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, ...

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. χ^2)

Outline Logistic Regression

Idea: Predict categorical variable using regression

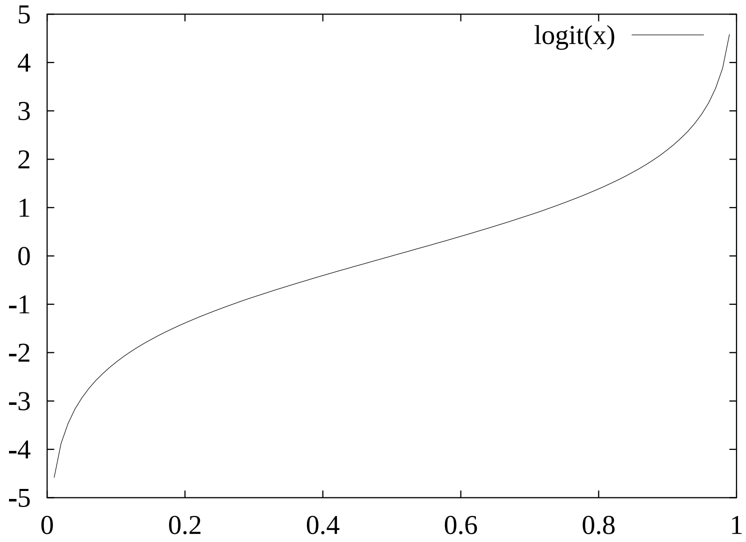
- core task: analyze dependency of categorical variable on others using regression
- problem: translating regression techniques to categorical domain
- key step: predict **chance of** categorical variable
  —transforming categorical to numeric variable
- note: independent variables may be numeric or 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) $= \ln \frac{p}{(1-p)}$

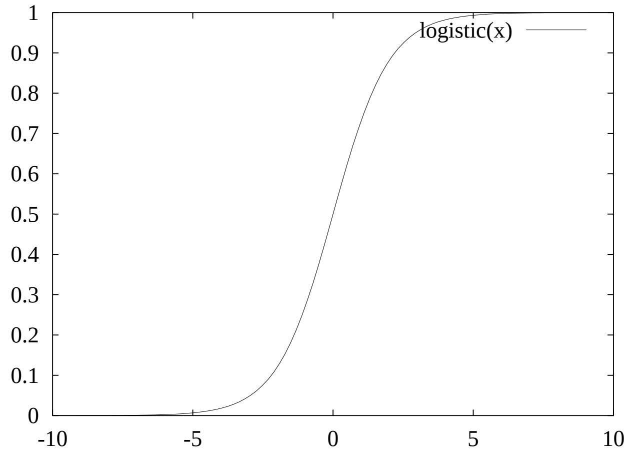


p	0.01	0.05	0.10	0.30	0.5	0.7	0.9	0.95	0.99
logit(p)	-4.6	-2.9	-2.2	-0.8	0.0	0.8	2.2	2.9	4.6

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

$$\begin{aligned}
\ln \tfrac{p}{1-p} &= \text{logit}(p) \\
\tfrac{p}{1-p} &= e^{\text{logit}(p)} \\
p &= e^{\text{logit}(p)}(1-p) \\
p &= e^{\text{logit}(p)} - pe^{\text{logit}(p)} \\
p + pe^{\text{logit}(p)} &= e^{\text{logit}(p)} \\
p(1+e^{\text{logit}(p)}) &= e^{\text{logit}(p)} \\
p &= \tfrac{e^{\text{logit}(p)}}{(1+e^{\text{logit}(p)})} (\times \tfrac{e^{-\text{logit}(p)}}{e^{-\text{logit}(p)}} \\
p &= \tfrac{1}{(1+e^{-\text{logit}(p)})}
\end{aligned}$$

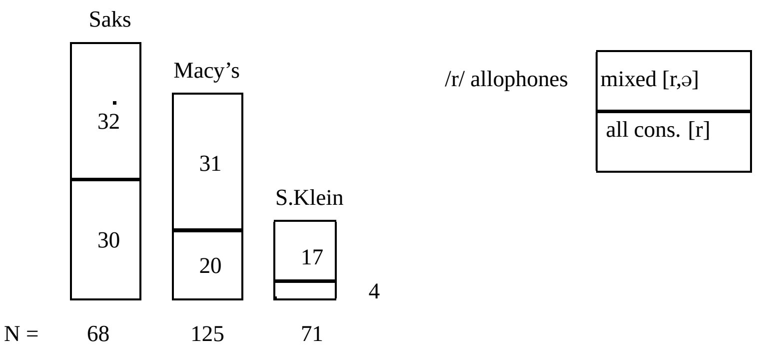
Strategy: Predict Logit Values

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

- try to find optimal β_0, β_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.



high social stratum Saks Macy's S.Klein low social stratum

Data on NYC /r/

Social Status	Pronunciation of /r/		
	cons. ([r])	vocalic ([ə])	mixed
high	30	6	32
medium	20	74	31
low	4	50	17

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/?

