Statistics II
Logistic Regression

Çağrı Çöltekin

University of Groningen
Information Science

April 22, 2014

Logistic regression: motivation

- For all methods discussed in this class, the response variable is numeric.
- Sometimes we want to analyze/predict categorical responses:
  - Alive or dead.
  - Grammatical or ungrammatical.
  - Correct or incorrect.
  - Pass or fail.
  - