Logistic Regression

Idea: Predict categorical variable using regression

Examples

- surgery survival dependent on age, length of surgery, ...
- whether purchase occurs dependent on age, income, web-site characteristics, ...
- whether speech error occur as alcohol level increases
- when linguistic rules apply (final [t] in Dutch) dependent on speed of utterance, stress, social group, ...

Ver very popular, especially in sociolinguistics.
Regression Techniques Attractive

- allow prediction of one variable value based on one or more others
- allow an estimation of the importance of various independent factors (cf. $\chi^2$)
Outline Logistic Regression

Idea: Predict categorical variable using regression

- core task: analyze dependency of categorical variable on others using regression
- problem: translating regression techniques to categorical domain
- key step: predict **chance of** categorical variable
  — transforming categorical to numeric variable
- note: independent variables may be numeric or categorical — as in regression in general, simple or multiple
Chance as Dependent Variable

Idea: Predict chance of categorical variable as dependent variable using regression

- real chances $p$ are positive numbers $0 \leq p \leq 1$
- problem: how to keep predicted values in correct bounds
- solution: don’t use chances directly, but rather a more complicated transformation
Logit\( (p) \) = \ln \frac{p}{(1-p)}

\[
\begin{array}{ccccccccccc}
 p & 0.01 & 0.05 & 0.10 & 0.30 & 0.5 & 0.7 & 0.9 & 0.95 & 0.99 \\
\text{logit}(p) & -4.6 & -2.9 & -2.2 & -0.8 & 0.0 & 0.8 & 2.2 & 2.9 & 4.6 \\
\end{array}
\]
Logit(p) vs. Logistic

- use of logit solves problems of bounds—we predict logit values \(-\infty \leq v \leq \infty\) (cf. chances \(0 \leq p \leq 1\))
- logit is easily interpretable as “odds”
  - “the odds of Real against Ajax are 4 to 1”
    - probability is 0.8, \(p/(1 - p) = 0.8/0.2 = 4/1\)
- why the name ‘logistic’?
Why ‘logistic’?

\[ f(x) = \frac{1}{1 + e^{-x}} \]

Similarly constrains predicted value \( v \): \( 0 \leq v \leq 1 \)
Logistic vs. Logit Functions

\[
\begin{align*}
\ln \frac{p}{1-p} & = \logit(p) \\
\frac{p}{1-p} & = e^{\logit(p)} \\
p & = e^{\logit(p)}(1 - p) \\
p & = e^{\logit(p)} - pe^{\logit(p)} \\
p + pe^{\logit(p)} & = e^{\logit(p)} \\
p(1 + e^{\logit(p)}) & = e^{\logit(p)} \\
p & = \frac{e^{\logit(p)}}{1 + e^{\logit(p)}} \times \frac{e^{-\logit(p)}}{e^{-\logit(p)}} \\
p & = \frac{1}{1 + e^{-\logit(p)}}
\end{align*}
\]
Strategy: Predict Logit Values

\[ \text{logit}(p) = \beta_0 + \beta_1 x, \text{ where } x \text{ is the independent variable} \]

- try to find optimal \( \beta_0, \beta_1 \) given data
- note that we’re seeking a **nonlinear** relationship
Example: Labov’s NYC /r/ study

William Labov examined variant pronunciations of syllable-final /r/ in American English ([r] vs [ə]). New York used to be like Boston, final /r/ is [ə], but it started changing in the 1950’s and 1960’s. Labov hypothesized a social basis for the change.

\[
\begin{array}{c|c|c|c}
 & Saks & Macy’s & S.Klein \\
\hline
N & 68 & 125 & 71 \\
\hline
32 & & 31 & 17 \\
30 & & 20 & \\
\end{array}
\]

\( /\ell / \) allophones

mixed [r,ə]

all cons. [r]

high social stratum Saks Macy’s S.Klein low social stratum
# Data on NYC /r/

<table>
<thead>
<tr>
<th>Social Status</th>
<th>Pronunciation of /r/</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cons. ([r])</td>
<td>vocalic ([ə])</td>
<td>mixed</td>
</tr>
<tr>
<td>high</td>
<td>30</td>
<td>6</td>
<td>32</td>
</tr>
<tr>
<td>medium</td>
<td>20</td>
<td>74</td>
<td>31</td>
</tr>
<tr>
<td>low</td>
<td>4</td>
<td>50</td>
<td>17</td>
</tr>
</tbody>
</table>

What stat. test is needed to ask **whether** soc. status influences pronunciation of /r/?
Analyzing Social Influence on /r/

What stat. test is needed to ask whether soc. status influences pronunciation of /r/?