- χ^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”:

Social Status	Pronunciation of /r/		
	cons. ([r])	vocalic ([ə])	mixed
high	30	6	32
medium	20	74	31
low	4	50	17

- 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:

Status	explanation	dummy-1	dummy-2
1	(high, Saks)	1	0
2	(mid, Macy’s)	0	1
3	(low, S.Klein)	0	0

SPSS Output—Coding

```
Dependent Variable Encoding:

Original        Internal
Value           Value
       0        0       [vocalic pronunciation]
       1        1       [consonantal    "       ]

                              Parameter
                 Value   Freq  Coding
                                 (1)    (2)
SOC_STAT
                     1      2  1.000   .000
                     2      2   .000  1.000
                     3      2   .000   .000
```


Inf. Stats

SPSS Output

```
-------------------- Variables in the Equation -----------------
 Variable              B     S.E.    Wald   df     Sig      R     Exp(B)

SOC_STAT                             43.90   2     .000    .42
 SOC_STAT(1)        4.13     .69    36.38   1     .000    .39    62.49
 SOC_STAT(2)        1.22     .58     4.44   1     .035    .10     3.38
Constant           -2.53     .52    23.63   1     .000
```

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

$$\begin{aligned} \text{logit}(p) &= \beta_0 + \beta_1 s_1 + \beta_2 s_2 \\ &= -2.5 + 4.1 s_1 \\ &= -2.5 + 1.2 s_2 \\ &= -2.5 \end{aligned}$$

Interpreting SPSS Output

$$
\begin{aligned}
\text{logit}(p) &= -2.5 + 4.1 s_1 && \text{Saks, } s_1 = 1 \\
&= -2.5 + 1.2 s_2 && \text{Macy's, } s_2 = 1 \\
&= -2.5 && \text{S.Klein, } s_1 = s_2 = 0 \\
\\
&= -2.5 + 4.1 = 1.6 && \text{Saks} \\
&= -2.5 + 1.2 = -1.3 && \text{Macy's} \\
&= -2.5 && \text{S.Klein}
\end{aligned}
$$

Checking Interpretation of Output

$$
\begin{aligned}
\ln \frac{p}{(1-p)} &= 1.6 && \text{Saks} \\
&= -1.3 && \text{Macy's} \\
&= -2.5 && \text{S.Klein}
\end{aligned}
$$

$\ln \frac{p}{(1-p)}$	$\frac{p}{(1-p)}$	p	
1.6	$30/6$	≈ 0.84	Saks
-1.3	$20/74$	≈ 0.21	Macy's
-2.5	$4/50$	≈ 0.07	S.Klein

These indeed match the data to be predicted.

SPSS Output

```
-------------------- Variables in the Equation -----------------
 Variable              B     S.E.   Wald   df    Sig      R    Exp(B)

SOC_STAT                            43.90   2    .000    .42
 SOC_STAT(1)         4.13    .69    36.38   1    .000    .39   62.49
 SOC_STAT(2)         1.22    .58     4.44   1    .035    .10    3.38
Constant            -2.53    .52    23.63   1    .000
```

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
- `Exp (B)` $= e^{\beta}$

Understanding SPSS Output

```
Classification Table for UITSPRK
The Cut Value is .50
                 Predicted
                 0       1      Percent Correct
                  0  I    1
Observed     +-------+-------+
   0     0   I  124  I    6  I    95.38%
             +-------+-------+
   1     1   I   24  I   30  I    55.56%
             +-------+-------+
                        Overall   83.70%
```

