
Tipo	Tipo de película presentada {Acción, Comedia, Romance, Drama, Terror}	-
Edad	Restricción de edad de la película {TE, TE+7, MA+14, MA+18}	-
Estudio	Estudio que produjo la película {Dis, WB, Para, Uni}	-
Pre supuesto	Presupuesto de la película [en CLP]	160 ,200,000

Variable	Descripción	Media
Ingreso	Ingreso recibido por las entradas vendidas en la función [en CLP]	714,500

Tabla: Resumen de ventas para cada función.

Para tener una mejor idea de cómo se ven los datos, se pueden observar estos 6 registros:

Id$_i$	Función$_i$	Fecha$_i$	Hora$_i$	Película$_i$	Duración$_i$	Tipo$_i$	Edad$_i$	Estudio$_i$	Presupuesto$_i$	Ingreso$_i$
14156	23/07/2022	19:00	Odín 4	105	Acción	TE+7	Dis	354,970,000	948,640
14157	23/07/2022	22:00	Chillido 5	132	Terror	MA+18	Para	135,600,000	804,300
14158	24/07/2022	19:00	Odín 4	105	Acción	TE+7	Dis	354,970,000	800,000
14159	24/07/2022	19:30	Desafiar Barreras	158	Romance	MA+14	WB	94,500,000	769,200
14160	24/07/2022	22:15	Odín 4	105	Acción	TE+7	Dis	354,970,000	802,150
14161	25/07/2022	19:00	Desafiar Barreras	158	Romance	MA+14	WB	94,500,000	752,900

Tabla: Ejemplo de 6 registro en Tabla de Ventas.

1. Nótese que una película puede tener varias funciones y que, debido a la naturaleza de las variables, el rango de valores y el tipo de variables son distintos. Debido a esto, elija un nivel de agregación y transforme las variables de forma que puedan modelar el éxito total en ventas para cada película.

2. El equipo escogió como variable dependiente el logaritmo de la suma de los ingresos de las funciones para cada película y generó dos modelos distintos, entregando los siguientes resultados:

	Modelo 1 Coeficiente	Modelo 1 p-valor	Modelo 2 Coeficiente	Modelo 2 p-valo
Intercepto	-15.2	0.704	3.1	0.161

	Modelo 1 Coeficiente	Modelo 1 p-valor	Modelo 2 Coeficiente	Modelo 2 p-valor
Duración	-1.04	0.452	-0.23	0.004
log(P resupuesto)	0.62	0.005	1.15	0.001
Edad.TE	4.5	0.058	-	-
Edad.TE+7	6.71	0.020	-	-
Edad.MA+14	5.16	0.017	-	-
Edad.MA+18	4.04	0.009	-	-
Tipo.Acción	-	-	4.05	0.520
T ipo.Comedia	-	-	1.63	0.017
T ipo.Romance	-	-	2.11	0.110
Tipo.Drama	-	-	-1.52	0.025
Tipo.Terror	-	-	6.72	0.001
Estudio.Dis	5.12	0.004	-	-
Estudio.WB	4.09	0.240	-	-
Estudio.Uni	3.14	0.104	-	-
E studio.Para	-0.63	0.186	-	-
LL		-644		-650
AIC		1,304		1,295
BIC		1,366		1,350

Tabla: Parámetros y métricas de ajuste para dos modelos alternativos.

A partir de los datos entregados elija cuál modelo es mejor para predecir el número de entradas vendidas para una película y entregue 3 argumentos de porqué lo considera mejor.

3. Interprete los coeficientes de cada modelo. Según cada modelo, ¿qué tipo de películas debería presentar el cine para tener mayores ingresos?

4. Suponga que en su base de datos posee también la variable independiente $NFun$ que representa el número de funciones que recibió la película en el cine. Proponga un modelo de regresión lineal que considere siguientes condiciones:

   a. Se debe considerar el efecto marginal por parte de las variables Presupuesto, Duración, Edad, Tipo, Estudio y $NFun$.

   b. Existe un efecto en las ventas que depende simultáneamente del estudio y el Tipo de película (Por ejemplo, el estudio WB puede tener más exito con películas de terror, mientras que Dis lo puede tener con aquellas de acción).

   c. El éxito de la películas apropiadas para niños (Edad $\in$ {TE, TE+7}) depende de su duración.

Solución

1. Nótese que una película puede tener varias funciones y que, debido a la naturaleza de las variables, el rango de valores y el tipo de variables son distintos. Para evaluar el ingreso total en ventas obtenido por una película, se deben considerar los ingresos a nível película. De modo que se define el ingreso para la película j como la suma del éxito de todas sus funciones i:

$$Ingreso_j = \sum_i Ingreso_i \cdot \mathbf{1}_{[Nombre_i = j]}$$

   Cabe notar que las características de la película $(Presupuesto_i, Duracin_i, Tipo_i, Edad_i, Estudio_i)$ se mantienen para cada función i de la película j de modo que $Presupuesto_j = Presupuesto_i$ para toda función de la película j (lo mismo para las demás características). De este modo, tenemos un vector de características por película $(Presupuesto_j, Duracin_j, Tipo_j, Edad_j, Estudio_j)$.

   Finalmente, notemos que los valores de $Ingreso_j$ y $Presupuesto_j$ son muy elevados en comparación con la variable $Duracin_j$. Similarmente, el rango no se compara con el valor de las variables dummy generadas por las categorías $Tipo_j$, $Edad_j$ y $Estudio_j$. Debido a esto, decidimos estudiar el comportamiento del logaritmo de estos montos: $\log(Ingreso)$ y $\log(Presupuesto_j)$.

   Para ser más claros, la tabla con las transformaciones apropiadas tendría los 6 registros mostrados anteriormente de la siguiente manera:

$Pelcula_j$	$Duracin_j$	\$ Tipo_j\$	\$ Edad_j\$	$Estudio_j$	$log(Pre_j)$	$log(Ing_j)$
Odín 4	105	Acción	TE+7	Dis	8.55	6.43
C hillido 5	132	Terror	MA+18	Para	8.13	5.91
D esafiar B arreras	158	Romance	MA+14	WB	7.97	6.01

2. El modelo que explica de manera más efectiva el comportamiento es el modelo 2. Se pueden considerara varias razones para esto, algunas de estas razones son:

   a. A pesar del modelo 1 tener una mejor log-verosimilitud, el BIC del modelo 2 es mejor que el modelo 1.
   b. La cantidad de parámetros del modelo 1 es mayor (11 vs 8), por lo que podríamos argumentar que preferimos el modelo 2 por su parsimonia (argumento se refuerza mirando la métrica BIC).

c. Una gran parte de los coeficientes del modelo 1 son no significativos a un 5% de confianza, mientras que los coeficientes del modelo 2 tienden a ser <5%
d. La significancia de las variables que ambos modelos comparten poseen un p-valor menor en el modelo 2, de modo que sus intervalos de confianza deberían ser más acotados. Lo anterior, a pesar de algunas variables no ser significativas al 5%, es un hecho positivo respecto al modelo 2.

e. Pensando en la interpretación, el modelo 1 posee algunos comportamientos discutibles: esperaríamos, por ejemplo, que la duración tenga un efecto negativo y significativo en los ingresos. Sin embargo, el modelo la considera no significativa.
f. Los valores de intercepto en el modelo 1 está demasiado desviados en magnitud respecto al resto de coeficientes. Lo anterior podría dar indicios de algún error en cuánto al escalamiento de las variables al momento de calibrar el modelo.

3. Incluye:

- Primero, observamos aquellos coeficientes con mayor significancia estadística. Para eso consideramos aquellos cuyo p-valor sea menor a 0.01.
- Para el modelo 1 se tienen las variables $\log(Presupuesto)$, $Edad.MA + 18$ y $Estudio.Dis$. De acuerdo al signo de los coeficientes, podemos concluir que las películas que generan un mayor ingreso son aquellas con un mayor presupuesto, para mayores de 18 años producidas por Dis.
- En cambio, para el modelo 2, las variables a considerar son $Duracin$, $\log(Presupuesto)$ y $Tipo.Terror$. Por el signo de los coeficientes, se deben proyectar películas de terror cortas con un alto presupuesto.
- **Observación**: En el modelo 1, sería deseable considerar la variable $Duracin$. Sin embargo, su p-valor alto refleja que no es una variable significativa en este modelo.

4. Incluye:

a. La condición (a) muestra que se deben incluir las variables Presupuesto, Duración, Edad, Género, Estudio y Nfun en el modelo, con sus propios coeficientes.
b. La condición (b) nos dice que se debe agregar además, las variables de estudio condicionada al género.
c. Finalmente, la condición (c) dice que queremos conocer el impacto de la duración en las películas cuya variable Edad es TE y TE+7.

Recordemos que, para variables categóricas, se deben generar variables *dummy* para cada categoría. Sea $E = \{TE, TE + 7, MA +$

$14, MA+18\}$, \$T = {Acción, Comedia, Romance, Drama, Terror} \$ y $S = \{Dis, WB, Uni, Para\}$. A partir de estos conjuntos, podemos definir las variables *dummy*:

- $Edad_{j,e} = 1$ si para la película j, $Edad_j = e$ para $e \in E$.
- $Tipo_{j,t} = 1$ si para la película j, $Tipo_j = t$ para $t \in T$.
- $Tipo_{j,s} = 1$ si para la película j, $Estudio_j = s$ para $s \in S$.

De este modo, el modelo que cumple las condiciones mencionadas sería:

$$\begin{aligned}\log(Ingreso_j) \;=\; & \beta_1 \log(Presupuesto_j) + \beta_2 Duracin_j + \sum_{e\in E} \beta_e Edad_{j,e} \\ & + \sum_{t\in T} \beta_t Tipo_{j,t} + \sum_{s\in S} \beta_s Estudio_{j,S} + \beta_3 Nfun_j \\ & + \sum_{t\in T}\sum_{s\in S} \gamma_{t,s} \left(Tipo_{j,t} \times Estudio_{j,s}\right) \\ & + \sum_{e\in\{TE,TE+7\}} \delta_e \left(Edad_{j,e} \times Duracin_j\right)\end{aligned}$$

Se puede realizar una notación vectorial de los coeficientes para tener una notación más concisa, pero se debe tener en cuenta los índices de estos coeficientes, de modo que el largo del vector sea la cantidad de coeficientes a buscar.

2.4.3 Problema 3

Debido a la gran variedad de alternativas que tiene el consumidor para satisfacer sus necesidades, empresas de distintos sectores han buscado formas de fidelizar a sus clientes mediante programas de recompensas. Estos programas proponen premiar al consumidor con relación a su nivel de actividad, con el fin de que estos sean más leales a la firma. A continuación, nos proponemos estudiar el comportamiento de los consumidores que están inmersos en un programa de recompensas en el cual se acumulan puntos al comprar en alguna de las tiendas de la cadena, al comprar en alguna alianza comercial asociada, o utilizando la tarjeta del banco aliado a la cadena. Cuando el consumidor tiene una cantidad suficiente de puntos acumulados puede canjear algún producto que se ofrece en un catálogo, o puede hacer uso de sus puntos como medio de pago al momento de comprar cualquier producto disponible en tienda.

Para modelar los comportamientos anteriores, se cuenta con información demográfica de cada consumidor que es actualmente parte del programa, y cada una de las transacciones que ha realizado desde su inicio, como se observa en las Tablas 1 y 2:

Variable	Descripción
IDCliente	Identificador del cliente
Ingreso	Ingreso del cliente
AñoInicio	Año en el que el cliente se une al programa
TamañoHogar	Número de personas que componen el hogar del cliente
Género	1 si el cliente es hombre
Edad	Edad del cliente
Categoría	Silver, Gold o Premium

Tabla: Variables Demográficas

Variable	Descripción
IDCliente	Identificador del cliente
Año	Año de la observación
Mes	Mes de la observación
Acreditación	Cantidad de puntos acreditados por el cliente en el mes t
Canje	Cantidad de productos canjeados por el cliente en el mes t
TipoAcreditación	Tiendas Propias, Comercios Asociados o Tarjeta Bancaria
TipoCanje	Producto o Descuento
Promoción	1 si hubo promo o descuentos adicionales en el mes t
Mail	1 si se avisó estado de cuenta al cliente via mail en el mes t

Tabla: Variables Transaccionales

1. Como primera aproximación, se propone modelar la acreditación (A_{it}) y canje (C_{it}) de cada cliente i en el mes t como:

$$A_{it} = \alpha + \alpha_i + \alpha_t + \beta X_{it} + \cdots + \epsilon_{it} \tag{2.1}$$
$$C_{it} = \gamma + \gamma_i + \gamma_t + \rho Z_{it} + \cdots + \xi_{it} \tag{2.2}$$

   Donde X_{it} y Z_{it} son covariables construidas a partir de la información disponible.

   a. ¿Cuál es la interpretación de los parámetros α_i, α_t, γ_i y γ_t?
   b. ¿Son los modelos anteriores identificables? En caso de ser su respuesta negativa, ¿cómo se corrige el problema?

2. Suponga que se han identificado los modelos anteriores (de requerirlo, con las restricciones necesarias para alcanzar identificación). Derive una expresión que permita calcular el número esperado del saldo de punto de un cliente i en cualquier período de tiempo t.

3. Proponga especificaciones de los modelos de regresión que permitan capturar que:

    a. El impacto que tiene informar el estado de cuenta de los puntos depende del género.
    b. Los niveles de actividad (acreditación y canje) de un cliente varían significativamente de acuerdo a su categoría.
    c. El nivel de actividad (acreditación y canje) depende de la antiguedad de los clientes en el programa, pero aquellos que llevan más años tienden a tener un comportamiento similar.

4. Suponga ahora que se han estimado tres especificaciones para entender el comportamiento de acreditación (A_{it}), donde f representa alguna transformación o función de la información disponible, y cuyos resultados se encuentra en la tabla siguiente:

Variable	Modelo 1	Modelo 2	Modelo 3
α_0	35.9 (12.8)	37.1 (8.6)	32.7 (11.5)
f(Ingreso)	12.7 (5.1)	–	10.7 (2.9)
f(Edad)	−1.5 (0.6)	−1.2 (2.3)	−0.6 (0.3)
f(Mail)	–	−0.3 (0.2)	−1.9 (5.1)
f(Promoción)	–	15.1 (6.7)	10.9 (4.2)
\$ f\$(TiendasPropias)	–	10.1 (3.7)	8.9 (2.1)
f(C omerciosAsociados)	–	3.6 (2.5)	2.9 (2.3)
R^2 Ajustado	0.34	0.53	0.55
MAPE	45.5%	20.9%	23.4%

Tabla: Resultados Modelos Regresión

    a. ¿Qué variables ayudan a explicar el comportamiento de acreditación de los consumidores?
    b. ¿Cuál es el medio en que se acreditan o acumulan, en promedio, menos puntos?
    c. Justifique cuál de los modelos anteriores es mejor.

Solución

1. Incluye:

    a. Los parámetros α_i y γ_i representan el efecto fijo de cada consumidor, o el nivel promedio de actividad en creditación y canje, respectivamente, independiente de las demás covariables. Los parámetros α_t y γ_t representan un efecto fijo temporal en los niveles de acreditación

y canje, respectivamente, cuya interpretación dependerá de la definición de las variable que acompañe (si las covariables son meses representaría estacionalidad, mientras que si es fecha o año representaría alguna tendencia en los nvieles anteriores).

b. Al estar las constantes α y γ, junto con los parámetros anteriores, se produce un problema de colinealidad y por lo tanto los parámetros no son identificables. Para corregirlo, se debe omitir α y γ, u omitir α_i y γ_i para un individuo y especificar dummies desde $i = 1$ hasta $i = I - 1$.

2. Sabemos que $\mathbb{E}(Y|X) = X'\hat{\beta}_{OLS}$. Luego, el balance promedio se obtiene como:

$$\mathbb{E}(\sum_{\tau=1}^{t} A_\tau - \sum_{\tau=1}^{t} C_\tau|X, Z) = \sum_{\tau=1}^{t}(\alpha_i + \alpha_t + X'\beta) - \sum_{\tau=1}^{t}(\gamma_i + \gamma_t + Z'\rho) + \epsilon_i$$

3. Incluye:

a. Se debe incorporar una interacción entre las variables de *Mail* y *Género*.

b. Se debe transformar la variable categórica en dummies y omitir uno de los niveles para no tener variables colineales.

c. Se debe construir la variable logaritmo de antigüedad como $ln(Ao_t - AoIngreso_t)$

$$X'\beta = \beta_1 Mail_i Genero_i + \beta_2 1_{[categoria_i = Gold]} + \beta_3 1_{[categoria_i = Premium]} + \beta_4 ln(Ao_t - AoIngreso_t) + \epsilon_i'$$

4. Detalle:

a. Las variables que son significativas independiente del modelo en el que fueron estimadas, y que permiten explicar el comportamiento de interés, es decir, *Ingreso*, *Promoción*, y *TiendasPropias*.

b. El parámetro asociado al medio *TarjetaBancaria* fue fijado en cero (para evitar problemas de colinealidad), y como el resto de los parámetros son positivos, es aquel en que menos puntos se acreditan.

c. La respuesta dependerá del criterio de evaluación. Si se quiere evaluar en cuanto a la capacidad de ajuste, el mejor es aquel con mayor R^2 Ajustado (M3), pero si se evalúa en cuanto a la capacidad de pronóstico, se prefiere aquel con menor MAPE (M2).

2.4.4 Problema 4 - P1 Control 1 Primavera 2025

En un contexto de globalización, movilidad laboral y teletrabajo, comparar el costo de vida entre países se ha vuelto una preocupación frecuente para profesionales y estudiantes. Aunque la fracción de ciudadanos chilenos que deciden

estudiar en el extranjero son aún limitadas, hoy en día existen amplias posibilidades tanto desarrollo académico como laboral más allá de nuestras fronteras. Ciertamente, la decisión de emigrar es de alto envolvimiento y requiere evaluar una multitud de factores. A continuación, nos concentraremos en describir el costo del precio de arriendo de propiedades inmobiliarias en distintas ciudades del mundo.

Para estudiar las diferencias de precio de arriendos, un grupo de investigadores construyó una base de datos internacional de precios de propiedades en más de 50 ciudades de América Latina, Norteamérica y Europa. La información fue recolectada haciendo un barrido histórico en un conjunto amplio de portales especializados de corretaje. La base considera los siguientes campos.

- **ID:** Identificador de la propiedad.
- **Ciudad**: ciudad.
- **Pais**: país.
- **Año**: Año en la que se regista el precio de la propiedad.
- **PrecioUSD**: precio publicado de la propiedad en USD.
- M^2: superficie en metros cuadrados.
- **Dorms**: número de dormitorios.
- **Baths**: número de baños.
- **Centro**: distancia (km) al centro de la ciudad.
- **Residencial**: 1 si la propiedad es residencial (0 si es comercial).
- **AñoConstrucción**: año de construcción de la propiedad.
- **CiudadIPC**: ingreso per cápita promedio de la ciudad (USD PPP).
- **CiudadPob**: Población total de la ciudad.
- **CiudadSup**: Superficie total de la ciudad.

El objetivo es comparar precios de propiedades entre países y explorar diferencias sistemáticas entre países, mientras controlamos por factores estructurales de las ciudades y características de las propiedades.

1. (1.0 puntos) Formule un modelo de regresión para describir el log-precio (en USD) de una propiedad en una ciudad dada, en un instante del tiempo y con una lista de característica definida.

   Para esto considere que:

   a) Existe un efecto fijo por país-año, que permite establecer si, condicional en las características de la propiedad y la ciudad dónde está emplazada, ciertos países tienen niveles de precios mayores que otros en un año dado.
   b) La superficie, la antigüedad, el número de dormitorios y el número de baños afectan linealmente en el log-precio, pero la magnitud en que aumenta el precio al variar estas características es distinta dependiendo de si la propiedad es residencial o comercial.

2. (1.0 puntos) Adicional a lo ya expuesto, se considera que la magnitud en que el precio por metro cuadrado aumenta es mayor para ciudades más densamente pobladas. ¿Cómo modificaría el modelo ya propuesto para considerar esta descripción?

3. (1.0 puntos) Considere ahora que las expectativas de variación de precio importan. Indique cómo modificaría el modelo anterior para considerar que el log-precio se ve afectado linealmente por la variación del precio en los dos años anteriores en esa ciudad. Si necesita crear variables adicionales, explicite cómo se calcularían a partir de la base de datos.

4. (1.0 puntos) En todos los modelos anteriores se ha usado como variable dependiente el log-precio en USD. Discuta qué limitaciones tiene esa variable para comparar qué tan abordable es la vivienda entre distintos países y proponga al menos una métrica alternativa que pudiera usarse para la comparación.

5. (1 punto) Considerando que los precios pueden variar año a año, interesa considerar estimadores agregados de varios años. Sea InP_{kt} es el log-precio esperado para una propiedad en el país k en el año t (que se obtiene a través del producto punto entre estimadores máximo-verosímiles y las características promedio de las propiedades del país i en el año t). Explique cómo testearía que los precios de Chile son distintos a los de Argentina considerando los últimos 3 años.

6. (1 punto) Disconformes con la capacidad de pronóstico de los modelos lineales, se postula que algunos modelos sencillos de aprendizaje de máquina podrían generar mejores resultados. Para ello se considera usar un modelo de k-vecinos más cercanos. Defina una métrica de distancia y determine qué si la propiedad B o la propiedad C es más cercana a la propiedad A.

Propiedad	Tiene Piscina	Tiene Calefacción Central	Tiene Esta cionamiento	Ciudad
A	Sí	Sí	Sí	Santiago
B	No	No	Sí	Concepción
C	Sí	Sí	No	Antofagasta

Tabla: Muestra de datos de 3 propiedades de arriendo en Chile, 2025

Solución

1. (1 punto) Indexemos cada precio como P_{ijkt} representando la observación i del país j en la ciudad k en el año t.

- Para el efecto fijo basta incluir un coeficiente α_{ij} asociado a cada combinación país-año. Aunque no es lo que dice el enunciado, se puede dar 0.4 puntos a quienes incluyan un par de coeficientes aditivos $\alpha_i + \alpha_j$.
- Para las otras variables agregamos términos lineales, más la interacción con el tipo de propiedad.

$$\begin{aligned}\ln(P_{ijkt}) \;=\; & \alpha_{it} + \beta_1 M2_{ijkt} + \beta_2 Dorms_{ijkt} + \beta_3 Bath_{ijkt} \\ & + \beta_4 M2_{ijkt} \cdot Residencial_{ijkt} + \beta_5 Dorms_{ijkt} \cdot Residencial_{ijkt} \\ & + \beta_6 Bath_{ijkt} \cdot Residencial_{ijkt} + \varepsilon_{ijkt}\end{aligned}$$

2. (1 punto) Tenemos que redefinir los índices β_1 y β_4 para que dependan de la densidad poblacional.

$$\beta_1 \;\to\; \gamma_1 + \gamma_2 \cdot \frac{CiudadPop_{ijkt}}{CiudadSup_{ijkt}} \qquad \text{y} \qquad \beta_4 \;\to\; \gamma_3 + \gamma_4 \cdot \frac{CiudadPop_{ijkt}}{CiudadSup_{ijkt}}$$

3. (1 punto) Para poder ingresar al modelo tenemos que calcular el precio de la ciudad de cada año.
   Definimos

   $$AVGP_{kt} = \sum_{h=k} P_{ijht},$$

   como el precio promedio de la ciudad k en el año t. Con esto, el modelo puede expandirse agregando un término adicional.

   $$\beta_7 \cdot (AVGP_{kt-1} - AVGP_{kt-2})$$

   También se considera correcto considerar la diferencia de los dos precios en t y $t-1$ respectivamente.

4. (1 punto) Hay al menos dos consideraciones a tener en cuenta, aunque con nombrar cualquiera es suficiente:

   - Los precios pueden estar explicadas por variaciones en el tipo de cambio y no en cambios del costo en las monedas corrientes. En este caso, se podría normalizar por un promedio de tipo de cambio en un horizonte más largo.

- Los precios pueden ser altos en valores absolutos, pero no serlo relativos a los ingresos. En este caso se podría normalizar por los ingresos, i.e usar:

$$\frac{P_{ijkt}}{Ciudad_{ijkt}}.$$

5. (1 punto) Por definición $lnP_{kt} = \theta' \bar{x}_{kt}$, en que $\bar{x}_{it}$ son las características promedio de las propiedades del país k en el año t. De acá se infiere que $\ln P_{kt}$ es una función lineal de los parámetros y por tanto podemos testear directamente si $\theta' \bar{x}_{k=Chile,t} - \theta' \bar{x}_{k=Argentina,t} = 0$ a través de un test F.

6. (1 punto) La clave de la pregunta es notar que las variables categóricas siempre pueden transformarse a numéricas a través de la generación de variables dummies (e.g. $Piscina_{ijkt} = 1$, si la propiedad tiene piscina, 0 en caso contrario). Con esto se puede utilizar cualquier medida de distancia (incluyendo euclidiana), aunque la más simple es una distancia de Manhattan o de Hamming que simplemente cuenta en cuantas dimensiones los puntos tienen distinto valor.

   En este caso:

   - $d(A, B) = 3$ (A y B difieren en tres atributos)
   - $d(A, C) = 1$ (A y C difieren en un atributo)

   por lo que C sería el vecino más cercano a A.

2.4.5 Problema 5 - P1 Control Otoño 2025

Las tecnologías tradicionales de búsqueda basadas en texto han sido durante mucho tiempo uno de los pilares del comercio electrónico, permitiendo a los usuarios ingresar palabras clave para encontrar productos. Sin embargo, en un entorno de comercio digital en rápida evolución, donde el comportamiento del consumidor se orienta hacia experiencias más interactivas y multimedia, están ganando terreno tecnologías de búsqueda más avanzadas. En particular, la búsqueda visual representa un avance innovador, que permite a los consumidores iniciar búsquedas utilizando imágenes en lugar de texto. Al cargar una imagen en la herramienta de búsqueda, los algoritmos más avanzados de aprendizaje profundo analizan sus características visuales para identificar y recuperar productos relevantes. Estos mecanismos sofisticados generan sugerencias amplias, que incluyen productos similares y ofertas relacionadas de otras marcas.

