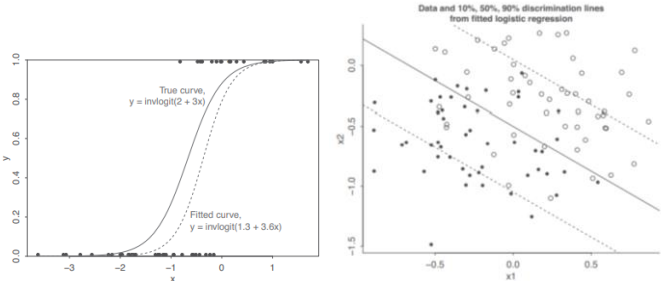
Chapter 14
Working with logistic regression

14.1 Graphing logistic regression and binary data

El capítulo inicia mostrando gráficas de la regresión logística (a la izquierda un ejemplo de ajuste del modelo y a la derecha líneas correspondientes a P(y=1)=0.1, 0.5, 0.9



14.2 Logistic regression with interactions

Se trabajará de ahora en adelante con el caso "well-switching example" sobre los niveles de arsénico en pozos de agua, cuya concentración es independiente entre las ubicaciones de los pozos. Se hizo una campaña para que la gente consumiera sólo agua del pozo seguro y no del más cercano que tenga, y un año después se pregunto a las personas si hicieron el cambio ("switch")

Se empieza con un modelo con 2 predictores e interacciones entre esos 2 predictores obteniendo los siguientes resultados e interpretaciones del modelo:

	Median	MAD_SD
(Intercept)	-0.15	0.12
dist100	-0.58	0.21
arsenic	0.56	0.07
dist100:arsenic	-0.18	0.10

- Para entender se requiere:
  - Evaluar predicciones e interacciones en la media de las variables
  - Dividir entre 4 para obtener las diferencias predictivas aproximadas en la escala de probabilidad
- Para interpretar el intercepto le sacamos la inversa de la función logística y evaluamos en los valores medios que es de 0.48 para dist100, 1.66 para arsénico como: logit^{-1}(-0.15 -0.58*0.48 + 0.56*1.66 -0.18*0.48*1.66)=0.59 obteniendo así la probabilidad de cambiar. La proba de cambiar si el pozo está a 0 metros del más cercano y el nivel de arsénico del pozo actual es 0, es de 0.47=$\log^{-1}$(-0.15)
- Respecto al coeficiente de distancia, al igual que con el intercepto, no se recomienda intentar interpretarlo así como está. En su lugar nos podemos fijar en los valores promedio y obtener: -0.58-0.18*1.66 igual a -0.88 en escala logarítmica, y para interpretarlo de forma rápida dividimos entre 4 obteniendo: -0.88/4= -0.22 [*recordar que $\beta/4$ es la diferencia máxima en Pr (y = 1) correspondiente a una diferencia unitaria en x*]. Entonces, a un nivel promedio de arsénico (1.66), cada 100 metros de distancia reduce en 22% la proba de cambiar. La interpretación del coeficiente de arsénico se hace en forma análoga
- Respecto al coeficiente de la interacción, para cada unidad adicional de arsénico, el valor -0.18 se suma al coeficiente de distancia. Considerando que el coeficiente de distancia es -0.88 a un nivel promedio de arsénico, entonces podemos entender la interacción diciendo que la importancia de la distancia como predictor aumenta para los hogares con niveles más altos de arsénico existente (tiene una interpretación análoga desde el punto de vista del arsénico)

Centering the input variables

Se hace para facilitar la interpretación del modelo, sin embargo las predicciones son las mismas que con el modelo sin centrar

	Median	MAD_SD
(Intercept)	0.35	0.04
c_dist100	-0.88	0.10
c_arsenic	0.47	0.04
c_dist100:c_arsenic	-0.18	0.10

Statistical significance of the interaction

El estimador de la interacción (-0.18) no está a dos errores estándar (0.10) de cero y por lo tanto, no es estadísticamente significativo. Comparando el LOO log score con el modelo que no incluia la interacción, nos señala que añadir la interacción no cambia el desempeño predictivo y no es necesario conservarla para hacer una predicción

```
Model comparison:
(negative 'elpd_diff' favors 1st model, positive favors 2nd)
elpd_diff        se
      0.6       1.9
```

