INSTITUTO DE SOCIOLOGÍA
FACULTAD DE CIENCIAS SOCIALES

Trabajo N°1

Análisis de Datos Categóricos - SOL3070

Profesor Mauricio Bucca
Ayudante Roberto Cantillan

Estudiante Andreas Laffert

lunes 07, octubre 2024

Enunciado I: Linear Probability Model (LPM)

1. Calcula las probabilidades de que un hombre con biografía en wikipedia sea político (p_h) y de que una mujer con biografía en wikipedia sea política (p_m). Calcula la diferencia entre ambas proporciones.

En la Tabla 1 se muestran las probabilidades condicionales de ser político (o no) según el género. Formalmente, la ecuación corresponde a: $\mathbb{P}_{j|i} = \frac{P(\text{politico}=j,\ \text{genero}=i)}{\text{genero}=i}$. Los resultados sugieren que los hombres con biografía en wikipedia tienen una mayor probabilidad de ser políticos en comparación con las mujeres con biografía en wikipedia, ya que (p_h) = 0.31 mientras que (p_m) = 0.19.

Implementación en R:

```
xtab <- table(db$genero, db$politico)

dimnames(xtab) <- list(
  sexo = c("Femenino", "Masculino"),
  politico = c("No politíco", "Político")
)

round(prop.table(xtab,1), 3) %>%
  kableExtra::kable(format = "latex",
                    col.names = c("No político", "Político"),
                    booktabs = T,
                    align = 'c',
                    caption = paste("(\\#tab:table1)",
                                    "Probabilidades condicionales de ser
  kableExtra::kable_styling(latex_options = "hold_position",
                            position = "center",
                            bootstrap_options = c("striped", "hover",
                                                  "condensed", "responsiv
                            full_width = F) %>%
  column_spec(1, width = "2cm", bold = T) %>%
  row_spec(0, bold = T)
```

Table 1: Probabilidades condicionales de ser político según género

	No político	**Político**
Femenino	0.811	0.189
Masculino	0.690	0.310

Además, la diferencia entre estas proporciones corresponde a: $(p_h) - (p_m) = 0.31 - 0.19 = 0.121$, lo que se traduce en que hay una mayor proporción de hombres políticos que de mujeres políticos cuya biografía se encuentra en wikipedia.

2. Usa un LPM para estimar la probabilidad de ser político en función del género. Escribe la ecuación de regresión correspondiente y presenta un `summary()` de los resultados. Explica el significado estadístico de cada coeficiente y su conexión con los resultados de la pregunta anterior.

Se estima un Modelo Lineal de Probabilidad (LPM) para evaluar la probabilidad de ser político en función del género. La ecuación de regresión correspondiente es:

$$\mathbb{P}(\text{político} \mid \text{género}) = \beta_0 + \beta_1(\text{género}) + \epsilon$$

Donde:

- β_0 es el intercepto que representa la probabilidad de que una mujer (categoría de referencia) sea política
- β_1 captura la diferencia en la probabilidad de ser político entre hombres y mujeres
- ϵ corresponde al término error.

Implementación en R:

```
model_1 <- lm(formula = politico ~ genero, data = db)
summary(model_1)
```

```
##
## Call:
## lm(formula = politico ~ genero, data = db)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3105 -0.3105 -0.3105  0.6895  0.8106
##
## Coefficients:
##                 Estimate Std. Error t value            Pr(>|t|)
## (Intercept)      0.18936    0.01065   17.78 <0.0000000000000002 ***
## generoMasculino  0.12114    0.01185   10.22 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.45 on 9284 degrees of freedom
## Multiple R-squared:  0.01113,    Adjusted R-squared:  0.01102
## F-statistic: 104.5 on 1 and 9284 DF,  p-value: < 0.00000000000000022
```

En este modelo, los resultados indican que el coeficiente asociado al género (β_1) es positivo, lo que sugiere que la probabilidad esperada de que los hombres sean políticos es 0.12 puntos porcentuales mayor en comparación con las mujeres. Este valor coincide con la diferencia de proporciones observada previamente ($dif = 0.121$), ya que refleja la magnitud de la diferencia

entre géneros en la probabilidad de ser político. El intercepto (β_0) representa la probabilidad de que una mujer sea política, que es 0.19, coincidiendo con el valor obtenido anteriormente en la tabla de contingencia (p_m = 0.19).