Las plataformas líderes han sido fundamentales en el desarrollo y perfeccionamiento de la búsqueda visual. Pinterest fue pionera en esta tecnología en 2014, permitiendo a los usuarios seleccionar segmentos de una imagen para encontrar contenido visualmente similar. Otros actores relevantes, como Google (con Google Lens) y Bing (con Bing Visual Search), han desarrollado tecnologías similares. Grandes minoristas como Amazon y Aliexpress también han adoptado esta funcionalidad para mejorar la precisión del emparejamiento. La reducción en los costos tecnológicos han facilitado su acceso, permitiendo que proveedores emergentes como Impresee y Visidea ofrezcan soluciones de búsqueda visual fácilmente integrables en sitios de comercio electrónico.

En este problema buscaremos comparar la búsqueda visual con la búsqueda tradicional basada en texto, en términos de comportamiento de clics y conversiones. El análisis empírico se basa en un conjunto de datos que abarca más de 2 millones de búsquedas realizadas por cerca de 50.000 usuarios. Complementamos esta base con etiquetas y métricas de similitud derivadas de avances recientes en visión computacional, lo que nos permite controlar por la complejidad de las imágenes y la similitud entre la imagen de entradas y los productos resultantes. Estadísticas preliminares indican que, en comparación con las búsquedas por texto, las búsquedas por imagen presentan una mayor tasa de clics (CTR), y que los clics se concentran más en los resultados mejor posicionados. Este patrón podría sugerir que quienes buscan por imagen estarían en etapas más avanzadas del proceso de compra, con menor necesidad de explorar alternativas. Sin embargo, al analizar las conversiones, el mayor CTR no parece traducirse en un aumento de ventas, de modo que las búsquedas por texto están asociadas con tasas de conversión más altas.

1. Plantee un modelo de regresión para describir el número de clics del usuario i en el producto j en la instancia de búsqueda k, que permita entender las diferencias dependiendo de si la búsqueda se realiza usando imágenes o texto. Para eso considere que:

    - El *ranking* o posición en el listado de búsqueda afecta en los clics de modo que los productos listados primero tienen muchos más clics que aquellos listados después. Notar que la tasa a la que se degradan los clics con el *ranking* es distinta dependiendo de que la búsqueda sea por imágenes o texto.
    - La complejidad de la búsqueda importa en los clics generados. Para las búsquedas por texto, la complejidad se puede capturar a partir del largo del texto. Para las imágenes se puede medir por el tamaño de la imagen.
    - Cada producto tiene una propensión a clics base distinta. Hay algunos productos populares que tienen a recibir más clics que otros menos populares, independientemente de los inputs de la búsqueda.

2. A partir de sus resultados de regresión, ¿Cómo determinaría (explique con sus palabras) si el número total de clics generado por un producto en la

quinta posición en una búsqueda por imágenes es mayor que los generados por el producto en la quinta posición de una búsqueda por texto?

3. Considere que los modelos de regresión confirman que, aunque haya un efecto en clics, este no se traduce en más conversiones. A la luz de este resultado, parece interesante estudiar si las búsquedas de imágenes quizás generan búsquedas adicionales que generen más conversiones. Esto es, que las imágenes sean usadas en una etapa temprana del embudo de compras, ayudando a encontrar la categoría de productos, las que después se refinarían haciendo búsquedas tradicionales de texto. Formule un modelo de regresión que permita identificar si una búsqueda de imagen aumenta o no el número de búsquedas siguientes. Para esto, suponga que para cada instancia de búsqueda k del usuario i, se puede calcular $NNextSearch_{ik}$, que cuenta el número total de búsquedas hechas por el mismo usuario, dentro de la misma categoría, durante el mismo día de la búsqueda k. Para su modelo de regresión considere que, junto con el tipo de búsqueda, debe controlar por:

   - La categoría de productos, considerando que cada búsqueda se asocia principalmente a una única categoría, y cada categoría de productos tiene una mayor o menor propensión a ser buscadas. Por ejemplo, la categoría *juguetes* en general tiene más búsquedas que la categoría *colchones*.
   - El número de resultados reportados en la búsqueda *NumResult* y el número de categorías distintas reportadas en la búsqueda *NumProdCat*. Considere que tanto *NumResult* como *NumProdCat* ya están calculados en la base de trabajo.
   - El número de búsquedas diarias que en promedio hace el cliente *AvgNumSearch* (que puede suponer conocido).

4. Suponga que, después de un par de iteraciones, se ha determinado que existen dos especificaciones de modelos lineales que funcionan relativamente bien, como indican los resultados de la tabla.

   a) ¿Cómo interpreta los resultados de la variable *NumResult* en el contexto de esta aplicación?

   b) ¿Cómo interpreta el dramático aumento en el valor de R^2_{adj} entre las especificaciones (1) y (2)?

Variable dependiente: ln(NumSubsequentSearch + 1)

Coefficient	**(1)**	**(2)**
IMAGE	-4.595*** (0.012)	-0.375*** (0.007)

Coefficient	**(1)**	**(2)**
NumResult	-0.081*** (0.000)	-0.003*** (0.000)
NumProdCategory	0.031*** (0.001)	0.00138** (0.000588)
LogImageSize	-0.075*** (0.018)	0.001 (0.010)
NumImageObject	0.023* (0.011)	-0.003 (0.007)
NumImageLabel	0.055*** (0.017)	0.003 (0.010)
KeywordLength	-0.018*** (0.000)	-0.001*** (0.000)
AvgNumSearch	1.85E-6*** (6.86E-8)	-1.03E-7** (3.28E-8)
Efecto fijo: User	No	Yes
Efecto fijo: Product Category	No	Yes
Observations	1,715,495	1,715,495
$R^2_{adj.}$	0.19	0.825

Tabla: Resultados de modelos de regresión próximas búsquedas

Solución

1. Detalles de variables:

   - Incluimos una variable con $Ranking_{ijk}$ y su interacción con el tipo de búsqueda. Notar que hay dos tipos de búsqueda de carácter dummy. Para modelarlo, se debe incorporar un coeficiente para cada tipo (Imagen y texto).
   - Hay que incluir el tamaño de la imagen $ImageSize_{ik}$ y el largo del texto $TextLength_{ijk}$.
   - Hay que incluir un efecto fijo por producto α_k.

   Así, si C_{ijk} es el númeroe de clics que recibe el producto j en la búsqueda k del cliente i, el modelo queda descrito como:

$$\begin{aligned} C_{ijk} \;=\; & \alpha_k + \beta_1 \,\text{Image}_{ik} + \beta_2 \,\text{Rank}_{ijk} + \beta_3 \,\text{Image}_{ik}\cdot\text{Rank}_{ijk} \\ & + \beta_4 \,\text{Image}_{ik}\cdot\text{ImageSize}_{ik} + \beta_5 \,(1-\text{Image}_{ik})\cdot\text{TextLength}_{ik} + \varepsilon_{ijk} \end{aligned}$$

2. Se puede utilizar un test F para evaluar que el grupo de variables de $Image$, $Ranking$ y $(1 - Image)$ sean distintos en su conjunto.

3. Si llamamos a CAT_{ihk} a las dummy que toman el valor 1 si la búsqueda k del usuario i está asociada a la categoría h, entonces el modelo de regresión puede escribirse como:

$$NNextSearch_{ik} = (\sum_h \alpha_h CAT_{ihk}) + \beta_1 Image_{ik} + \beta_2 NumResult_{ik} + \beta_3 NumProdCat_{ik} + \beta_4 AvgNumSea$$

4. Detalle:

    a. Los coeficientes son significativamente negativos en ambas especificaciones. En el contexto de la aplicación, esto sugiere que aquellas búsquedas que generan más resultados, están relacionadas a un menor número de búsquedas posteriores.
    b. La diferencia entre las especificaciones (1) y (2) es que la segunda tiene efectos fijos, tanto por usuario como por categoría de producto. Como esta segunda especificación tiene un R^2_{adj} mucho mayor, concluimos que la categoría y el usuario explican gran parte de la variabilidad en el número de búsquedas posteriores.

2.4.6 Problema 6 - P1 Control Primavera 2024

Una de las características más relevantes de los ambientes de negocio contemporáneos, es la disponibilidad de grandes cantidades de información respecto al comportamiento de clientes. Es así como día a día aparecen nuevas fuentes de información que las compañías están constantemente incorporando a sus plataformas de analítica avanzada. *Whileus* es un conglomerado que tiene la representación de varias marcas internacionales con un centenar de tiendas en el país y está interesado en investigar si los datos derivados del uso de dispositivos móviles pueden ayudar a entender el volumen de ventas de las distintas tiendas. Para eso ha firmado un acuerdo de colaboración con una importante compañía de telecomunicaciones que posee datos de tráfico, posición y uso de aplicaciones de varios de millones de clientes en todo el territorio nacional, que podría ser informativa respecto la demanda de las marcas de *Whileus*. El propósito del acuerdo es evaluar, en el plazo de un mes y a través de un prototipo a cargo de la empresa de telecomunicaciones, si la incorporación de estas nuevas variables permite generar aprendizajes valiosos para el estudio de la demanda de las tiendas de *Whileus*.

En la primera semana del piloto y después de una reunión técnica entre los equipos de *Whileus* con la empresa de telecomunicaciones se acuerda extraer los datos de los data lakes de las respectivas compañías a nivel de zona censal. Esto se fundamenta porque la provisión de datos de movilidad de individuos específicos podría poner en entredicho el uso de información privada de los usuarios. Así, luego de consolidar las bases de datos, se han propuesto los siguientes modelos de regresión para describir el número de visitas a la tienda k desde la zona censal i en la semana t q_{ikt}.

$$q_{ikt} = \alpha^1_i + \beta^1_t + \gamma^1 d_{ik} + \theta^1 z_k + \varepsilon^1_{ikt} \tag{1}$$

$$q_{ikt} = \alpha^2_k + \beta^2_t + \gamma^2 d_{ik} + \theta^2 w_i + \varepsilon^2_{ikt} \tag{2}$$

En estas especificaciones, d_{ik} corresponde al logaritmo de la distancia entre la tienda k y el centroide de la zona censal i, mientras que z_k y w_i son un conjunto de características de la tienda k y la zona censal i respectivamente. Para esta primera iteración se ha considerado que z_k incluye la superficie de la tienda, un conjunto de variables binarias para denotar la marca y una última variable binaria para indicar si la tienda se encuentra dentro de un centro comercial o no. Para las variables de la zona censal, se ha considerado que w_i incluye el número de habitantes, la comuna en la que habita la zona censal, la fracción de usuarios que tienen instalada una app de las marcas y una variable binaria para indicar si la zona censal está a menos de 250 metros de alguna estación de metro.

1. Comente las ventajas teóricas de cada una de las formulaciones. Si prefiere, puede responder la pregunta argumentando en qué escenarios prefería una formulación y en cuales preferiría la otra.

2. En la Tabla 1 se despliegan los resultados de un primer ejercicio de regresión para los dos modelos anteriormente planteados. Por simplicidad, se ha considerado solo una muestra de tiendas y zonas censales. Como el modelo se ha estimado usando una técnica de descentrado, los efectos fijos no se reportan en la Tabla 1.

    a. ¿Cómo es posible que, usando los mismos datos, las ecuaciones 1 y 2 nos reporten distintos estimadores del efecto de la distancia en la demanda de cada zona censal por cada marca?

    b. ¿Qué información necesita para cnstruir un intervalo de confianza para el efecto que tiene la superficie de la tienda en su demanda? Si fuera posible, reporte el valor numérico de este intervalo de confianza usando los datos de la tabla.

    c. Si el objetivo del análisis sería tener el pronóstico más preciso de la demanda en cada tienda, ¿cómo elegiría entre estos dos modelos?

    d. ¿Cómo testearía si hay diferencias significativas en la demanda por marca?

Tabla: Resultados modelos alternativos de regresión

Variable	Ecuación 1	Ecuación 2
Log-Distancia	-2.123 ***	-2.817 ***
Superficie	0.234 *	
Marca A	0.245 .	
Marca B	1.567 **	
Marca C	0.453 .	
Mall	2.145 ***	

Variable	Ecuación 1	Ecuación 2
Nhabitantes		3.234 ***
Fracusuarios		0.881 **
Providencia		1.233 **
Santiago		0.029
Macul		-0.913 *
La Florida		-0.372 .
Estación Central		-1.409 *
Metro		1.199 ***
Efectos fijos: Zona censal	Yes	No
Efectos fijos: Tienda	No	Yes
N. Obs	34,207	34,207
R^2	0.234	0.489
$R^2_{adj.}$	0.233	0.482

3. Para evaluar la capacidad de los modelos para apoyar la gestión del negocio, el equipo de desarrollo quiere generar algunos pronósticos con la demanda para la próxima semana para cuatro tiendas seleccionadas. Para esto, necesita ajustar la propuesta inicial de modo que los efectos temporales sean controlados por una lista de variables temporales z_t que incluyen un pronóstico del clima, tendencia, estacionalidad mensual, la intensidad de la actividad promocional de la marca, así como la intensidad de la actividad promocional de otras marcas competidoras. Considerando que variables como el clima y la actividad promocional son inciertas, se ha generado un reporte con tres escenarios: uno pesimista, uno neutro y otro optimista. Mientras en el escenario neutro se considera que las variables independientes toman valores típicamente observados, los escenarios pesimistas y optimistas se evalúan en combinaciones más extremas que favorecen y desfavorecen la demanda respectivamente. Los pronósticos generados se despliegan en la siguiente tabla.

Tabla: Pronóstico de demanda para cuatro tiendas seleccionadas

Tienda	Ecuación 1			Ecuación 2		
	Pesimista	**Neutro**	**Optimista**	**Pesimista**	**Neutro**	**Optimista**
Tienda A01	316,6 (11,8)	345,1 (7,2)	392,1 (12,1)	318,5 (12,2)	352,1 (6,7)	387,7 (10,9)
Tienda A12	555,8 (12,2)	578,8 (8,1)	602,2 (11,8)	547,6 (11,0)	566,3 (9,1)	607,7 (11,6)

Tienda		E cuación 1			E cuación 2	
Tienda B09	834,4 (14,3)	888,9 (9,2)	965,4 (13,3)	811,4 (15,9)	874,9 (8,9)	954,7 (13,8)
Tienda B14	733,9 (13,6)	786,6 (8,4)	822,3 (12,8)	728,6 (11,5)	791,9 (8,7)	817,8 (12,6)

En la tabla anterior, junto con los pronósticos se despliegan los errores estándar de pronóstico (*sef*). Explique por qué los *sef* son en general menores en los escenarios neutros.

4. Habiendo revisado la primera ronda de resultados, el equipo de desarrollo se inclina por concentrarse en el modelo dado por la ecuación 1. Sin embargo, a partir de la retroalimentación de los gerentes de tienda, se reconoce que hay varios elementos adicionales que deben incorporarse. Formule un modelo de regresión lineal que expanda la ecuación (1) pero que considere los siguientes factores:

   - Las tiendas en los centros comerciales suelen ser más grandes, por lo que es esperable que la superficie de la tienda tenga una influencia distinta dependiendo si la tienda está o no en un mall.
   - Para zonas censales relativamente cercanas, la distancia es un buen indicador del costo de acceder. Sin embargo, si la distancia es mayor que $d*$, entonces es mejor considerar el tiempo de viaje v_{ik}.
   - El efecto de la cercanía al metro es distinto si la tienda está en un mall o no.
   - En la muestra hay marcas que son de destino u otras de conveniencia. Una marca de destino se caracteriza por clientes muy leales dispuestos a viajar con el propósito principal de visitar la tienda de esa marca. Por su parte, las marcas de conveniencia son visitadas principalmente por clientes que se encuentran con la tienda al paso. Considere que, en el modelo, el efecto del tiempo de viaje depende de si la marca es de destino o conveniencia.
   - La variable $Fracusuarios_i$ se define como la fracción de habitantes de la zona censal i que tiene instalada alguna app de cualquiera de las marcas de *Whileus* al momento de hacer el pronóstico. Los resultados de la Tabla 2 indican que, entre mayor es el número de usuarios de las aplicaciones móviles, mayor es la demanda de esa zona censal. Sin embargo, es esperable que esta variable sea endógena. Es decir, que no sea el uso de las aplicaciones las que empujan la demanda, sino que es la demanda la que impulsa el uso de las aplicaciones. Sugiera alguna modificación al modelo para aminorar la endogeneidad del uso de las aplicaciones.

Solución

1. La diferencia principal es que uno describe el efecto de características de las tiendas (Ecuación 1), mientras que la otra describe el efecto de las características de la zona censal (Ecuación 2). Es preferible (1) si el foco está en entender qué características de la tienda inciden en el tráfico, y (2) si el foco está en entender qué características de las zonas censales son las que generan más tráfico.

2. Cada pregunta tiene respuestas bastante directas:

   a. Los estimadores dan cuenta del valor esperado de la variable dependiente **condicional** en la realización de las otras variables. Cada vez que cambiamos el conjunto de las variables sobre las que se condiciona, es esperable que cambien los estimadores.
   b. Necesitamos (una estimación de) los errores estándares. Como no están reportados en la tabla, no podemos calcular su valor numérico.
   c. Si el objetivo es pronóstico, habría que calibrar el modelo en una muestra de calibración y evaluarlo en una muestra retenida de validación. La evaluación típicamente consideraría métricas como MAE o MAPE.
   d. Por medio de una prueba F. Por ejemplo, si queremos evaluar si hay diferencia entre las marcas A y C, podemos testear la hipótesis lineal $\beta_A - \beta_C = 0$.

3. Los errores estándares de pronóstico dependen de las variables explicativas. Técnicamente el error de pronóstico considera tanto un error residual como uno de sampleo. Este último crece al alejarse de los valores típicos de sampleo.

4. La ecuación 1, por extensión puede escribirse como:

$$q_{ikt} = \alpha_k^1 + \beta_t^1 + \gamma^1 d_{ik} + \theta^{1'}(\text{Superficie}_k, B_k, C_k, Mall_k) + \varepsilon_{ikt}^1$$

   Donde el 1 es un superíndice para indicar que provienen del primer modelo, B_k y C_k son variables binarias que indican la marca de cada tienda (para que sea identificable, tenemos que eliminar una marca). Sobre este modelo debemos:

   - Agregar $\text{Superficie}_k \cdot Mall_k$ en la función que integra $\theta^{1'}$.
   - Reemplazar el término de la distancia por $\gamma_1^1 d_{ik} 1_{[d_{ik} < d^*]} + \gamma_2^1 v_{ik} 1_{[d_{ik} > d^*]}$.
   - Agregar $Metro_i \cdot Mall_k$.
   - Agregar $v_{ik} \cdot Conveniencia_k \cdot 1_{[d_{ik} > d^*]}$ (donde $Conveniencia_k = 1$ si marca es de conveniencia) y $v_{ik} \cdot (1 - Conveniencia_k) \cdot 1_{[d_{ik} \leq d*]}$.

5. Aunque hay otras ideas admisibles, el fraseo de la pregunta parece sugerir la solución más directa que es agregar la dinámica de adopción. Esto es, agregar un índice temporal a la variable $Fracusuarios_{it}$ de modo de representar como varía en el tiempo. Así, podemos ver cómo cambia la demanda en periodos de mayor o menor uso de las aplicaciones. Observaciones:

    a. Un nivel mayor de sofisticación podría sugerir usar $Fracusuarios_{it-1}$ y/o $Fracusuarios_{it-2}$. La inclusión de rezagos presupone que la adopción antecede las compras.
    b. Esta especificación puede ayudar a aminorar los problemas de endogeneidad, pero no los soluciona del todo. Si hay un momento específico en que se lanza la app, o si hay algunas campañas específicas que solo se promocionan a través de las apps, podrían dar elementos adicionales para reducir la endogeneidad del constructo.

2.4.7 Problema 7 - P1 Control 1 Otoño 2024

Uno de los cambios más relevantes de los últimos años, en ambientes de negocio y en la sociedad en su conjunto, ha sido la irrupción de las inteligencias artificiales (IA) generativas[1]. En simple, las inteligencias artificiales generativas son un tipo de inteligencia artificial que puede crear contenidos nuevos, incluyendo conversaciones, historias, imágenes, videos e incluso música. Entre las ventajas de estas tecnologías es su capacidad de comprender matices del lenguaje no estructurado que les permite dar respuestas a entornos complejos. A continuación, exploraremos el potencial uso de estas tecnologías para generar sistemas automatizados de atención de clientes en la industria bancaria.

Muchas de las aplicaciones de estas tecnologías de IA generativas se basan en modelos fundacionales[2] que son modelos de machine learning entrenados en un amplio espectro de datos generalizados y por tanto son capaces de realizar una amplia variedad de tareas generales. Los modelos de generative pre-trained transformer (GPT) de OpenAI son precisamente un ejemplo de un modelo fundacional. Aunque los modelos fundacionales tienen capacidades generales de lenguaje, en general no logran el nivel de especificidad requeridos para dar servicio a los clientes. Por ejemplo, un modelo de GPT no sabrá cuáles son las políticas internas de cada banco para resolver cada tipo de requerimiento de sus clientes. Para dotar a las IA generativas de este conocimiento contextual, se puede hacer una etapa liviana de entrenamiento o de *fine tuning*[3].

En este proyecto, investigamos el rendimiento de diferentes estrategias de entrenamiento sobre el desempeño de un asistente virtual basado en IA generativas en el contexto de la industria bancaria, donde tener el control de las respuestas es especialmente sensible para las empresas. En específico se consideraron tres estrategias de entrenamiento:

- **Prompt 1**: Se instruye al agente con una lista breve y resumida de cómo debe reaccionar a los requerimientos de los usuarios.
- **Prompt 2**: Se instruye al agente con una lista detallada de cómo debe reaccionar a los requerimientos de los usuarios.
- **Prompt 3**: Se instruye al agente con una lista detallada de cómo debe reaccionar a los requerimientos de los usuarios y se le indica que solo puede responder desde un conjunto acotado de respuestas.

Para entender el desempeño de distintas estrategias de entrenamiento se diseñó un experimento en que se le pidió a un grupo de usuarios de tarjetas de crédito que, en una serie de escenarios, interactuara libremente con un agente basado en IA generativas para resolver preguntas asociada a un desconocimiento de transacción. Un desconocimiento de transacción ocurre cuando un cliente revisa su cartola de cuentas e identifica una transacción que no reconoce como propia. Al terminar cada interacción se les preguntó a los participantes una serie de preguntas para entender cómo evaluaban la interacción con el agente digital.

Como resultado de estas interacciones se generó una base de datos con las siguientes columnas:

- **customer_id**: Identificador del participante dentro del experimento.
- **task_id**: Identificador del escenario.
- **age**: Edad del participante.
- **sex**: Sexo del participante.
- **NSE**: Nivel socioeconómico del participante.
- **prompt_type**: Tipo de prompt asignado en el experimento.
- **resolutivity**: Resolutividad percibida por el participante.
- **confidence**: Confianza percibida por el participante en la correctitud de la información.
- **experience**: Satisfacción percibida por el participante.
- **efficacy_input**: Eficacia de comprensión de mensajes percibida por el participante.
- **efficacy_output**: Eficacia de generación de respuestas percibida por el participante.
- **interaction_time**: Duración total de la interacción del participante con el agente.
- **cus_res_time**: Tiempo promedio que tarda en responder el participante en la interacción.
- **agt_res_time**: Tiempo promedio que tarda en responder el agente en la interacción.
- **n_messages**: Número de mensajes en la interacción.
- **n_words**: Número de palabras promedio por mensaje.

Adicionalmente, considerando las normativas del banco, cada interacción fue examinada para determinar el porcentaje de errores cometido por el agente. El experimento consideró la participación de 643 clientes a cada uno de los cuales

se les pidió interactuar con el agente digital en tres escenarios distintos con lo que se tiene un total de 1,929 filas en la base de datos.

1. (1 punto) La compañía está considerando usar un(os) modelo(s) de regresión para medir el impacto de distintas estrategias de entrenamiento. Considerando que el sistema está diseñado para dar servicio de atención a los clientes del banco, ¿cuál cree usted que debieran ser las variables dependientes para analizar en un modelo de regresión?

2. (2 puntos) Sea y_{ik} la variable de desempeño que usted considera relevante para el participante i ($i \in \{1, \ldots, 643\}$) en el escenario k ($k \in \{1, 2, 3\}$). Formule un modelo de regresión que permita evaluar el desempeño de las distintas estrategias y que considere las siguientes componentes:

   - Algunos escenarios podrían ser más fáciles o más difíciles por lo que el agente podría tener desempeños sistemáticamente diferentes por tareas.
   - Los usuarios de mayor edad tienden a tener una peor actitud hacia este tipo de tecnología y por tanto hacer peores evaluaciones de desempeño.
   - Si la interacción con el agente virtual es muy corta, probablemente los usuarios evalúen el desempeño como insuficiente. Del mismo modo, si la interacción es muy larga los usuarios en general tenderán a tener malas evaluaciones de desempeño.
   - Aunque no hay diferencias en los niveles generales de la evaluación de desempeño entre género, en general los hombres tienden a preferir interacciones más cortas que el resto de los géneros.

3. (1 punto) ¿Qué limitaciones tiene el modelo anteriormente planteado para el entendimiento de la pregunta de investigación? ¿Como recomendaría corregir estas limitaciones?

4. Para una parte del análisis, se consideraron tres modelos de regresión que se reportan en la Tabla 1.

   a) (1 punto) De acuerdo a los resultados de estos modelos de regresión, ¿qué rol juega las distintas estrategias de entrenamiento en el desempeño del agente virtual?

   b) (1 punto) ¿Cómo podríamos determinar el impacto de la introducción de restricciones en la estrategia de entrenamiento?

Variable	Experiencia (Coeficiente)	Resolutividad (Coeficiente)	Porcentaje Errores (Coeficiente)
Intercepto	0.025	0.146	4.675

Variable	Experiencia (Coeficiente)	Resolutividad (Coeficiente)	Porcentaje Errores (Coeficiente)
Prompt 2	-0.079 **	-0.157 ***	-4.965 **
Prompt 3	-0.005	0.109 **	-0.262
Eficacia entrada	0.174 ***	0.155 ***	0.541
Eficacia salida	0.100 ***	-0.030	1.400
Confianza	0.720 ***	0.786 ***	0.508
Tiempo	-0.006	-0.024 ***	0.094
Interacción			
Num Mensajes	0.007 **	0.020 ***	0.224
Edad	-0.002	-0.003 **	0.044
Mujer	0.027	-0.038	0.030
N. obs	1,929	1,929	126
R^2	0.798	0.709	0.220
$R^2_{adj.}$	0.796	0.707	0.105

Tabla: Resultados modelos alternativos de regresión.Fuente: Hernández, F. (2024): "Agentes inteligentes basa

[1] https://aws.amazon.com/es/what-is/generative-ai/
[2] https://aws.amazon.com/es/what-is/foundation-models/
[3] https://platform.openai.com/docs/guides/fine-tuning

Solución

1) La respuesta más clara es la satisfacción de los usuarios (Experience). Sin embargo, resulta a todas luces insuficiente. Es necesario tener un acercamiento hacia la reflexión de los siguientes puntos:

   a) El agente debe seguir un conjunto de procedimientos definidos institucionalmente y por tanto es necesario mirar los errores cometidos. Los agentes de IA son conocidos por alucinar y podrían dejar satisfecho al usuario entregándole respuestas que son inconvenientes por la firma.
   b) Uno de los motivos de la automatización es que el agente pueda resolver sin la necesidad de derivar los casos a un agente humano. Desde el punto de vista de la costo-efectividad, también es importante mirar su resolutividad.

