Multiple Logistic Regression

Response: $Y$ binary, $\pi = P(Y = 1)$

Explanatory variables: $x_1, x_2, \ldots, x_k$
can be quantitative, qualitative (dummy variables), or both.

Model form is
\[ \logit(\pi) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \]
or equivalently
\[ \pi = \frac{\exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}{1 + \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)} \]

$\beta_i$ = partial effect of $x_i$ controlling for other variables in model
$e^{\beta_i}$ = conditional odds ratio at $x_i + 1$ vs at $x_i$ keeping other $x$'s fixed
= multiplicative effect on odds of 1-unit increase in $x_i$

$w/ \text{ other } x$'s fixed

Example (Horseshoe Crabs)

Model 1:
\[ \logit(\pi) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \]

$\pi = \frac{\exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}{1 + \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}$ $\pi$ = odds

Below “odds” = odds having at least one satellite

For female crabs of the same width,

- $e^{\beta_3}$ = odds ratio of (med. dark vs. med. light)
- $e^{\beta_4}$ = odds ratio of (dark vs. med. light)

Example (Horseshoe Crabs)

In addition to Width ($X$), consider adding a categorical predictor
- Color, coded 1-4 as

\[ 1 = \text{medium light}, 2 = \text{medium}, 3 = \text{medium dark}, 4 = \text{dark} \]

For a categorical predictor, need to create a dummy variable (= indicator variable) for each category:

- $c_1 = \begin{cases} 1 & \text{medium light} \\ 0 & \text{o/w} \end{cases}$
- $c_2 = \begin{cases} 1 & \text{medium} \\ 0 & \text{o/w} \end{cases}$
- $c_3 = \begin{cases} 1 & \text{medium dark} \\ 0 & \text{o/w} \end{cases}$
- $c_4 = \begin{cases} 1 & \text{dark} \\ 0 & \text{o/w} \end{cases}$

Model: $\logit(\pi) = \alpha + \beta_1 c_1 + \beta_2 c_2 + \beta_3 c_3 + \beta_4 c_4 + \beta x$

- $c_1 + c_2 + c_3 + c_4 = 1$ always true, so one of them is redundant.
- To account for redundancies, most software set one of $
\beta_1, \beta_2, \beta_3, \beta_4$ to 0

For female crabs of the same width,

- $e^{\beta_3}$ = odds ratio of (med. dark vs. med. light)
- $e^{\beta_4}$ = odds ratio of (dark vs. med. light)

Observe $e^{\beta_1}$'s are odds ratios of a category vs. the baseline category (medium light), for crabs of the same width.

Observe the effect of Color does not change with Width
Example (Horseshoe Crabs)

Model 1: \[ \text{odds} = \frac{\pi}{1 - \pi} = e^{\alpha + \beta_2 c_2 + \beta_3 c_3 + \beta_4 c_4 + \beta x} \]

For female crabs of same color but different width \( x_1, x_2 \),

odds for crabs of Width \( x_1 \):
\[ \text{odds for Width } x_1 = e^{\alpha + \beta_2 c_2 + \beta_3 c_3 + \beta_4 c_4 + \beta x_1} \]

odds for crabs of Width \( x_2 \):
\[ \text{odds for Width } x_2 = e^{\alpha + \beta_2 c_2 + \beta_3 c_3 + \beta_4 c_4 + \beta x_2} \]

\( \Rightarrow \) Width have the same effect for all colors.

As neither the effect of color change with width, nor the effect of width change with color, we said Model 1 assumes no interaction between color and width effects.

\[ \logit(\hat{\pi}) = -11.39 + 0.07 c_2 - 0.22 c_3 - 1.33 c_4 + 0.468 x \]

\begin{align*}
\logit(\hat{\pi}) &= \begin{cases} 
-11.39 + 0.468 x & \text{if medium light} \\
-11.32 + 0.468 x & \text{if medium} \\
-11.61 + 0.468 x & \text{if medium dark} \\
-12.72 + 0.468 x & \text{if dark}
\end{cases}
\end{align*}

Observe the four curves have the same shape because they have identical coefficient for Width.

Chapter 4 - 5

R regards Color (coded 1-4) as a numeric variable. The R command \texttt{as.factor()} can create the dummy variables.

\begin{verbatim}
> C = as.factor(Color)
> crabs.fit1 = glm(has.sate ~ C + Weight, family = binomial)
> crabs.fit1$coef

(Intercept) C2 C3 C4 Width
-11.38519276 0.07241694 -0.22379766 -1.32991913 0.46795598
\end{verbatim}

The fitted model is

\[ \logit(\hat{\pi}) = -11.39 + 0.07 c_2 - 0.22 c_3 - 1.33 c_4 + 0.468 x \]

For a medium light female \( (c_2 = c_3 = c_4 = 0) \) of width \( x = 25 \) cm,

\[ \hat{\pi} = \frac{\exp(-11.39 + 0.468 \times 25)}{1 + \exp(-11.39 + 0.468 \times 25)} \approx 0.58 \]

For a dark female \( (c_2 = c_3 = 0, c_4 = 1) \) of width \( x = 25 \) cm,

\[ \hat{\pi} = \frac{\exp(-11.39 + (-1.33)(1) + 0.468 \times 25)}{1 + \exp(-11.39 + (-1.33)(1) + 0.468 \times 25)} \approx 0.265. \]

Chapter 4 - 6

Medium v.s. Medium Light Crabs

\begin{verbatim}
Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.3852  2.87346 -3.962  7.43e-05 ***
C2        0.0724   0.07989   0.908     0.369
C3       -0.2238  0.77708  -0.288     0.773
C4       -1.3299  0.85252  -1.560     0.119
Width     0.4679   0.10554   4.434  9.26e-06 ***
\end{verbatim}

\begin{itemize}
  \item Interpretation of \( \beta_2 \): estimated odds of having satellite(s) for medium crabs are \( e^{\hat{\beta}_2} = e^{0.07} \approx 1.07 \) times the estimated odds for medium light crabs of the same width.
  \item \( H_0 : \beta_2 = 0 \) means medium and medium light crabs do not differ in their chance of having satellite(s) given width. To test
  \[ H_0 : \beta_2 = 0 \quad \text{v.s.} \quad H_a : \beta_2 \neq 0 \]
  Wald statistic \( z = \frac{\hat{\beta}_2}{SE} = \frac{0.072}{0.074} = 0.098, \ P\text{-value} = 0.922. \]
  \item Conclusion: Medium light and medium crabs of the same width don’t differ significantly in the prob. of having satellites.
\end{itemize}

Chapter 4 - 7
95% LR CI for $\beta_2$ is $(-1.54, 1.45)$, which contains 0. So LR test also fail to reject $H_0$: $\beta_2 = 0$.

> confint(crabs.fit1, test="Chisq")

<table>
<thead>
<tr>
<th></th>
<th>2.5 %</th>
<th>97.5 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-17.3084</td>
<td>-5.985952</td>
</tr>
<tr>
<td>C2</td>
<td>-1.5397</td>
<td>1.2396</td>
</tr>
<tr>
<td>C3</td>
<td>-1.8919</td>
<td>1.3157</td>
</tr>
<tr>
<td>C4</td>
<td>-3.1357</td>
<td>0.2780</td>
</tr>
<tr>
<td>Width</td>
<td>0.2713</td>
<td>0.6870</td>
</tr>
</tbody>
</table>

What about (medium dark v.s. medium light) crabs? What about (dark v.s. medium light) crabs?

What about Medium v.s. Dark Crabs?

For medium and dark crabs of the same width, the odds ratio is

$$\frac{\text{odds for a medium crab}}{\text{odds for a dark crab}} = e^{\alpha + \beta_2 + \beta_x} / e^{\alpha + \beta_4 + \beta_x} = e^{\beta_2 - \beta_4}.$$ 

Estimated odds of having satellite(s) for a medium crab is

$$e^{\hat{\beta}_2 - \hat{\beta}_4} = e^{0.07 - (1.33)} = e^{1.4} \approx 4.06$$

times the estimated odds for a dark crabs of the same width.

However, to test $H_0 : \beta_2 = \beta_4$, need SE for $\hat{\beta}_2 - \hat{\beta}_4$, which is not provided in R.

The simplest solution is to change the baseline category. Say, use dark color as the baseline and model as

Model 1a : $\logit(\pi) = \alpha' + \beta'_1 c_1 + \beta'_2 c_2 + \beta'_3 c_3 + \beta_x$

The two models are equivalent, just a change of parameters.

$$\alpha' = \alpha + \beta_4, \quad \beta'_i = \beta_i - \beta_4 \quad \text{for } i = 1, 2, 3$$

Testing $\beta_2 = \beta_4$ in Model 1 is equivalent to testing $\beta'_2 = 0$ in Model 1a.
> drop1(crabs.fit1a,test="Chisq")
Single term deletions
Model:
  has.sate ~ C1 + C2 + C3 + Width
<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Deviance</th>
<th>AIC</th>
<th>LRT</th>
<th>Pr(&gt;Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td>1</td>
<td>187.46</td>
<td>197.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>1</td>
<td>190.07</td>
<td>198.07</td>
<td>2.6154</td>
<td>0.105831</td>
</tr>
<tr>
<td>C2</td>
<td>1</td>
<td>194.37</td>
<td>202.37</td>
<td>6.9101</td>
<td>0.008571 **</td>
</tr>
<tr>
<td>C3</td>
<td>1</td>
<td>191.11</td>
<td>199.11</td>
<td>3.6518</td>
<td>0.056010 .</td>
</tr>
<tr>
<td>Width</td>
<td>1</td>
<td>212.06</td>
<td>220.06</td>
<td>24.6038</td>
<td>7.041e-07 ***</td>
</tr>
</tbody>
</table>

LR test of $\beta_2 = 0$ gives $P$-value 0.0086, same conclusion as Wald test

> confint(crabs.fit1a)
Waiting for profiling to be done...

<table>
<thead>
<tr>
<th></th>
<th>2.5 %</th>
<th>97.5 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-18.46</td>
<td>-7.58</td>
</tr>
<tr>
<td>C1</td>
<td>-0.27</td>
<td>3.13</td>
</tr>
<tr>
<td>C2</td>
<td>0.35</td>
<td>2.53</td>
</tr>
<tr>
<td>C3</td>
<td>-0.03</td>
<td>2.32</td>
</tr>
<tr>
<td>Width</td>
<td>0.27</td>
<td>0.69</td>
</tr>
</tbody>
</table>

95% for $\beta_2$ is (0.35, 2.52) $\Rightarrow$ estimated odds for medium crabs are at least $e^{0.35} \approx 1.42$, at most $e^{2.52} \approx 12.5$ times the est. odds for dark crabs of the same width.

Chapter 4 - 13

Likelihood Ratio Test for Model Comparison

Likelihood Ratio Test can be used to do model comparison between a **simpler model** and a **more complex model**.

- The simpler model must be a **special case** of the more complex model. If not, CANNOT use LRT to do model comparison

- $H_0$: the simpler model is correct
  $H_a$: the complex model is correct, the simpler model is not

- Rejecting $H_0$ means the simpler model doesn’t fit the data well, compared to the more complex model

- Not rejecting $H_0$ means the simpler model fits the data nearly as well as the more complex model

Chapter 4 - 14

Likelihood Ratio Test for Model Comparison

Rather than reporting the max. log-likelihood for a model, R reports

$$\text{Deviance} = -2(\max \text{ log-likelihood} + C)$$

in which $C$ is a constant depends only on the data but not the model. So

$$\text{LR statistic} = -2(L_0 - L_1)$$

$$= -2(L_0 + C) - [-2(L_1 + C)]$$

$$= \text{diff. in deviance for the two models}$$

- We will introduce deviance in Chapter 5
- d.f. for a deviance is
  $$(\text{num. of observations}) - (\text{num. of parameters})$$
  so d.f. for a LR statistic = diff. in d.f. for the two deviances
- LR test for model comparison is also called “**analysis of deviance**”
Example (Horseshoe Crabs)

Do We Need Color in the Model?

H₀ : β₂ = β₃ = β₄ = 0  (given width, Y indep. of color)

i.e.,

H₀ : logit(π) = α + βₓ

Hₐ : logit(π) = α + β₂c₂ + β₃c₃ + β₄c₄ + βₓ  (complex model)

> anova(crabs.logit, crabs.fit1, test="Chisq")

Analysis of Deviance Table

Model 1: has.sate ~ Width
Model 2: has.sate ~ C + Width

Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1 171 194.45
2 168 187.46 3 6.9956 0.07204 .

The LR statistic = diff. of deviance = 194.45 - 187.46 = 6.99
with df = 171 - 168 = 3, P-value = 0.072

⇒ Some evidence (not strong) of a color effect given width.

Chapter 4 - 18

Other simpler models might be adequate.

Plot of the four curves on Slide 9 suggests that maybe only dark crabs are different from others.

Model 2: logit(π) = α + β₄c₄ + βₓ, where c₄ = \begin{cases} 1 & \text{dark} \\ 0 & \text{o/w} \end{cases}

Fitting gives \( \hat{\beta}_4 = -1.300 \) (SE = 0.5259).

Odds of satellites for a dark crab is estimated to be \( e^{-1.300} = 0.27 \) times the odds a non-dark crab of the same width.

> crabs.fit2 = glm(has.sate ~ I(Color==4) + Width, family = binomial)
> summary(crabs.fit2)

```
Call: glm(formula = has.sate ~ I(Color==4) + Width, family = binomial)

Deviance Residuals:
        Min         1Q     Median         3Q        Max
-2.1124     -0.9848      0.5243      0.8513      2.1413

Coefficients:                 Estimate Std. Error z value  Pr(>|z|)
(Intercept)                -11.67903    2.69252   -4.338  1.44e-05 ***
I(Color == 4)TRUE               -1.30049    0.52593    -2.473     0.0134 *
Width                         0.478201    0.10410    4.592   4.39e-06 ***
```
Compare model with 1 dummy for color to full model with 3 dummies.

\[
H_0: \logit(\pi) = \alpha + \beta_4 c_4 + \beta x \quad \text{(simple model)}
\]

\[
H_a: \logit(\pi) = \alpha + \beta_2 c_2 + \beta_3 c_3 + \beta_4 c_4 + \beta x \quad \text{(more complex model)}
\]

Note \(H_0\) is \(\beta_2 = \beta_3 = 0\) in more complex model.

\[
> \text{anova(crabs.fit2, crabs.fit1, test="Chisq")}
\]

Analysis of Deviance Table

<table>
<thead>
<tr>
<th>Model 1: has.sate ~ I(Color == 4) + Width</th>
<th>Model 2: has.sate ~ C + Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resid. Df</td>
<td>Resid. Dev</td>
</tr>
<tr>
<td>1 170</td>
<td>187.96</td>
</tr>
<tr>
<td>2 168</td>
<td>187.46</td>
</tr>
</tbody>
</table>

LR stat = diff. in deviances = 187.96 - 187.46 = 0.50
\(df = 170 - 168 = 2\), \(P\)-value = 0.7785

Simpler model is adequate.

Chapter 4 - 21

Does model treating color as nominal fit as well as model treating it as qualitative?

\[
H_0: \logit(\pi) = \alpha + \gamma c + \beta x \quad \text{(simpler (ordinal) model)}
\]

\[
H_a: \logit(\pi) = \alpha + \beta_2 c_2 + \beta_3 c_3 + \beta_4 c_4 + \beta x \quad \text{(more complex model)}
\]

\[
> \text{anova(crabs.fit3, crabs.fit1, test="Chisq")}
\]

Analysis of Deviance Table

<table>
<thead>
<tr>
<th>Model 1: has.sate ~ Color + Width</th>
<th>Model 2: has.sate ~ C + Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resid. Df</td>
<td>Resid. Dev</td>
</tr>
<tr>
<td>1 170</td>
<td>189.12</td>
</tr>
<tr>
<td>2 168</td>
<td>187.46</td>
</tr>
</tbody>
</table>

LR stat = diff. in deviances = 189.12 - 187.46 = 1.66
\(df = 170 - 168 = 2\), \(P\)-value = 0.4351

Simpler model is adequate.

Chapter 4 - 22

> \text{anova(crabs.fit3, crabs.fit1, test="Chisq")}

Analysis of Deviance Table

<table>
<thead>
<tr>
<th>Model 1: has.sate ~ Color + Width</th>
<th>Model 2: has.sate ~ C + Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resid. Df</td>
<td>Resid. Dev</td>
</tr>
<tr>
<td>1 170</td>
<td>189.12</td>
</tr>
<tr>
<td>2 168</td>
<td>187.46</td>
</tr>
</tbody>
</table>

LR stat = diff. in deviances = 189.12 - 187.46 = 1.66
\(df = 170 - 168 = 2\), \(P\)-value = 0.4351

Simpler model is adequate.

Chapter 4 - 23

Models Allowing Interactions

\[
\logit(\pi) = \alpha + \beta_2 c_2 + \beta_3 c_3 + \beta_4 c_4 + \beta x + \gamma_2 c_2 x + \gamma_3 c_3 x + \gamma_4 c_4 x
\]

\[
= \begin{cases} 
\alpha + \beta x & \text{if medium light} \\
\alpha + \beta_2 + (\beta + \gamma_2) x & \text{if medium} \\
\alpha + \beta_3 + (\beta + \gamma_3) x & \text{if medium dark} \\
\alpha + \beta_4 + (\beta + \gamma_4) x & \text{if dark}
\end{cases}
\]

Different colors have different coefficient for “Width.”

Chapter 4 - 24
> crabs.fit4 = glm(has.sate ~ C + Width + C:Width, family = binomial)
> summary(crabs.fit4)

Call:
glm(formula = has.sate ~ C + Width + C:Width, family = binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.75261  11.46409  -0.153   0.878
C2         -8.28735  12.00363  -0.690   0.490
C3        -19.76545  13.34251  -1.481   0.139
C4         -4.10122  13.27532  -0.309   0.757
Width      0.10600   0.42656   0.248   0.804
C2:Width   0.31287   0.44794   0.698   0.485
C3:Width   0.75237   0.50435   1.492   0.136
C4:Width   0.09443   0.50042   0.189   0.850

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 225.76 on 172 degrees of freedom
Residual deviance: 183.08 on 165 degrees of freedom

Testing H₀: no interaction (γ₂ = γ₃ = γ₄ = 0)

> anova(crabs.fit1, crabs.fit4, test="Chisq")

Analysis of Deviance Table

Model 1: has.sate ~ C + Width
Model 2: has.sate ~ C + Width + C:Width

<table>
<thead>
<tr>
<th>Resid. Df</th>
<th>Resid. Dev</th>
<th>Df</th>
<th>Deviance</th>
<th>Pr(&gt;Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>168</td>
<td></td>
<td>187.46</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>165</td>
<td>3</td>
<td>4.3764</td>
<td>0.2236</td>
</tr>
</tbody>
</table>

LR stat = diff. in deviances = 187.46 − 183.08 = 4.3764
df = 168 − 165 = 3, P-value = 0.2236

Simpler model is adequate (no interaction).