Examen Parcial - Métodos Lineales Generalizados

Gerardo Rocha Benigno

2025-03-22

1. Contexto

El Banco de México es el responsable de emitir los billetes que circulan en la economía mexicana. Se cuenta con la información del número de billetes en circulación (C) y la cantidad de billetes falsos (Y), ambas en millones de piezas, para los años de 2000 a 2011. Para identificar la denominación del billete definimos variables indicadoras x20, x50, x100, x200 y x500.

2. Análisis exploratorio

a) Realiza un análisis exploratorio de los datos. Crea gráficas y encuentra las estadísticas que mejor describan la información y comentalas. Obten conclusiones por tipo de información.

En el Gráfico 1 es posible observar que el número de billetes en circulación para todas las denominaciones presentó una tendencia creciente entre los años 2000 a 2011, con el número de billetes de 500 pesos exhibiendo el mayor crecimiento, ya que el núemro de billetes de esta denopminación fue el menor en 2000 pero el mayor en 2011.

Por su parte, dependiendiendo de la denminación de los billetes, presentaron tendencias heterogeneas. Por ejemplo, el número de billetes de 20 y 100 pesos presentaron una tendencia decreciente, mientras que el número de billetes de 200 y 500 pesos presentaron una tendencia creciente; el número de billetes falsos en circulación de 50 pesos no presentó una tendencia clara.

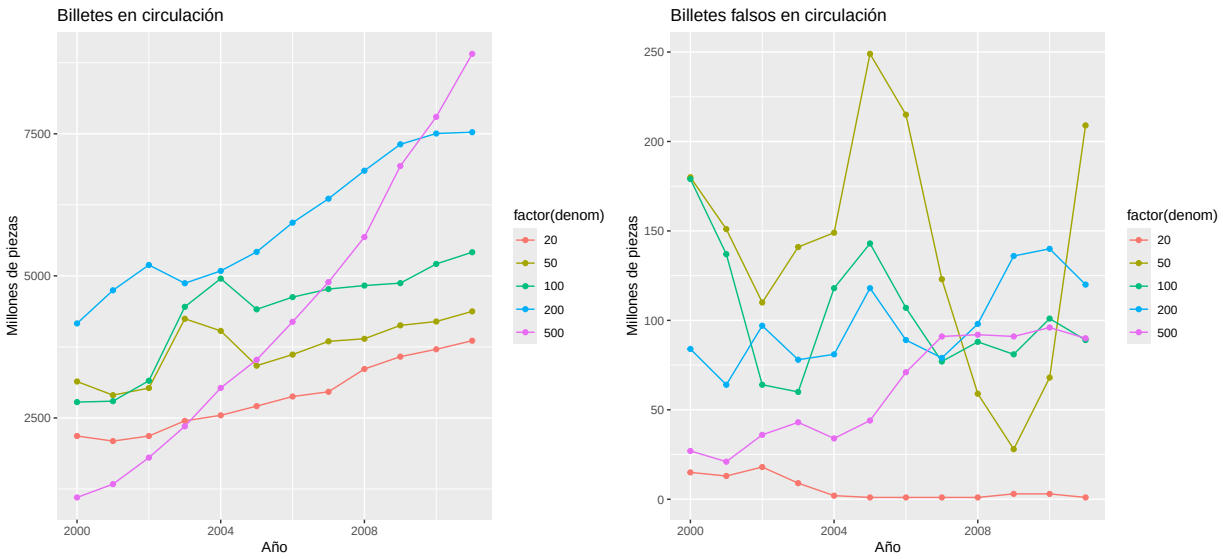

Gráfico 1: Evolución de los billetes en circulación

Como porcentaje del número total de billetes en circulación, la proporción de billetes falsos de 20, 100 y 500

pesos decreció, mientras que la proporción de billetes falsos de 50 y 200 pesos no presentó una tendencia clara.

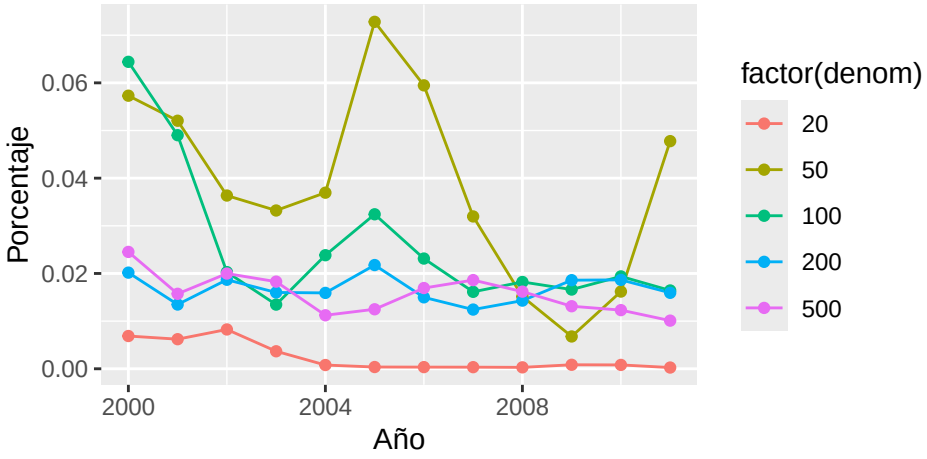

Gráfico 2: Evolución de los billetes falsos

Relación entre número total de billetes y número de billetes falsos en circulación

Para el billete de 500 pesos se de una relación positiva entre el número de billetes en circulación y el número de billetes falsos en circulación; este mismo patrón se observa con el billete de 100 pesos. Para el billete de denominación de 20 pesos se observa una relación negativa. Para el billete de 50 y 100 pesos no es posible observar una tendencia clara.

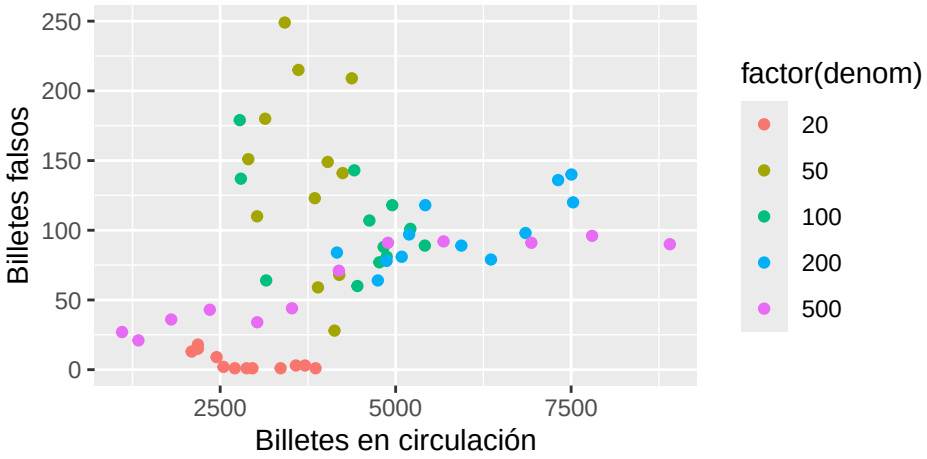

Gráfico 3: Relación entre número total de billetes y billetes falsos en circulación

3. Modelado

Ignorando la dependencia temporal que pudiera existir entre las observaciones de distintos años, considera un modelo de regresión binomial de la forma:

$$Y_i \sim Binomial(C_i, \pi_i)$$

para $i = 1, ..., 60$, y define el siguiente predictor líneal:

$$\eta_i = \alpha + \beta_1 x20_i + \beta_2 x50_i + \beta_3 x100_i + \beta_4 x200_i + \beta_5 x500_i$$

Incluye alguna restricción de estimabilidad adecuada, ya sea i. (i) $\alpha = 0$, ii. (ii) $\beta_j = 0$ para algún $j = 1, ..., 5$ o, iii. (iii) $\sum_{j=1}^{5} \beta_j = 0$.

Queremos comparar el desempeño del modelo con dos funciones liga, la logística y la complementaria log-log.

b) Ajuste el modelo de regresión binomial con liga logística, i.e., $logit(\pi_i) = \eta$ y usa distribuciones iniciales vagas. Calcula los indicadores de ajuste DIC y $Pseudo\ R^2$. Encuentra los estimadores puntuales y por intervalo de los parámetros del modelo, interprétalos y comenta qué tan bueno es el modelo.