Adding social predictors

Se hizo un modelo con las variables dist100, arsenic, aedu (años de educación entre 4, para separarlo en bloques). Encontrando que las gente con más educación tiene más chance de cambiarse (porque 0.17/4=4% diferencia positiva en la probabilidad de cambio cuando se comparan hogares que difieren en 4 años de educación), y que se mejora el log score predictivo (8.5), aunque hay incertidumbre considerable (4.9). Cuando los predictores tienen efectos considerables, como práctica general se incluyen las interacciones

```
            Median MAD_SD
(Intercept) -0.22   0.09
dist100     -0.90   0.11
arsenic      0.47   0.04
educ4        0.17   0.04
elpd_diff         se
      8.5        4.9
```

Adding further interactions

Se agrgó la interacción entre distancia y educación, así como arsénico y educación, encontrando que las personas con más educación están sí están dispuestas a recorrer mayor distancia (por quizá tener acceso a otros recursos), y que entre más educados más propensos a cambiarse ante un mayor nivel de arsénico

Standardizing predictors

Cuando se usan interacciones en los modelos se debe considerar seriamente en standarizar los predictores

14.3 Predictive simultaion

Se puede usar simulaciones para hacer pronósticos de probabilidad

Simulating the uncertainity in the estimated coefficients

Puedes hacer las simulaciones con:

```
fit <- stan_glm(switch ~ dist100, family=binomial(link="logit"), data=wells)
sims <- as.matrix(fit)
n_sims <- nrow(sims)
```

Predictive simulation using the binomial distribution

Para predecir el comportamiento de cambio de nuevos jefes de familia dada una matriz de predictores puedes usar la distribución binomial como sigue:

```
n_new <- nrow(X_new)
y_new <- array(NA, c(n_sims, n_new))
for (s in 1:n_sims){
  p_new <- invlogit(X_new %*% sims[s,])
  y_new[s,] <- rbinom(n_new, 1, p_new)
}
```

Predictive simulation using the latent logistic distribution

Una forma alternativa de simular predicciones de regresión logística utiliza la formulación de datos latentes [*la variable latente es* $z_i = X_i\beta + \varepsilon_i$], agrega errores independientes al predictor lineal y luego convierte a datos binarios.

14.4 Average predictive comparisons on the probability scale

Si se tienen muchos predictores o el graficar no es conveniente, puedes plantear calcular/observar cuanto cambia la probabilidad de P(y=1) si se mueve solamente una variable dejando a las demás fijas, con:

$E(y|u^{hi}, v, \theta) - E(y|u^{lo}, v, \theta)$, donde: **u** la unidad con la que quieres ver cómo cambia la proba
u^{hi} el valor más alto de u y **u^{lo}** el valor más bajo de u
v todos los otros predictores
θ los coeficientes

Problems with evaluating predictive comparisons at a central value

Para la aproximación mencionada arriba, se pueden dejar fijar los otros predictores en su media o mediana, aunque podríamos encontrar problemas si nuestro espacio de predictores está muy disperso o si son variables binarias o bimodales

Demostration with the well-swithching example
Con un modelo del pozo donde sólo considera distancia, arsénico y educación (sin interacciones) se ejemplifica como sacar las "diferencias predictivas" cuando cambiamos la distancia de 0 a 1

	Median	MAD_SD
(Intercept)	-0.22	0.09
dist100	-0.90	0.10
arsenic	0.47	0.04
educ4	0.17	0.04

$$\delta(\text{arsenic}, \text{educ4}) = \text{logit}^{-1}(-0.22 - 0.90 * 1 + 0.47 * \text{arsenic} + 0.17 * \text{educ4}) - \text{logit}^{-1}(-0.22 - 0.90 * 0 + 0.47 * \text{arsenic} + 0.17 * \text{educ4}). \quad = -0.21$$

En los datos promedio, el dueño de la casa que está a 100 metros del pozo seguro más cercano es 21% menos dispuesto a cambiarse.

Comparing probabilities of swithching for households differing in arsenic levels
Se hace la "diferencia predictiva" con cálculos análogos para arsénico de 0.5 y 1.0

Average predictive difference in probability of switching, comparing householders with 0 and 12 years of education
Se hace la "diferencia predictiva" con cálculos análogos para años de educación de 0 y 12

Average predictive comparison in the presence of interactions
Se hace la "diferencia predictiva" con cálculos análogos para distancia en 1 y 0