2) El modelo de regresión puede expresarse como:

$$y_{ik} = \sum_{e\in\{2,3\}} \theta_e 1_{ei} + \alpha_k + \beta_1 Age_i + \beta_2 NW_{ik} + \beta_3 NW_{ik}^2 + \beta_4 Sex_i + \beta_5 NW_{ik} Sex_i + \varepsilon_{ik}$$

Donde

a) La variable 1_{ei} toma el valor 1 si el usuario i es asignado a la estrategia de entrenamiento e y por tanto θ_e captura el efecto que tiene la estrategia de entrenamiento e. Notar que, para que el modelo sea identificable necesitamos excluir una estrategia.
b) El efecto fijo α_k controla por diferencias entre tipos de escenarios.
c) El término β_1 permite controlar que la evaluación dependa de la edad.
d) Los términos β_2 y β_3 permiten capturar el efecto no lineal del largo de la interacción. En esta especificación usamos el número de palabras (NW_{ik}), pero el número de mensajes es una alternativa posible.
e) El término β_4 captura el efecto de ser hombre, sumando al efecto de β_5 que captura la interacción entre el largo de la interacción y el sexo hombre.

3) Hay varias limitaciones, pero probablemente se puedan agrupar en las dos siguientes líneas:

a) Respecto al diseño de investigación. Por ejemplo, el diseño solo compara tres estrategias específicas, pero no sabemos cuál sería el desempeño de una nueva estrategia de entrenamiento.
b) Respecto al modelo de regresión. Por ejemplo, hay factores que no controlamos como el NSE o, si se considera más de una variable de desempeño, cuál es la relación entre estos factores.

4) Las respuestas se derivan directo de las tablas.

a) Los coeficientes del Prompt 2 es significativo en las tres variables de desempeño consideradas y por tanto esa estrategia de entrenamiento tiene peor resolutividad y satisfacción, pero también un menor número de errores procedurales. Similarmente, el Prompt 3 tiene una mayor resolutividad que la estrategia de entrenamiento base, pero las diferencias no son significativas en la satisfacción ni en los errores procedurales.
b) Habría que testear si $\theta_2 = \theta_3$ (por ejemplo, si $\theta_2 > \theta_3$ en la ecuación de satisfacción usuaria, implicaría que incluir restricciones en la estrategia de entrenamiento empeora la evaluación. Como el objeto a testear son restricciones lineales, entonces puede hacerse con una prueba F.

Nota: Para los curiosos, los p-valores de los test para las tres ecuaciones son (0.066, 0.000 y 0.016) por lo que si consideramos un 95% las restricciones solo tendrían un impacto en la resolutividad.

2.4.8 Problema 8 - P1 Control Primavera 2023

Con el desarrollo de los canales digitales, muchas industrias han *desintermediarizado* la distribución de productos y servicios. Este es precisamente el caso de la industria musical en que tanto artistas consagrados[1] como independientes[2] pueden subir directamente su música a plataformas como Spotify, Deezer o Youtube Music. Junto con el cambio en la estructura de distribución, las plataformas digitales proveen información detallada para entender los factores que afectan la popularidad de cada canción y artista. En esta pregunta usaremos un enfoque de regresión para cuantificar la relación entre distintos factores en la audiencia de cada pieza. Para ello, se cuenta con los registros históricos de una muestra de 123.456 canciones en una plataforma de distribución musical, en la que se observan las siguientes variables:

- **IdSong**: Identificador único para cada canción.
- **Date**: Fecha del registro.
- **Title**: Título de la canción.
- **Artist**: Intérprete de la canción.
- **Genre**: Género musical al que pertenece la canción (e.g. Rock, Pop, Indie, etc.)
- **Length**: Duración de la canción (medido en segundos).
- **LaunchDate**: Fecha en la que fue lanzada la canción.
- **Nplay**: Número de veces que ha sido escuchada la canción.
- **Tprom**: Tiempo de escucha promedio cada vez que es escuchada una canción.
- **Danceability**: Puntaje [0-1] de qué tan bailable es una canción (basada en tiempo y regularidad).
- **Energy**: Puntaje [0-1] de qué tan energética es una canción (basada en timbre y volumen).
- **PromoExpend**: Monto del gasto en promoción dentro de la plataforma para esa fecha.
- **Nfollow**: Seguidores en redes sociales de artista.
- **Top50**: 1 si el artista ha tenido alguna vez una canción en el Top 50 del Billboard.

1. Considere un modelo de regresión lineal genérico

$$y_i = \beta' x_i + \varepsilon_i$$

que cuantifica la relación de una medida de la popularidad de una canción i (y_i), con un conjunto de variables observables de esa canción (x_i). Proponga dos métricas de popularidad. Para cada una de ellas indique una limitación de usar esa variable como indicador del éxito de audiencia de una canción.

Observación: Notar que la base está indexada por la canción y el tiempo, mientras la métrica pedida depende solo de la canción.

2. Describa un modelo de regresión lineal para estimar la popularidad de una canción. Para la construcción de este modelo de regresión considere que:

   a) Es importante controlar por la popularidad que típicamente tiene cada artista y que las canciones lanzadas más recientemente tienen acceso a un mayor número de usuarios activos en la plataforma.
   b) Las promociones aumentan la popularidad de las canciones, pero este efecto es acumulativo en el tiempo.
   c) El efecto de la *danceability* de una canción puede ser positivo para unos géneros, pero negativo para otros.
   d) El efecto del largo de una canción es no lineal en la popularidad: mientras temas muy cortos raramente son muy populares, los temas demasiado largos también tienen dificultades en generar audiencias.

   Defina cuidadosamente las variables e indique cómo se obtienen a partir de los registros disponible

3. Suponga que se estiman tres modelos simples de regresión lineal para describir la popularidad de las canciones dentro de la plataforma. Los resultados de estos dos modelos se reportan en la siguiente tabla

Tabla: Resultados modelos alternativos de regresión

V ariable	Modelo 1 Coef iciente	Modelo 1 SE	Modelo 2 Coef iciente	Modelo 2 SE	Modelo 3 Coef iciente	Modelo 3 SE
ln(*textlength*)	0.453	0.001	-0.462	0.023	-0.3983	0.024
Promo	5.341	0.562	4.341	0.011	5.101	0.012
Energy	0.138	0.453	3.453	0.003	3.238	0.001
Dance ability	2.345	0.002	0.091	0.645	0.192	0.345
NFollow	2.323	0.011	2.421	0.016	1.933	0.008
Top50					1.498	0.011

Fixed Effects

	Modelo 1	Modelo 2	Modelo 3
Artist	Yes	Yes	Yes
Year	Yes	Yes	Yes
Genre	No	Yes	Yes

Métricas

	Modelo 1	Modelo 2	Modelo 3
Std. Error	5.355	4.307	4.278
R^2	0.296	0.445	0.455
R^2_{adj}	0.286	0.398	0.401

A partir de los resultados anteriores:

a. ¿Qué aprendizaje puede derivar a partir de observar los R^2?

b. ¿Cuál es el intervalo de confianza del 95% para el coeficiente $NFollow$ en el modelo 2?

c. ¿Qué puede explicar el cambio de significancia de *Energy* y *Danceability* en Modelos 1 y 2?

d. ¿Cómo se puede testear que *Danceability* y *Energy*, de manera conjunta tienen un efecto nulo?

Solución

1. Algunas posibilidades:

   - $y_i^0 = \max\{NPlay_{it}\}_t$, que corresponde al máximo de veces que ha sido escuchada una canción.
     Una limitación de esta métrica es que las canciones se lanzan en instantes distintos del tiempo y las canciones más antiguas tendrán más escuchas, aunque no sean tan populares.
   - $y_i^1 = \max\{NPlay_{it}\}_t/(LaunchDate_i)$. Acá, el denominador corresponde al tiempo transcurrido desde el lanzamiento y, por lo tanto, la métrica mide el número de reproducciones por unidad de tiempo. Una limitación de esta métrica es que no considera el ciclo de vida del producto (ciclo de reproducción de las canciones). Las canciones recién lanzadas suelen ser muy escuchadas, pero algunas pasan rápidamente de moda.
   - $y_i^2 = NPlay_{i,LaunchDate_i+365}$, que corresponde al número de reproducciones un año después de haber sido lanzado la canción. Una limitación de esta métrica es que las plataformas aumentan sus audiencias y, por tanto, favorece a canciones que han sido lanzadas más recientemente.

     Hay varias otras posibilidades. Por ejemplo, usar el *share* de reproducciones dentro de la categoría, la desviación de reproducciones con respecto a la canción anterior del artista, etc. Notar que hay algunas métricas que no capturan precisamente popularidad. Por ejemplo:
   - $y_i^3 = \max\{NPlay_{it}\}/\sum_t Promo_{it}$, que mide eficiencia promocional y se indefine para canciones que no han sido promocionadas.

- $y_i^4 = \sum\{TProm_{it}\}/\text{mean}\{Length_i\}$, que mide el *envolvimiento o lealtad* con la canción . Una canción podría ser escuchada una vez completa y aunque no diríamos que es popular, tendría el máximo en esta métrica.

2. Todas las componentes se agregan más o menos directo en la ecuación

   - Hay que agregar controles de artista ($Artist_i$) y tiempo de lanzamiento ($LaunchDate_i$).
   - Hay que agregar el gasto acumulado en promoción ($\sum_t Promo_{it}$).
   - Hay que agregar las interacciones entre danzabilidad y género ($Danceability_i\, Genero_i$).
   - Hay que agregar tanto efectos lineales como cuadráticos ($Danceability_i + Danceability_i^2$).

   Con esto el modelo puede expresarse como:

$$\begin{aligned} y_i \; &= \alpha^1_{Artist_i} + \alpha^2_{LaunchDate_i} + \beta_1 \left(\sum_t PromoExpend_{it} \right) \\ &\quad + \beta_2\, Danceability_i + \beta_3\, Genre_i \\ &\quad + \sum_{g \in genero} \alpha_g\, Danceability_i\, Genre_i + \beta_4\, Length_i + \beta_5\, Length_i^2 \end{aligned}$$

3. Cada una de las partes tiene respuestas bastante directas:

   a. Al comparar el Modelo 1 con el Modelo 2, vemos que el género explica bastante variabilidad en la popularidad. Al comparar el Modelo 2 con el Modelo 3, vemos que la distinción de $Top50$ ayuda a explicar una fracción muy menor de la varianza en popularidad.
   b. El intervalo de confianza viene dado por:

$$(2.421 - 1.96 \cdot 0.016,\; 2.421 + 1.96 \cdot 0.016)$$

   c. En ambos casos una de las variables es significativa, pero la otra no. Aunque hay otras explicaciones posibles, la más directa es que estas dos variables estén altamente correlacionadas y, por tanto, el efecto conjunto es identificable. Es difícil separar el efecto de una con respecto a la otra.
   d. Queremos testear que $\beta_{dance} = 0$ y $\beta_{Energy} = 0$. Este es un sistema de restricciones lineales que puede testearse directamente a través de una prueba F.

2.4.9 Problema 9 - P1 Control Otoño 2023

KornerK introdujo hace un par de años un modelo de negocio innovador en que, a través de una aplicación móvil, los clientes de la industria supermercadista podían hacer sus compras remotamente. Para llevar a cabo sus compras, los clientes listan el conjunto de productos que desean adquirir y ponen una orden a la plataforma, la que contacta a alguno de los numerosos *shoppers* que tiene afiliados, los que se encargan de recoger los productos desde la sala de supermercado y transportarlo hasta el domicilio del cliente. La asignación de *shoppers* se hace considerando la cercanía al domicilio del usuario, por lo que la compañía se ha preocupado de tener suficientes afiliados disponibles para cubrir la demanda en cada geografía y bloque horario.

Después de un par de años de un crecimiento constante, lo que incluye un período de crecimiento explosivo durante los confinamientos de la pandemia, en los últimos meses se ha visto afectado por importantes cambios tanto regulatorios como de los patrones de la demanda. A partir de esta coyuntura, *KornerK* se ha visto obligado a rediseñar su dotación y ha considerado hacer una reducción de hasta un 20% de los *shoppers* en aquellas zonas con mayor caída en la demanda. Con este propósito, se han propuesto analizar los patrones de demanda históricos para hacer un pronóstico de la demanda.

Para sostener el análisis, se ha preparado un set de datos con las ventas agregadas por bloque horario en cada una de las zonas en que opera *KornerK*. En esta base, cada fila se compone de las siguientes variables:

- **IdZip**: Identificador de la zona geográfica.
- **Date**: Fecha en formato dd/mm/yyyy.
- **Hour**: Un entero que representa la hora que define el bloque horario (8, 9, 10, ..., 22).
- **NOrdersC**: Número de órdenes por servicio que fueron exitosamente completadas.
- **NOrdersN**: Número de órdenes por servicio que no pudieron ser completadas por falta de *shoppers*.
- **Day**: Día de la semana que se registra la discrepancia (con lunes=1, hasta domingo=7).
- **Mes**: Correlativo del número de meses desde inicio de operación (1, 2, 3, ...).
- **Rain**: 1 si llovió en el bloque horario.
- **IMACEC**: Variación del índice mensual de actividad económica.

Como ilustración, se muestran las primeras seis entradas se presentan en la siguiente tabla.

Tabla: Muestra de los primeros 6 registros de la base de discrepancias

I dZip	Date	Hour	NOrd ersC	NOrd ersN	Day	Mes	Rain	IM ACEC
1 5217	02 /01/ 2018	8	248	0	2	1	0	0.02
1 5217	02 /01/ 2018	9	325	0	2	1	0	0.02
1 5217	02 /01/ 2018	10	331	13	2	1	0	0.02
2 3156	03 /01/ 2018	8	1345	1	3	1	0	0.02
2 3156	03 /01/ 2018	9	1221	0	3	1	0	0.02
2 3156	03 /01/ 2018	10	1436	54	3	1	0	0.02

1. Describa un modelo de regresión lineal para estimar la **demanda total** por servicios en cada zona y bloque horario que considere **al menos cuatro** de los siguientes elementos:

   a) Efectos fijos geográficos para capturar que gran parte de la variabilidad de la demanda se explica por diferencias en los patrones de consumo de las distintas zonas.

   b) El patrón de demanda a lo largo de día es altamente estacional, por lo que es esperable que la inclusión de efectos fijos por hora sea relevante.

   c) Aunque hay alguna variación de la demanda dependiendo del día de semana, es despreciable en relación con la diferencia que hay entre días de semana y fin de semana, la que es muy relevante.

   d) El nivel de actividad económica parece estar positivamente correlacionado con la demanda. Sin embargo, este efecto parece estar rezagado, por lo que parece razonable considerar los valores de actividad económica de los últimos 3 meses.

   e) La demanda ha ido creciendo sostenidamente a lo largo del tiempo, por lo que se necesita incluir un factor que capture una tendencia de crecimiento lineal de la demanda como función del tiempo.

   f) La demanda tiende a crecer en los bloques en que llueve. Sin embargo, la magnitud del crecimiento es distinta dependiendo de que si es un día de semana o un fin de semana.

   Defina cuidadosamente las variables dependientes e independientes e indique cómo se obtienen a partir de los registros disponibles en la base de datos.

2. Suponga que se estiman dos modelos simples de regresión lineal para describir la demanda por servicios de *KornerK*. Los resultados de estos dos modelos se reportan en la siguiente tabla:

Tabla: Resultados modelos alternativos de regresión

V ariable	Modelo 1 Coef iciente	Modelo 1 p-valor	Modelo 2 Coef iciente	Modelo 2 p-valor	Modelo 3 Coef iciente	Modelo 3 p-valor
Int ercepto	589.415	0.000	-	-	-	-
FinD eSem- ana	120.732	0.001	89.462	0.023	88.559	0.024
Rain	20.341	0.562	14.341	0.011	11.101	0.012
IMACEC	65.323	0.011	59.021	0.016	62.774	0.019

Fixed Effects

	Modelo 1	Modelo 2	Modelo 3
Zip	No	No	Yes
Hour	No	Yes	Yes

Resumen

Métrica	Modelo 1	Modelo 2	Modelo 3
Std. Error	5.355	4.307	4.278
R^2	0.196	0.445	0.544
R^2_{adj}	0.186	0.392	0.417

A partir de los resultados anteriores:

a) ¿Por qué no se reporta un intercepto para los modelos 2 y 3?
b) ¿Qué modelo explica la mayor varianza de la demanda observada?
c) ¿Qué podemos afirmar con respecto al rol de los efectos fijos regionales para describir la variabilidad en la demanda?
d) ¿Qué rol juega la actividad económica en la demanda por servicios de *KornerK*?

3. Volviendo al problema de gestión al que se enfrenta la compañía, *KornerK* está interesado en evaluar cómo podría impactar la reducción de *shoppers* en la capacidad de responder a la demanda. Para esto, ha elaborado un plan preliminar en que quedarían N_{ith} *shoppers* disponibles para trabajar en el bloque horario h del día t en la zona i.

   a) Usando el modelo 1, calcule el valor esperado de la demanda para un bloque en un día de semana, en que no llueve y en que hay una variabilidad en la actividad económica del 10% ($IMACEC = 0.1$).

b) Para el mismo bloque horario de la parte anterior, escriba una expresión para calcular la probabilidad de que la demanda total exceda la capacidad disponible expresada para ese bloque $N_{ith} = 600$.

Solución

1. Primero que todo, tenemos que clarificar que la variable dependiente es la demanda total en una zona i, fecha t y hora h, que corresponde tanto a las órdenes completadas como aquellas que excedieron la capacidad:

$$DT_{iht} = NOrdersC_{iht} + NOrdersN_{iht}$$

En el lado derecho de la regresión se pueden considerar cualquiera de los siguientes puntos:

- Efectos fijos geográficos: α_i^1
- Efectos fijos por hora: α_h^2
- Diferencia entre día de semana y fin de semana: $\beta_1 FDS_t$ donde $FDS_t = 1_{\{Day \in \{Sbado, Domingo\}\}}$
- El nivel de actividad económica de los últimos 3 meses $\beta_2 IMACEC_t + \beta_3 IMACEC_{t-1} + \beta_4 IMACEC_{t-2}$. Un modelo admisible es suponer que $\beta_2 = \beta_3 = \beta_4$. En vez de tener un parámetro por cada mes, se tiene un único para el promedio de los tres.
- La demanda ha ido creciendo sostenidamente: $\beta_5 MES_t$
- Lluvia y fin de semana: $\beta_6 RAIN_{th} + \beta_7 RAIN_{th} FDS_t$

Así, la ecuación de regresión que considera todos los factores viene dada por:

$$DT_{iht} = \alpha_i^1 + \alpha_h^2 + \beta_1 FDS_t + \beta_2 IMACEC_t + \beta_3 IMACEC_{t-1} \quad (2.3)$$
$$+ \beta_4 IMACEC_{t-2} + \beta_5 MES_t + \beta_6 RAIN_{th} + \beta_7 (RAIN_{th} \cdot FDS_t) \quad (2.4)$$

2. Incluye:

   a. Los modelos 2 y 3 tienen efectos fijos y, por tanto, hay un intercepto distinto para cada hora (Modelo 2) o para cada (geografía-hora). En general, los efectos fijos no se reportan porque suelen ser una secuencia larga de dummies por las que solo queremos controlar.

   b. El R^2 captura la varianza explicada. En este caso el modelo con mayor R2 es el modelo 3.

   c. Al agregar efectos fijos por región, el R^2 (varianza explicada) aumenta bastante, por lo que concluimos que una parte relevante de la variación en la demanda se explica porque hay regiones que hay más demanda que en otras.

d. En los tres modelos es positiva y significativa p-valor < 0.05, por lo que pode mos decir que mayor actividad económica correlaciona positivamente con mayor demanda para KornerK.

3. Detalle:

a. El valor esperado viene dado por:

$$\mu = 589.4 + 65.3 \cdot 0.1 = 595.9$$

b. La probabilidad viene dada por:

$$Pr(DT_{iht} \geq 600)$$

Donde DT_{iht} se distribuye normal con media $\mu = 595.9$ y varianza $\sigma^2 = (5.355)^2$. Este cálculo se puede hacer directamente en cualquier paquete estadístico. En R:

$$1 - pnorm(600, mean = 595.9, sd = 5.355)$$

Chapter 3

Modelos Probabilisticos

3.1 Introducción

En este capítulo se abarcarán problemas con un enfoque probabilístico, esto es, que se asume que los tomadores de decisiones se comportan aleatoriamente. Esto permite abordar una gran cantidad de problemas asociados al Marketing, los cuales se pueden caracterizar en tres tipos de modelos básicos:

- *Duración*: La pregunta clave es ¿Cuándo? Son situaciones ligadas a la duración de una determinada conducta del cliente, como por ejemplo: tiempo de permanencia en una compañía, tiempo de adopción de un cierto producto innovador, entre otros. Puede ser con tiempo continuo, o discreto.

- *Conteo*: La pregunta clave es ¿Cuántos? Son situaciones ligadas al estudio de llegadas de clientes y contabilización de una determinada conducta, como por ejemplo: número de visitas a un portal web y la cantidad de productos comprados en una tienda de retail.

- *Elección*: La pregunta clave es ¿Cuál? Son situaciones asociadas a las decisiones de elección de un determinado cliente, como por ejemplo: clientes que eligen responder (o no) a una campaña publicitaria y la elección de cambiar (o no) de canal de televisión.

Cada uno se estos modelos tiene muchas aplicaciones dentro de diversas situaciones en la vida real. Comportamientos más complejos pueden ser descritos usando combinaciones de los modelos básicos.

3.2 Metodología

Dicho enfoque posee una metodología de modelamiento sugerida, que comparten los modelos vistos a lo largo del curso.

La metodología consiste en:

1. Determinar el problema de decisión a estudiar y la información requerida.
2. Identificar el comportamiento observable (heterogeneidad) de interés a nivel individual. Típicamente, se denota con una x_i.
3. Seleccionar la distribución de probabilidad que caracterice el comportamiento individual. Se consideran los parámetros de esta distribución, como características latentes a nivel individual. Típicamente, se denota con $f(x|\theta)$.
4. Escoger la distribución que caracterice cómo las características latentes están distribuidas en la población. Se le llama distribución mixta o heterogénea. Típicamente, se denota con $g(\theta)$
5. Derivar la distribución agregada, o distribución observable, del comportamiento de interés.

$$f(x) = \int f(x|\theta)\, g(\theta)\, d\theta$$

$$p(x) = \sum_i f(x|\theta)\, Pr(\theta = \theta_i)$$

6. Estimar los parámetros del modelo (de la distribución mixta), mediante el ajuste de la distribución agregada a los datos observados.
7. Usar los resultados para tomar una decisión sobre el problema de marketing en cuestión.

3.3 Problemas Teóricos

3.3.1 E1: Tasa de respuesta

Una compañía de venta de ropa por catálogo busca decidir a que segmento enviar los catálogos de la próxima colección. Para ello analiza la tasa de respuesta de una muestra de clientes de cada segmento (esto es el ratio entre número de clientes que compra y número de catálogos enviados). Al mirar el primer segmento observa que de los 18 clientes a quienes se les envió el catálogo, ninguno compró y por tanto decide no enviar catálogos a ningún cliente de ese segmento. Esta compañía:

(a) Muy probablemente esté subestimando la tasa de respuesta de ese segmento.
(b) Debiera redefinir los criterios de segmentación para hacer grupos más grandes y accionables.
(c) Ha definido una política que da cuenta de su aversión al riesgo.
(d) Debiera usar un modelo de duración en tiempo continuo con dependencia en la duración.
(e) Debiera incorporar variables explicativas en su modelo predictivo.

3.3.2 E2: Distribución Weibull

En relación a la distribución Weibull:

a. Es un caso particular de la Poisson.
b. Es una generalización de la Poisson.
c. Sólo tiene un parámetro c.
d. Es muy flexible y permite incluso generar distribuciones bimodales.
e. Ninguna de las anteriores.

3.3.3 E3: Distribución Gamma y Heterogeneidad

¿Cuál de los siguientes factores motivan la utilización de una distribución Gamma para modelar la heterogeneidad de las tasas de adopción en un modelo de duración en tiempo continuo? (Puede elegir más de un factor).

i. Consistencia con el dominio de la probabilidad de abandono en cada período.
ii. Para generar una fórmula recursiva de fácil implementación.
iii. Flexibilidad para acomodar distintas formas de la distribución.
iv. Para generar una fórmula cerrada que pueda ser calculada de manera computacionalmente eficiente.

3.3.4 E4: Modelos de Duración

¿Cuáles de los siguientes modelos NO describe la duración de la relación de los clientes con una firma?

a. Beta-geométrica desplazada.
b. Beta-geométrica NBD.
c. Gamma-Weibull.
d. Binomial Negativa.
e. Ninguna de las anteriores.

3.3.5 E5: Modelo NBD

Un modelo NBD describe la demanda de botellas de *oporto* en los últimos 6 meses. Si con este modelo se estima la demanda del próximo mes:

a. Hay que hacer un modelo de regresión que considere la dinámica del problema.
b. El cálculo no se puede hacer directamente, sin embargo podemos recalibrar el modelo considerando un horizonte de un mes.
c. El histograma del número de botellas consumidas por cliente se moverá a la izquierda.
d. La probabilidad de comprar x botellas resulta ser simplemente 1/6 veces las probabilidades calculadas para el primer semestre.
e. Ninguna de las anteriores.