3. Usa un LPM para estimar la probabilidad de ser político en función del género, controlando por el año de nacimiento de los individuos. Escribe la ecuación de regresión correspondiente y presenta un `summary()` de los resultados. Explica el significado estadístico de cada coeficiente y provee una breve interpretación sustantiva.

Se estima un Modelo Lineal de Probabilidad (LPM) para evaluar la probabilidad de ser político en función del género y el año de nacimiento. La ecuación de regresión correspondiente es:

$$\mathbb{P}(\text{político} \mid \text{género, año nacimiento}) = \beta_0 + \beta_1(\text{género}) + \beta_2(\text{año nacimiento}) + \epsilon$$

Donde:

- β_0 es el intercepto, que representa la probabilidad estimada de que una mujer (categoría de referencia) nacida en el año cero sea política.
- β_1 captura la diferencia en la probabilidad de ser político entre hombres y mujeres, controlando por año de nacimiento.
- β_2 corresponde al efecto del año de nacimiento en la probabilidad de ser político, controlando por el género.
- ϵ corresponde al término error.

Implementación en R:

```
model_2 <- lm(formula = politico ~ genero +
                agno_nacimiento, data = db)
summary(model_2)
```

```
##
## Call:
## lm(formula = politico ~ genero + agno_nacimiento, data = db)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4715 -0.2683 -0.1543  0.3486  0.9604
##
## Coefficients:
##                     Estimate  Std. Error t value             Pr(>|t|)
## (Intercept)      6.37212330  0.16343985  38.988 < 0.0000000000000002 *
## generoMasculino  0.06090585  0.01114238   5.466         0.0000000472 *
## agno_nacimiento -0.00316628  0.00008355 -37.899 < 0.0000000000000002 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.4188 on 9283 degrees of freedom
## Multiple R-squared:  0.1436, Adjusted R-squared:  0.1434
## F-statistic: 778.5 on 2 and 9283 DF,  p-value: < 0.00000000000000022
```

Interpretación de los coeficientes

- El coeficiente del género es positivo (β = 0.061) , lo que se traduce en que los hombres obtienen, en promedio, una probabilidad de 0.06 puntos porcentuales mayor en ser político en comparación con las mujeres, manteniendo constante el año de nacimiento.
- El coeficiente del año de nacimiento es negativo (β = -0.003), lo indica que, por cada año adicional, la probabilidad de ser político disminuye en 0.003 puntos porcentuales promedio, controlando por el género.
- El intercepto es positivo (β = 6.37), y representa la probabilidad estimada de que una mujer nacida en el año cero sea política, lo que es claramente un valor no interpretable dentro del rango de probabilidades [0-1] y refleja una limitación del modelo.

Estos resultados sugieren que tanto el género como el año de nacimiento influyen en la probabilidad de ser político. Los hombres presentan una mayor probabilidad de ser políticos, mientras que las personas nacidas más recientemente tienen una menor probabilidad de serlo. Sin embargo, como se observa con el intercepto y el coeficiente del año de nacimiento, el LPM no es adecuado para estimar probabilidades en variables binarias, ya que puede producir valores fuera del rango [0, 1], como se evidencia en este caso. Este sesgo del LPM en contexto de variables categóricas dicotómicas será solventado en lo que sigue con modelos logit, que es una aproximación adecuada para este tipo de variables.

4. De acuerdo al modelo estimado en la pregunta anterior, ¿cuál es el efecto marginal del "año de nacimiento" sobre la probabilidad esperada de ser político?

En la Tabla 2 se presenta el efecto marginal del año de nacimiento sobre la probabilidad esperada de ser político. El efecto estimado es de -0.003, lo que indica que, manteniendo los otros factores constantes, por cada año adicional la probabilidad esperada de ser político disminuye en 0.003 puntos porcentuales. Este resultado sugiere una relación negativa entre el año de nacimiento y la probabilidad de ser político, controlando por el género.