La Tabla 1 muestra el resumen de la estimación usando 120 mil iteraciones, con un burn-in del 20% y un thinning de 1, monitoreando dos cadenas por parámetro. Los resultados indican que el modelo tiene un buen ajuste, ya que el valor de DIC es menor al del modelo nulo y el $Pseudo\ R^2$ es cercano 0.5612. Sin embargo, el coeficiente de β_5 no es estadísticamente significativo.

Al imponer una restricción de estimabilidad en la que la suma de las β_i es cero, el intercepto α se interpreta como el logit de la probabilidad promedio de que un billete sea falso, sin importar la denominación. Por su parte, cada coeficiente β_i se interpreta como la desviación en la escala logit de la probabilidad promedio de que un billete en circulación sea falso dada la denominación. Dicho lo anterior, para calcular las probabilidades de que un billete de una denominación dada sea falso debemos realizar la siguiente transformación:

$$\pi_i = \frac{e^{\alpha+\beta_i}}{1 + e^{\alpha+\beta_i}}$$

Tabla 1. Resultados del modelo de regresión binomial con liga logística

Parámetro	Media.posterior	Percentil.2.75.	Percenrtil.97.5.
alpha	-4.2976	-4.3490	-4.2480
beta_1	-1.9187	-2.1011	-1.7450
beta_2	1.0534	0.9910	1.1171
beta_3	0.5835	0.5163	0.6499
beta_4	0.2206	0.1520	0.2877
beta_5	0.0613	-0.0138	0.1373
DIC	1296.3207	NA	NA
Pseudo R^2	0.5617	NA	NA

c) En el modelo de regresión binomial con liga logística, interpreta todos los coeficientes de tu modelo.

La Tabla 2 muestra los valores de los coeficientes del modelo de regresión binomial con liga logística, así como las probabilidades. La interpretación de los coeficientes es la siguiente:

- i. Sin importar la denominación, la probabilidad promedio de que un billete sea falso es de 0.0133.
- ii. El coeficiente del billete de 20 pesos es negativo, lo que implica una menor probabilidad de falsificación en comparación con el promedio. La probabilidad es de 0.0019, menos de una quinta parte que la probabilidad promedio.
- iii. El coeficiente del billete de 50 pesos es positivo, lo que implica una mayor probabilidad de falsificación en comparación con el promedio. La probabilidad es de 0.0375, 2.8 veces la probabilidad promedio.
- iv. El coeficiente del billete de 100 pesos es positivo, lo que implica una mayor probabilidad de falsificación en comparación con el promedio. La probabilidad es de 0.0238, 1.8 veces la probabilidad promedio.

- v. El coeficiente del billete de 200 pesos es positivo, lo que implica una mayor probabilidad de falsificación en comparación con el promedio. La probabilidad es de 0.0167, 1.25 veces la probabilidad promedio.
- vi. El coeficiente del billete de 500 pesos es positivo, lo que implica una mayor probabilidad de falsificación en comparación con el proomedio. La probabilidad de que un billete de 500 pesos sea falso es de 0.0143, casi la misma que la probabilidad promedio.

Tabla 2. Interpretación de los resultados del modelo de regresión binomial con liga logística

Denominación	Coeficiente	Probabilidad	Razón.respecto.al.promedio
Cualquiera	-4.2976	0.0134	1.0000
20	-1.9187	0.0020	0.1485
50	1.0534	0.0375	2.7972
100	0.5835	0.0238	1.7734
200	0.2206	0.0167	1.2427
500	0.0613	0.0143	1.0623

d) En el modelo de regresión binomial con liga logística define "la tasa de billetes falsos por mil circulando" para cada denominación como $1000p_j$, con p_j = proporción de billetes falsos para cada denominación $j = 1, 2, ..., 5$, donde $j = 1$ corresponde al billete de \$20, $j = 1$ al billete de \$50, etc. Nota que p_j debe de estar definido en términos de la liga y de los parámetros de tu modelo. Estima estas tasas mediante un intervalo de 95% de probabilidad y coméntalas.