3.3.6 E6: Modelos de Conteo

Un analista propone el uso de modelos de conteo para describir la intensidad de la participación de los usuarios de la red social X. Los parámetros que describen el comportamiento de los usuarios han sido estimados usando datos de 4 días de actividad. El modelo ajusta extremadamente bien. Sin embargo, al usar las estimaciones para pronosticar la actividad del quinto día el modelo no ajusta bien. Al respecto, se debiera concluir que:

a. Se debe agregar la data para considerar la actividad agregada en todo el horizonte.
b. Probablemente el comportamiento de los usuarios en el quinto día esté afectado por factores que no están presentes en los primeros cuatro días.
c. Hay que calcular esperanzas condicionales.
d. Los supuestos de comportamiento son errados y hay que desechar el modelo.
e. Todas las anteriores.

3.3.7 E7: Tiempo discreto en Modelos de Duración

En la práctica, siempre podemos discretizar el tiempo y por lo tanto no hay motivos para usar modelo de duración en tiempo continuo. Discuta respecto de la veracidad de esta afirmación.

- Para eventos con gran variabilidad de tiempos de ocurrencia, un modelo de tiempo discreto podría requerir un número muy grande de periodos para describir el apropiadamente comportamiento bajo estudio.
- Si el comportamiento tiene dependencia en la duración, dichas dependencias son fáciles de incluir en un modelo de tiempo continuo.

Solucionario.

E1: a
E2: e
E3: iii, iv
E4: d
E5: c
E6: b

E7: Tiempo discreto en Modelos de Duración

- **Correcto.** Esto muestra que en fenómenos con tiempos muy dispersos, el modelo discreto se vuelve poco práctico, mientras que el continuo es más parsimonioso.
- **Correcto.** En continuo se pueden modelar directamente con funciones de riesgo (hazard) dependientes del tiempo, lo cual puede ser más complejo de representar en discreto.

3.4 Problemas Aplicados

3.4.1 Problema 1 (Control 2 2025-2)

VTV es una empresa aseguradora, en la que una de sus principales líneas de negocio consiste en la provisión de cobertura de segunda línea para clientes corporativos. En simple, VTV ofrece un seguro de salud que se gatilla cuando ocurre un gasto médico, pero que se activa después de que operen las coberturas obligatorias de Fonasa o Isapre. El servicio se vende a empresas de distintos rubros para que sirvan como beneficio complementario para sus trabajadores.

Recientemente, VTV ha lanzado un nuevo servicio llamado *TuBienestar*. Este servicio ofrece que los empleados de las empresas clientes de VTV puedan acceder a una batería de exámenes preventivos que aspiran a detectar tempranamente problemas de consideración y así reducir los gastos en salud a largo plazo. A continuación, usaremos modelos probabilísticos para tener una primera aproximación del valor del programa *TuBienestar*. Para eso, haremos uso de la base histórica de gastos que contiene la valorización de todos los eventos de salud, incluyendo consultas médicas, hospitalizaciones, exámenes y compra de medicamentos. Aunque existe información más detallada de cada evento, nos concentraremos en n_{ijkt}, correspondiente al número de eventos de salud en el periodo t del beneficiario i, que pertenece al grupo j de la empresa k. Notar que en esta base, los eventos están agrupados por periodos trimestrales y los beneficiarios de cada empresa están separados por segmento, ya que distintos tipos de empleados podrían acceder a distintos beneficios.

1. (2.0 puntos) Actualmente, VTV describe el número de eventos de cada beneficiario usando un modelo de Poisson. En este modelo, como los eventos de salud no son tan frecuentes, una gran proporción de los beneficiarios no registra eventos en un trimestre dado y, por tanto, el caso de cero eventos merece un tratamiento especial. Para modelar los eventos de gasto, se considera que, si hay un evento, el número de eventos se distribuye Poisson, pero que la probabilidad de no observar ningún evento es mayor que la de un modelo de Poisson tradicional. Escriba la log-verosimilitud de este modelo de conteo con ceros inflados. ***Hint:*** *Le puede resultar útil conceptualizar el modelo pensando que con probabilidad π no hay ningún evento y que con probabilidad $1-\pi$ el proceso sigue distribución de Poisson standard.*
2. Suponga ahora que, en vez de describir el número de eventos de cada beneficiario, nos concentramos en el número de beneficiarios que registra al menos un evento dentro de cada segmento-compañía. Así, como alternativa al modelo existente, se postula describir la ocurrencia de eventos de salud como un modelo de elección binaria.
    a) (1.0 puntos) Suponiendo que puede haber segmentos con mayor propensión a registrar eventos, escriba la log-verosimilitud de un modelo binomial con dos clases latentes.
    b) (1.0 puntos) Suponga que ha estimado el modelo y los estimadores máximos verosímiles vienen dados por $(\theta_1, \theta_2, \pi) = (0.1, 0.6, 0.5)$. Considere un segmento con 824 beneficiarios que, transcurrido un trimestre, evidencia 17 eventos de salud. Escriba una expresión para la probabilidad que este segmento corresponda a la clase de mayor propensión a registrar eventos.

Para las descripciones anteriores considere que n_{jkt} es el número total de beneficiarios que tuvieron algún gasto de salud en el trimestre t dentro del segmento j de la empresa k y que m_{jkt} es el número total de beneficiarios dentro de ese segmento, empresa y trimestre.

3. Como último enfoque, se estiman tres versiones de modelos beta-binomiales. En los modelos que se incluye heterogeneidad observable, se consideran variables asociadas tanto a características del segmento como a variables estacionales y de tendencia.

Variable	Modelo 1	Modelo 2	Modelo 3
α	0,127 **	0,151 **	0,149 *
β	3,503 ***	3,729 ***	3,852 ***
Sector:Minería		1,234 ***	1,108 ***
Sector:Construcción		2,122 ***	1,432 **
Sector:Salud		1,784 ***	1,665 ***
Tamaño-Mediana		0,145	-0,004

Variable	Modelo 1	Modelo 2	Modelo 3
Tamaño-Grande		0,372 *	0,401 *
AntiguedadContrato		0,188 **	0,193 **
TuBienestar			-0,089 **
Invierno		2,921 ***	2,955 ***
Verano		-0,539 *	-0,564 *
Tendencia		0,278 **	0,305 **
N. Obs	34.825	34.825	34.825
LL	-51.771,8	-45.282,1	-41.321,9
AIC	107.543,6	90.580,2	82.665,8

Tabla 1. Estimadores máximo verosímiles de dos modelos beta-binomiales con variables explicativas.

a. (1.0 puntos) Discuta si el programa *TuBienestar* ayuda a describir el comportamiento de los beneficiarios y si el programa *TuBienestar* es bueno para la salud de los beneficiarios.
b. (1.0 puntos) En otros mercados el valor esperado de la probabilidad que un beneficiario exhiba un evento de salud en un trimestre dado es 0.16. ¿Cómo testearía si los datos de Chile son distintos a ese valor medio internacional?

Solución

1. Siguiendo la indicación, la probabilidad de los eventos viene dada por:

$$\Pr(N_{ijkt} = n \mid \lambda, \pi) = \begin{cases} \pi + (1-\pi)e^{-\lambda} & n = 0 \\ (1-\pi)\dfrac{e^{-\lambda}\lambda^n}{n!} & n > 0 \end{cases}$$

   Entonces, la función log-verosimilitud viene dada por:

$$LL = \sum_i \sum_j \sum_k \sum_t \left[1_{n_{ijkt}=0} \ln\left(\pi + (1-\pi)e^{-\lambda}\right) + 1_{n_{ijkt}>0} \ln\left((1-\pi)\frac{e^{-\lambda}\lambda^n}{n!} \right) \right]$$

2. Sean θ_1 y θ_2 las probabilidades de registrar un evento de las clases 1 y 2 respectivamente.

   a. La probabilidad de observar exactamente n eventos en el segmento de clientes j de la empresa k en el trimestre t viene dada por:

$$\Pr(n_{jkt} = n \mid \theta_1, \theta_2, \pi) = \pi \binom{m}{n_{jkt}} \theta_1^{n_{jkt}} (1-\theta_1)^{m-n_{jkt}} + (1-\pi) \binom{m}{n_{jkt}} \theta_2^{n_{jkt}} (1-\theta_2)^{m-n_{jkt}}$$

Entonces, la log-verosimilitud viene dada por:

$$LL \;=\; \sum_j \sum_k \sum_t \ln\Big(\Pr(n_{jkt} \mid \theta_1, \theta_2, \pi)\Big)$$

b. Aplicamos la fórmula de bayes directamente:

$$\Pr(\lambda = \lambda_2 \mid n_{jkt} = 17) = \frac{\Pr(n_{jkt} = 17 \mid \lambda = \lambda_2)\ \Pr(\lambda = \lambda_2)}{\Pr(n_{jkt} = 17 \mid \lambda = \lambda_1)\ \Pr(\lambda = \lambda_1) + \Pr(n_{jkt} = 17 \mid \lambda = \lambda_2)\ \Pr(\lambda =}$$

Donde:

$$\Pr(n_{jkt} = 17 \mid \lambda = \lambda_1) = \binom{824}{17} (0.1)^{17}\,(1-0.1)^{824-17},$$

$$\Pr(n_{jkt} = 17 \mid \lambda = \lambda_2) = \binom{824}{17} (0.6)^{17}\,(1-0.6)^{824-17},$$

$$\Pr(\lambda = \lambda_1) = \Pr(\lambda = \lambda_2) = \tfrac{1}{2}.$$

3. La parte a) requiere mirar los resultados de la tabla. La parte b) puede contestarse independientemente.

   a. El AIC del modelo 2 es mayor que el AIC del modelo 3 que solo agrega la variable tuBienestar, y por tanto, la variable tiene valor explicativo. Se puede argumentar también que el coeficiente mismo de la variable es significativo y por tanto ayuda a explicar.

   Respecto al impacto en la salud, como el coeficiente es significativo y negativo, diríamos que aquellos que participan del programa tienen una menor probabilidad de generar eventos y, por tanto, el programa sería beneficioso. Sería deseable discutir que este efecto no es necesariamente causal, pero no lo consideraremos necesario.

   b. Recordar que la esperanza de una $Beta(\alpha, \beta)$ viene dada por $\frac{\alpha}{\alpha+\beta}$. Por lo tanto, lo que necesitamos testear es que $\alpha = 0.16(\alpha + \beta)$, que es una restricción lineal que puede testearse usando un test de ratios de verosimilitud.

3.4.2 Problema 2

En el contexto del marketing B2B (Business to Business) para servicios profesionales, las empresas han comenzado a incorporar el marketing de contenidos (Content Marketing, CM) como parte integral de sus estrategias comerciales. Estas iniciativas buscan, principalmente, generar oportunidades de venta (leads) y fidelizar clientes existentes, mediante la entrega de información de valor. El marketing de contenidos puede adoptar múltiples formas, tanto en entornos presenciales como digitales. Las actividades presenciales incluyen eventos cara

a cara, como seminarios, conferencias o talleres; mientras que las digitales abarcan webinars, seminarios virtuales y la distribución de contenido a através de plataformas digitales, como los sitios web corporativos.

En los mercados B2B, donde un número reducido de cuentas clave suele concentrar una parte importante de las ventas, la gestión del marketing de contenidos debe articularse cuidadosamente con los procesos tradicionales de vents. Es habitual que estas organizaciones estructuren el proceso de ventas en dos grandes etapas: generación de leads y conversión de leads. Un lead representa una oportunidad de negocio que se manifiesta a través de señales tempranas de interés por parte de una cuenta clave, como una consulta por correo electrónico, una llamada comercial, la solicitud de una cotización o la descarga de un folleto técnico. Un lead se considera convertido cuando se concreta la venta y se formalizan los compromisos de pago.

Un desafío relevante para la gestión analítica del marketing de contenidos consiste en identificar qué tipos de actividaeds -presenciales o digitales- generan un mayor número de leads, cuáles de ellas contribuyen de forma más efectiva a su conversión. Asimismo, se plantea la hipótesis de que el nivel de particiapción (engagement) de los empleados de las cuentas clave podría desempeñar un rol mediador en estos procesos.

En este contexto, Vectorix, una empresa B2B que ofrece servicios profesionales, busca comprender cuáles componentes de su estrategia de marketing de contenidos resultan más efectivos. La empresa sospecha que la participación digital (por ejemplo, en webinars o mediante el consumo de contenido online) tiene un impacto positivo en la generación y conversión de leads. Sin embargo, la efectividad relavita de las actividades presenciales frente a las digitales sigue siendo una incógnita.

Con base a sus registros históricos de actividades de marketing y resultados comerciales, Vectorix se propone desarrollar un análisis cuantitativo que le permita evaluar la efectividad comparada de las distintas acciones de marketing de contenido. El objetivo final es orientar la asignación presupuestaria para los próximos dos trimestres, priorizando aquellas tácticas que maximicen tanto la generación como la conversión de oportunidades comerciales.

Los registros históricos se han condesando en una única base de datos que contiene las siguientes columnas:

- Idaccount: identificador de la cuenta clave.
- Zone: área geográfica en que se encuentra la sede central de la cuenta clave.
- Industry: industria a la que pertenece la cuenta.
- Tenure: Número de años transcurridos desde que la empresa comenzó su relación con la cuenta.

- Year: año fiscal.
- Leads: Número de oportunidades generadas por la cuenta durante el año fiscal correspondiente.
- ConvertedLeads: Número de oportunidades cerradas para la cuenta durante el año fiscal correspondiente.
- NEventOffline: Número de eventos presenciales en los que partició la cuenta clave durante el año fiscal.
- NeventOnline: Número de eventos digitales en los que partició la cuenta clave durante el año fiscal.
- NEmployeeAccess: Número de empleados de la cuenta que ha tenido al menos algún acceso o consumo de contenido digital durante el año fiscal.
- NAccessMean: Número promedio de accesos y consumos de contenido digital para los empleados de la cuenta que interactúan durante el año fiscal.

1. Proponga un modelo de conteo (debe definir la variable dependiente utilizando la base de datos disponible, junto a sus índices) para describir el número de oportunidades ganadas mediante un modelo que integre heterogeneidad observable. Escriba la log-verosimilitud del modelo anterior y especifique cuántos parámetros deben estimarse. Si necesita, suponga conocida la cardinalidad de todos los índices del modelo y que hay un año de desfase, por lo que debe considerar el t anterior. Para la construcción del modelo considere:

    a. La industria y la zona a la que pertenecen cada cuenta incide en la posibilidad de generar y convertir más o menos oportunidades.
    b. El tiempo de relación afecta las oportunidades ganadas. El primer año de relación, en general, hay un numero grande de oportunidades, el que cae marcadamente el segundo año para repuntar y mantenerse relativamente estable a partir del tercer año.
    c. Se espera que la participación en eventos presenciales y digitales tienen un efecto en conversión de oportunidades, pero se hipotetiza que estos dos canales son sustitutos entre sí. Considere que los datos tienen un año de desfase, por lo que debe considerar el período anterior.
    d. El número total de interacciones de los empleados debiera aumentar las oportunidades convertidas, pero el efecto adicional de cada nueva interacción tiende a ser menor a medida que el número de interacción crece.
    e. Los procesos de conversión son relativamente lentos, por lo que las oportunidades convertidas dependen principalmente de la actividad del año anterior y no del año en curso. Considere que los datos tienen un año de desfase, por lo que debe considerar el período anterior.

2. Usando los mismos supuestos del modelo anterior, construya un modelo de elección binaria con heterogeneidad observable para describir el número de oportunidades ganadas. Escriba la log-verosimilitud y especifique cuántos parámetros más deben estimarse con respecto al modelo anterior. ¿Qué ventajas puede tener un modelo de elección binario por sobre no de conteo (aplicado a este problema)? ***Hint:*** *Si le resulta conveniente en su notación, puede asumir que la tasa del modelo de conteo que captura todas las componentes relevantes del problema se denota* $\lambda_{it}(x_{it})$*, con* x_{it} *un conjunto de variables observables de la cuenta* i *en el año* t*.*

3. Considerando que, quizás, la heterogeneidad observable no es suficiente para capturar toda la variabilidad del fenómeno, se consideró extender el modelo para incluir heterogeneidad no observable. Aunque en este nuevo modelo el conjunto de variables explicativas es algo distinto al del modelo sin heterogeneidad no observable, el ajuste en términos de la log-verosimilitud es bastante mejor, como indica la **??**.

Table 3.2: Log-verosimilitud modelos alternativos. {#tbl:loglik}

Modelo	Log-Likelihood
Regresión Binomial	-10.145,0
Regresión Beta-Binomial	-9.855,4

En relación al modelo binomial, ¿cuántos parámetros debiera tener el modelo de regresión beta-binomial para que sea preferido en términos de AIC?

Solución

1. De acuerdo a las solicitudes:

   a. Necesitamos incluir dummies para zonas (Zone_{ih_1}) y para industrias (Industry_{ih_2}), donde h_1 corresponde a una zona existente de la base de datos, al igual que h_2, que corresponde a una industria existente.

   b. Se pueden generar dos variables para los niveles de antigüedad (porque una de las tres debe ser la de referencia). Se genera Tenure2_{it}, que es equivalente a $\text{Tenure}_{it}\mathbf{1}\begin{cases}1 & \text{si } t=2\\ 0 & \text{en otro caso}\end{cases}$ y Tenure3_{it}, equivalente a $\text{Tenure}_{it}\mathbf{1}\begin{cases}1 & \text{si } t\geq 3\\ 0 & \text{en otro caso}\end{cases}$.

   c. Hay que incluir el número de eventos físicos, digitales y su interacción.

   d. El número total de interacciones es la multiplicación del número de empleados por el promedio de interacciones por usuario. Dado los

rendimientos decrecientes, se debe transformar la variable a logaritmo.

$$\begin{aligned}\lambda_{it} = &(\sum_{h_1} \alpha_{h_1} \text{Zone}_{ih_1}) + (\sum_{h_2} \alpha_{h_2} \text{Industry}_{ih_2}) + \beta_1 Tenure2_{it} + \beta_2 Tenure3_{it} \\ &+ \beta_3 NEventOnline_{it-1} + \beta_4 NEventOffline_{it-1} \\ &+ \beta_5 NEventOnline_{it-1} NEventOffline_{it-1} \\ &+ \beta_6 ln(NEmployeeAccess_{it-1} NAccessMean_{it-1})\end{aligned}$$

Este modelo tiene $Z + I + 6$ parámetros a estimar, donde $Z = \sum_{h_1}$ e $I = \sum_{h_2}$.
La log-verosimilitud resulta de aplicar una distribución de Poisson:

$$LL = \sum_i \sum_t n_{it} \ln \left(\frac{(\lambda_{it} t)^{x_{it}} e^{\lambda_{it} t}}{x_{it}!} \right)$$

Donde x_{it} es el número de oportunidades convertidas por la cuenta i en el año t y n_{it} es el número de individuos que tuvieron esas x_{it} oportunidades convertidas. Acá la heterogeneidad observada fue incluida en la regresión de Poisson descrita.

2. El modelo es idéntico al anterior, pero en vez de describir la verosimilitud como un modelo de Poisson, ahora es un modelo binomial. Entonces:

$$LL = \sum_i \sum_t \binom{m_{it}}{n_{it}} p_{it}^{n_{it}} (1 - p_{it})^{m_{it} - n_{it}}$$

Aquí, m_{it} es el número de *leads* generados por la cuenta en el año y la probabilidad de conversión heterogenea viene dada por

$$p_{it} = \frac{e^{\lambda_{it}(x_{it})}}{1 + e^{\lambda_{it}(x_{it})}}$$

El cual tiene el mismo número de parámetros que el caso anterior.

3. Para preferir el modelo beta binomial, el AIC debiera ser menor que el de la regresión binomial. Entonces:

$$(AIC_B = 2k_B - 2LL_B) > (2k_{BB} - 2LL_{BB} = AIC_{BB})$$

Despejando y reemplazando los valores, $k_{BB} - k_B < 290$ (es decir, el modelo de regresión beta-binomial podría tener hasta 290 parámetros más y seguiría siendo más preferido que el modelo de regresión de Poisson.

3.4.3 Problema 3

Harcor es una importante firma latinoamericana que se concentra en la producción y venta de caramelos y snacks. Aunque la compañía tiene una estrategia multicanal, gran parte de sus ventas se concentra en el canal tradicional de almacenes y kioscos. Este segmento es muy atomizado con decenas de puntos de venta en el país y con una gran heterogeneidad entre los almaceneros. Mientras algunos almacenes tienen un desarrollo bien establecido con tiendas bien organizadas y sistemas automáticos de registros de venta, otros operan en espacios informales y con malas prácticas de gestión.

Para incentivar las ventas en el canal tradicional, Harcor ha creado un programa de capacitación que ha denominado *El Club del Almacenero*, que junto con proveer soporte en infraestructura de góndolas y estanterías con las marcas de Harcor, ofrece también un plan de capacitación con las mejores prácticas del comercio minorista, incluyendo temas de publicidad, administración de inventarios, control de costos y planificación de surtido. El plan de capacitación está inicialmente pensado en 4 niveles de módulos, desde el más básico hasta al más avanzado, de modo que para acceder a un nivel debe haberse completado el nivel anterior.

Aunque el nivel de satisfacción de los participantes del programa de capacitación es alto, el porcentaje de almaceneros que participa y completa el programa ha sido más bien bajo en los primeros dos años de operación por lo que Harcor está considerando hacer algunos ajustes al programa. Para que los ajustes sean realizados basados en evidencia, el equipo de analítica comercial se ha propuesto analizar los datos de participación en las ediciones pasadas del programa. Para eso, a partir de los registros históricos se ha generado una tabla, que, junto con un conjunto de variables demográficas de los almaceneros, se registra x_{ikt} que toma el valor 1 si el almacenero i participó en el nivel k en el programa de capacitación realizado en el año t.

1) Suponga que el número de niveles de los cuales participa un almacenero queda descrito por un modelo beta geométrico desplazado de parámetros α y β.

    1) (1 punto) Calcule una expresión para la probabilidad que un almacenero termine el programa de capacitación. Notar que la expresión debe poder calcularse a partir de los datos observados x_{ikt} y los parámetros α y β.
    2) (1 punto) Escriba la log-verosimilitud del problema. Al igual que en el caso anterior, exprese la log-verosimilitud en términos de x_{ikt}, α y β.

2) Suponga que se estima el modelo usando el método de la máxima verosimilitud resultando los valores de la Tabla:

Table 3.3: **Tabla:** Resumen de ventas para cada función.

Variable	Coeficiente
α	0.999
β	4.001
N	1,245
LL	-2.583

a) La Figura reporta la distribución beta evaluada en los estimadores máximos verosímiles de la Tabla. Por inspección de la figura, provea una aproximación numérica a la probabilidad de que un cliente no avance al siguiente nivel sea mayor que 0.8 (usando áreas de triángulos y rectángulos).

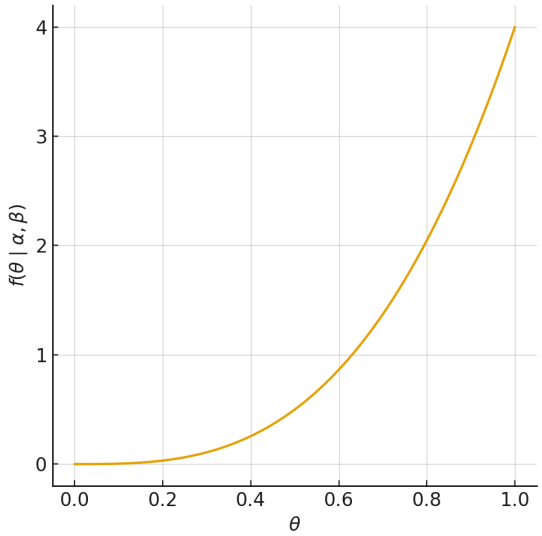


Figure 3.1: **Figura:** Distribución posterior de la probabilidad de no continuar

b) Calcule el intervalo de confianza al 95% del segundo parámetro β. Considere que el inverso del hessiano de la log-verosimilitud, evaluada en el estimador máximo verosímil, viene dado por Σ^{-1}:

$$\Sigma^{-1} = \begin{bmatrix} 0.16 & 0.04 \\ 0.04 & 0.36 \end{bmatrix}.$$

3) Suponga ahora que se rediseña el programa de modo que los contenidos de un módulo son independientes de los otros y, por tanto, un almacenero puede atender cualquier módulo sin necesidad de haber completado los anteriores. Suponga que ya se han desarrollado tres módulos y falta

uno por realizar. Escriba la verosimilitud de observar que un almacenero atiende a los tres módulos y otro atiende solo a uno, suponiendo que la asistencia se describe por un modelo beta-binomial.

4) Si los estimadores máximos verosímiles de la beta binomial son $\alpha = 3.543$ y $\beta = 1.231$, escriba una expresión para el valor esperado de la probabilidad que un almacenero atienda al último módulo, si ya atendió a los tres primeros.

Solución

1) Sea D_{it} el número de niveles cursado por el almacenero i en el año de ejecución t y sea s_{ikt} una variable binaria que toma el valor 1 si el almacenero cumple el curso $s_{ikt} = \begin{cases} \mathbb{1} & \text{si } \sum_h x_{iht} = 4 \\ 0 & \text{en otro caso} \end{cases}$.

   1) La expresión es directa:

$$\Pr(D_{it} = k \mid \alpha, \beta) = \frac{B(\alpha + 1,\ \beta + k - 1)}{B(\alpha, \beta)}$$

   2) Usando las definiciones de arriba, la log-verosimilitud:

$$LL = \sum_t \sum_i \sum_k s_{ikt} \log\Big(\mathbb{P}(D_{it} = 4 \mid \alpha, \beta)\Big) = \sum_t \sum_i \sum_k s_{ikt} \log\left(\frac{B(\alpha + 1,\ \beta + k - 1)}{B(\alpha, \beta)}\right)$$

   O bien, se puede absorber la suma para cada individuo i si $n_{kt} = \sum_i s_{ikt}$, correspondiente al número de almaceneros que cursaron k niveles. En dicho caso:

$$LL = \sum_t \sum_k n_{kt} \log\Big(\mathbb{P}(D_{it} = k \mid \alpha, \beta)\Big) = \sum_t \sum_k n_{kt} \log\left(\frac{B(\alpha + 1,\ \beta + k - 1)}{B(\alpha, \beta)}\right)$$

2) Para el primer caso:

   a) La probabilidad p corresponde al área bajo la curva, que puede aproximarse como

$$p \approx 0.2 \cdot 2 + 0.2 \cdot 2/2 = 0.6.$$

   b. Usando el inverso del hessiano, el ientervalo de confianza viene dado por $4.001 \pm 1.96\sqrt{0.36}$, donde el coeficiente β corresponde la componente Σ^{-1}_{22}.