Implementación en R:

```
marginaleffects::avg_slopes(model = model_2,
                            type = "response",
                            conf_level = .95) %>%
  as_tibble() %>%
  dplyr::mutate(
    term = if_else(term == "agno_nacimiento", "Año nacimiento",
                   "Género"),
    estimate = round(estimate, 3),
```

```
    ic_95 = paste0("[", round(conf.low, 3), ",",
                   round(conf.high, 3), "]")) %>%
  dplyr::select(term, contrast, estimate, std.error, ic_95) %>%
  slice_head() %>%
  kableExtra::kable(format = "latex",
                    col.names = c("Término", "Contraste",
                                  "Estimado", "ES", "IC 95%"),
                    booktabs = T,
                    align = 'c',
                    caption = paste("(\\#tab:table2)",
                                    "Efecto marginal de año de nacimiento
                                    de ser político")) %>%
  kableExtra::kable_styling(latex_options = "hold_position",
                            position = "center",
                            bootstrap_options = c("striped", "hover",
                                                  "condensed", "responsiv
                            full_width = F) %>%
  column_spec(1, width = "2cm",) %>%
  row_spec(0, bold = T)
```

Table 2: Efecto marginal de año de nacimiento sobre p de ser político

Término	**Contraste**	**Estimado**	**ES**	**IC 95%**
Año nacimiento	mean(dY/dX)	-0.003	0.0000835	[-0.003,-0.003]

5. En base al modelo usado en I.3., calcula las probabilidades esperadas de ser políticos para un hombre y una mujer que cuentan con una biografía en wikipedia y nacieron en 1973. Expresa formalmente las ecuaciones correspondiente a estas predicciones.

En base al Modelo 2 empleado en I.3, para calcular las probabilidades esperadas de que un hombre y una mujer nacidos en 1973 sean políticos, las ecuaciones serían:

Para hombre:

$$\mathbb{P}(\text{político} \mid \text{género = hombre, año nacimiento = 1973 }) = \beta_0 + \beta_1 * 1 + \beta_2$$

Para mujer:

$$\mathbb{P}(\text{político} \mid \text{género = mujer, año nacimiento = 1973 }) = \beta_0 + \beta_1 * 0 + \beta_2$$

Donde β_1 toma el valor de 0 para mujeres y 1 para hombres.

Implementación en R:

```
ggpredict(model_2, terms = c("genero", "agno_nacimiento [1973]")) %>%
  as_tibble() %>%
  dplyr::mutate(
        predicted = round(predicted, 3),
        std.error = round(std.error, 3),
        ic_95 = paste0("[", round(conf.low, 3), ",",
                       round(conf.high, 3), "]")) %>%
  dplyr::select(x, predicted, std.error, ic_95) %>%
  kableExtra::kable(format = "latex",
                    col.names = c("Género", "P esperada", "ES",
                                  "IC 95%"),
                    booktabs = T,
                    align = 'c',
                    caption = paste("(\\#tab:table3)",
                                    "Probabilidades predichas para ser po
                                    según género para personas nacidas en
  kableExtra::kable_styling(latex_options = "hold_position",
                            position = "center",
                            bootstrap_options = c("striped", "hover",
                                                  "condensed", "responsiv
                            full_width = F) %>%
  column_spec(1, width = "2cm",) %>%
  row_spec(0, bold = T)
```

Table 3: Probabilidades predichas para ser político según género para personas nacidas en 1973

Género	P esperada	ES	IC 95%
Femenino	0.125	0.010	[0.105,0.145]
Masculino	0.186	0.006	[0.174,0.197]

En la Tabla 3 se muestran las probabilidades predichas de ser político para un hombre y una mujer nacidos en 1973. Los resultados sugieren que la p esperada para los hombres (0.19) es mayor que para las mujeres (0.13). Esta diferencia sugiere que, manteniendo constante el año de nacimiento, el género sigue siendo un factor relevante para la probabilidad de ser político, con los hombres mostrando una probabilidad esperada superior a la de las mujeres.

6. Agrega una interacción entre `genero` y `agno_nacimiento` al modelo estimado en I.3. Escribe la ecuación de regresión y presenta un `summary()` de los resultados. Interpreta el efecto del año de nacimiento estimado en términos estadísticos y sustantivos.

Se estima un Modelo Lineal de Probabilidad (LPM) para la probabilidad de ser político en función del género, año de nacimiento y una interacción entre estas covariables. La ecuación de

regresión se resume de la siguiente forma:

$$\begin{aligned}\mathbb{P}(\text{político} \mid \text{género, año nacimiento }) = \beta_0 &+ \beta_1(\text{género}) \\ &+ \beta_2(\text{año nacimiento}) \\ &+ \beta_3(\text{género}) * (\text{año nacimiento}) + \epsilon\end{aligned}$$

Donde:

- β_0: Intercepto que representa la probabilidad de que una mujer (categoría de referencia) nacida en el año 0 sea política.
- β_1: Efecto del género, es decir, la diferencia en la probabilidad de ser político entre hombres y mujeres, controlando por el año de nacimiento.
- β_2: Efecto del año de nacimiento sobre la probabilidad de ser político, pero únicamente para mujeres (cuando género = femenino).
- β_3: Efecto de la interacción entre género y año de nacimiento, que captura cómo cambia el efecto del año de nacimiento en la probabilidad de ser político según el género.
- ϵ: Término de error.

Implementación en R:

```
model_3 <- lm(formula = politico ~ genero*agno_nacimiento,
              data = db)
summary(model_3)
```

```
##
## Call:
## lm(formula = politico ~ genero * agno_nacimiento, data = db)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6192 -0.2415 -0.1665  0.3052  0.9263
##
## Coefficients:
##                                   Estimate Std. Error t value
## (Intercept)                     -0.5557115  0.5052994  -1.100
## generoMasculino                  7.7678573  0.5327049  14.582
## agno_nacimiento                  0.0003816  0.0002587   1.475
## generoMasculino:agno_nacimiento -0.0039508  0.0002730 -14.471
##                                            Pr(>|t|)
## (Intercept)                                   0.271
## generoMasculino                 <0.0000000000000002 ***
## agno_nacimiento                               0.140
## generoMasculino:agno_nacimiento <0.0000000000000002 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4142 on 9282 degrees of freedom
## Multiple R-squared:  0.1625, Adjusted R-squared:  0.1623
## F-statistic: 600.4 on 3 and 9282 DF,  p-value: < 0.00000000000000022
```

De acuerdo con el modelo estimado, los resultados sugieren que el año de nacimiento no tiene un efecto sobre la probabilidad de ser político para las mujeres, controlando por las demás variables (β = 0.00, ES = 0.00). Este coeficiente refleja un efecto condicional, ya que la interpretación de su efecto principal aplica cuando la variable de interacción, género, es igual a cero (es decir, cuando el género es femenino). Esto quiere decir que no hay un único efecto del año de nacimiento, si no dos. Para los hombres, el año de nacimiento tiene un efecto relevante. El coeficiente negativo de la interacción indica que, a medida que aumenta el año de nacimiento, la probabilidad de ser político disminuye de manera más pronunciada para los hombres que para las mujeres (β = -0.003, ES = 0.00), lo que sugiere que las cohortes más jóvenes de hombres tienen una menor probabilidad de ser políticos en comparación con las cohortes más antiguas.

Enunciado II: Regresión Logística

1. Calcula la odds de que un hombre con biografìa en wikipedia sea político (odd_h) y de que una mujer con biografìa en wikipedia sea política (odd_m). Calcula el ratio entre ambas odds (hombre vs mujer) e interpreta el odds ratio resultante.

Para calcular las odds de que un hombre (Odd_H) o una mujer (Odd_M) con biografía en Wikipedia sea político/a, utilizamos la fórmula:

$$\text{Odds} = \frac{p}{1-p}$$

Donde p es la probabilidad de ser político/a. Las probabilidades estimadas para hombres y mujeres fueron previamente calculadas y presentadas en la Tabla 1, siendo 0.31 para hombres y 0.19 para mujeres.

$$\text{Odd}_H = \frac{0.31}{1-0.31} = \frac{0.31}{0.69} = 0.449$$

$$\text{Odd}_M = \frac{0.19}{1-0.19} = \frac{0.19}{0.81} = 0.233$$

La Odds Ratio θ que compara las odds de ser político entre hombres y mujeres, se calcula como:

$$\theta = \frac{\text{Odd}_H}{\text{Odd}_M} = \frac{0.449}{0.233} = 1.928$$

Este resultado indica que las odds de que un hombre sea político son aproximadamente 1.93

veces mayores que las odds para una mujer. En términos interpretativos, esto sugiere que, entre personas con biografía en Wikipedia, los hombres tienen una mayor probabilidad relativa de ser políticos en comparación con las mujeres.

2. Usa una regresión logística para estimar la log-odds de ser político en función del género. Escribe la ecuación de regresión correspondiente y presenta un `summary()` de los resultados. Explica el significado estadístico de cada coeficiente y su conexión con los resultados de la pregunta anterior.