Es posible calcular la probabilidad de que un billete sea falso para cada denominación, dada la liga logística y los coeficientes estimados. Usando esta información podemos calcular la tasa de billetes falsos por cada mil billetes en circulación para cada denominación, así como sus intervalos de probabilidad del 95% de la siguiente forma:

$$\#\ de\ billetes\ falsos\ por\ cada\ mil\ billetes\ en\ circulacin = 100 * \frac{e^{\alpha+\beta_{j,i}}}{1+e^{\alpha+\beta_{j,i}}},\ j = 1, ...5,\ i = \{2.5\%\,,\ media,\ 97.5\%\}$$

La Tabla 3 muestra la tasa de billetes falsos por cada mil billetes en circulación para cada denominación. De acuerdo con la información presentada, en promedio:

- Existen 1.9 billetes falsos de 20 pesos por cada mil billetes en circulación de esta denominación.
- Existen 37.5 billetes falsos de 50 pesos por cada mil billetes en circulación de esta denominación.
- Existen 20.9 billetes falsos de 100 pesos por cada mil billetes en circulación de esta denominación.
- Existen 16.7 billetes falsos de 200 pesos por cada mil billetes en circulación de esta denominación.
- Existen 14.3 billetes falsos de 500 pesos por cada mil billetes en circulación de esta denominación.

Tabla 3. Número de billetes falsos por cada mil billetes en circulación

Denominación	Percentil.2.5.	Media	Percentil.97.5.
20 pesos	1.6	2.0	2.5
50 pesos	35.8	37.5	39.3
100 pesos	22.5	23.8	25.1
200 pesos	15.8	16.7	17.6
500 pesos	13.3	14.3	15.3

e) Ajuste el modelo de regresión binomial con liga complementaria $log{-}log$, i.e., $log{-}log(1-\pi_i) = \eta_i$ y usa también distribuciones iniciales vagas para todos los parámetros del modelo. Calcula los indicadores de ajuste DIC y $pseudoR^2$. Encuentra los estimadores puntuales y por intervalo de los parámetros del modelo, interprétalos y comenta qué tan bueno es el modelo.

La Tabla 4 muestra los resultados de la estimación del modelo binomial con liga complementaria log-log y se observa que el modelo tiene un ajuste comparable al modelo con liga logística, ya que el valor de DIC y el $Pseudo\ R^2$ son similares. Respecto a los coeficientes, estos también son similares y esto es esperado, dado que nuestro conjunto de datos contiene probabilidades pequeñas, donde ambas ligas se comportan de manera similar. Sin embargo, los coeficientes se interpretan de la siguiente manera:

- Bajo la condición de estimabilidad cada coeficiente β_i representa un cambio en la transformación complementaria log-log para la denominación j, con respecto al promedio de todas las denominaciones. Si el coeficiente es positivo, entonces implica un mayor riesgo de falsificación y viceversa.

Tabla 4. Resultados del modelo de regresión binomial con liga complementaria log-log

Parámetro	Media.posterior	Percentil.2.75.	Percenrtil.97.5.
alpha	-4.3108	-4.3649	-4.2589
beta_1	-1.9250	-2.1220	-1.7384
beta_2	1.0469	0.9830	1.1123
beta_3	0.5845	0.5169	0.6535
beta_4	0.2254	0.1570	0.2948
beta_5	0.0683	-0.0087	0.1457
DIC	1296.6916	NA	NA
Pseudo R^2	0.5615	NA	NA

f) En el modelo de regresión binomial con liga complementaria log-log, interpreta todos los coeficientes de tu modelo.

La Tabla 5 muestra los valores de los coeficientes del modelo de regresión binomial con liga complementaria log-log, así como las probabilidades. La interpretación de los coeficientes es la siguiente:

- i. Sin importar la denominación, la probabilidad promedio de que un billete sea falso es de 0.0133.
- ii. El coeficiente del billete de 20 pesos es negativo, lo que implica el menor riesgo de falsificación en comparación con el promedio. La probabilidad es de 0.0020 menos de una séptima. parte de la probabilidad promedio.
- iii. El coeficiente del billete de 50 pesos es positivo y es el mayor, lo que implica que este billete está en mayor riesgo de ser falsificado en comparación con el promedio. La probabilidad es de 0.0375, 2.8 veces la probabilidad promedio.
- iv. El coeficiente del billete de 100 pesos es positivo, lo que implica también un mayor riesgo de ser falsificado en comparación con el promedio. La probabilidad es de 0.0238, 1.8 veces la probabilidad promedio.
- v. El coeficiente del billete de 200 pesos es positivo, lo que implica un mayor riesgo de ser falsificado en comparación con el promedio. La probabilidad es de 0.0167, 1.25 veces la probabilidad promedio.
- vi. El coeficiente del billete de 500 pesos es positivo, lo que implica un mayor riesgo de falsificación en comparación con el proomedio. La probabilidad es de 0.0143, casi la misma que la probabilidad promedio.

Tabla 5. Interpretación de los resultados del modelo de regresión binomial con liga complementaria log-log

Denominación	Coeficiente	Probabilidad	Razón.respecto.al.promedio
Cualquiera	-4.3108	0.0133	1.0000
20	-1.9250	0.0020	0.1467
50	1.0469	0.0375	2.8137
100	0.5845	0.0238	1.7845
200	0.2254	0.0167	1.2507
500	0.0683	0.0143	1.0702

g) En el modelo de regresión binomial con liga complementaria log-log define "la tasa de billetes falsos por mil circulando" para cada denominación como $1000p_j$, con p_j = proporción de billetes falsos para cada denominación $j = 1, 2, ..., 5$, donde $j = 1$ corresponde al billete de \$20, $j = 1$ al billete de \$50, etc. Nota que p_j debe de estar definido en términos de la liga y de los parámetros de tu modelo. Estima estas tasas mediante un intervalo de 95% de probabilidad y coméntalas.

Es posible calcular la probabilidad de que un billete sea falso para cada denominación, dada la liga complementaria log-log con ayuda de los coeficientes estimados. Usando esta información podemos calcular la tasa de billetes falsos por cada mil billetes en circulación para cada denominación, así como sus intervalos de probabilidad del 95% de la siguiente forma:

$$\#\ de\ billetes\ falsos\ por\ cada\ mil\ billetes\ en\ circulacin = 100*(1-\exp(-exp(\alpha+\beta_i))),\ j = 1, ...5,\ i = \{2.5\%,\ media,\ 97.5\%\}$$

La Tabla 3 muestra la tasa de billetes falsos por cada mil billetes en circulación para cada denominación. De acuerdo con la información presentada, en promedio:

- Existen 2.0 billetes falsos de 20 pesos por cada mil billetes en circulación de esta denominación.
- Existen 37.5 billetes falsos de 50 pesos por cada mil billetes en circulación de esta denominación.
- Existen 23.8 billetes falsos de 100 pesos por cada mil billetes en circulación de esta denominación.
- Existen 16.7 billetes falsos de 200 pesos por cada mil billetes en circulación de esta denominación.
- Existen 14.3 billetes falsos de 500 pesos por cada mil billetes en circulación de esta denominación.

Tabla 6. Número de billetes falsos por cada mil billetes en circulación

Denominación	Percentil.2.5.	Media	Percentil.97.5.
20 pesos	1.5	2.0	2.5
50 pesos	35.8	37.5	39.3
100 pesos	22.5	23.8	25.1
200 pesos	15.8	16.7	17.6
500 pesos	13.3	14.3	15.3

h) Compara los modelos de regresión binomial con las dos ligas, logística y complementaria log-log. De acuerdo con sus medidas de ajuste determina cuál de los dos es el mejor. Con el mejor modelo realiza una gráfica de predicción del número de billetes falsos y compáralo con los datos observados. Comenta los puntos importantes de esta gráfica. En particular comenta sobre los billetes de \$20 y de \$50.

De acuerdo con los resultados del DIC y el $Pseudo\ R^2$, no es posible determinor el mejor modelo basado en estos valores. Con lo anterior, elegimos el modelo logit debido a que su interpretabilidad es más sencilla.