3) Solo tenemos que multiplicar la probabilidad de los dos eventos:

$$\begin{aligned} L(\theta) &= \mathbb{P}(D_{it} = 3 \mid \alpha, \beta)\mathbb{P}(D_{it} = 3 \mid \alpha, \beta) \\ &= \binom{3}{3} \frac{\mathbb{B}(\alpha + 3, \beta)}{\mathbb{B}(\alpha, \beta)} \binom{3}{1} \frac{\mathbb{B}(\alpha + 1, \beta + 2)}{\mathbb{B}(\alpha, \beta)} \end{aligned}$$

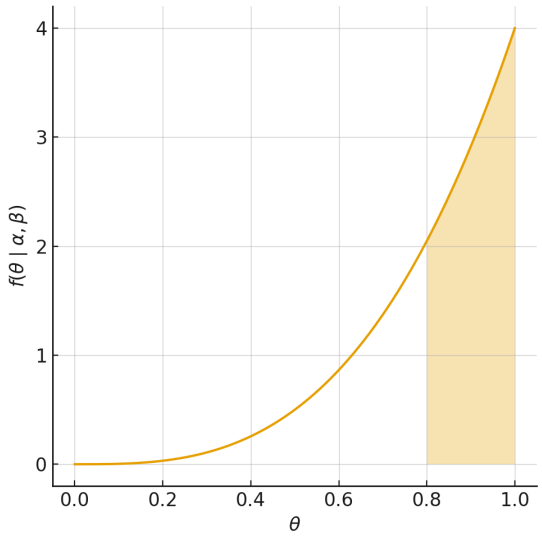


Figure 3.2: **Figura:** Distribución posterior con área sombreada

4) La probabilidad solicitada es simplemente la esperanza condicional del parámetro. En este caso, la distribución condicional de observar tres visitas positivas en tres módulos es $\mathbb{B}(\alpha+3,\beta)$. La esperanza condicional de la probabilidad viene dada por

$$\frac{\alpha+3}{\alpha+\beta+3}=0.842$$

3.4.4 Problema 4

En 1972, la *National Association of Food Chains* (NAFC) impulsó el uso de códigos de barras para administrar los precios en las góndolas de los supermercados en Estados Unidos. Desde aquel entonces, ha existido una continua preocupación respecto a la precisión de estos mecanismos para mantener información precisa respecto de los precios. Uno de los problemas potenciales más relevantes del sistema es que los precios cargados en el sistema podrían no estar bien coordinados con la información de precios desplegada a los clientes en los flejes de las góndolas. En simple, como la información de precios se realiza manualmente, es posible que se generen diferencias entre lo que se despliega de cara al cliente al momento de la elección, con lo que efectivamente se cobra en caja.

Motivados por este problema, se ha decidido investigar la situación actual de la precisión con que se despliegan los precios en la industria supermercadista chilena. Más allá de la existencia de discrepancias, nos interesa investigar la

duración de estos eventos en que el precio informado difiere del precio efectivamente cobrado. Para investigar estas preguntas, se ha construido una base de datos que contiene alrededor de 75.000 observaciones de precios realizadas durante 28 días en una sala de supermercado de la Región Metropolitana, que considera tanto el precio registrado en el sistema de cajas como el precio exhibido en góndola. Mientras que los precios de cajas se obtienen directamente de los sistemas transaccionales de la compañía, los precios en góndola resultan de una inspección visual en sala, por un equipo de revisores entrenados para ese propósito. La medición visual se hace sobre una lista predefinida de productos sobre los que se registra el precio exhibido.

Habiendo realizado la medición, se identifica que más de un 10% de los precios registrados tiene asociada alguna discrepancia, es decir, que el precio cobrado en caja no coincide con el precio exhibido en el fleje de la góndola. Notar que las discrepancias pueden ser favorables para el cliente o favorables para el supermercado. Por convención, llamaremos discrepancias de tipo I a aquellas que favorecen al cliente y de tipo II a las que favorecen al supermercado. Para entender con más detalle este fenómeno de discrepancias, se ha seleccionado una base con todos los registros con precios discrepantes. En esta base, cada fila se compone de las siguientes variables:

- **Id**: Identificador de un evento de discrepancia de precios.
- **Day**: Día de la semana que se registra la discrepancia (con lunes = 1 y domingo = 7).
- **Type**: Variable categórica con el tipo de discrepancia (Tipo I o Tipo II).
- **Section**: Categoría del supermercado a la que pertenece el producto que presenta un diferencial (Abarrotes, Limpieza, Bebidas, Galletas y Cocktail).
- **Duration**: Número de días que dura la discrepancia de precios entre el fleje y el sistema transaccional.
- **Tag_price**: Precio que aparece en el fleje.
- **Real_price**: Precio cobrado en caja.
- **diff**: Diferencia entre precio fleje y precio real.

Para ilustrar la estructura de la base, las primeras entradas se presentan en la Tabla 1:

Id	Day	Type	Section	Duration	Tag_price	Real_price	diff
1	1	I	limpieza	1	1668	1467	201
2	1	II	abarrotes	7	1162	9183	-8021
3	2	I	bebidas	1	7517	6616	901
4	1	I	bebidas	1	1319	1160	158
5	5	I	galletas	3	1411	1241	169
6	3	II	galletas	2	1411	1641	-231

Tabla: Muestra de los primeros 6 registros de la base de discrepancias

1. Describa un modelo de duración probabilístico que permita describir cuántos días se mantienen las discrepancias y escriba la log-verosimilitud correspondiente. Defina cuidadosamente las variables de la función, especificando qué es parámetro y qué es dato. Para los datos, indique cómo se obtienen de los datos ilustrados en la Tabla 1.

2. Suponga que complementario a los modelos de duración, se estiman dos modelos de regresión lineal para describir la duración de las discrepancias como variable dependiente. Los resultados de estos dos modelos se reportan en la Tabla 2.

Variable	Coef. M1	p-valor M1	Coef. M2	p-valor M2
intercepto	3.55	0.0901	-1.353	0.1607
diff	-0.0002	0.7830	–	–
log(Tag_price)	-0.2688	0.3661	–	–
log(diff)	–	–	0.563	0.0019
type.I	1.51	0.0210	2.168	0.0000
type.I:diff	0.0046	0.1165	–	–
day.1	–	–	-1.013	0.0061
section.limpieza	–	–	4.394	0.0000
section.cocktail	–	–	-1.329	0.0221
R^2	**0.096**		**0.136**	
Adj. R^2	**0.086**		**0.120**	

Tabla: Resultados de dos modelos de regresión lineal para la duración de la discrepancia.

Más allá de la varianza explicada por cada modelo, ¿qué modelo le parece más adecuado? Provea al menos dos razones por las cuales prefiere el modelo elegido.

3. Suponga que ahora se estiman tres modelos para describir la duración de discrepancias tipo I (aquellas que favorecen al cliente): un modelo geométrico desplazado sin heterogeneidad, un modelo beta-geométrico desplazado y un modelo de regresión geométrico desplazado en que el parámetro del modelo se asume dependiendo de la magnitud de la discrepancia (Diff) y si corresponde a la categoría de limpieza. Para garantizar que el parámetro θ esté en el rango $[0, 1]$ para incluir variables explicativas se aplica la transformación

$$\theta_i \;=\; \frac{\exp(\beta \cdot x_i)}{1 + \exp(\beta \cdot x_i)}.$$

Los estimadores máximo-verosímiles se muestran a continuación. Elija dos de los modelos y, para cada uno, calcule la probabilidad de que una discrepancia para un producto de limpieza, cuya diferencia de precios es 201, dure más de 3 días (deje la expresión simbólica; no es necesario valor numérico).

Parámetros	Geométrico desplazado	Beta geométrico	Regresión geométrica
θ_0	0.33	–	0.46
α	–	2.08	–
β	–	2.62	–
Diff	–	–	-0.43
limpieza	–	–	-0.87
LL	-788	-761	-566
AIC	1278	1526	1142

Tabla: Estimadores y criterios de información de tres modelos de duración.

Solución

1. Dado que la duración la observamos en días, parece más natural considerar un modelo de duración en tiempo discreto (la muestra de datos también sugiere que el proceso podría ser bastante discreto). Sea:

   - θ = Probabilidad de que un producto deje de estar en discrepancia en un día dado.
   - T_i = la duración de la discrepancia de la observación i.
   - τ = el máximo período de observación (28 días).
   - n_t = el número de casos en que observamos que la discrepancia dura t días.

   Entonces, la log-verosimilitud viene dada por:

   $$LL \;=\; n_\tau \,\ln((1-\theta)^\tau) \;+\; \sum_{t=1}^{\tau-1} n_t \,\ln((1-\theta)^{t-1}\,\theta)$$

   Acá, θ es el único parámetro del modelo. Los datos son n_t y n_τ y se obtienen directamente a través de un conteo condicional en los valores de la columna Duration.

2. Hay varias consideraciones que sugieren que el **Modelo 2** podría ser preferible.

- `diff` no es significativo en modelo 1, pero `log(diff)` sí lo es en modelo 2.
- El p-valor del tipo es menor en el modelo 2.
- La interacción que agrega el modelo 1 (`type.I × diff`) no es significativa.
- Las variables adicionales del modelo 2 sí son significativas.

3. Las expresiones se derivan directamente de las definiciones y vienen dadas por:

   - **Geométrica:** $\Pr(T > 3 \mid \theta_0) = (1 - 0.33)^3$.
   - **Beta-Geométrica:**

   $$\Pr(T > 3 \mid \alpha, \beta) \;=\; \frac{B(2.08,\ 2.62 + 3)}{B(2.08,\ 2.62)}.$$

   - **Regresión Geométrica:**

   $$\Pr(T > 3 \mid \theta_0, \theta_{\text{Diff}}, \theta_{\text{lim}}) \;=\; \left(\frac{1}{1 + \exp(0.46 + 0.43 \cdot 201 + 0.87)}\right)^3.$$

3.4.5 Problema 5

Se acerca el verano (sí, hay que usar la imaginación) y tu última práctica profesional. Como has sido una estudiante excepcional, pudiste elegir con cuidado dónde realizarla y después de descartar varias ofertas, decidiste aceptar la oferta de NotComida, una startup chilena exitosa que basa su propuesta de valor en el diseño y manufactura de productos alimenticios basados en plantas que funcionan como sucedáneos cercanos a otros productos de origen animal, como la leche, el helado o las hamburguesas. Cada vez que te piensas en lo que viene, se te escapa un suspiro de satisfacción. Piensas en lo terrible que sería para ti trabajar en una compañía que no esté alineada con tus ideales. A fin de cuentas, entraste a estudiar ingeniería para cambiar el mundo. Para ti este es solo el primer paso.

Todavía quedan dos semanas para que empiece la práctica, pero ya estás ansiosa, por lo que recibes con entusiasmo y sorpresa un correo electrónico de Kiara, quien será tu supervisora directa durante tu estadía en la compañía. *"¡Hola! Esperando que estés muy bien, te escribo para contarte los próximos desafíos en los que vamos a estar trabajando en el equipo en los próximos seis meses. Por supuesto que no es necesario que hagas absolutamente nada, pero como queremos que te sientas integrada te vamos a ir copiando en las comunicaciones internas del equipo".* Antes de cerrar el mensaje, te das el tiempo de darle un vistazo rápido al diccionario de datos que viene adjunto:

Variable	Descripción	Media
IdCliente	Identificador de cliente	-
Mes	Mes {enero, febrero, ..., diciembre}	-
Año	Año	-
Categoría	Categoría de producto {Hamburguesas, Leche, Helados}	-
Tienda	Identificador de la tienda dónde se realizó la compra	-
Edad	Edad del panelista	31.3
Mujer	1 si la panelista se identifica como mujer	0.51
Ingreso	Ingreso mensual declarado por el panelista [millones CLP]	0.56
Zona Oriente	1 si el panelista vive en la zona oriente de Santiago	0.11
NCompras	Número de compras en esa categoría y mes	2.34

Tabla 1: Diccionario de datos del panel (variables y media cuando corresponde).

Tienes súper claro que, de verdad, no tienes que hacer nada. Sin embargo, te dan unas ganas enormes de empezar ya a pensar en cómo puedes contribuir. El equipo se ha embarcado en una serie de proyectos para entender el comportamiento de compra de los clientes y tú sabes que tienes las herramientas para hacer una contribución. Aunque no estás tan segura de en qué subproyecto te quieres involucrar, ya has delineado un plan bastante claro de lo que te gustaría hacer. Considerando que la fuente de datos principal es un panel de compras con el número de unidades y montos comprados por cliente, se te ocurre que tu primer modelo puede estar basado en un modelo de conteo.

1. Proponga un modelo de regresión de Poisson en que la tasa de compra depende de cada individuo, categoría, tienda y semana. Escriba la log-verosimilitud del problema considerando que n_{icst} es el número de unidades compradas por el cliente i en la categoría c en la tienda s el mes t. Para esto considere que:

    a. Existe un efecto fijo por categoría y uno por tienda.
    b. El género induce diferencias relevantes en las ventas y el efecto del género es distinto para la zona oriente de la capital.
    c. Existe una tendencia creciente en las ventas. Como el crecimiento es relativamente pequeño, puede modelarse con un efecto fijo a nivel de año.
    d. Aunque en el consumo es relativamente constante a lo largo del año, hay dos categorías que presentan una estacionalidad relevante. Mientras los helados tienen un patrón marcadamente diferente en los meses de diciembre, enero y febrero, las hamburguesas declinan sus ventas en el mes de septiembre.

2. La investigación exploratoria de los datos sugiere que, restringiendo el análisis a la categoría $c =$ Hamburguesas, podría haber dos clases latentes. Un segmento de alto consumo y otro de consumo más bien ocasional. Para acomodar esta regularidad se propone un modelo sencillo en que la tasa de compra de un cliente i, independiente del mes, viene dado por:

$$\log(\lambda_i) \;=\; \beta_{0i} + \beta_1 \text{MUJER}_i$$

Donde β_{0i} toma el valor -1 con probabilidad 0.4 y el valor 0.5 con probabilidad 0.6. ¿Cuál es la probabilidad de que un cliente hombre, con MM$1 de ingreso que no compra en un mes ($n_i = 1$) pertenezca al segmento 1? Deje computada la probabilidad.

3. Después de revisar los resultados de los modelos anteriores, te preguntas si dado el bajo número de compras registrados cada mes, quizás podría existir un modelo alternativo que pudiera complementar los aprendizajes anteriores. Rápidamente se te ocurre que quizás puedes definir una variable y_{icst} que toma el valor 1 si el cliente i compra en la categoría c en la tienda s el mes t (sí o no). Con esto, te resulta evidente que puedes estimar un modelo de elección discreta que describa la decisión de comprar o no de los clientes de la base.

    a. Considerando los datos de la Tabla 2, que describen todas las compras observadas por el cliente 321045, escriba la log-verosimilitud del historial de compra de este cliente.

IdCliente	Mes	Año	Tienda	Categoría	Edad	Mujer	Ingresos	Zona Oriente	N
321045	Julio	2021	S816	Hamburguesa	26	0	1.3	0	
321045	Septiembre	2021	S816	Hamburguesa	26	0	1.3	0	

Tabla 2: Historial resumido de compras del cliente 321045.

Solución

1. En un modelo de regresión de Poisson, la distribución ya viene dada y solo tenemos que especificar la tasa del proceso. Sea λ_{icst} la tasa de compras del cliente i en la categoría c en la tienda s el mes t. Además, definamos:

    - $\delta_{ta} = 1$ si el mes t pertenece al año a.
    - $\Delta^1_{ct} = 1$ si $c =$ Helado y $t \in \{$diciembre, enero, febrero$\}$.
    - $\Delta^2_{ct} = 1$ si $c =$ Hamburguesa y $t =$ septiembre.

Entonces, la tasa de compra puede escribirse como

$$\lambda_{icst} = \alpha_c + \alpha_s + \beta_1 \,\text{MUJER}_i + \beta_2 \,\text{MUJER}_i \cdot \text{ZORIENTE}_s + \sum_a \gamma_a \,\delta_{ta} + \theta_1 \,\Delta^1_{ct} + \theta_2 \,\Delta^2_{ct}.$$

Tanto $\Delta^1_{ct} = 1$ como $\Delta^2_{ct} = 1$ pueden ser obtenidos usando una interacción entre Categoría, Tiempo y una indicatriz, que indique $\mathbb{1}$ cuando t pertenece al mes indicado, y 0 en caso contrario.

Con esto, la **log-verosimilitud** viene dada por:

$$LL = \sum_{i,c,s,t} \ln\big(\,\Pr(N_{icst} = n_{icst})\big) = \sum_{i,c,s,t} \ln\left(\frac{\lambda_{icst}^{\,n_{icst}}\; e^{-\lambda_{icst}}}{n_{icst}!}\right).$$

2. Las tasas de compra de los dos segmentos vienen dadas por:

$$log(\lambda_{i1}) = -1 + \beta_1 \text{MUJER}_i = -1 + 0 \implies \lambda_{i1} = e^{-1}$$

$$log(\lambda_{i2}) = 0.5 + \beta_1 \text{MUJER}_i = 0.5 + 0 \implies \lambda_{i1} = e^{-0.5}$$

Por tanto, la probabilidad de comprar una vez en cada segmento es:

$$\mathbb{P}(n_i = 1 \mid s_1) = \lambda_{i1} e^{-\lambda_{i1}} = e^{-1} e^{-e^{-1}}$$

$$\mathbb{P}(n_i = 0 \mid s_2) = \lambda_{i2} e^{-\lambda_{i2}} = e^{-0.5} e^{-e^{-0.5}}$$

Aplicando Teorema de Bayes:

$$\mathbb{P}(s_1 \mid n_i = 1) = \frac{\mathbb{P}(n_i = 0 \mid s_1)\mathbb{P}(s_1)}{\mathbb{P}(s_1 \mid n_i = 0)\mathbb{P}(s_1) + \mathbb{P}(s_2 \mid n_i = 0)\mathbb{P}(s_2)}$$

Donde $\mathbb{P}(s_1) = 0.4$ y $\mathbb{P}(s_2) = 0.6$.

3. La utilidad que deriva el cliente i por comprar en la categoría c en la tienda s en el mes t viene dado por:

$$u_{icst} = \alpha_c + \alpha_c + (\beta_1 \text{MUJER}_i = 0) + (\beta_2 \text{MUJER}_i \text{ZORIENTE}_s = 0) + \sum_a \gamma_a \delta_{ta} + (\theta_1 \Delta^1_{ct} = 0) + \theta_2 \Delta^2_{ct}$$

Donde $\text{MUJER}_i = 0$ porque el registro es de un hombre, y $\theta_1 \Delta^1_{ct} = 0$ porque aplica para helado y en los meses de verano, no septiembre y hamburguesa. Con esto, la log verosimilitud es:

$$LL(\hat{\theta}) = \sum_{icst} y_{icst} ln\left(\frac{e^{u_{icst}}}{1 + e^{u_{icst}}}\right) + (1 - y_{icst}) ln\left(\frac{1}{1 + e^{u_{icst}}}\right)$$

3.4.6 Problema 6

En los últimos 20 años, hemos visto un crecimiento explosivo en el uso de redes sociales, con más de 4.4 billones de usuarios a nivel mundial y con un promedio de más de 8 cuentas por usuario en las distintas plataformas que han surgido en los últimos años. En este ecosistema, hay un creciente interés por convertirse en *creador de contenido*, que les permita generar audiencia que pueda eventualmente ser monetizada. Cualquier persona con un dispositivo móvil puede crear contenido que potencialmente puede volverse viral y captar niveles importantes de atención de los usuarios. Ya sea tocando la guitarra y cantando, jugando videojuegos o mostrando la ejecución de sofisticadas recetas de cocina, muchos creadores buscan alcanzar altos niveles de visitas y generar interacciones significativas con sus audiencias.

Considerando que existen un alto potencial comercial en la industria de generación de contenido, se ha propuesto investigar en profundidad cuáles son los factores que inciden en el nivel de interacciones entre creadores y usuarios. Para estos efectos, se ha aliado con un nuevo portal en que creadores de contenido pueden subir su material, el que es exhibido a los usuarios de la plataforma usando un algoritmo interno que personaliza el contenido de acuerdo con la probabilidad de generar interacciones. Aunque hay varios niveles de interacción con el contenido (ver, comentar, compartir), la plataforma considera que el número de *me gusta* es la principal métrica para capturar el éxito relativo de una publicación.

El portal tiene un módulo de analítica de datos que permite recolectar algunas métricas claves de cada uno de los contenidos publicados en su primer año de operación para todos los creadores registrados en el portal. Entre los datos disponibles se encuentran las llaves para identificar cada contenido y algunas características como la categoría del contenido (información, humor o noticia), el número de caracteres, el número de imágenes y la duración del video, si existiese. Adicionalmente, y gracias a una alianza con una herramienta de inteligencia artificial, se cuenta con el puntaje de cada contenido en tres dimensiones distintos: producción, contenido y novedad.

Variable	Descripción	Media
IdCC	Identificador de cliente	-
IdPost	Identificador del contenido	-
DatePost	Fecha de publicación del contenido	-
Category	Categoría de producto {Information, Humour, News}	-
Nchar	Número de caracteres en el texto del contenido publicado	84.4
Npic	Número de fotos en el contenido publicado	0.4
Tvid	Duración [seg] de los videos que acompañan el contenido	18.5
AIproduction	AI score en la calidad de la producción [0,1]	0.5
AIcontent	AI score en la calidad del contenido [0,1]	0.3
AInovelty	AI score en la novedad del contenido [0,1]	0.2

Variable	Descripción	Media
Nlikes	Número de *me gusta* que ha recibido la publicación	12.5

Tabla 1: Descripción base de contenidos en la plataforma de creadores.

1. Modele el número de *me gusta* que recibe una publicación como un proceso de conteo con heterogeneidad observable. Para este modelo:

    a. Justifique la inclusión de al menos dos variables explicativas indicando cuál es el signo esperado del coeficiente correspondiente.
    b. Escriba la log-verosimilitud del problema. Indique cuáles son los parámetros para estimar y explicite cuántos parámetros tiene su modelo.
    c. Suponiendo que ha estimado el modelo y los parámetros proveen una buena descripción de la distribución del número de *me gusta* que recibirá un post. Usando dichos estimadores, escriba una expresión para la probabilidad de que un cliente reciba al menos un *me gusta*.

2. Una de las limitaciones del modelo anterior es que el uso del algoritmo interno de la plataforma implica una distribución muy heterogénea de exposiciones. Esto es, algunas publicaciones son desplegadas miles de veces, mientras que otras reciben muy poca exposición. Para hacer frente a este problema, se propone usar un modelo de elección.

    a. Describa el proceso como un proceso de elección binaria con heterogeneidad no observable y escriba la log-verosimilitud del problema. Para eso, suponga que la base de datos también contiene una variable $Ndisp$ que, para cada post, indica el número de veces que fue expuesto.
    b. Si al estimar el modelo, se encuentra que $\alpha = 1$ y $\beta = 2$ son los estimadores máximo-verosímiles. En promedio, ¿con qué probabilidad a un cliente le gusta una publicación en la plataforma?
    c. Considere una publicación que ha sido desplegada 100 veces y ha recibido 10 *me gusta*. Considerando los valores de $(\alpha, \beta) = (1, 2)$, de la parte anterior, ¿cuántos *me gusta* esperaría observar luego de mostrar el contenido 200 veces?

Solución

1. Se tiene:

    a. Cualquiera de las variables listadas puede justificarse de manera más o menos directa:

- **Efecto fijo por creador:** hay creadores más populares que otros.
- **Efecto fijo por fecha:** hay fechas en que los usuarios son más activos.
- **Dummies por categoría:** hay categorías que pueden generar más *me gusta*.
- **Nchar:** textos más largos o más cortos pueden ser más atractivos.
- **Npic:** agregar más fotos puede ser más atractivo.
- **Tvid:** videos más largos o más cortos pueden ser preferidos.
- **AIproduction:** material mejor producido puede generar más *me gusta*.
- **AIcontent:** contenido de mejor calidad puede generar más *me gusta*.
- **AInovelty:** material más novedoso puede generar más *me gusta*.

b. Aunque cualquier combinación de las variables arriba descritas es admisible, para efectos ilustrativos consideraremos un modelo en que la tasa de *me gusta* del contenido i viene dada por:

$$\lambda_i \;=\; \lambda_0 \cdot \exp(\beta_1 \,\mathrm{Nchar}_i + \beta_2 \,\mathrm{Npic}_i + \beta_3 \,\mathrm{Tvid}_i).$$

Con esto, la log-verosimilitud viene dada por:

$$LL \;=\; \sum_i \ln\left(\frac{\lambda_i^{\mathrm{Nlikes}_i}\, e^{-\lambda_i}}{\mathrm{Nlikes}_i!} \right).$$

c. La probabilidad de recibir al menos un *me gusta* viene dada por:

$$\Pr(\mathrm{Nlikes}_i > 0 \mid \lambda_0, \beta) \;=\; 1 - \Pr(\mathrm{Nlikes}_i = 0 \mid \lambda_0, \beta) \;=\; 1 - e^{-\lambda_i}.$$

2. El modelo es una beta-binomial estándar. Recordar que la esperanza de una $\mathrm{Beta}(\alpha, \beta)$ es $\alpha/(\alpha+\beta)$.

a. La log-verosimilitud de una beta-binomial:

$$LL \;=\; \sum_i \ln\left(\binom{\mathrm{NDisp}_i}{\mathrm{Nlikes}_i} \frac{B(\alpha + \mathrm{Nlikes}_i,\; \beta + \mathrm{NDisp}_i - \mathrm{Nlikes}_i)}{B(\alpha, \beta)} \right).$$

b.

$$\mathbb{E}(\theta \mid \alpha, \beta) \;=\; \frac{\alpha}{\alpha + \beta} \;=\; \frac{1}{3}.$$

c. Se tiene que despejar la esperanza condicional

$$\mathbb{E}(\theta_i \mid \alpha, \beta, \mathrm{NDisp}_i, \mathrm{Nlikes}_i) = \frac{\alpha + \mathrm{Nlikes}_i}{\alpha + \beta + \mathrm{NDisp}_i} = \frac{11}{103}.$$

Entonces, el número esperado de *me gusta* al mostrar el contenido 200 veces es:

$$200 \times \frac{11}{103} \approx 21.36.$$

Notar que $\frac{11}{103}$ es el número esperado de likes de una persona, dada su verosimilitud (información previa) por su comportamiento esperado (distribución beta). Por eso, para 200 personas simplemente se multiplica.