Se estima un Modelo de Regresión Logística para evaluar la probabilidad de ser político en función del género. La ecuación de regresión que resume esta estimación se define como:

$$\ln\left[\frac{\mathbb{P}(\text{político} \mid \text{género})}{1-\mathbb{P}(\text{político} \mid \text{género})}\right] = \beta_0 + \beta_1(\text{género}) + \epsilon$$

Donde:

- β_0 es el intercepto que representa el logit de la probabilidad de ser político cuando género es mujer
- β_1 captura la diferencia en el logit de la probabilidad de ser político entre hombres y mujeres
- ϵ corresponde al término error.

Implementación en R:

```
model_4 <- glm(formula = politico ~ genero,
               data = db, family=binomial(link="logit"))
summary(model_4)
```

Call: glm(formula = politico ~ genero, family = binomial(link = “logit”), data = db)

Deviance Residuals: Min 1Q Median 3Q Max
-0.8623 -0.8623 -0.8623 1.5294 1.8243

Coefficients: Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.45420 0.06041 -24.07 <0.0000000000000002 ***generoMasculino 0.65638 0.06536 10.04 <0.0000000000000002*** — Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 11136  on 9285  degrees of freedom
```

Residual deviance: 11026 on 9284 degrees of freedom AIC: 11030

Number of Fisher Scoring iterations: 4

Los resultados del modelo sugieren que las log-odds de ser político para los hombres son 0.66 puntos mayores en comparación con las mujeres. Este coeficiente se puede convertir en un Odds Ratio (OR) mediante la exponenciación, resultando en un valor de 1.93. Esto indica

que las chances de que un hombre sea político son 1.93 veces mayores que las de una mujer, coincidiendo con el resultado obtenido en el punto anterior.

Por otro lado, el coeficiente del intercepto ($\beta = -1.45$, $ES = 0.06$) indica que las log-odds de ser política para las mujeres (la categoría de referencia) son negativas. Exponenciando este coeficiente, obtenemos que las chances de que una mujer sea política son 0.233, lo cual también coincide con el resultado anterior.

La razón por la cual los Odds Ratios de los coeficientes logit son equivalentes a las odds de ser político para hombres y mujeres es que el logit es el logaritmo natural de las odds. Al realizar la transformación inversa mediante la exponenciación, obtenemos las odds correspondientes a cada género.

3. Usa una regresión para estimar las log-odds de ser político en función del género, controlando por el año de nacimiento de los individuos. Escribe la ecuación de regresión correspondiente y presenta un `summary()` de los resultados. Explica el significado estadístico de cada coeficiente y provee una breve interpretación sustantiva.

Se estima un Modelo de Regresión Logística para evaluar la probabilidad de ser político en función del género y el año de nacimiento. La ecuación de regresión que resume esta estimación se define como:

$$\ln\left[\frac{\mathbb{P}(\text{político} \mid \text{género, año nacimiento})}{1-\mathbb{P}(\text{político} \mid \text{género, año nacimiento})}\right] = \beta_0 + \beta_1(\text{género}) + \beta_2(\text{año nacimiento}) + \epsilon$$

Donde:

- β_0 es el intercepto que representa el logit de la probabilidad de ser político cuando género es mujer y el año de nacimiento es 0
- β_1 captura la diferencia en el logit de la probabilidad de ser político entre hombres y mujeres, controlando por el año de nacimiento.
- β_2 mide cómo cambia el logit de la probabilidad de ser político con cada año adicional de nacimiento, manteniendo constante el género.
- ϵ corresponde al término error.

Implementación en R:

```
model_5 <- glm(formula = politico ~ genero + agno_nacimiento,
               data = db, family=binomial(link="logit"))
summary(model_5)
```

Call: glm(formula = politico ~ genero + agno_nacimiento, family = binomial(link = “logit”), data = db)

Deviance Residuals: Min 1Q Median 3Q Max
-3.1166 -0.7614 -0.5879 0.8611 2.1822

Coefficients: Estimate Std. Error z value Pr(>|z|)
(Intercept) 28.9710802 0.9664043 29.978 < 0.0000000000000002 ***generoMasculino 0.3660260 0.0688208 5.319 0.000000105*** agno_nacimiento -0.0156276 0.0004964 -31.484 < 0.0000000000000002 *** — Signif. codes: 0 *'' 0.001 '' 0.01 ''* 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

`Null deviance: 11136.4  on 9285  degrees of freedom`

Residual deviance: 9839.2 on 9283 degrees of freedom AIC: 9845.2

Number of Fisher Scoring iterations: 4

Interpretación de los coeficientes

- El coeficiente del género es positivo (β = 0.37), lo que sugiere que, manteniendo constante el año de nacimiento, las log-odds de ser político para los hombres son 0.37 puntos mayores en comparación con las mujeres. En términos de Odds Ratio (OR), las probabilidades de que un hombre sea político son aproximadamente 1.44 veces mayores que las de una mujer, controlando por el año de nacimiento.
- El coeficiente del año de nacimiento es negativo (β = -0.016), lo que indica que, manteniendo constante el género, por cada año adicional las log-odds de ser político disminuyen en 0.016 puntos. Esto implica que las personas más jóvenes tienen menores probabilidades de ser políticos en comparación con las personas mayores.
- El intercepto es positivo (β = 28.97), lo cual indica las log-odds de que una mujer nacida en el año 0 (un año irreal en el contexto de la muestra) sea política. Aunque este valor no tiene un significado sustantivo directo, funciona como punto de referencia para los cálculos de log-odds del modelo.

Con todo, los resultados sugieren que ser hombre incrementa las probabilidades de ser político, incluso al controlar por el año de nacimiento. Al mismo tiempo, el año de nacimiento tiene un efecto negativo, lo que sugiere que las cohortes más jóvenes tienen menos probabilidades de ser políticos.

4. De acuerdo al modelo estimado en la pregunta anterior, ¿cuál es la fórmula para el efecto marginal del "año de nacimiento" sobre la probabilidad esperada de ser político?

La fórmula para el efecto marginal del año de nacimiento sobre la probabilidad esperada de ser político p_i puede expresarse de la siguiente manera:

$$\frac{\partial p_i}{\partial \text{año nacimiento}} = \beta_{\text{año nacimiento}} * p_i(1 - p_i)$$

Donde:

- $\beta_{\text{año nacimiento}}$ es el coeficiente estimado del año de nacimiento en el modelo.
- p_i es la probabilidad predicha de que el individuo i sea político.

- $1 - p_i$ representa la probabilidad de que el individuo i no sea político.

Este efecto marginal indica cómo cambia la probabilidad de ser político para un cambio infinitesimal en el año de nacimiento, manteniendo constantes las demás variables del modelo.

5. De acuerdo al modelo estimado en II.3., ¿cual es el efecto marginal del año de nacimiento sobre la la probabilidades esperadas de ser políticos para un hombre y una mujer que cuentan con una biografía en wikipedia y nacieron en 1973. Expresa formalmente las ecuaciones correspondiente dichos efectos. Compara los resultados con la respuesta dada en I.4.

El efecto marginal para el año de nacimiento sobre las probabilidades predichas de ser político/a para un hombre y una mujer nacidos en 1973 se puede expresar de la siguiente manera:

Para hombre:

$$\mathbb{P}(\text{político} \mid \text{género = hombre, año nacimiento = 1973 }) = \frac{1}{1 + e^{\beta_0 + \beta_1 * 1 + \beta_2}}$$

Mientras que el efecto marginal sería:

$$\frac{\partial p_{\text{hombre}}}{\partial \text{año nacimiento}} = \beta_{\text{año nacimiento}} * p_{\text{hombre}}(1 - p_{\text{hombre}})$$

Para mujer:

$$\mathbb{P}(\text{político} \mid \text{género = mujer, año nacimiento = 1973 }) = \frac{1}{1 + e^{\beta_0 + \beta_1 * 0 + \beta_2}}$$

Mientras que el efecto marginal sería:

$$\frac{\partial p_{\text{mujer}}}{\partial \text{año nacimiento}} = \beta_{\text{año nacimiento}} * p_{\text{mujer}}(1 - p_{\text{mujer}})$$

Implementación en R:

```
newdata <- data.frame(
  genero = c("Masculino", "Femenino"),
  agno_nacimiento = 1973
)

marginaleffects::avg_slopes(model = model_5,
                            variables = "agno_nacimiento",
                            by = "genero",
                            conf_level = .95,
                            newdata = newdata) %>%
  as_tibble() %>%
  dplyr::mutate(
```

```
    estimate = round(estimate, 4),
    ic_95 = paste0("[", round(conf.low, 4), ",",
                   round(conf.high, 4), "]")) %>%
  dplyr::select(genero, contrast, estimate, std.error, ic_95) %>%
  kableExtra::kable(format = "latex",
                    col.names = c("Género", "Contraste",
                                  "Estimado", "ES", "IC 95%"),
                    booktabs = T,
                    align = 'c',
                    caption = paste("(\\#tab:table4)",
                                    "Efecto marginal de año de nacimiento
                                    de ser político según género")) %>%
  kableExtra::kable_styling(latex_options = "hold_position",
                            position = "center",
                            bootstrap_options = c("striped", "hover",
                                                  "condensed", "responsiv
                            full_width = F) %>%
  column_spec(1, width = "2cm",) %>%
  row_spec(0, bold = T)
```

Table 4: Efecto marginal de año de nacimiento sobre p de ser político según género

Género	Contraste	Estimado	ES	IC 95%
Femenino	mean(dY/dX)	-0.0018	0.0000916	[-0.002,-0.0016]
Masculino	mean(dY/dX)	-0.0023	0.0000564	[-0.0024,-0.0022]

En la Tabla 4 se muestran los efectos marginales del año de nacimiento sobre la probabilidad de ser político para un hombre y una mujer nacidos en 1973. Los resultados muestran que el efecto marginal estimado del año de nacimiento es negativo tanto para hombres como para mujeres: manteniendo constantes los demás factores, por cada año adicional, los log-odds de ser político disminuyen en 0.0023 puntos para los hombres y en 0.0018 puntos para las mujeres. Esto sugiere una relación negativa entre el año de nacimiento y la log-odds de ser político, siendo más pronunciada para los hombres.

En la sección I.4, se sostuvo que el efecto marginal del año de nacimiento sobre la probabilidad esperada de ser político era de -0.003 puntos porcentuales en un modelo de probabilidad lineal (LPM). Al comparar estos resultados con los efectos marginales derivados del modelo logit en esta sección, se observa que, aunque ambos indican una relación negativa con el año de nacimiento, la magnitud y el signo de los efectos marginales en el modelo logit son más sutiles y muestran que el efecto negativo es menos pronunciado en términos de log-odds.

Enunciado III: Bonus

1. Agrega una interacción entre `genero` y `agno_nacimiento` al modelo estimado en II.3 y presenta un `summary()` de los resultados.

Implementación en R:

```
model_6 <- glm(formula = politico ~ genero*agno_nacimiento,
               data = db, family=binomial(link="logit"))
summary(model_6)
```

Call: glm(formula = politico ~ genero * agno_nacimiento, family = binomial(link = “logit”), data = db)

Deviance Residuals: Min 1Q Median 3Q Max
-3.3726 -0.7111 -0.6001 0.7800 2.1007

Coefficients: Estimate Std. Error z value (Intercept) -6.638093 3.332294 -1.992 generoMasculino 40.455108 3.498909 11.562 agno_nacimiento 0.002653 0.001704 1.557 generoMasculino:agno_nacimiento -0.020607 0.001792 -11.500 Pr(>|z|)
(Intercept) 0.0464 *
generoMasculino <0.0000000000000002 ***agno_nacimiento 0.1195***
generoMasculino:agno_nacimiento <0.0000000000000002 — Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ’’ 0.05 ‘.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 11136.4  on 9285  degrees of freedom
```

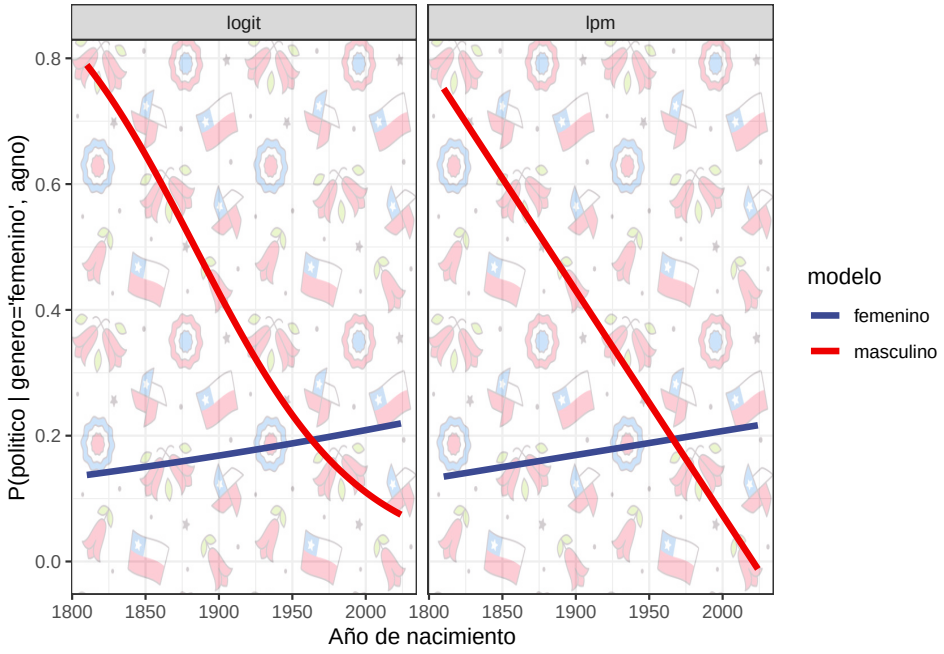
Residual deviance: 9678.1 on 9282 degrees of freedom AIC: 9686.1