Predictions, Correctness


```
                   Predicted
                  [@]       [r]      Percent Correct
                Macy's I
                /Klein I  Saks
Observed       +-------+-------+
   0    [@]    I  124  I    6  I   95.38%
               +-------+-------+
   1    [r]    I   24  I   30  I   55.56%
               +-------+-------+
                          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).

Log Likelihood

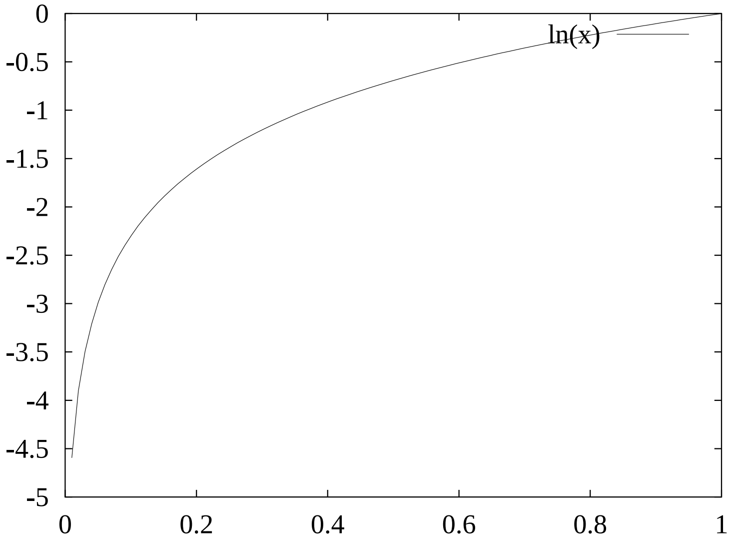
Variance in the binomial case is $p(1-p)$, and variance of the number of observations is $p^k(1-p)^{(n-k)}$ where the positive value [r] was seen k times and the null value $(n-k)$ times. From this we derive the **log likelihood** L:

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

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 χ^2 distribution with $(n-1)$ degrees of freedom.

Log Probabilities



Very likely events ($p \approx 1$) contribute little to log likelihoods.

Log Likelihood

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:

Social Status	Pronunciation of /r/ cons. ([r])	vocalic ([ə])
high	30	6
medium	20	74
low	4	50
totals	54	130
best guess	0.293	0.707

Simplest Model—No Social Class

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

$$\begin{aligned}
L &= k \ln p + (n-k)\ln(1-p) \\
&= 54\ln(0.293) + 130\ln(0.707) \\
&= 54(-1.23) + 130(-0.35) \\
&= -66.4 + -45.1 = -111.5 \\
-2L &= 223
\end{aligned}$$

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 dsitinguishes two classes, for which distinct "best guesses" are obtained, again using the empirical frequencies:

Social Status	Pronunciation of /r/		
	cons. ([r])	vocalic ([ə])	**prop. r**
high	30	6	0.833
nonhigh	24	124	0.162

$-2L$ in New (Two-Class) Model

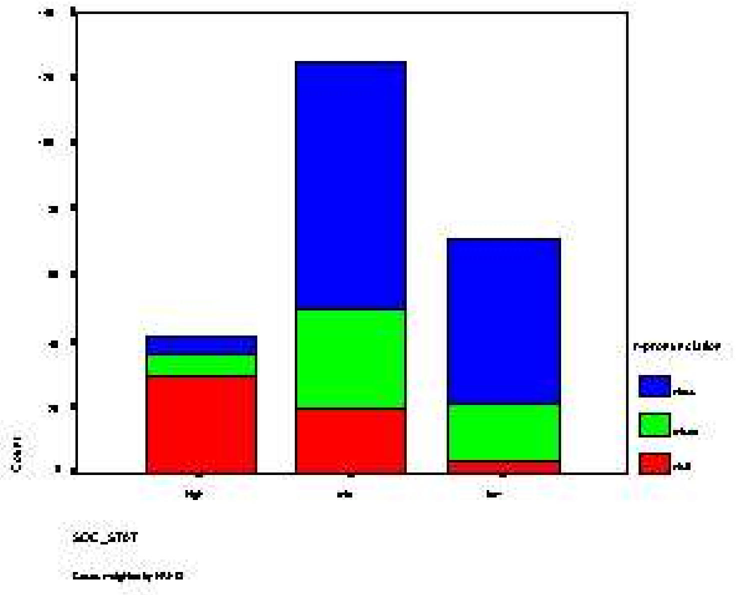
$$\begin{array}{rcll} L & = & k \ln p + (n-k)\ln(1-p) & \\ & = & 30\ln(0.833) + 6\ln(0.167) & \\ & = & 30(-0.183) + 6(-1.79) & \\ & = & -5.5 + -10.7 & = -16.2 \\ \hline L & = & k \ln p + (n-k)\ln(1-p) & \\ & = & 24\ln(0.162) + 124\ln(0.838) & \\ & = & 24(-1.82) + 124(-0.177) & \\ & = & -43.7 + -21.9 & = -65.6 \\ \hline \textbf{sum} & & & = -81.8 \\ & & & \times -2 \\ \cline{4-4} -2L & & & = 161.6 \end{array}$$

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

                    Chi-Square    df Significance
Model                   64.461     2        .0000
```

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 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

RuG

Interpreting SPSS Output

RuG