3.4.7 Problema 7

Hippi es una empresa nacional dedicada a la fabricación de productos para alta montaña. La empresa cuenta con una base de datos que registra la compras que sus clientes en cada temporada y busca estudiar el numero de parkas que los clientes comprarán en la temporada. Para esto, se ha formulado un modelo probabilístico donde el comportamiento de los clientes se describe como:

- Cada cliente compra un numero de parkas X que sigue una distribución geométrica. Esto es, condicional en el parámetro p de cada cliente, la función de probabilidad esta dada por:

$$Pr(X = x|p) = p(1-p)^x$$

- Los clientes son heterogéneos en su parámetro p el que esta distribuido en la población de acuerdo a una distribución Beta:

$$g(p|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

Se puede mostrar que la distribución del número de compras para un cliente en la población, y su respectiva esperanza están dadas por:

$$Pr(X = x) = \frac{B(\alpha+1, \beta+x)}{B(\alpha, \beta)} \qquad \mathbb{E}(X) = \frac{\alpha}{\beta - 1}$$

Siguiendo la convención adoptada en el curso, llamamos a este modelo Beta-Geométrico.

1. Muestre que la penetración de mercado (fracción de clientes que compró al menos una unidad) puede expresarse como $\frac{\beta}{\alpha+\beta}$.

2. Para ello considere que la penetración de mercado es del 60 % y que el numero promedio de parkas que compra un cliente (incluyendo aquellos que no compraron) es igual a 2.5. Estime los parámetros del modelo. *Hint: Realice un sistema de ecuaciones.*

3. Se desea extender el modelo para que además incluya un segmento de clientes que nunca compra. Para este segmento de clientes, $Pr(X = 0) = 1$. Sea π la proporción de estos clientes en la población, y $(1 - \pi)$ el segmento de clientes que se comporta de acuerdo al modelo Beta-Geométrico descrito anteriormente. Suponga que conoce x_i, el numero de compras hechas por cada cliente i en la base de datos (i = 1, ..., n). Deduzca la función de verosimilitud que utilizaría para estimar el modelo.

Solución

1. Despejando p_0:

$$\begin{aligned} p_0 &= 1 - \Pr(X = 0) \\ &= 1 - \frac{B(\alpha, \beta + 1)}{B(\alpha, \beta)} \\ &= 1 - \frac{\Gamma(\alpha)\Gamma(\beta + 1)}{\Gamma(\alpha + \beta + 1)} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \\ &= 1 - \frac{\beta\,\Gamma(\beta)}{(\alpha + \beta)\,\Gamma(\beta)} \\ &= \frac{\alpha}{\alpha + \beta}. \end{aligned}$$

2. A partir de la penetración de compra se despeja α:

$$0.6 = \frac{\beta}{\alpha + \beta} \quad \Rightarrow \quad \beta = 0.6(\alpha + \beta) \;\Rightarrow\; 0.4\beta = 0.6\alpha \;\Rightarrow\; \alpha = \tfrac{2}{3}\beta$$

Despejando β utilizando el número promedio de parkas comprado por un cliente:

$$2.5 = \frac{\alpha}{\beta - 1} \quad \Rightarrow \quad 2.5 = \frac{\frac{2}{3}\beta}{\beta - 1} \Rightarrow\; \tfrac{2}{3}\beta = 2.5(\beta - 1) \Rightarrow\; \beta = \tfrac{15}{11}$$

Entonces, sustituyendo:

$$\alpha = \tfrac{2}{3}\beta = \tfrac{2}{3} \cdot \tfrac{15}{11} = \tfrac{10}{11}$$

3. Sea n_x el número de clientes que compra x parkas. Entonces, la verosimilitud puede escribirse como:

$$L(\alpha, \beta, \pi|x) = \prod_{x=0}^{\infty} \mathbb{P}(X = x|\alpha, \beta, \pi)^{n_x}$$

Donde,

$$Pr(X = x \mid \alpha, \beta, \pi) = \pi\, Pr(X = x \mid \text{no adopta}) + (1 - \pi)\, Pr(X = x \mid \text{puede adoptar})$$

$$= \begin{cases} \pi + (1 - \pi)\dfrac{B(\alpha + 1, \beta)}{B(\alpha, \beta)}, & x = 0 \\ (1 - \pi)\dfrac{B(\alpha + 1, \beta + x)}{B(\alpha, \beta)}, & x > 0 \end{cases}$$

3.4.8 Problema 8

Para revitalizar el interés de los usuarios, Pokémon Go está pensando introducir 7 nuevos personajes Pokémon en las comunas de Melipilla, Talagante, Pirque, Colina, Alhue, Paine, Lampa y Buin. Para estudiar el impacto en el atractivo del juego, los desarrolladores han estado realizando pruebas para ver el interés de los jugadores en estos nuevos personajes. Las pruebas se han desarrollado sobre una muestra de jugadores de acuerdo a los números desplegados en la siguietne tabla:

Pokemon	Melipilla	Talagante	Pirque	Colina	Alhue	Paine	Lampa	Buin	Total
Wigglytuff	5	834	504	117	763	424	387	683	3,717
Slowbrow	24	975	8	342	715	362	788	793	4,007
Lickitung	101	651	200	993	562	614	691	595	4,407
Chansley	198	517	960	788	441	844	836	438	5,022
Tangela	79	990	756	868	210	944	509	233	4,589
Ditto	69	587	96	365	182	634	373	419	2,725
Phanpy	138	969	971	890	586	764	585	361	5,264

Tabla: Número de veces que un determinado pokemón fue mostrado a un jugador.

Existe dos parámetros relevantes para entender el grado de interacción entre jugadores y personajes. El primer parámetro corresponde a q_k que corresponde a la probabilidad de atrapar a un personaje, condicional de que un jugador intente atraparlo. Aunque esta probabilidad es relevante para saber en qué

medida los usuarios aumentan su mazo, el parámetro es ajustado internamente en la plataforma y, por tanto, no hay incertidumbre desde el punto de vista de los usuarios. El segundo parámetro que denominamos como p_k, mide la probabilidad que un usuario intente atrapar un personaje k si este le aparece disponible en la aplicación móvil. Desde el punto de vista de la jugabilidad es importante disponibilizar personajes que los usuarios quieran atrapar. Para entender el comportamiento de esta intención de atrapar personajes, la tabla siguiente despliega el número de intentos para los personajes y comunas que forman parte del estudio piloto:

Pokemon	Melipilla	Talagante	Pirque	Colina	Alhue	Paine	Lampa	Buin	Total
Wigglytuff	0	218	98	39	31	63	26	43	518
Slowbrow	8	186	4	110	393	79	240	113	1,133
Lickitung	8	47	13	98	183	180	219	60	808
Chansley	49	31	274	24	48	91	205	42	764
Tangela	14	294	71	35	59	304	116	2	895
Ditto	14	180	14	110	33	84	0	10	445
Phanpy	4	19	66	189	9	37	92	33	449

Tabla: Número de veces que un jugador intentó atrapar a un determinado pokemón.

A partir de los datos presentados:

1. Use el método de los momentos para estimar un modelo binomial sin heterogeneidad que caracterice la probabilidad que un usuario intente atrapar a wigglytuff. ***Hint:*** *Con usar el método de los momentos, se refiere a que el primer momento de una distribución es el valor esperado, donde el valor esperado ed una distribución binomial es np.*
2. Escriba ahora un modelo que incorpore heterogeneidad para caracterizar la probabilidad que un número de usuarios intente atrapar a un pokemón k en una comuna c.
3. Escriba la log-verosimilitud del modelo anterior, para entender la popularidad de cada uno de los personajes del estudio.

Solución

1. Definimos una variable aleatoria que signifique lo buscado

$$X_w := \#\text{ de usuarios que intentan atrapar a wigglytuf}$$

Tenemos un número de experimentos Bernoulli (# de personas a las que se les muestra wigglytuf) y un número de "éxitos" entre estos experimentos

(# de personas que intentan atrapar a wigglytuf cuando se les muestra) que ocurren con alguna probabilidad desconocidad p_w fija, ya que nos piden el caso homogéneo (único valor). Por lo tanto, tenemos que la probabilidad que queremos caracterizar sería de la forma:

$$P(X_w = x_w) = \binom{n_w}{x_w} p_w^{x_w} (1 - p_w)^{n_w - x_w}$$

Donde n_w es la cantidad de personas a las que se les mostró el pokemón (dato en la primera Tabla). Esdecir, $X_w \sim Bin(n_w, p_w)$. Sabemos que el primer momento de una distribución es su media y que la media de una binomial es $n_w p_W$. Por otro lado, estimamos el valor de la media con algún promedio muestras (con muestras del mismo tamaño) de número de personas que intentaron atrapar a wigglytuf, pero en este caso este número es único, por lo que usamos este valor (total de personas que intentaron atrapar a wigglytuf de la última tabla) que, en este caso, es 518. Por lo tanto, podemos obtener la probabilidad p_w con la igualdad del primer momento:

$$518 = n_w p_w 518 = 3717 p_w p_w \approx 0,1393$$

Esto tiene mucho sentido, ya que estamos obteniendo la probabilidad de manera empírica con la fracción de personas que intentaron atraparlo y el número de personas a las que se les mostró.

2. Como debemos considerar heterogeneidad, ahora el parámetro de la distribución que nos intersa, p_k, no lo asumimos como único, sino que sigue una distribución de probabilidad. Como el parámetro es una probabilidad ($p_k \in [0, 1]$), utilizamos la distribución a priori $Beta(\alpha_k, \beta_k)$, con α_k y β_k parámetros a estimar ($p_k \sim Beta(\alpha_k, \beta_k)$, por eso el sub k, ya que cada pokemón tiene un probabilidad p_k asociada y no tienen porqué seguir la misma distribución). Luego, tenemos que cada comuna nos va a estar aportando en realidad datos sobre la heterogeneidad existente para cada pokemón (sería el equivalente a tener distintas observaciones para distintos individuos). Así, la probabilidad quedaría, para cada pokemón y comuna, con la v.a. anterior, pero segmentando por comuna:

$$P(X_{k,c} = x_{k,c}) = \int_0^1 P(X_{k,c} = x_{k,c}|p_k) Beta(p_k|\alpha_k, \beta_k) dp_k = \int_0^1 \left(\binom{n_{k,c}}{x_{k,c}} p_k^{x_{k,c}} (1 - p_k)^{n_{k,c} - x_{k,c}} \right) \left(\frac{p_k^{\alpha_k - 1} (1 - p_k)^{\beta_k - 1}}{B(\alpha_k, \beta_k)} \right.$$

3. Nos interesa analizar cada personaje por separado, por lo tanto vamos a tener una expresión de la log-verosimilitud del parámetro p_k para cada pokemón. Luego, la muestra que tenemos de la v.a. en cuestión para cada pokemón k es de la forma $\{x_{k,1}, x_{k,2}, ..., x_{k,7}\}$, donde el segundo índice es

para cada una de las comunas. Sigue que la log-verosimilitud es el logaritmo de la función de verosimilitud con esta muestra para cada pokemón:

$$\begin{aligned}
\ell(p_k|x_{k,1},...,x_{k,8}) &= \ln(\mathcal{L}(p_k|x_{k,1},...,x_{k,8})) \\
&= \ln(P(X_{k,1}=x_{k,1},...,X_{k,8}=x_{k,8}|p_k)) \\
&= \ln\left(\prod_{c=1}^{8} P(X_{x,c}=x_{x,c}|p_k)\right) \\
&= \sum_{c=1}^{8} \ln(P(X_{k,c}=x_{k,c}|p_k)) \\
&= \sum_{c=1}^{8} \ln\left(\binom{n_{k,c}}{x_{k,c}} \frac{B(\alpha_k+x_{k,c},\beta_k+n_{k,c}-x_{k,c})}{B(\alpha_k,\beta_k)}\right)
\end{aligned}$$

Esta es la expresión pedida de la log-verosimilitud para cada pokemón k. En el tercer paso asumimos indepndencia de las v.a. entre columnas y en el último paso reemplazamos por la expresión encontrada en la parte anterior.

3.4.9 Problema 9

Inmersos en la era digital, los consumidores están cada vez conectados en distintos dispositivos los cuales pueden ser usados como canales durante el proceso de compra. En este contexto, algunas empresas han decidido lanzar aplicaciones móviles que funcionan en base a información georreferenciada de los dispositivos de sus clientes, y que permiten enviarles promociones personalizadas cuando identifica que este se encuentra cercano a un punto de venta.

La aplicación que más éxito ha tenido en el mercado, quiere evaluar la efectividad de las campañas que envían a sus usuarios, y así poder ofrecer un mejor servicio a las empresas que lo contratan. Para esto, cuenta con información de los m_i mensajes promocionales que envía al cliente i, y el número x_i de esos que son efectivamente utilizados por el cliente.

1. Plantee un modelo de elección para determinar la probabilidad de respuesta de cada cliente, en cada ocasión de compra. En particular,

   a. Escriba la log-verosimilitud del problema.
   b. Sea N el número de clientes. Si se observa el comportamiento por T períodos y en promedio se envían M promociones por cliente, ¿cuántos parámetros deben estimarse?

2. Considere que se cuenta con información demográfica de los usuarios de la aplicación móvil, como edad, Ingreso y frecuencia de uso de la app. Reformule el modelo anterior incorporando variables explicativas considerando

además que existe un segmento de clientes que realiza pocos canjes y otro que realiza muchos (es decir, un segmento canjeador C y otro no-canjeador N). ***Hint**: Recuerde que si $x \in \mathbb{R}$, entonces $\frac{exp(x)}{1+exp(x)} \in (0,1)$.*

3. Suponga que se ha estimado el modelo anterior y se encuentra que los segmento canjeadores y no-canjeadores tienen el mismo tamaño y que los parámetros asociados a las variables explicativas son tales que para un usuario i con un vector de variables explicativas z_i, $\beta_C' x_i = ln(2)$ y $\beta_N' x_i = ln(1/2)$. Si al usuario le han enviado dos promociones y ha canjeado las dos, ¿cuál es la probabilidad que este usuario pertenezca al segmento de canjeadores?

4. Suponga ahora que del total de m_i mensajes que recibe un usuario este lee n_i de ellos. De los n_i mensajes que efectivamente lee, decide canjear x_i de ellos. Modele el comportamiento anteriormente descrito como un modelo integrado elección-elección. Plantee explícitamente la probabilidad de canje y la log-verosimilitud del problema.

Solución

1. Por partes:

    1. Sea X_i el número de veces que la persona i hace efectivo un cupón. La probabilidad de respuesta es

    $$\mathbb{P}(X_i = x_i \mid p_i) = \binom{m_i}{x_i} p_i^{x_i} (1-p_i)^{m_i - x_i}$$

    donde la log-verosimilitud viene dada por

    $$LL(p) = \sum_i \ln\Big(\binom{m_i}{x_i} p_i^{x_i} (1-p_i)^{m_i - x_i} \Big)$$

    2. Se deben estimar N parámetros, uno para cada cliente (lo que sugiere introducir heterogeneidad).

2. Modelo de elección con dos segmentos. Sea p_{iC} la probabilidad de canje si i pertenece al segmento canjeador y p_{iN} si pertenece al segmento no canjeador. Sea π la fracción de clientes canjeadores. Entonces

$$\mathbb{P}(X_i = x_i \mid \beta, \pi) = \pi \binom{m_i}{x_i} p_{iC}^{x_i} (1-p_{iC})^{m_i - x_i} + (1-\pi) \binom{m_i}{x_i} p_{iN}^{x_i} (1-p_{iN})^{m_i - x_i}$$

donde

$$p_{iC} = \frac{\exp(\beta_C' z_i)}{1 + \exp(\beta_C' z_i)}, \qquad p_{iN} = \frac{\exp(\beta_N' z_i)}{1 + \exp(\beta_N' z_i)}$$

y la log-verosimilitud es

$$LL = \sum_i \ln\Big(\pi \binom{m_i}{x_i} p_{iC}^{x_i} (1-p_{iC})^{m_i - x_i} + (1-\pi) \binom{m_i}{x_i} p_{iN}^{x_i} (1-p_{iN})^{m_i - x_i} \Big)$$

3. Probabilidad posterior de pertenecer al segmento canjeador dado $X_i = 2$ (usando probabilidades condicionales):

$$\mathbb{P}(C \mid X_i = 2) = \frac{\mathbb{P}(X_i = 2 \mid C)\,\mathbb{P}(C)}{\mathbb{P}(X_i = 2 \mid C)\,\mathbb{P}(C) + \mathbb{P}(X_i = 2 \mid N)\,\mathbb{P}(N)}$$

donde, para $m_i = 2$,

$$\mathbb{P}(X_i = 2 \mid C) = \binom{2}{2}\left[\frac{\exp(\beta_C' z_i)}{1+\exp(\beta_C' z_i)}\right]^2\left[\frac{1}{1+\exp(\beta_C' z_i)}\right]^{2-2} = \frac{4}{9}$$

$$\mathbb{P}(X_i = 2 \mid N) = \binom{2}{2}\left[\frac{\exp(\beta_N' z_i)}{1+\exp(\beta_N' z_i)}\right]^2\left[\frac{1}{1+\exp(\beta_N' z_i)}\right]^{2-2} = \frac{1}{9}$$

y, tomando $\mathbb{P}(C) = \mathbb{P}(N) = \frac{1}{2}$,

$$\mathbb{P}(C \mid X_i = 2) = \frac{\frac{4}{9}\cdot\frac{1}{2}}{\frac{4}{9}\cdot\frac{1}{2} + \frac{1}{9}\cdot\frac{1}{2}} = \frac{4}{5}$$

4. Lectura y canje con dos etapas. La probabilidad de leer un mensaje es:

$$\begin{aligned}\mathbb{P}(X_i = x_i \mid p_{\text{canjear}}, p_{\text{leer}}) &= \sum_{n_i = x_i}^{m_i} \mathbb{P}(X_i = x_i \mid p_{\text{canjear}}, n_i)\,\mathbb{P}(N_i = n_i \mid p_{\text{leer}}) \\ &= \sum_{n_i = x_i}^{m_i} \binom{n_i}{x_i} p_{\text{canjear}}^{x_i}(1-p_{\text{canjear}})^{\,n_i - x_i}\binom{m_i}{n_i} p_{\text{leer}}^{n_i}(1-p_{\text{leer}})^{\,m_i - n_i}\end{aligned}$$

y la log-verosimilitud se describe como:

$$LL = \sum_i \ln(\mathbb{P}(X_i = x_i \mid p_{\text{canjear}}, p_{\text{leer}}))$$

Chapter 4

Modelos Estructurales

4.1 Introducción

Los modelos de elección discreta buscan modelar **comportamiento** (toma de decisiones), describiendo la probabilidad de que un agente n elija la alternativa i, es decir, P_{ni}. Se asume que el individuo decide entre las alternativas tal que éste maximiza sus utilidades u_{ni}.

En general, los modelos de decisión discreta se definen como $u_{ni} = v_{ni} + \varepsilon_{ni}$, con v_{ni} la componente determinista y ε_{ni} la componente estocástica.

A continuación se presentan los modelos estructurales abordados en el curso:

4.1.1 Modelo Logit

El modelo logit es un modelo de elección discreta ampliamente utilizado para analizar decisiones en las que un individuo selecciona una alternativa entre un conjunto finito de opciones. Este modelo asume que los individuos maximizan su utilidad u_{ni}, la cual se divide en una componente determinista (observable) v_{ni} y una estocástica (no observable) ε_{ni}, con la componente estocástica siguiendo una distribución de valor extremo Gumbel tipo 1.

La probabilidad de que un agente n elija la alternativa i viene dada por:

$$P_{ni} = \frac{e^{v_{ni}}}{\sum_j e^{v_{nj}}}$$

Donde los parámetros del modelo pueden ser estimados usando el método de máxima verosimilitud con $v_{ni} = \beta' x_{ni}$:

$$LL(\beta) = \sum_n \sum_i y_{ni} \ln(P_{ni})$$

Con x_{ni} un conjunto de covariables y $y_{ni} = 1$ si n elige la alternativa i, 0 en caso contrario.

4.1.1.1 Persistencia de Elección (Lealtad) y Precios de Referencia

En modelos de panel con observaciones repetidas, es posible incorporar efectos dinámicos como la **lealtad** y los **precios de referencia**:

Lealtad: Se introduce una variable de lealtad latente z_{nit} que captura la propensión del individuo a repetir elecciones previas:

$$z_{nit} = \lambda z_{nit-1} + (1 - \lambda) y_{nit-1}$$

donde $0 \leq \lambda \leq 1$ es un parámetro de decaimiento.

Precios de Referencia: Los consumidores comparan precios actuales con un precio de referencia interno formado a partir de precios observados en el pasado:

$$RP_{nit} = \lambda RP_{nit-1} + (1 - \lambda) P_{nit-1}$$

La utilidad se especifica incorporando la desviación respecto al precio de referencia:

$$u_{nit} = \alpha_i + \beta_1 P_{nit} - \beta_2 (P_{nit} - RP_{nit}) + \delta x_{nit} + \varepsilon_{nit}$$

4.1.2 Modelo Probit

El modelo probit es un modelo de elección discreta utilizado para analizar decisiones en las que un individuo selecciona una alternativa entre un conjunto finito de opciones. Este modelo asume que los individuos maximizan su utilidad u_{ni}, la cual se divide en una componente determinista (observable) v_{ni} y una estocástica (no observable) ε_{ni}, con la componente estocástica siguiendo una **distribución normal estándar**.

Asumiendo que el término del error ε_{ni} se distribuye normal multivariado centrado en 0:

$$\phi(\varepsilon_n) = \frac{1}{(2\pi)^{I/2} |\Sigma|^{1/2}} e^{-\frac{1}{2} \varepsilon_n' \Sigma^{-1} \varepsilon_n}$$

La matriz de covarianza Σ tiene una dimensión $I \times I$, con I la cantidad de alternativas de elección. Como la matriz es simétrica, esta tiene $I(I+1)/2$ parámetros.

La probabilidad de elección se aproxima mediante simulación:

$$\hat{P}_{ni} \approx \frac{1}{R} \sum_{r=1}^{R} \mathbb{1}[\varepsilon_{nj}^r - \varepsilon_{ni}^r < v_{ni} - v_{nj}, \forall j \neq i]$$

Los parámetros del modelo se estiman usando el **método simulado de máxima verosimilitud (SMLE)**:

$$\hat{\theta} = \arg\max \sum_n \sum_i y_{ni} \ln(\hat{P}_{ni})$$

4.1.3 Modelo Nested Logit

El modelo Nested Logit es una extensión del modelo Logit que organiza las alternativas en 'nidos' o grupos que comparten características comunes, permitiendo modelar correlaciones entre opciones dentro de un mismo nido. Este modelo relaja la suposición de independencia de alternativas irrelevantes (IIA) al introducir un parámetro de correlación para cada nido.

La probabilidad de que un agente n elija la alternativa j se descompone como el producto de la probabilidad de elegir la alternativa j dado que eligió el nodo B_k que contiene la alternativa j, multiplicado por la probabilidad de elegir el nodo B_k:

$$P_{nj} = P_{nj|B_k} \cdot P_{nB_k}$$

Donde:

$$P_{nj|B_k} = \frac{e^{v_{nj}/\lambda_k}}{\sum_{i \in B_k} e^{v_{ni}/\lambda_k}}$$

$$P_{nB_k} = \frac{e^{w_{nk} + \lambda_k IV_{nk}}}{\sum_h e^{w_{nh} + \lambda_h IV_{nh}}}$$

Con $IV_{nk} = \ln\left(\sum_{j \in B_k} \exp\left(\frac{v_{nj}}{\lambda_k}\right)\right)$ el valor inclusivo del nido.

El conjunto de parámetros $\{\lambda_k\}_{k=1}^K$ mide el grado relativo de independencia en la parte no observable de la utilidad entre las alternativas del nido k. Probar que $\lambda_k = 1$ para todos los nidos podría llevarnos de vuelta al modelo logit simple.

4.1.4 Modelo Mixed Logit

El modelo Mixed Logit es una extensión del modelo Logit que permite modelar la heterogeneidad de preferencias entre los individuos. A diferencia del modelo Logit tradicional, los parámetros de la utilidad u_{ni} no son constantes, sino que varían aleatoriamente en la población.

La utilidad de un individuo n para una alternativa i se define como:

$$u_{ni} = \beta_n' x_{ni} + \varepsilon_{ni}$$

donde β_n es un vector de coeficientes específicos de cada individuo, que sigue una distribución probabilística $f(\beta|\theta)$ en la población.

La probabilidad de elección de la alternativa i por el individuo n es:

$$P_{ni} = \int \frac{e^{\beta' x_{ni}}}{\sum_j e^{\beta' x_{nj}}} f(\beta|\theta) d\beta$$

donde $f(\beta|\theta)$ describe la distribución de los parámetros en la población, comúnmente asumida como normal multivariada: $\beta \sim N(b, V_b)$.

Estimación: Para estimar los parámetros, se utiliza el **método de máxima verosimilitud simulada (SMLE)**. La probabilidad simulada se calcula como:

$$\hat{P}_{ni} = \frac{1}{R} \sum_{r=1}^{R} \frac{e^{\beta_r' x_{ni}}}{\sum_j e^{\beta_r' x_{nj}}}$$

con β_r siendo una muestra de $f(\beta|\theta)$ y R el número de simulaciones.

4.1.5 Logit con Clases Latentes

El modelo de Clases Latentes asume que la población está compuesta por un número finito de segmentos o clases, cada uno con parámetros específicos de preferencia. A diferencia del Mixed Logit, aquí se considera que las preferencias dentro de cada clase son homogéneas, pero varían entre clases.

La utilidad de un individuo n para una alternativa i en la clase m se define como:

$$u_{ni|m} = \beta_m' x_{ni} + \varepsilon_{ni}$$

La probabilidad de que el individuo n elija la alternativa i se define como:

$$P_{ni} = \sum_m s_m \frac{e^{\beta'_m x_{ni}}}{\sum_j e^{\beta'_m x_{nj}}}$$

donde s_m es la proporción de la población en la clase m y β_m son los parámetros específicos de la clase m.

Estimación: Para estimar los parámetros, se maximiza la función de verosimilitud:

$$LL(\{\beta_m, s_m\}_{m=1}^M) = \sum_n \sum_i y_{ni} \ln \left(\sum_m s_m \frac{e^{\beta'_m x_{ni}}}{\sum_j e^{\beta'_m x_{nj}}} \right)$$

con restricciones sobre s_m para asegurar que representen proporciones válidas: $0 \leq s_m \leq 1$ y $\sum_m s_m = 1$.