- $\chi^2$ test of independence (see that section)
  —is one nominal variable dependent on another?
- we exercise logistic regression for two reasons:
  - to measure the degree of dependence
  - to combine with questions of further dependence
Simplifying the Question

Eliminate the “mixed-r reports”:

<table>
<thead>
<tr>
<th>Social Status</th>
<th>Pronunciation of /r/</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cons. ([r])</td>
<td>vocalic ([ə])</td>
</tr>
<tr>
<td>high</td>
<td>30</td>
<td>6</td>
</tr>
<tr>
<td>medium</td>
<td>20</td>
<td>74</td>
</tr>
<tr>
<td>low</td>
<td>4</td>
<td>50</td>
</tr>
</tbody>
</table>

- now we’re predicting a **dichotomous** (two-valued) variable (instead of a polytomous one). Note that the predictor is still polytomous.
- this step would be questionable if the category being eliminated dominated
Coding

- we code /r/ as '0, vocalic' and '1, consonantal'
- remember the “weight by frequency” command
- SPSS offers several alternatives for the Independent Variable (Status)
- “dummy” coding (SPSS: “indicator”) is recommended:

<table>
<thead>
<tr>
<th>Status</th>
<th>explanation</th>
<th>dummy-1</th>
<th>dummy-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(high, Saks)</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>(mid, Macy’s)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>(low, S.Klein)</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
## SPSS Output—Coding

### Dependent Variable Encoding:

<table>
<thead>
<tr>
<th>Original Value</th>
<th>Internal Value</th>
<th>[vocalic pronunciation]</th>
<th>[consonantal &quot; ]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Parameter

<table>
<thead>
<tr>
<th>Value</th>
<th>Freq</th>
<th>Coding 1</th>
<th>Coding 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOC_STAT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1.000</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>
### SPSS Output

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
<th>Sig</th>
<th>R</th>
<th>Exp(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOC_STAT</td>
<td>43.90</td>
<td>2.000</td>
<td>.42</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOC_STAT(1)</td>
<td>4.13</td>
<td>.69</td>
<td>36.38</td>
<td>1</td>
<td>.000</td>
<td>.39</td>
<td>62.49</td>
</tr>
<tr>
<td>SOC_STAT(2)</td>
<td>1.22</td>
<td>.58</td>
<td>4.44</td>
<td>1</td>
<td>.035</td>
<td>.10</td>
<td>3.38</td>
</tr>
<tr>
<td>Constant</td>
<td>-2.53</td>
<td>.52</td>
<td>23.63</td>
<td>1</td>
<td>.000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Recall that we’re finding the parameters to the following equation:

\[
\logit(p) = \beta_0 + \beta_1 s_1 + \beta_2 s_2 \\
= -2.5 + 4.1 s_1 \\
= -2.5 + 1.2 s_2 \\
= -2.5
\]
Interpreting SPSS Output

\[
\begin{align*}
\text{logit}(p) &= -2.5 + 4.1s_1 \\
&= -2.5 + 1.2s_2 \\
&= -2.5 \\
&= -2.5 + 4.1 = 1.6 \\
&= -2.5 + 1.2 = -1.3 \\
&= -2.5
\end{align*}
\]

Saks, \( s_1 = 1 \)
Macy’s, \( s_2 = 1 \)
S.Klein, \( s_1 = s_2 = 0 \)
Saks
Macy’s
S.Klein
Checking Interpretation of Output

\[
\ln \left( \frac{p}{1-p} \right) = 1.6 \quad \text{Saks} \\
= -1.3 \quad \text{Macy’s} \\
= -2.5 \quad \text{S.Klein}
\]

<table>
<thead>
<tr>
<th>(\ln \left( \frac{p}{1-p} \right))</th>
<th>(\frac{p}{1-p})</th>
<th>(p)</th>
<th>(\approx)</th>
<th>\text{Store}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.6</td>
<td>30/6</td>
<td>(\approx 0.84)</td>
<td>Saks</td>
<td></td>
</tr>
<tr>
<td>-1.3</td>
<td>20/74</td>
<td>(\approx 0.21)</td>
<td>Macy’s</td>
<td></td>
</tr>
<tr>
<td>-2.5</td>
<td>4/50</td>
<td>(\approx 0.07)</td>
<td>S.Klein</td>
<td></td>
</tr>
</tbody>
</table>

These indeed match the data to be predicted.
### SPSS Output

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
<th>Sig</th>
<th>R</th>
<th>Exp(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOC_STAT</td>
<td>43.90</td>
<td>2</td>
<td>.000</td>
<td>.42</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOC_STAT(1)</td>
<td>4.13</td>
<td>.69</td>
<td>36.38</td>
<td>1</td>
<td>.000</td>
<td>.39</td>
<td>62.49</td>
</tr>
<tr>
<td>SOC_STAT(2)</td>
<td>1.22</td>
<td>.58</td>
<td>4.44</td>
<td>1</td>
<td>.035</td>
<td>.10</td>
<td>3.38</td>
</tr>
<tr>
<td>Constant</td>
<td>-2.53</td>
<td>.52</td>
<td>23.63</td>
<td>1</td>
<td>.000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note that:

- **all** variables are significant
- a kind of $r (-1 \leq R \leq 1)$ is being estimated
  —without the **certainty** that $r^2, R^2$ indicates explained variance
- $\text{Exp (B)} = e^\beta$
## Understanding SPSS Output

Classification Table for UITSPRK

The Cut Value is .50

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Percent Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

Overall 83.70%
# Predictions, Correctness

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Percent Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>[@]</td>
<td></td>
</tr>
<tr>
<td>[r]</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted</th>
<th>Percent Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>124 [I] 6</td>
<td>95.38%</td>
</tr>
<tr>
<td>1</td>
<td>24 [I] 30</td>
<td>55.56%</td>
</tr>
</tbody>
</table>

Overall 83.70%

This shows the prediction of the variable coded for status.