	Median	MAD_SD
(Intercept)	0.35	0.04
c_dist100	-0.92	0.10
c_arsenic	0.49	0.04
c_educ4	0.19	0.04
c_dist100:c_educ4	0.33	0.11
c_arsenic:c_educ4	0.08	0.04

```
b <- coef(fit_8)
hi <- 1
lo <- 0
delta <- invlogit(b[1] + b[2]*hi + b[3]*wells$c_arsenic + b[4]*wells$c_educ4 +
            b[5]*hi*wells$c_educ4 + b[6]*wells$c_arsenic*wells$c_educ4) -
          invlogit(b[1] + b[2]*lo + b[3]*wells$c_arsenic + b[4]*wells$c_educ4 +
            b[5]*lo*wells$c_educ4 + b[6]*wells$c_arsenic*wells$c_educ4)
round(mean(delta), 2)
```

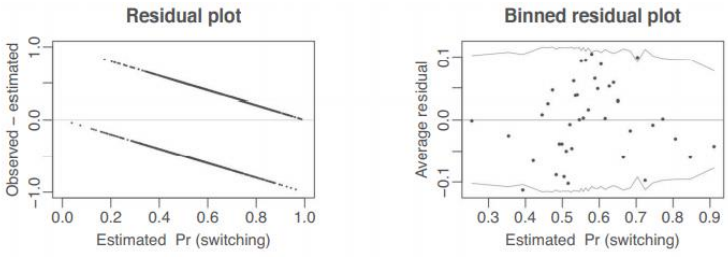
which comes to −0.21.

14.5 Residuals for discrete-data regresion

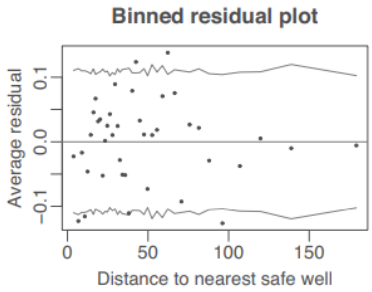
Se puede definir residuales para la regresión logística como sigue:

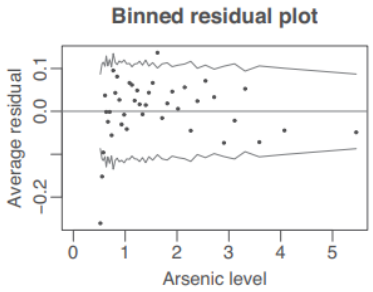
$$\text{residual}_i = y_i - \text{E}(y_i|X_i) = y_i - \text{logit}^{-1}(X_i\beta).$$

Si $\text{logit}^{-1}(X_i\beta)=0.7$ entonces $\text{residual}_i=-0.7$ o 0.3 si $y_i=0$ o $y_i=1$; por lo anterior no es útil graficar los residuales. En su lugar se hacen categorías (bins) establecidos a partir de los valores ajustados



Plotting binned residuals versus inputs of interest
Se puede agrupar y graficar los residuos con respecto a variables individuales o combinadas



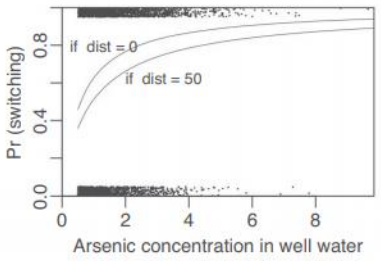


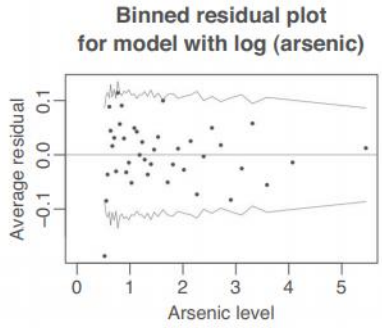
Improving a model by transformation

Al observar en la gráfica de residuales patrones de alza y baja (como la gráfica anterior -derecha-) se sugiere aplicar escala logaritmica sobre arsénico (dado que son todos valores positivos) o agregar un termino cuadrático (en el caso de distancia no es recomendable aplicar logaritmo dado que la gráfica muestra un buen ajuste)

Se hizo un modelo que incluyó distacia, log_arsénico, educación, distancia:educación, log_arsénico:educación

La gráfica de residuales se observa mejor que la anterior aunque mantiene problemas al final





Error rate and comparison to the null model

La tasa de error se define como la proporción de casos en los que la predicción determinista es incorrecta ($y_i = 1$ if $\text{logit}^{-1}(X_i\beta) > 0.5$ and guessing $y_i = 0$ if $\text{logit}^{-1}(X_i\beta) < 0.5$). El error debe ser siempre menor a 0.5 (de lo contrario, se podría establecer todos los $\beta=0$ y obtener un modelo de mejor ajuste)