4.2 Metodología

La estimación de los modelos estructurales comparte una metodología:

1. **Especificación de la utilidad:** Definir la función de utilidad u_{ni} con sus componentes determinísticas v_{ni} y estocásticas ε_{ni}.

2. **Derivación de probabilidades:** Calcular las probabilidades de elección P_{ni} según el modelo especificado (Logit, Probit, Nested Logit, Mixed Logit, o Clases Latentes).

3. **Construcción de la verosimilitud:** Escribir la función de (log-)verosimilitud basada en las observaciones y_{ni}.

4. **Estimación:** Para modelos con solución cerrada (Logit, Nested Logit, Clases Latentes), se utiliza **máxima verosimilitud directa**. Para modelos sin solución cerrada (Probit, Mixed Logit), se emplea **máxima verosimilitud simulada (SMLE)**.

5. **Evaluación:** Calcular métricas de ajuste como ρ de McFadden, AIC, BIC y validar la significancia estadística de los parámetros.

6. **Interpretación:** Analizar los parámetros estimados, calcular elasticidades y evaluar implicaciones para la toma de decisiones.

4.3 Problemas Teóricos

4.3.1 Pregunta 1

Suponga que dispone de una muestra de tamaño R de los parámetros que definen la utilidad de elección en un modelo logit. Describa cómo estimar la probabilidad que un consumidor n elija la alternativa i.

4.3.2 Pregunta 2

¿Cómo se calcula el Akaike Information Criterion (AIC) para un modelo cuya verosimilitud $f(X|\theta)$ es maximizada en un único valor de $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$?

4.3.3 Pregunta 3

Suponga que la componente determinística de la utilidad de los consumidores que pertenecen al segmento i puede calcularse como $v(\theta_i)$. ¿Cuál es la probabilidad de elección en un modelo logit binario si asumimos que en la población solo existen dos clases caracterizadas por los set de parámetros θ_1 y θ_2 y además asumimos que la utilidad de no elegir está normalizada a 0?

4.3.4 Pregunta 4

Describa cómo incluir que los consumidores eligen usando precios de referencia en un modelo logit.

4.3.5 Pregunta 5

Explique una situación en la que un modelo mixed logit (o de heterogeneidad continua) podría preferirse en comparación a un modelo logit con clases latentes.

4.3.6 Pregunta 6

Suponga que estamos interesados en estudiar el efecto del ingreso de los hogares en la elasticidad precio en las compras de una categoría. Explicar por qué para estudiar este efecto es informativo poner atención al coeficiente δ que acompaña a $Precio \times Ingreso$ en la definición en la función de utilidad.

4.3.7 Soluciones

1. Dicha probabilidad puede estimarse como:

$$\hat{P}_{ni} = \frac{1}{R}\sum_{r=1}^{R}\frac{\exp(v_i(\theta^{(r)}))}{\sum_k \exp(v_k(\theta^{(r)}))}$$

   donde $v_i(\theta)$ es la componente determinística de la utilidad que el tomador de decisión deriva al elegir la alternativa i.

2. Aplicamos la definición directamente:

$$AIC = -2\ln(f(X|\hat{\theta})) + 2k = -2\ln(f(X|\hat{\theta})) + 4$$

3. Sea π la probabilidad de que un cliente pertenezca al segmento 1. Entonces, la probabilidad de elección viene dada por:

$$P_{ni} = \pi\frac{\exp(v(\theta_1))}{1+\exp(v(\theta_1))} + (1-\pi)\frac{\exp(v(\theta_2))}{1+\exp(v(\theta_2))}$$

4. Definimos la componente determinística de la utilidad como:

$$v_{nit} = \beta_0 + \beta_1(P_{nit} - RP_{nit})$$

   donde P_{nit} es el precio de la alternativa i al que se enfrenta el cliente n en la ocasión de compra t y RP_{nit} es el precio de referencia que típicamente definimos como:

$$RP_{nit} = \lambda RP_{nit-1} + (1-\lambda)P_{nit-1}$$

   Con esta especificación de la utilidad, la probabilidad de elección se calcula usando la fórmula estándar del modelo logit.

5. Cuando una distribución discreta no describe bien la heterogeneidad de los parámetros en la población. Por ejemplo, si esperamos que la distribución de un parámetro tenga colas pesadas (i.e. una proporción significativa de la población se aleja de la media), entonces se requerirá demasiadas clases para aproximar la heterogeneidad en la población.

6. El efecto del ingreso en la elasticidad precio depende directamente del parámetro γ_1 de la siguiente ecuación:

$$\gamma_{precio} = \gamma_0 + \gamma_1 Ingreso$$

   Al reemplazar dicha ecuación en la definición de la función de utilidad obtenemos:

$$u_{nit} = \beta_0 + \gamma_0 Precio + \gamma_1 Ingreso \cdot Precio$$

4.4 Problemas Aplicados

4.4.1 Pregunta 1

Un administrador de la categoría aceites de una sala del centro de Santiago de una importante retailer nacional, se dispone a analizar datos de panel que reportan y_{ijt} que toma el valor 1 si el hogar i compra la marca-tamaño j en la semana t y 0 en caso contrario ($i = 1, \ldots, N$, $j = 1, \ldots, J$, $t = 1, \ldots, T$). Junto con las elecciones semanales de cada hogar, el panel provee $PRICE_{jt}$ correspondiente al precio por centímetro cúbico de la alternativa j en la semana t y $COUPON_{ijt}$ que toma el valor 1 si el cliente i dispone de un cupón de descuento para la alternativa j en la semana t (aunque el valor de descuento de los cupones varía entre alternativas y de semana a semana, los valores típicamente rondan el 10% del precio de lista).

Después de una reunión de trabajo con administradores de categoría en otras salas se ha acordado que un buen punto de partida es usar un modelo estructural en que la componente determinística de la utilidad de compra viene dada por:

$$v_{ijt} = \alpha_{ij} + \beta_{i1} PRICE_{jt} + \beta_{i2} COUPON_{ijt}$$

Sin embargo, hasta el momento no hay demasiada claridad respecto a cómo introducir heterogeneidad en el modelo, por lo que se barajan varias alternativas.

a) Escriba la log-verosimilitud de un modelo logit con dos clases latentes. La expresión debe quedar expresada directamente en función de los parámetros que ingresarán como variables de decisión en la maximización irrestricta de la log-verosimilitud.

b) Escriba una (aproximación a) la log-verosimilitud de un modelo logit mezclado. Al igual que en el caso anterior, la expresión debe quedar expresada directamente en función de los parámetros que ingresarán como variables de decisión en la maximización irrestricta de la log-verosimilitud. Para ello puede asumir que dispone de una función que le permite samplear desde cualquier distribución de probabilidad generando una muestra de tamaño R del vector de parámetros α_{ij}, β_1 y β_2.

Solución:

a) La función de log-verosimilitud viene dada por:

$$LL = \sum_i \sum_j \sum_t y_{ijt} \ln \left[\frac{\exp(\lambda)}{1+\exp(\lambda)} \frac{\exp(v^1_{ijt})}{\sum_k \exp(v^1_{ikt})} + \frac{1}{1+\exp(\lambda)} \frac{\exp(v^2_{ijt})}{\sum_k \exp(v^2_{ikt})} \right]$$

donde v^1_{ijt} y v^2_{ijt} son las componentes sistemáticas de la utilidad de los clientes pertenecientes a las clases 1 y 2 respectivamente y vienen dadas por:

$$v^1_{ijt} = \alpha^1_{ij} + \beta^1_{i1} PRICE_{jt} + \beta^1_{i2} COUPON_{ijt}$$
$$v^2_{ijt} = \alpha^2_{ij} + \beta^2_{i1} PRICE_{jt} + \beta^2_{i2} COUPON_{ijt}$$

Los parámetros α^1_{ij}, β^1_{i1} y β^1_{i2} corresponden a los parámetros caracterizando a la clase 1 y α^2_{ij}, β^2_{i1} y β^2_{i2} los que caracterizan a la clase 2. El parámetro λ determina el tamaño relativo de cada clase y la transformación logística sobre λ garantiza que las proporciones estén en el rango [0,1] y sumen 1.

b) La (aproximación de la) función de log-verosimilitud viene dada por:

$$LL = \sum_i \sum_j \sum_t y_{ijt} \ln \left(\frac{1}{R} \sum_r \frac{\exp(v^r_{ijt})}{\sum_k \exp(v^r_{ikt})} \right)$$

donde v^r_{ijt} son las componentes sistemáticas de la utilidad de los clientes si los parámetros toman los valores α^r_{ij}, β^r_1 y β^r_2.

4.4.2 Pregunta 2

1. Suponga que un analista propone usar un modelo logit para estudiar el comportamiento de compra de 5 marcas de margarina las que pueden ofrecerse en formato pan o pote.

   a) Escriba la función de utilidad si los clientes tienen preferencias dependientes exclusivamente de precio y formato.

   b) Escriba la función de utilidad si los clientes tienen preferencias dependientes de precio, marca y formato.

   c) Escriba la función de utilidad si se quiere evaluar cuánto influye la interacción entre el nivel de ingreso ING_n y el precio en la utilidad del cliente n.

2. Considere un agente que se enfrenta a la elección entre dos alternativas A y B de modo que las componentes determinísticas de su utilidad vienen dadas por $v_A = v_B = 1$. Usted como analista ha decidido modelar la elección usando un modelo probit con una matriz de varianza-covarianza completamente general, es decir sobre la cual no se ha impuesto ninguna restricción más allá de la simetría. Calcule explícitamente:

$$P_A = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbb{1}(v_A + \varepsilon_A > v_B + \varepsilon_B) \phi(\varepsilon_A, \varepsilon_B) d\varepsilon_A d\varepsilon_B$$

Justifique su respuesta e interprete el resultado.

3. Considere un cliente comprando por tres periodos en una categoría con dos marcas A y B como lo indica la siguiente tabla:

Periodo	Marca A	Marca B
1	0	1
2	0	1
3	1	0

Calcule una medida de lealtad para cada marca y periodo usando un modelo de suavización exponencial con parámetro de factor de alisamiento λ. Para ello inicialice las medidas de lealtad en $\frac{1}{2}$ para el periodo 0 en ambas marcas.

4. Para analizar el comportamiento de compra de los clientes en un mercado de alimentos enlatados, una importante empresa consultora ha implementado una rutina de maximización de verosimilitud para estimar los parámetros de un modelo logit que se aplica sobre una base de 324 clientes que hacen 3615 compras en la categoría. En la rutina propuesta, los parámetros máximo verosímiles se obtienen usando el comando `mymle = optim(par=0.01*rep(1,3), fn=loglikel, hessian=TRUE, method="BFGS")`.

Al terminar la rutina, un analista de la empresa consultora quiere explorar los resultados de la estimación para lo que ejecuta una serie de comandos sencillos en la consola obteniendo los siguientes resultados:

```
> mymle$par
[1] 0.124 0.486 -2.351
> mymle$value
[1] 1344.35
```

¿Cuál es el BIC del modelo estimado por la empresa consultora?

Solución:

1. a) La utilidad que experimenta un cliente n respecto a la marca i en la ocasión de compra t viene dada por:

$$u_{nit} = \beta_{PRECIO} PRECIO_{it} + \beta_{PAN} PAN_i + \varepsilon_{nit}$$

donde PAN_i toma el valor 1 si la alternativa i corresponde al formato pan, $PRECIO_{it}$ es el precio de la alternativa i en la oportunidad de compra t y ε_{nit} independientes e idénticamente distribuidas valor extremo tipo I. Notar que por motivos de identificación, este modelo hemos fijado en cero la utilidad intrínseca de la alternativa pote.

b) Hay al menos dos formas de modelar este problema. La primera en que los interceptos se definen al nivel marca-formato y otra en que los interceptos se definen en términos de marca y formato:

i. La utilidad que experimenta un cliente n respecto a la marca i en la ocasión de compra t viene dada por:

$$u_{nit} = \beta_{PRECIO} PRECIO_{it} + \beta_{PAN} PAN_i + \sum_{j=1}^{4} \alpha_j M_{ij} + \varepsilon_{nit}$$

donde PAN_i toma el valor 1 si la alternativa i corresponde al formato pan, M_{ij} toma el valor 1 si la alternativa i es de la marca j, $PRECIO_{it}$ es el precio de la alternativa i en la oportunidad de compra t y ε_{nit} independientes e idénticamente distribuidas valor extremo tipo I. Notar que por motivos de identificación, este modelo hemos fijado en cero la utilidad intrínseca de la alternativa pote y de la última marca (Es por esto que la sumatoria llega hasta 4 y no hasta 5).

ii. La utilidad que experimenta un cliente n respecto a la marca i en la ocasión de compra t viene dada por:

$$u_{nit} = \beta_{PRECIO} PRECIO_{it} + \sum_{j=1}^{N-1} \alpha_j MF_{ij} + \varepsilon_{nit}$$

donde MF_{ij} toma el valor 1 si la alternativa i corresponde a la marca-formato j, $PRECIO_{it}$ es el precio de la alternativa i en la oportunidad de compra t y ε_{nit} independientes e idénticamente distribuidas valor extremo tipo I. Si todas las marcas están disponibles en todos los formatos, entonces tenemos $N = 2 \times 5$ marca-formatos, de los cuales solo podemos identificar $N - 1$ interceptos.

c) Asumiendo que la utilidad que experimenta un cliente n respecto a la marca i en la ocasión de compra t depende exclusivamente del precio (si dependiera de otros factores simplemente tendríamos que agregar los términos correspondientes), la función de utilidad puede expresarse como:

$$u_{nit} = \beta_{0p} + \beta_{1p} ING_n + \beta_{2p} PRECIO_{it} + \beta_{3p} ING_n + PRECIO_{it} + \varepsilon_{nit}$$

donde $PRECIO_{it}$ es el precio de la alternativa i en la oportunidad de compra t y ε_{nit} independientes e idénticamente distribuidas valor extremo tipo I.

2. P_A no es más que la probabilidad de elegir la alternativa A. La clave para calcular su valor viene de observar que $v_A = v_B$. Si las dos alternativas

tienen el mismo valor de la componente sistemática, entonces el cálculo de P_A corresponde a la probabilidad de que una componente sea mayor que la otra en una distribución normal bivariada centrada en 0, la que evidentemente toma el valor $\frac{1}{2}$.

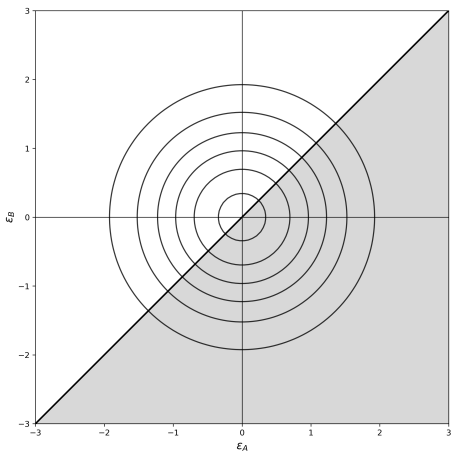


Figure 4.1: Plot de variables transformadas logarítmicamente.

La interpretación es sencilla. Estamos describiendo una situación de un tomador de decisión que, a excepción de una perturbación normal centrada en cero. Dada la simetría de la normal las perturbaciones son igualmente probables de inclinar la balanza en favor de la alternativa A como de la alternativa B y por tanto la probabilidad final que el tomador de decisión elija la alternativa A es simplemente $\frac{1}{2}$.

3. Podemos calcular las medidas de lealtad con que el cliente se enfrenta a las siguientes 3 ocasiones de compra:

Periodo	Marca A	Marca B
2	$(1-\lambda)/2$	$\lambda + (1-\lambda)/2$
3	$(1-\lambda)^2/2$	$\lambda(1-\lambda) + (1-\lambda)^2/2$
4	$\lambda + (1-\lambda)^3/2$	$(1-\lambda)\lambda + (1-\lambda)^2 + (1-\lambda)^3/2$

4. Por definición:

$$BIC = -2LL(\hat{\theta}) + k\ln(n)$$

donde $LL(\hat{\theta})$ es la log verosimilitud evaluada en el estimador máximo verosímil, k el número de parámetros en el modelo y n el número de observaciones sobre la que se calcula la verosimilitud. Entonces:

$$BIC = -2(-1344.35) + 3\ln(3615) = 2713.279$$

4.4.3 Pregunta 3

Una compañía de telefonía móvil de reciente ingreso al país ha iniciado una fuerte campaña en algunas redes sociales y busca estudiar el perfil de los clientes que han evaluado favorablemente la campaña. Para eso se ha construido una base de datos de 5243 usuarios de facebook que vieron al menos uno de los elementos considerados en la campaña (video, afiche, álbum de fotos, etc.). Junto con información de los usuarios considerados en la muestra, la base de datos contiene una variable y_n que toma el valor 1 si el usuario n le dio un "me gusta" a alguno de las componentes de la campaña. La siguiente tabla contiene una breve descripción de la información disponible en la base.

Variable	Descripción	Promedio
y_n	1 si a usuario n le gusta alguna componente de la campaña	0.093
$Hombre_n$	1 si usuario n es hombre	0.421
$Twitter_n$	1 si usuario n es usuario de twitter	0.134
$Edad_n$	Edad usuario n	23.18

a) Suponga que para perfilar a los clientes se propone un modelo logit en que el usuario n deriva una utilidad u_n por darle un "me gusta" a alguna de las componentes de la campaña y 0 si no. Escriba la log-verosimilitud del problema asumiendo que la componente determinística de la utilidad depende linealmente de $x_n = (Hombre_n, Twitter_n, Edad_n, Edad_n^2)$.

Suponga que analistas dentro de la compañía han codificado la verosimilitud anterior en R en una función `loglikel` que recibe como primer argumento el intercepto de la función de utilidad y los cuatro siguientes los otros parámetros de la utilidad en el orden antes planteado. Para encontrar los estimadores máximo verosímiles se ha ejecutado el siguiente comando:

```
mymle = optim(par=0.01*rep(1,5), fn=loglikel, hessian=TRUE, method="BFGS")
```

Al terminar la rutina, los analistas quieren hacer inferencia sobre los parámetros para lo que ejecutan el siguiente comandos en la consola generando los resultados correspondientes:

```
> cbind(mymle$par, sqrt(diag(solve(mymle$hessian))))
     [,1]      [,2]
[1,] -0.9623  0.2069
[2,]  3.4340  0.0518
```

```
[3,]  0.0047  0.0522
[4,]  0.0365  0.0043
[5,] -0.0048  0.0025
```

b) ¿Cuál es la probabilidad que una mujer de 20 años que no es usuaria de twitter le de un "me gusta" a la campaña?

c) ¿Cómo varía la probabilidad de elección anterior si la mujer sí es usuaria de twitter? Describa que implicancias podría tener su resultado en la elaboración de próximas campañas.

Solución:

a) La componente determinística de la utilidad viene dada por:

$$v_n = \beta_0 + \beta_1 Hombre_n + \beta_2 Twitter_n + \beta_3 Edad_n + \beta_4 Edad_n^2$$

Luego, la log-verosimilitud puede escribirse como:

$$LL = \sum_n y_n \ln \frac{\exp(v_n)}{1+\exp(v_n)} + (1-y_n) \ln \frac{1}{1+\exp(v_n)}$$

b) Para esta mujer, la componente sistemática de la utilidad viene dada por:

$$v_n = -0.9623 + 3.4340 \times 0 + 0.0047 \times 0 + 0.0365 \times 20 - 0.0048 \times 400 = -2.1523$$

Entonces:

$$Pr(y_n = 1) = \frac{\exp(v_n)}{1+\exp(v_n)} = 0.10412$$

c) Si fuere usuaria de twitter, basta modificar la componente determinística de la utilidad $v_n \to v_n + 0.047 = -2.1476$ y recalcular la probabilidad:

$$Pr(y_n = 1) = \frac{\exp(v_n)}{1+\exp(v_n)} = 0.10456$$

Como el cambio de probabilidad es muy bajo, podemos concluir que ser usuario de twitter no es muy relevante para explicar si una persona le dará un "me gusta" a alguna componente de la campaña (notar que esto se puede inferir directamente de los estimadores máximo verosímiles). Para los próximos elementos de la campaña, probablemente convenga concentrarse en el género y edad de los usuarios.

4.4.4 Pregunta 4

En el departamento de estudios de una importante multitienda buscan analizar simultáneamente los montos de compra en la tienda y el uso de la tarjeta de crédito de la casa comercial. Para eso la empresa ha recopilado, para cada cliente n, una serie de variables que caracterizan su relación con la firma en el último año, como describe la siguiente tabla.

Variable	Descripción	Promedio
y_n	Monto total gastado por cliente n con la tienda	124.045
w_n	1 si cliente n tiene la tarjeta de crédito de la tienda	0.64
$Mujer_n$	1 si cliente n es mujer	0.462
$Edad_n$	Edad cliente n	36.21

Como primera aproximación, los analistas del departamento de estudio han propuesto describir el gasto en tienda usando el siguiente modelo:

$$y_n = \begin{cases} \beta_0 + \beta_1 Mujer_n + \beta_2 Edad_n + \varepsilon_{1n}, & \text{si } w_n = 1 \\ \beta_0' + \beta_1' Mujer_n + \beta_2' Edad_n + \varepsilon_{2n}, & \text{si } w_n = 0 \end{cases}$$

Y para la decisión de adquisición de tarjeta:

$$u_n = \begin{cases} \gamma_0 + \gamma_1 Mujer_n + \gamma_2 Edad_n + \varepsilon_{3n}, & \text{adquiere tarjeta} \\ \varepsilon_{4n}, & \text{no adquiere} \end{cases}$$

1. Suponga que ε_{3n} y ε_{4n} son independientes y están idénticamente distribuidos valor extremo tipo 1. Calcule una expresión para la probabilidad que el cliente n elija adquirir la tarjeta de la tienda.

2. Suponga que ε_{1n} y ε_{2n} son independientes y están normalmente distribuidos con media 0 y varianzas σ_1 y σ_2. Calcule una expresión para la verosimilitud de observar un gasto y_n condicional en que el cliente tiene tarjeta y otra para el gasto en que el cliente no tiene tarjeta.

3. Suponga que ε_{1n} y ε_{2n} son independientes de ε_{3n} y ε_{4n}. Escriba la log-verosimilitud de observar simultáneamente $\{y_n\}_{n=1}^N$ y $\{w_n\}_{n=1}^N$.

4. La empresa está especialmente interesada en identificar si hay alguna diferencia en los patrones de gasto entre aquellos clientes que tienen tarjeta con respecto a aquellos que no tienen. Para ello se estima el modelo antes planteado obteniendo una log-verosimilitud $LL_0 = -2.332,57$ y luego otro en que se impone $\beta_0 = \beta_0'$, $\beta_1 = \beta_1'$, $\beta_2 = \beta_2'$ con una log-verosimilitud $LL_1 = -2.418,42$. Basado en el test de ratios de verosimilitud, determine si la tenencia de tarjeta afecta los patrones de gasto en tienda.

ν	1	2	3	4	5	6	7	8
χ^2_ν	3.84	5.99	7.81	9.49	11.07	12.59	14.07	15.51

La tabla indica los valores de χ^2_ν tales que $Pr(\chi^2_\nu \leq \chi^2_\nu) = 0.95$.

5. Si las componentes ε_{1n}, ε_{2n}, ε_{3n}, ε_{4n} son todos independientes, discuta brevemente si hay algún beneficio de estudiar simultáneamente los comportamientos de gasto y tenencia de tarjeta.

Solución:

1. Si ε_{3n} y ε_{4n} son independientes y distribuidos valor extremo, entonces la probabilidad que el cliente elija adquirir la tarjeta viene dada directamente por la fórmula del logit, donde la utilidad de no adquisición se ha normalizado en 0:

$$P_n = \frac{\exp(\gamma_0 + \gamma_1 Mujer_n + \gamma_2 Edad_n)}{1 + \exp(\gamma_0 + \gamma_1 Mujer_n + \gamma_2 Edad_n)}$$

2. Dada la distribución de los errores, los montos gastados para ambos casos estarán normalmente distribuidos. Entonces, las verosimilitudes condicionales vienen dadas por:

$$L(y_n|w_n = 1) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{1}{2\sigma_1^2}(y_n - (\beta_0 + \beta_1 Mujer_n + \beta_2 Edad_n))^2\right)$$

$$L(y_n|w_n = 0) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{1}{2\sigma_2^2}(y_n - (\beta_0' + \beta_1' Mujer_n + \beta_2' Edad_n))^2\right)$$

3. La respuesta se deriva directamente de la definición de verosimilitud y la fórmula de probabilidades totales:

$$LL = \sum_n \ln(P_n L(y_n|w_n = 1) + (1 - P_n)L(y_n|w_n = 0))$$

4. La restricción propuesta implicaría que los patrones de gastos entre clientes con y sin tarjeta serán equivalentes. Para testear esta hipótesis se puede usar el test de ratios de verosimilitud directamente:

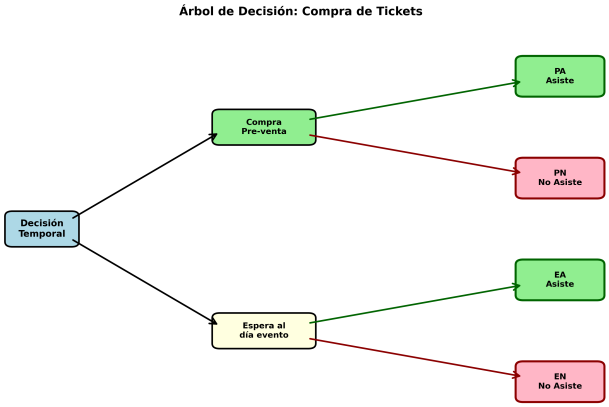
$$LR = 2(LL_0 - LL_1) = 2(-2332.57 - (-2418.42)) = 171.7$$

De acuerdo al número de restricciones, el LR se distribuye $\chi^2_{k=3}$. Por tanto, como el estadístico es mayor que el valor crítico $\chi^2_{0.05,3} = 7.81$, rechazamos la hipótesis nula que el comportamiento de gasto es igual en los casos de clientes con y sin tarjeta.

5. El enfoque de modelación es útil principalmente para estudiar la interacción entre las decisiones de apertura de tarjeta y gasto. Si asumimos que los errores son independientes entonces podríamos enfrentar el problema simplemente como dos problemas separados. La respuesta esperada es que no hay beneficio en términos de la captura del fenómeno de comportamiento. Argumentar en favor de un modelo integrado en virtud de su escalabilidad y posibilidad de comparar con modelos más complejos son también respuestas correctas.