Tabla 7. Comparación de los modelos de regresión binomial con liga logística y complementaria log-log

Modelo	DIC	Pseudo.R.2
Logística	1296.32	0.56
Complementaria log-log	1296.69	0.56

El Gráfico 4 muestra la predicción del número de billetes falsos en circulación para cada denominación, comparado con los datos observados. Se observa que el modelo logit predice de manera adecuada el número de billetes falsos en circulación para los billetes de mayor denominación (100, 200 y 500 pesos).

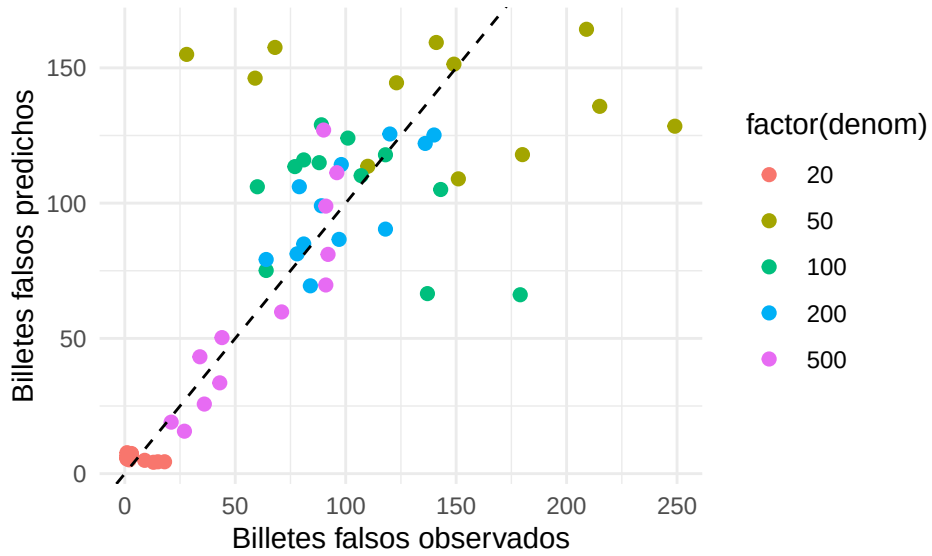

Gráfico 4: Valores reales y predichos para todas las denominaciones

Sin embargo, las predicciones para los billetes de denominaciones de 20 y 50 pesos no son tan precisas, ya que el modelo logit sobrestima la mayoría de observacviones del número de billetes falsos en circulación para los billetes de 20 pesos y tiende a sobrestimar algunas observaciones de los billetes de 50 pesos y a subestimar algunas otras (Gráfica 5).

i) Con el mejor modelo, compara las estimaciones de "las tasas de billetes falsos por mil circulando" para las cinco denominaciones. Determina cuales de ellas son estadísticamente diferentes justificando tu respuesta con las estimaciones obtenidas.

Para que las tasas de billetes falsos por cada mil billetes en circulación sean estadísticamente diferentes, los intervalos de probabilidad del 95% no deben traslaparse para indicar que con alta probabilidad las tasas verdaderas difieren entre sí. Lo anterior debe cumplirse debido a que cada denominación es un grupo independiente y no se espera que las tasas de billetes falsos sean iguales entre las denominaciones. Como señalan Gelman et al. (2013), Gelman et al. [2013] " El traslape general en los intervalos posteriores basados en análisis independientes sugiere que todos los experimentos podrían estar estimando la misma cantidad".

Tabla 8. Intervalos de credibilidad el número de billetes falsos por cada mil billetes en circulación

Denominación	Percentil.2.5.	Percentil.97.5.

20 pesos	1.6	2.5
50 pesos	35.8	39.3
100 pesos	22.5	25.1
200 pesos	15.8	17.6
500 pesos	13.3	15.3

Adicional al modelo de regresión binomial, podrías haber usado otro modelo para ajustar estos datos. Escribe cuál sería tu selección de modelo alternativo definiendo la verosimilitud y la función liga. Ajusta tu nueva propuesta de modelado, compara con los dos modelos anteriores y comenta.

Debido a que tenemos datos de conteo, una alternativa al modelo de regresión binomial es el modelo de Poisson. La verosimilitud de este modelo es:

$$L(\theta) = \prod_{i=1}^{n} \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}$$

y la función liga es la función logarítmica:

$$\log(\lambda_i) = \alpha + \beta_1 x20_i + \beta_2 x50_i + \beta_3 x100_i + \beta_4 x200_i + \beta_5 x500_i$$

La Tabla 9 muestra los resultados de la estimación del modelo de Poisson.