Note that we’re predicting that Saks’s pronunciations should be all [r] and the others all [@] (schwa).
Variance in the binomial case is $p(1 - p)$, and variance of the number of observations is $p^k(1 - p)^{(n-k)}$ where the positive value $[r]$ was seen $k$ times and the null value $(n - k)$ times. From this we derive the log likelihood $L$:

$$L = \ln p^k (1 - p)^{(n-k)} = k \ln p + (n - k) \ln(1 - p)$$

We measure the quality of the model using log likelihood and estimating the parameters to obtain the optimal value:

It also turns out that $-2L$ has a $\chi^2$ distribution with $(n - 1)$ degrees of freedom.
Very likely events \((p \approx 1)\) contribute little to log likelihoods.
We measure the quality of the model using log likelihood and estimating the parameters to obtain the optimal value. We obtain the **optimal** value by using the overall frequencies as a best guess:

<table>
<thead>
<tr>
<th>Social Status</th>
<th>Pronunciation of /r/</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cons. ([r])</td>
</tr>
<tr>
<td>high</td>
<td>30</td>
</tr>
<tr>
<td>medium</td>
<td>20</td>
</tr>
<tr>
<td>low</td>
<td>4</td>
</tr>
<tr>
<td><strong>totals</strong></td>
<td><strong>54</strong></td>
</tr>
<tr>
<td>best guess</td>
<td>0.293</td>
</tr>
</tbody>
</table>
Simplest Model—No Social Class

We measure the quality of the model using log likelihood and estimating the parameters to obtain the optimal value.

\[
L = k \ln p + (n - k) \ln(1 - p) \\
= 54 \ln(0.293) + 130 \ln(0.707) \\
= 54(-1.23) + 130(-0.35) \\
= -66.4 + -45.1 = -111.5 \\
-2L = 223
\]

This is the simplest model.

We then turn to the model which distinguishes Saks from everything else.
Parameters in New Model

We examine the new model, which distinguishes two classes, for which distinct “best guesses” are obtained, again using the empirical frequencies:

<table>
<thead>
<tr>
<th>Social Status</th>
<th>Pronunciation of /r/</th>
<th></th>
<th></th>
<th>prop. r</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cons. ([r])</td>
<td>vocalic ([ɜ])</td>
<td></td>
<td></td>
</tr>
<tr>
<td>high</td>
<td>30</td>
<td>6</td>
<td></td>
<td>0.833</td>
</tr>
<tr>
<td>nonhigh</td>
<td>24</td>
<td>124</td>
<td></td>
<td>0.162</td>
</tr>
</tbody>
</table>
\(-2L\) in New (Two-Class) Model

\[
L = k \ln p + (n - k) \ln(1 - p)
= 30 \ln(0.833) + 6 \ln(0.167)
= 30(-0.183) + 6(-1.79)
= -5.5 + -10.7 = -16.2
\]

\[
L = k \ln p + (n - k) \ln(1 - p)
= 24 \ln(0.162) + 124 \ln(0.838)
= 24(-1.82) + 124(-0.177)
= -43.7 + -21.9 = -65.6
\]

\[
\text{sum} = -81.8 \\
\times -2 = 161.6
\]
SPSS Report on Explained Variance

Beginning Block Number 0. Initial Log Likelihood Function
-2 Log Likelihood 222.7

[...]

Estimation terminated at iteration number 4 because L decreased ...
-2 Log Likelihood 158.3

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>df</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>64.461</td>
<td>2</td>
<td>.0000</td>
</tr>
</tbody>
</table>

Reduction in \(-2L\): 222.7 – 158.3 = 64.4 is the best measure of the quality of the model. 64.4 is 29% of the variance (222.7).
Visualizing Relations
Analysis of Residuals

- Just as in linear regression, useful in order to see where predictions go wrong, where other/additional ideas might be useful.
- SPSS can save residuals (false predictions).
- Labov’s data is not available except in the tabular form used, so we cannot examine the residuals here.
Logistic Regression

Idea: Predict categorical variable using regression

- Example: whether linguistic rules apply, e.g., syllable-final [r] in NYC
- key step: predict **chance of** categorical variable
  — transforming categorical to numeric variable
  — logit (log-odds) transformation used

\[
\text{logit}(x) = \ln \frac{p}{1 - p}
\]

- independent variables may be numeric or categorical