Number of Fisher Scoring iterations: 4

2. Reproduce el siguiente gráfico que muestra – en base a los LMP y logit más complejos (con interacción) – las probabilidad predichas de ser polìtico para hombres y mujeres con biografìa en wikipedia nacidos entre los años 1810 y 2024. Compara principales resultados arrojados por ambos modelos.

```
p_load("tinytex","tidyverse","modelr","httr","ggsci","png","grid","margin

wiki_chileans <- rio::import(file = "https://github.com/mebucca/cda_soc30

lpm_3 <- lm(politico ~ factor(genero)*agno_nacimiento, data=wiki_chileans

logit_3 <- glm(politico ~ factor(genero)*agno_nacimiento, family=binomial

# crea un nuevo set de datos sobre los cuales crear predicciones
```

```r
newx <- expand.grid(
  genero = c('femenino', 'masculino'),
  agno_nacimiento = 1810:2024
)

newx$pred_logit <- predict(logit_3, newdata = newx, type = "response")

# crea valores predichos para el nuevo set de datos
xb_lpm = predict(lpm_3 , newdata = newx)
xb_logit = predict(logit_3, newdata = newx)
prob_lpm = xb_lpm
prob_logit = 1/(1 + exp(-xb_logit))

newy <- newx %>% mutate(prob_lpm = prob_lpm, prob_logit = prob_logit) %>%
          pivot_longer(c(prob_lpm,prob_logit), names_to = "model", names_


# Load the Chilean flag image (make sure to have the image in your workin
flag <- readPNG(here("homework/chilean_flag.png"))

# If the image doesn't already have an alpha channel (transparency), you
# Assuming the image is RGB, create an RGBA version by adding an alpha ch
flag_with_alpha <- array(0, dim = c(dim(flag)[1], dim(flag)[2], 4))
flag_with_alpha[,,1:3] <- flag[,,1:3]  # Copy RGB channels
flag_with_alpha[,,4] <- 0.2  # Set alpha to 0.2 for transparency (adjust

# Add the image as a background to the plot with transparency
newy %>% ggplot(aes(x = agno_nacimiento, y = value, group=genero, colour
  annotation_custom(rasterGrob(flag_with_alpha,
                               width = unit(1, "npc"),
                               height = unit(1, "npc")),
                    -Inf, Inf, -Inf, Inf) + # Ensures the image covers th
  geom_line(linewidth = 1.5) +
  labs(y = "P(politico | genero='femenino', agno)", x = "Año de nacimient
  facet_grid( . ~ model ) +
  scale_color_aaas() +
  theme_bw() +
  theme(panel.background = element_blank()) # Ensures a blank background
```



3. ¿En la regresión logística, cuál es el mayor efecto marginal posible de año de nacimiento sobre la probabilidad de ser político para hombres y mujeres? Compara con el respectivo efecto marginal en el LPM.

El efecto marginal de año de nacimiento en el modelo logit estimado en el apartado anterior era de -0.00182 para mujeres y de -0.00234 para hombres. Para obtener el mayor efecto marginal posible, debemos evaluar el coeficiente en su valor máximo, es decir, cuando $p = 0.5$. Un atajo matemático para obtener el mismo resultado es dividir por 4 el efecto marginal obtenido:

Para mujeres:

$$\text{Max marginal mujer} = \frac{-0.00182}{4} = -0.000455$$

Para hombres:

$$\text{Max marginal hombre} = \frac{-0.00234}{4} = -0.000585$$

En el LPM del apartado I.2, el efecto marginal del año de nacimiento era de -0.003, el cual es constante o no tiene una diferenciación según género.

De este modo, el efecto marginal en el modelo logístico es más pequeño que en el LPM, pero depende de la probabilidad p y varía según los valores predichos de p, mientras que en el LPM es constante.