Respecto a los coeficientes este caso los coeficientes se interpretan de la siguiente manera:

- Bajo la condición de estimabilidad cada coeficiente β_i representa la diferencia relativa en escala logarítmica del número esperado de billetes falsos para la denominación i, con respecto al promedio de todas las denominaciones.

Tabla 9. Resultados del modelo de Poisson

Parámetro	Media.posterior	Percentil.2.75.	Percenrtil.97.5.
alpha	4.0327	3.9802	4.0839
beta_1	-2.2997	-2.4892	-2.1173
beta_2	0.9984	0.9346	1.0629
beta_3	0.6217	0.5554	0.6892
beta_4	0.5604	0.4928	0.6282
beta_5	0.1192	0.0415	0.1980
DIC	1221.3707	NA	NA
Pseudo R^2	0.6197	NA	NA

Para una interpretabilidad, se puede realizar la transformación siguiente:

$$\lambda_i = \exp(\alpha + \beta_i)$$

La Tabla 10 los DIC y $Pseudo\ R^2$ de los tres modelos. Se observa que el modelo tiene un ajuste superior al modelo binomial con liga logística y complementaria log-log, ya que la $Pseudo\ R^2$ es mayor que la de los otros dos modelos.

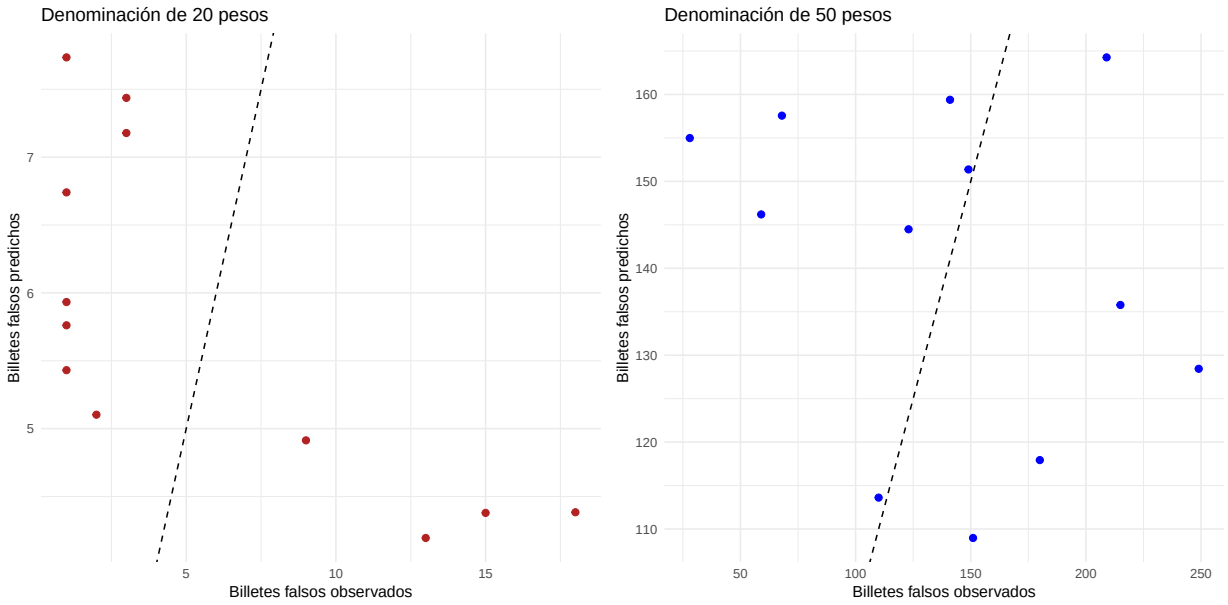

Gráfico 5: Valores reales y predichos para billetes de 20 y 50 pesos

Tabla 10. Comparación de los modelos de regresión binomial con liga logística, complementaria log-log y Poisson

Modelo	DIC	Pseudo.R.2
Logística	1296.32	0.56
Complementaria log-log	1296.69	0.56
Poisson	1221.37	0.62

Apéndice

Aquí se presentan el codigo de JAGS para los modelos de regresión binomial con liga logística y complementaria log-log, así como el modelo de Poisson.