4.4.5 Pregunta 5

Con el avenimiento de los espectáculos masivos, las productoras que organizan los eventos han puesto en práctica múltiples tácticas de discriminación de precios. A continuación nos enfocaremos en el estudio del comportamiento de clientes respecto a las ventas anticipadas de tickets o *pre-venta*, donde los clientes pueden comprar anticipadamente a un precio p_1 menor que p_2, el precio de las entradas el día del evento. Para ello, nos enfocaremos exclusivamente en los clientes interesados en comprar tickets, cuyo comportamiento se describirá usando un modelo logit anidado. En este modelo, los clientes primero deciden si comprar anticipadamente o si esperan al día del evento. En el día del evento, los clientes deciden si asistir o no, generando cuatro escenarios como queda descrito en la siguiente figura.



Para completar el modelo se ha propuesto las siguientes especificaciones para las componentes sistemática de la utilidad:

$$v_{PA} = \alpha_1 - \beta p_1$$
$$v_{PN} = \alpha_2 - \beta p_1$$

$$v_{EA} = \alpha_3 - \beta p_2$$
$$v_{EN} = 0$$

1. Escriba una expresión para P_{EA}, la probabilidad que un cliente espere sin comprar en la pre-venta y luego decida asistir al evento.

2. Suponga que observa el comportamiento de N clientes, para cada uno de los cuales se observa y_{ni} que toma valor 1 si el cliente decide la alternativa $i \in \{PA, PN, EA, EN\}$ (0 en caso contrario). Escriba una expresión para la log-verosimilitud $LL(\theta = (\alpha_1, \alpha_2, \alpha_3, \beta, \lambda))$. Para eso asuma conocidas las expresiones de $P_i(\theta)$ $\forall i \in I$.

3. Como analista, le interesa investigar las restricciones de comportamiento $\alpha_1 = \alpha_3$ y $\alpha_2 = 0$. Explique qué interpretación podría tener dichas hipótesis y cómo podrían testearse.

Solución:

1. Partimos por calcular la probabilidad condicional de asistir al evento condicional en que esperó. Esta probabilidad viene dada directamente por la fórmula del logit:

$$P_{A|E} = \frac{\exp(\alpha_3 - \beta p_2)}{\exp(\alpha_3 - \beta p_2) + 1}$$

Luego calculamos la probabilidad de esperar no comprando en la preventa:

$$P_E = \frac{\exp(\lambda IV_E)}{\exp(\lambda IV_P) + \exp(\lambda IV_E)}$$

donde IV_P y IV_E son los valores inclusivos de comprar y no comprar en la preventa respectivamente:

$$IV_P = \ln(\exp(\alpha_1 - \beta p_1) + \exp(\alpha_2 - \beta p_1))$$
$$IV_E = \ln(\exp(\alpha_3 - \beta p_2) + 1)$$

Finalmente, la probabilidad total viene dada por la multiplicación de las dos probabilidades anteriores:

$$P_{EA} = P_{A|E} \cdot P_E$$

2. Aplicamos directamente la definición de la verosimilitud:

$$LL(\theta) = \sum_n \sum_{i \in I} y_{ni} \ln(P_i(\theta))$$

3. Imponer que $\alpha_1 = \alpha_3$ implicaría que la utilidad intrínseca que los clientes derivan por asistir al evento es la misma independiente de si compraron en la preventa o no. Del mismo modo, imponer que $\alpha_2 = 0$ implicaría que no hay una pérdida de utilidad por comprar los boletos y no asistir más allá del precio pagado. Para testear cada hipótesis, usamos test de ratios de verosimilitud, lo que implica (i) el cálculo de la log-verosimilitud de los modelos con (LL_A) y sin (LL_B) restricciones, (ii) el cálculo del estadístico $LR = 2(LL_A - LL_B)$ y (iii) la comparación de dicho estadístico contra el valor crítico una χ^2_1 con un grado de libertad.

4.4.6 Pregunta 6

Una empresa busca estudiar el impacto del gasto publicitario en su producto estrella, donde tiene una posición monopólica en un mercado de tamaño conocido. Para ello ha reunido las series de precios (p_t), ventas (q_t) y gasto publicitario (ADV_t) para las últimas 205 semanas y ha propuesto un modelo de elección binaria basado en la siguiente función de utilidad $u_t = \alpha + \beta \ln(p_t) + \gamma ADV_t + \nu_t + \varepsilon_t$, donde ν_t están normalmente distribuidos con media 0 y ε_t están distribuidos valor extremo.

1. Muestre que si s_t es la participación de mercado del producto, los parámetros α, β y γ pueden estimarse usando un enfoque de mínimos cuadrados ordinarios sobre la siguiente ecuación:

$$\ln(s_t) - \ln(1 - s_t) = \alpha + \beta \ln(p_t) + \gamma ADV_t$$

2. Usando un enfoque de regresión se ha estimado el modelo, obteniendo los siguientes resultados:

Parámetro	MLE	s.e
α	-1.76	0.27
β	-1.23	0.02
γ	0.31	0.09

Si se mantiene el precio constante en 1 y el gasto publicitario sube de 10 a 20, ¿cuántos puntos porcentuales aumenta la demanda por el producto?

Solución:

1. Aplicando la definición del modelo logit:

$$s_t = \frac{\exp(\alpha + \beta \ln(p_t) + \gamma ADV_t + \nu_t)}{1 + \exp(\alpha + \beta \ln(p_t) + \gamma ADV_t + \nu_t)}$$

$$1 - s_t = \frac{1}{1 + \exp(\alpha + \beta \ln(p_t) + \gamma ADV_t + \nu_t)}$$

Como los denominadores son iguales:

$$\frac{s_t}{1 - s_t} = \exp(\alpha + \beta \ln(p_t) + \gamma ADV_t + \nu_t)$$

Aplicando logaritmo y reordenando términos:

$$\ln(s_t) - \ln(1 - s_t) = \alpha + \beta \ln(p_t) + \gamma ADV_t + \nu_t$$

Como ν_t está normalmente distribuido, se satisfacen las condiciones para aplicar mínimos cuadrados ordinarios.

2. Evaluamos directamente:

$$s_t(ADV_t = 10) = \frac{\exp(-1.76 - 1.23 \ln(1) + 0.31 \times 10)}{1 + \exp(-1.76 - 1.23 \ln(1) + 0.31 \times 10)} = 0.792$$

$$s_t(ADV_t = 20) = \frac{\exp(-1.76 - 1.23 \ln(1) + 0.31 \times 20)}{1 + \exp(-1.76 - 1.23 \ln(1) + 0.31 \times 20)} = 0.988$$

Por lo tanto hay un incremento de $\Delta = 19.6$ puntos porcentuales.

4.4.7 Pregunta 7

Pizza a Pieza es un servicio de pizzas a domicilio que lleva varios años operando en el sector sur de la capital. Los administradores de la tienda quieren estudiar el comportamiento de compra de sus clientes. Para eso, se proponen calibrar modelos de elección discreta usando una base de clientes registrados en la compañía y que hayan comprado al menos una vez en los últimos 6 meses. La base de datos está compuesta por 763 clientes de los que se conoce su número telefónico, la edad y género del jefe de hogar y si han comprado en cada semana t durante los últimos 6 meses de operación ($y_{nt} = 1$ si el cliente n compró en la semana t). Además, para cada semana se conoce el precio de lista de las pizzas (que por políticas de la empresa es el mismo independiente de la variedad de la pizza) y si se repartieron volantes con publicidad.

1. Suponga que la utilidad que un cliente n deriva por comprar en un semana t puede describirse como $u_{nt} = \beta_0' x_{nt} + \varepsilon_{nt}$. Si $\{\varepsilon_{nt}\}$ con $n = 1, \ldots, N$ y $t = 1, \ldots, T$ están independiente e idénticamente distribuidos valor extremo tipo I, escriba la log-verosimilitud de un modelo que describa las compras de la base de clientes.

2. Al revisar la data, los analistas se dan cuenta que en los últimos 6 meses de operación, los precios de las pizzas ha sido exactamente los mismos. ¿Es posible estimar la sensibilidad al precio de la demanda con datos de esta naturaleza? Si no fuera posible, describa qué otra información podría recolectarse para hacer la estimación.

3. Un largo debate en el ámbito de la planificación comercial es el rol que juega la publicidad en afectar el comportamiento del consumidor. El primer argumento indica que la publicidad afecta directamente la utilidad de elección. Bajo este argumento un consumidor obtendrá un mayor beneficio al consumir un producto sobre el que ha sido expuesto a avisaje comercial con respecto a uno en que no ha sido expuesto. Un segundo argumento indica que la publicidad no implica una mayor utilidad en el consumo, pero hace a los consumidores menos sensibles al precio. Siguiendo los pasos que se indican, proponga un procedimiento para determinar cuál de estos argumentos describe mejor el comportamiento de los clientes de Pizza a Pieza.

   (a) Resulta plausible suponer que el efecto de la publicidad se manifiesta en el mediano plazo. Esto es, el avisaje publicitario hoy no afecta sólo el comportamiento de compra hoy sino que también en los próximos días. Proponga una métrica para capturar la intensidad acumulada de la publicidad.

   (b) Usando una métrica de intensidad acumulada de publicidad, escriba la función de utilidad que ejemplifique la situación en que la publicidad afecta directamente la utilidad del consumo.

   (c) Usando una métrica de intensidad acumulada de publicidad, escriba la función de utilidad que ejemplifique la situación en que la publicidad afecte indirectamente a través de la sensibilidad al precio.

   (d) Describa cómo evaluaría cuál de los dos modelos describe mejor el comportamiento de compra de los clientes.

4. Después de explorar la data se ha acordado estimar dos modelos sin heterogeneidad. El primero (M1) considera solo un intercepto y una variable de lealtad que permita diferenciar de aquellos clientes que compran mucho de aquellos que compran menos. El segundo, junto con las variables anteriores consideran además un coeficiente de precio, otro de promoción y finalmente el efecto (directo) de la publicidad (M2). Los estimadores máximo verosímiles de estos dos modelos se encuentran disponibles en la siguiente tabla:

Parámetros	M1	M2
Intercepto	5.23	5.54
Lealtad	3.17	X

Parámetros	M1	M2
Precio	-	-3.25
Promoción	-	4.73
Publicidad	-	0.04
$-2LL$	20.17	Y
BIC	3145	Z

Discuta si valores de X, Y y Z debieran ser, mayores o menores que los estimadores máximo verosímiles del M1.

Solución:

1. Como la componente aleatoria de la probabilidad se distribuye valor extremo tipo I, entonces resulta un modelo logit, cuya log verosimilitud viene dada por:

$$LL = \sum_n \sum_t y_{nt} \ln \frac{\exp(\beta_0' x_{nt})}{1 + \exp(\beta_0' x_{nt})} + (1 - y_{nt}) \ln \frac{1}{1 + \exp(\beta_0' x_{nt})}$$

2. Si no hay variabilidad entonces no podemos saber cuál es el efecto del precio. Otras variables que podrían recolectarse son algunas actividades promocionales que hagan variar el precio de lista como uso de cupones o promociones. También podría saber variabilidad de precio de alternativas cercanas como otras pizzerías cercanas o de incluso otros restaurantes con cobertura de reparto similar.

3. (a) Al igual como hemos usado para medir el impacto acumulado de algunos comportamientos, para capturar la publicidad podemos usar una media móvil. Sea $a_t = 1$ si la firma repartió volantes publicitarios en la semana t y $\lambda \in (0, 1)$, entonces la métrica acumulada de publicidad en t (ADV_t) puede definirse como:

$$ADV_t = \lambda ADV_{t-1} + (1 - \lambda) a_{t-1}$$

   (b) Aunque otras formas funcionales son posibles, lo más sencillo es imponer que la publicidad acumulada entra aditivamente en la función de utilidad:

$$u_{nt} = \beta_0' x_{nt} + \delta ADV_t + \varepsilon_{nt}$$

   donde x_{nt} incluye cualquier otra variable explicativa que quiera usarse para describir la utilidad de elección.

(c) Aunque otras formas funcionales son posibles, lo más sencillo es imponer que la publicidad acumulada entra aditivamente en la sensibilidad al precio:

$$u_{nt} = \beta_0' w_{nt} + (\beta_1 + \delta ADV_t) PRICE_{nt} + \varepsilon_{nt}$$

donde w_{nt} incluye cualquier otra variable explicativa que quiera usarse para describir la sensibilidad al precio.

(d) Habría que estimar ambos modelos y evaluar cuál de los dos ajusta mejor a la data. Para eso se deben comparar métricas como pseudo-R^2, AIC, BIC y MAPE.

4. Como hay más variables explicativas, $-2LL$ debiera ser menor o igual (es decir Y debiera ser menor que 20.17). Como BIC penaliza usar más variables, Z puede subir o bajar. Por último, tal como vimos en clases, X debiera subir. Esto es porque al no controlar por promoción y publicidad, no podremos identificar que a veces clientes relativamente leales dejan de comprar por variaciones en la oferta.

4.4.8 Pregunta 8

Un retailer nacional está interesado en determinar el impacto de las publicaciones en su sitio web en sus ventas online. Esto es, la evaluación de cómo la promoción de un determinado producto afecta el volumen de sus ventas. Para esto, se ha propuesto un modelo estructural en donde la componente determinística de la utilidad de comprar un producto i de cada cliente n en el día t está dada por (note que es plausible que un cliente compre más de un producto por día):

$$v_{nit} = \alpha_i + \beta_{PR} Precio_{it} + \beta_S Size_{it} + \beta_{POS} Posicion_{it} + \beta_M Mail_{nt}$$

1. Plantee un modelo logit para describir el comportamiento de compra de los usuarios. Si hay I productos disponibles y cada día t llegan a la tienda N_t clientes, escriba la log-verosimilitud del modelo.

2. Si en vez de plantear un modelo logit, se prefiriera un modelo probit, ¿qué dimensión tendría la matriz de varianzas-covarianzas?

3. Para el modelo logit, determine el número esperado de ventas del producto i en un día t.

4. Escriba la expresión aproximada de la log-verosimilitud de un mixed logit, cuyos parámetros son generados a partir de un sampleo de R valores de una distribución normal.

Solución:

1. Bajo un modelo logit, la probabilidad que el cliente n elija compre el producto i en el día t viene dada por:

$$p_{nit} = \frac{\exp(v_{nit})}{1 + \exp(v_{nit})}$$

   Luego, si y_{nit} toma el valor 1 si el cliente n compró el producto i en t, entonces la log verosimilitud viene dada por:

$$LL = \sum_i \sum_t \sum_{n=1}^{N_t} y_{nit} \ln(p_{nit}) + (1 - y_{nit}) \ln(1 - p_{nit})$$

2. La elección es binaria por lo que matriz de varianza covarianza tiene dos filas y dos columnas. Se puede argumentar también que por temas de identificación hay solo un parámetro identificable.

3. El número esperado es directamente $\sum_{n=1}^{N_t} p_{nit}$.

4. La expresión es idéntica a la anterior, pero ahora no tenemos una fórmula cerrada para p_{nit} y debemos aproximarlo:

$$p_{nit} \approx \frac{1}{R} \sum_{r=1}^{R} \frac{\exp(v_{nit}(\theta^r))}{1 + \exp(v_{nit}(\theta^r))}$$

4.4.9 Pregunta 9

Con el reciente sorteo de la Copa América Chile 2015 (cómo pasó tan rápido el tiempo), el comité organizador está interesado en estudiar cómo los hinchas deciden a qué partido asistir. Como primera aproximación se busca un modelo que indique si un hincha decide asistir o no a los partidos que juega su equipo preferido. Suponga que usted tiene total certeza que su equipo favorito llegará a la final. A continuación se presenta los datos con los que dispone.

Variable	Descripción
$DISTANCIA_{ij}$	Distancia del hincha i para el partido j de su equipo favorito
$FASE_{ij}$	Fase del partido j del equipo favorito del hincha i (eg: semifinal)
$RIVAL_{ij}$	Rival en el partido j del equipo favorito del hincha i

Variable	Descripción
$PRECIO_{ij}$	Precio de la entrada del partido j del equipo favorito del hincha i

a) Escriba la utilidad para un modelo de elección discreta que considere que (i) cada fase (grupos, cuartos, semi, final) agrega más público, (ii) la distancia afecta de manera no lineal (discuta qué forma funcional podría ser adecuada) y que (iii) los rivales Brasil y Argentina atraen más público que el resto.

b) Los datos que muestra la tabla a continuación son de un espectador que podría haber asistido a 2 partidos, pero solo asiste a uno. Escriba explícitamente la log-verosimilitud de un modelo logit para este hincha.

Partido	Compra	Fase	Rival	Precio Entrada	Distancia
1	0	Grupos	Colombia	\$5.000	5.5 km
2	1	Cuartos de Final	Brasil	\$7.000	474 km

c) Proponga un modelo de decisión anidado en dos etapas especificando las probabilidades que permitan describir el proceso de compra de abonos para la fase grupal. En esta modalidad, los hinchas deben pagar por anticipado P_{abono} por los tres partidos de la fase de grupos. Una vez comprado el abono el hincha debe decidir si asistir a uno, dos o a los tres partidos de la etapa de grupos. Por simplicidad suponga que las componentes determinísticas de las utilidades de asistir a k partidos es v_k.

d) Suponga que el gobierno instaura un programa para estudiantes que les da un descuento para que pueda asistir a un único partido de la fase de grupos. En este programa, los estudiantes pueden elegir de entre los tres partidos programados. Suponga que para estudiar este comportamiento se estima un modelo probit sobre datos de encuestas a potenciales beneficiarios y se encuentra una matriz varianza-covarianza (desnormalizada) como sigue. Interprete los resultados.

$$\Sigma = \begin{bmatrix} 0.198 & 0.533 & 0 \\ 0.533 & 0.360 & 0.258 \\ 0 & 0.258 & 0.852 \end{bmatrix}$$

Solución:

a) Un modelo que soporta la situación descrita viene dado por:

$$v_{ij} = \beta_0 + \beta_1 \delta(cuartos_j) + \beta_2 \delta(semis_j) + \beta_3 \delta(final_j) + \beta_4 distancia_{ij} + \beta_5 distancia_{ij}^2 + \beta_6 \delta(ArgBra_j)$$

Al respecto, es útil comentar algunos puntos:

1. En este modelo la asistencia diferenciada por etapa queda incluida a través de variables dummies (una etapa debe excluirse por identificación). Alternativamente, podría crearse una variable $NumEtapa_{ij}$ que tome el valor 1 para grupos, 2 para cuartos, 3 para semis y 4 para la final e ingresarse directamente en la definición de la utilidad. Notar sin embargo que esto asume un aumento lineal.
2. En esta especificación se propone una forma polinomial flexible para el efecto de la distancia. Si por ejemplo se asume una forma logarítmica se debiera explicar porque se espera que la distancia tenga efectos decrecientes.

b) Basta reemplazar la especificación anterior con los datos de la tabla. Sean:

$$w_1 = \beta_0 + \beta_4 \times 5.5 + \beta_5 \times 5.5^2$$
$$w_2 = \beta_0 + \beta_1 + \beta_4 \times 474 + \beta_5 \times 474^2 + \beta_6$$

Entonces, la contribución a la log-verosimilitud viene dada por:

$$LL = \ln\left(\frac{1}{1+e^{w_1}}\right) + \ln\left(\frac{e^{w_2}}{1+e^{w_2}}\right)$$

c) Las probabilidades de los números de partidos en la segunda etapa vienen dados directamente por:

$$p_{1|A} = \frac{e^{v_1}}{e^{v_1}+e^{v_2}+e^{v_3}}$$
$$p_{2|A} = \frac{e^{v_2}}{e^{v_1}+e^{v_2}+e^{v_3}}$$
$$p_{3|A} = \frac{e^{v_3}}{e^{v_1}+e^{v_2}+e^{v_3}}$$

Finalmente, la decisión de si abonarse o no, viene dada por:

$$p_a = \frac{e^{\lambda IV - Precio_{Abono}}}{1+e^{\lambda IV - Precio_{Abono}}}$$

donde el valor inclusivo viene dado por $IV = \ln(e^{v_1}+e^{v_2}+e^{v_3})$.

d) Aunque se puede elaborar más, es importante reconocer al menos dos factores:

- Los términos de la diagonal son crecientes indicando que la utilidad del tercer partido tiene mayor variabilidad que el segundo y este más que el primero.
- El 0 en la esquina superior derecha indica que no hay correlación entre las componentes no observables del primer con el tercer partido.

4.4.10 Pregunta 10

En cada periodo t, una agencia de medios está encargada de seleccionar los soportes publicitarios en que se emitirán las campañas de cada uno de sus K clientes de acuerdo a los requerimientos que ellos indiquen. Por ejemplo una marca puede indicar que quiere distribuir su presupuesto para que las piezas publicitarias sean exhibidas solo en TV o en una mezcla de radio y vallas públicas. Aunque cada cliente puede requerir presencia en cualquier tipo de medio, el 86% de los clientes elige en alguno de los siguiente paquetes:

- **Masivo multimedia:** en este paquete la agencia se compromete a alcanzar un cierto nivel de GRP para lo cuál puede hacer uso de cualquier medio. Los promedios históricos sugieren que un 40% se gasta en TV, un 25% en prensa escrita, un 20% en radio y un 10% en medios digitales.
- **Nicho multimedia:** en este paquete la agencia se compromete a alcanzar un cierto nivel de GRP, pero concentrado en medios especializados. Así, dependiendo del rubro, la marca puede elegir entre algunas de las variantes temáticas disponible como deportes, infantil, outdoors, entre otros.
- **Nicho radio y escrito:** Similar al paquete Nicho multimedia, pero restringiendo la parrilla a prensa escrita y a radio. A raíz de esta restricción el paquete conlleva un importante descuento en el precio.
- **Solo online:** En este paquete, todo el avisaje se realiza a través de medios digitales incluyendo sponsored links, banners y sitios de redes sociales. Junto con el bajo costo del paquete, esta propuesta permite menor granularidad en la selección de los segmentos objetivos y un incipiente módulo de analytics para reportar la efectividad de la propuesta.

a) Escriba la utilidad total (determinista y aleatoria) que cada cliente k deriva por la elección de cada alternativa i en cada periodo t, considerando que:

- Marcas de distintas industrias tienen distintas valoraciones intrínsecas por cada paquete. Por ejemplo, la industria bancaria tiene una gran valoración por el paquete Masivo Multimedia, mientras que los proveedores de servicios tecnológicos prefieren el Solo online.

- Cada paquete i tiene una fracción de su GRP en el segmento ABC1 (F_{ABC1i}) y la valoración que cada cliente k tiene respecto este porcentaje depende tanto de la penetración de la marca (PEN_{kt}) cómo de su coeficiente de exclusividad (EX_k) los que son definidos por una consultora externa.
- Considerando la baja sofisticación técnica de la industria y que existen costos relevantes en la evaluación de cada propuesta, en la práctica se evidencia una alta inercia en la elección del mix de medios (i.e, marcas tienden a elegir lo que han elegido en periodos anteriores).

b) Derive una expresión para la probabilidad que cada cliente elija cada alternativa. Para eso, considere que el paquete Solo online solo estuvo disponible en un set de semanas τ.

c) Escriba la log-verosimilitud del problema si existen dos clases latentes. Escriba de modo que todos los parámetros puedan ser estimados directamente a través de la maximización irrestricta de la log-verosimilitud.

d) Al estimar el modelo anterior con dos clases, el tamaño de la primera clase queda determinado por $\lambda = 0$, estimado directamente de la maximización de la log-verosimilitud. ¿Qué fracción de clientes pertenece a la clase 2?

e) Se propone un modelo mixed logit. Para estimarlo se implementa el método de la máxima verosimilitud simulada con tres simulaciones por iteración ($R = 3$). La tabla describe los valores de la última iteración para el periodo 53 para el cliente 114. ¿Cuál es la probabilidad que dicho cliente elija Nicho Multimedia en dicho periodo?

Paquete	R=1	R=2	R=3
Masivo multimedia	0.3	0.3	0.4
Nicho multimedia	0.2	0.3	0.3
Nicho radio y escrito	0.4	0.3	0.2
Solo online	0.1	0.1	0.1

Solución:

a) Sea $y_{ikt} = 1$ si el cliente k elige el paquete i en la semana t y $\delta_{kj} = 1$ si el cliente k pertenece al rubro j. Entonces, la utilidad total viene dada por $u_{ikt} = v_{ikt} + \varepsilon_{ikt}$ donde ε_{ikt} distribuye valor extremo tipo I y v_{ikt} viene dada por:

$$v_{ikt}(\beta) = \sum_j \beta_{ikj}\delta_{kj} + \gamma_{kt}F_{ABC1ikt} + \omega y_{ikt-1}$$

$$v_{ikt}(\beta) = \sum_j \beta_{ikj}\delta_{kj} + (\gamma_0 + \gamma_1 PEN_{kt} + \gamma_2 EX_k)F_{ABC1ikt} + \omega y_{ikt-1}$$

b) Aplicamos directamente la fórmula del logit, con la única salvedad que el conjunto de alternativas difiere en algunos casos:

$$p_{ikt}(\beta) = \begin{cases} 0 & t \notin \tau \wedge i = 4 \\ \frac{\exp(v_{ikt}(\beta))}{\sum_{j=1}^{3} \exp(v_{jkt}(\beta))} & t \notin \tau \wedge i < 4 \\ \frac{\exp(v_{ikt}(\beta))}{\sum_{j=1}^{4} \exp(v_{jkt}(\beta))} & t \in \tau \end{cases}$$

c) Escribimos directamente:

$$LL(\beta_1, \beta_2, \lambda) = \sum_k \sum_i \sum_t \left[\frac{\exp(\lambda)}{1 + \exp(\lambda)} y_{ikt} \ln(p_{ikt}(\beta_1)) + \frac{1}{1 + \exp(\lambda)} \ln(p_{ikt}(\beta_2)) \right]$$

d) La probabilidad de pertenecer al segmento 2 viene dada por:

$$s_2 = \frac{1}{1 + \exp(\lambda)} = \frac{1}{2}$$

e) Simplemente tomamos el promedio de las 3 muestras:

$$p_{ikt} = \frac{1}{R} \sum_r p_{ikt}^r = \frac{0.2 + 0.3 + 0.3}{3} = 0.2667$$