Ch 4.3.3 and 4.3.4 - Multiple and Multinomial Logistic Regression

Lecture 7 - CMSE 381

Michigan State University
::
Dept of Computational Mathematics, Science & Engineering

January 31, 2024

Covered in this lecture

Last Time:

- Logistic Regression

This time:

- More on Logistic Regression
- Multiple Logistic Regression
- Multinomial Logistic Regression

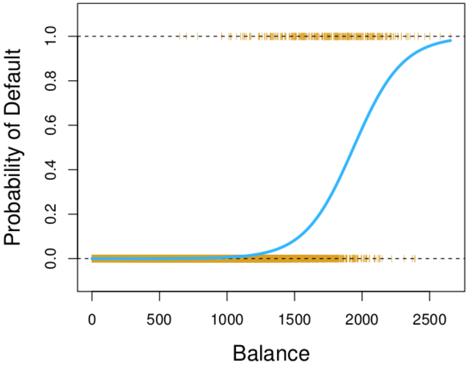
Section 1

Review of Logistic Regression from last time

Logistic regression

- Assume single input $\mathtt{X}$
- Output takes values $\mathtt{Y} \in \{\mathtt{Yes}, \mathtt{No}\}$

$$p(\mathtt{X}) = \Pr(\mathtt{Y} = \mathtt{yes} \mid \mathtt{balance})$$



$$p(\mathtt{x}) = \frac{e^{\beta_0 + \beta_1 \mathtt{x}}}{1 + e^{\beta_0 + \beta_1 \mathtt{x}}}$$

prob of default given X

How to get logistic function

Assume the (natural) log odds (logits) follow a linear model

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$$

Solve for $p(x)$:

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

Playing with the logistic function: desmos.com/calculator/cw1pyzzqci

How to perform logistic regression?

gradient descent

Given $p(\mathrm{x}) = \frac{e^{\beta_0+\beta_1 \mathrm{x}}}{1+e^{\beta_0+\beta_1 \mathrm{x}}}$ and the training data $\{(x_i, y_i)\}_{i=1}^m$. How to estimate β_0,β_1?

Maximum Likelihood:
The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to maximize the likelihood function.

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\max_{\beta_0, \beta_1} \ell(\beta_0, \beta_1) = \prod_{i: y_i = 1} p(x_i) \prod_{i': y_{i'} = 0} (1 - p(x_{i'}))$$

if $y_i = 1$, $p(x_i) \approx 1$
if $y_i = 0$, $p(x_i) \approx 0$

β_0 and β_1 are such that the predicted conditional probability is as close as possible to the individual's observed default status.

Example

	Balance	Prediction
1.	0	*No*
2.	500	*No*
3	1000	*No*
4	1500	*Yes*
5	2000	*Yes*
6	2500	*Yes*

$$\prod_{i,\, y_i = 1} p(x_i) = p(x_4)\, p(x_5)\, p(x_6)$$

$$= \frac{e^{\beta_0 + \beta_1 1500}}{1 + e^{\beta_0 + \beta_1 \cdot 1500}} \cdot \quad \cdot$$

$$\prod_{i,\, y_i = 0} 1 - p(x_i) = (1 - p(x_1))(1 - p(x_2))(1 - p(x_3))$$

$$= \left(1 - \frac{e^{\beta_0}}{1 + e^{\beta_0}}\right) \cdot \quad \cdot$$

Section 2

Multiple Logistic Regression

New assumption

$p \geq 1$ input variables

Y output variable has only two levels

$X_1, X_2, \cdots, X_p$

Multiple Logistic Regression

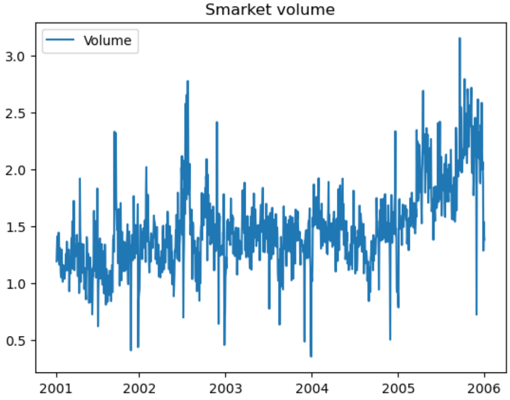
Multiple features:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

Equivalent to:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Example from Smarket data



	Lag1	Lag2	Volume	Direction
1	0.381	-0.192	1.19130	Up
2	0.959	0.381	1.29650	Up
3	1.032	0.959	1.41120	Down
4	-0.623	1.032	1.27600	Up
5	0.614	-0.623	1.20570	Up
...	...	...	...	...
1246	0.422	0.252	1.88850	Up
1247	0.043	0.422	1.28581	Down
1248	-0.955	0.043	1.54047	Up
1249	0.130	-0.955	1.42236	Down
1250	-0.298	0.130	1.38254	Down

1250 rows × 4 columns

Goal in lab was predicting direction from three input variables

Our Results

```
X = smarket[['Lag1','Lag2','Volume']]
Y = smarket.Direction

clf = LogisticRegression(random_state=0)
clf.fit(X,Y)
```

```
LogisticRegression
LogisticRegression(random_state=0)
```

$$p(X) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}$$

```
print(clf.coef_)
print(clf.intercept_)
```

```
[[-0.07302967 -0.04272162  0.12862433]]
[-0.1158254]
```

$$p(X) = \frac{\exp(-0.115 - 0.073 \cdot \texttt{Lag1} - 0.043 \cdot \texttt{Lag2} + 0.129 \cdot \texttt{Volume})}{1 + \exp(-0.115 - 0.073 \cdot \texttt{Lag1} - 0.043 \cdot \texttt{Lag2} + 0.129 \cdot \texttt{Volume})}$$

Section 3

Multinomial Logistic Regression

New assumption

$p \geq 1$ input variables

$X_1, X_2, \cdots, X_p$

Y output variable has K levels

symmetric

Region:

	x_{i1}	x_{i2}	X_{i1}	X_{i2}	X_{i3}
South	1	0	1	0	0
West	0	1	0	1	0
East	0	0	0	0	1

Create spare dummy variables:

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person from South} \\ 0 & \text{if } i\text{th person not from South} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person from West} \\ 0 & \text{if } i\text{th person not from West} \end{cases}$$

Baseline is the level we're not using

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

$\hat{y}_1$ $\qquad$ $\hat{y}_2$ $\qquad$ $\hat{y} = w\hat{y}_1 + (1-w)\hat{y}_2$, $w \in [0,1]$

Predict $Y \in \{\texttt{stroke}, \texttt{overdose}, \texttt{seizure}\}$ for hospital visits based on some input(s) X

$$\Pr(Y = \texttt{stroke} \mid X = x) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \qquad 0$$

$$\Pr(Y = \texttt{overdose} \mid X = x) = 0 \qquad \frac{e^{\tilde{\beta}_0 + \tilde{\beta}_1 x}}{1 + e^{\tilde{\beta}_0 + \tilde{\beta}_1 x}}$$

baseline $\quad \Pr(Y = \texttt{seizure} \mid X = x) = 1 - \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \qquad 1 - \frac{e^{\tilde{\beta}_0 + \tilde{\beta}_1 x}}{1 + e^{\tilde{\beta}_0 + \tilde{\beta}_1 x}}$

- We're going to figure out three numbers for any given input x, then pick the one with the highest probability
- Note that if we know two we can figure out the third

$$w = \frac{1 + e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X} + e^{\tilde{\beta}_0 + \tilde{\beta}_1 X}}$$

Example

Predict $Y \in \{\texttt{stroke, overdose, seizure}\}$ for hospital visits based on Xp

$$\Pr(Y = \texttt{stroke} \mid X = x) = \frac{\exp(\beta_{\texttt{str},0} + \beta_{\texttt{str},1}x)}{1 + \exp(\beta_{\texttt{str},0} + \beta_{\texttt{str},1}x) + \exp(\beta_{\texttt{OD},0} + \beta_{\texttt{OD},1}x)}$$

$$\Pr(Y = \texttt{overdose} \mid X = x) = \frac{\exp(\beta_{\texttt{OD},0} + \beta_{\texttt{OD},1}x)}{1 + \exp(\beta_{\texttt{str},0} + \beta_{\texttt{str},1}x) + \exp(\beta_{\texttt{OD},0} + \beta_{\texttt{OD},1}x)}$$

$$\Pr(Y = \texttt{seizure} \mid X = x) = \frac{1}{1 + \exp(\beta_{\texttt{str},0} + \beta_{\texttt{str},1}x) + \exp(\beta_{\texttt{OD},0} + \beta_{\texttt{OD},1}x)}$$

Note that using seizure is the baseline

Multinomial Logistic Regression

Plan A

- Assume `Y` has K levels
- Make K (the last one) the baseline

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}$$

$$\Pr(Y = K|X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}.$$

Log odds

Calculated so that log odds between *any pair of* classes is linear.
Specifically, for $Y = k$ vs $Y = K$, we have

$$\log\left(\frac{\Pr(Y = k \mid X = x)}{\Pr(Y = K \mid X = x)}\right) = \beta_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kp}x_p$$

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}$$

$e^{\beta_{k0}+\beta_{k1}\cdots}$

$$\Pr(Y = K|X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}.$$

Plan B: Softmax coding

Treat all levels symmetrically

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{\sum_{l=1}^{K} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}.$$

Calculated so that log odds between two classes is linear

$$\log\left(\frac{\Pr(Y = k|X = x)}{\Pr(Y = k'|X = x)}\right) = (\beta_{k0} - \beta_{k'0}) + (\beta_{k1} - \beta_{k'1})x_1 + \cdots + (\beta_{kp} - \beta_{k'p})x_p.$$

Softmax example

$$\Pr(Y = \texttt{stroke} \mid X = x)$$
$$= \frac{\exp(\beta_{\texttt{str},0} + \beta_{\texttt{str},1}x)}{\exp(\beta_{\texttt{str},0} + \beta_{\texttt{str},1}x) + \exp(\beta_{\texttt{OD},0} + \beta_{\texttt{OD},1}x) + \exp(\beta_{\texttt{seiz},0} + \beta_{\texttt{seiz},1}x)}$$

$$\Pr(Y = \texttt{overdose} \mid X = x)$$
$$= \frac{\exp(\beta_{\texttt{OD},0} + \beta_{\texttt{OD},1}x)}{\exp(\beta_{\texttt{str},0} + \beta_{\texttt{str},1}x) + \exp(\beta_{\texttt{OD},0} + \beta_{\texttt{OD},1}x) + \exp(\beta_{\texttt{seiz},0} + \beta_{\texttt{seiz},1}x)}$$

$$\Pr(Y = \texttt{seizure} \mid X = x)$$
$$= \frac{\exp(\beta_{\texttt{seiz},0} + \beta_{\texttt{seiz},1}x)}{\exp(\beta_{\texttt{str},0} + \beta_{\texttt{str},1}x) + \exp(\beta_{\texttt{OD},0} + \beta_{\texttt{OD},1}x) + \exp(\beta_{\texttt{seiz},0} + \beta_{\texttt{seiz},1}x)}$$