Modelo binomial con liga logística

```
# Código de JAGS para un modelo de regresión binomial con
# liga logística

model {
  # y: Número de billetes falsos
  # C: Número de billetes en circulacion
  # p: Probabilidad de que un billete sea falso (pi en el ejercicio)

  # Verosimilitud
  for (i in 1:n) {
    y[i] ~ dbin(p[i], C[i])

    # Liga logística
    logit(p[i]) <- alpha +
    beta[1]*x20[i] +
```

```
    beta[2]*x50[i] +
    beta[3]*x100[i] +
    beta[4]*x200[i] +
    beta[5]*x500[i]

    # Predicción
    yf[i] ~ dbin(p[i], C[i])
  }

  # Priors
  # alpha
  alpha ~ dnorm(0, 0.001)

  #beta
  for (j in 1:5) {
    beta[j] ~ dnorm(0, 0.001)
  }

  # tasa de falsificación
  for(j in 1:5){
    pi[j] <- exp(alpha.est + beta.est[j])
    tasa[j] <- 1000 * pi[j] / (1 + pi[j])
  }

  # Restricciones de estimabilidad
  alpha.est <- alpha + mean(beta[])
  for (j in 1:5) {
    beta.est[j] <- beta[j] - mean(beta[])
  }

}
```

Modelo binomial con liga complementaria log-log

```
# Código de JAGS para un modelo de regresión binomial con
# liga log-log: cloglog(p) = log(-log(1 - p))

model {
  # y: Número de billetes falsos
  # C: Número de billetes en circulacion
  # p: Probabilidad de que un billete sea falso (pi en el ejercicio)

  # Verosimilitud
  for (i in 1:n) {
    y[i] ~ dbin(p[i], C[i])

    # Liga complementaria log-log
    cloglog(p[i]) <- alpha +
        beta[1]*x20[i] +
        beta[2]*x50[i] +
        beta[3]*x100[i] +
        beta[4]*x200[i] +
        beta[5]*x500[i]

    # Predicción
```

```
    yf[i] ~ dbin(p[i], C[i])
  }

  # Priors
  # alpha
  alpha ~ dnorm(0, 0.001)
  # beta
  for (j in 1:5) {
    beta[j] ~ dnorm(0, 0.001)
  }
  # tasa de falsificación clog-log
  for(j in 1:5){
    pi[j] <- 1 - exp(-exp(alpha.est + beta.est[j]))
    tasa[j] <- 1000 * pi[j]
  }

  # Restricciones de estimabilidad
  alpha.est <- alpha + mean(beta[])
  for (j in 1:5) {
    beta.est[j] <- beta[j] - mean(beta[])
  }

}
```

Modelo Poisson con liga logarítmica

```
# Código de JAGS para un modelo de regresión poisson con
# liga logarítmica

model {
  # y: Número de billetes falsos
  # C: Número de billetes en circulacion
  # p: Probabilidad de que un billete sea falso (pi en el ejercicio)

  # Verosimilitud
  for (i in 1:n) {
    y[i] ~ dpois(lambda[i])

    # Liga logarítmica
    log(lambda[i]) <- alpha +
      beta[1]*x20[i] +
      beta[2]*x50[i] +
      beta[3]*x100[i] +
      beta[4]*x200[i] +
      beta[5]*x500[i]

    # Predicción
    yf[i] ~ dpois(lambda[i])
  }

  # Priors
  # alpha
  alpha ~ dnorm(0, 0.001)

  #beta
```

```
  for (j in 1:5) {
    beta[j] ~ dnorm(0, 0.001)
  }

  # tasa de falsificación
  for (j in 1:5) {
  lambda_denom[j] <- exp(alpha.est + beta.est[j])
  tasa[j] <- 1000 * lambda_denom[j] / C_denom[j]
  }

  # Restricciones de estimabilidad
  alpha.est <- alpha + mean(beta[])
  for (j in 1:5) {
    beta.est[j] <- beta[j] - mean(beta[])
  }
}
```

References

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. CRC press, 2013.