Chapter 4 – Classification

Wenjing Liao
School of Mathematics
Georgia Institute of Technology

Math 4803
Fall 2019
Outline

1. 4.2 Why not linear regression?

2. 4.3 Logistic regression
Binary qualitative response

Example: predict the medical condition of a patient in the emergency room on the basis of their symptoms

Binary response: stroke and drug overdose

\[ Y = \begin{cases} 
0 & \text{if stroke;} \\
1 & \text{if drug overdose.} 
\end{cases} \]

Prediction: linear regression \( X\hat{\beta} \) as an estimate of \( \Pr(\text{drug overdose}|X) \) and predict drug overdose if \( \hat{Y} > 0.5 \).

Invariant to coding: If we flit the coding above, linear regression will produce the same prediction.

Problem: \( \hat{Y} \) may not belong to \([0, 1]\).
Qualitative response with more than two levels

Three responses: stroke, drug overdose and epileptic seizure

\[ Y = \begin{cases} 
1 & \text{if stroke;} \\
2 & \text{if drug overdose;} \\
3 & \text{if epileptic seizure.}
\end{cases} \]

Problem:

- Different coding would produce fundamentally different linear models that would ultimately lead to different sets of predictions on test data.

- The dummy variable cannot be easily extended to qualitative variables with more than two levels.
4.2 Why not linear regression?

4.3 Logistic regression
Probability model for binary response

\[ \text{Default} = \text{yes or no} \]
\[ \Pr(\text{default} = \text{Yes}|\text{balance}) \]

**FIGURE 4.2.** Classification using the Default data. Left: Estimated probability of default using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for default (No or Yes). Right: Predicted probabilities of default using logistic regression. All probabilities lie between 0 and 1.
Logistic function

Logistic function:

\[ p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}. \]

Odds:

\[ \frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X}. \]

Take values between 0 and \( \infty \) indicating low or high probabilities of default. For example, \( p(X) = 0.2 \) implies an odds of \( 1/4 \) and \( p(X) = 0.9 \) implies an odds of 9.

Log-odds (logit):

\[ \log \left( \frac{p(X)}{1 - p(X)} \right) = \beta_0 + \beta_1 X. \]
Coefficient estimation

Maximum likelihood:

\[ \ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})). \]

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Std. error</th>
<th>Z-statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-10.6513</td>
<td>0.3612</td>
<td>-29.5</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>balance</td>
<td>0.0055</td>
<td>0.0002</td>
<td>24.9</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

**TABLE 4.1.** For the Default data, estimated coefficients of the logistic regression model that predicts the probability of default using balance. A one-unit increase in balance is associated with an increase in the log odds of default by 0.0055 units.

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Std. error</th>
<th>Z-statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-3.5041</td>
<td>0.0707</td>
<td>-49.55</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>student[Yes]</td>
<td>0.4049</td>
<td>0.1150</td>
<td>3.52</td>
<td>0.0004</td>
</tr>
</tbody>
</table>

**TABLE 4.2.** For the Default data, estimated coefficients of the logistic regression model that predicts the probability of default using student status. Student status is encoded as a dummy variable, with a value of 1 for a student and a value of 0 for a non-student, and represented by the variable student[Yes] in the table.
### Prediction

#### Making predictions: balance $X = 1,000$

$$
\hat{p}(X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}} = \frac{e^{-10.6513 + 0.0055 \times 1,000}}{1 + e^{-10.6513 + 0.0055 \times 1,000}} = 0.00576,
$$

$$
X = 2,000 \rightarrow \hat{p}(X) = 58.6\%.
$$

If we predict default from student,

$$
\hat{\Pr}(\text{default}=\text{Yes}|\text{student}=\text{Yes}) = \frac{e^{-3.5041 + 0.4049 \times 1}}{1 + e^{-3.5041 + 0.4049 \times 1}} = 0.0431,
$$

$$
\hat{\Pr}(\text{default}=\text{Yes}|\text{student}=\text{No}) = \frac{e^{-3.5041 + 0.4049 \times 0}}{1 + e^{-3.5041 + 0.4049 \times 0}} = 0.0292.
$$
Multiple logistic regression

Log-odds and odds:

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

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

Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Std. error</th>
<th>Z-statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>−10.8690</td>
<td>0.4923</td>
<td>−22.08</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>balance</td>
<td>0.0057</td>
<td>0.0002</td>
<td>24.74</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>income</td>
<td>0.0030</td>
<td>0.0082</td>
<td>0.37</td>
<td>0.7115</td>
</tr>
<tr>
<td>student[Yes]</td>
<td>−0.6468</td>
<td>0.2362</td>
<td>−2.74</td>
<td>0.0062</td>
</tr>
</tbody>
</table>

**TABLE 4.3.** For the Default data, estimated coefficients of the logistic regression model that predicts the probability of default using balance, income, and student status. Student status is encoded as a dummy variable student[Yes], with a value of 1 for a student and a value of 0 for a non-student. In fitting this model, income was measured in thousands of dollars.
Interpretation

Contradiction? The coefficient for student becomes negative.

FIGURE 4.3. Confounding in the Default data. Left: Default rates are shown for students (orange) and non-students (blue). The solid lines display default rate as a function of balance, while the horizontal broken lines display the overall default rates. Right: Boxplots of balance for students (orange) and non-students (blue) are shown.

- Multiple and single logistic regression
- Student and balance are correlated.
Student versus non-student

A student with a credit card balance of 1,500 and income 40,000

\[ \hat{p}(X) = \frac{e^{-10.869+0.00574\times1,500+0.003\times40-0.6468\times1}}{1 + e^{-10.869+0.00574\times1,500+0.003\times40-0.6468\times1}} = 0.058. \]

A non-student with a credit card balance of 1,500 and income 40,000

\[ \hat{p}(X) = \frac{e^{-10.869+0.00574\times1,500+0.003\times40-0.6468\times0}}{1 + e^{-10.869+0.00574\times1,500+0.003\times40-0.6468\times0}} = 0.105. \]

However, students on average have a higher credit balance.
Reference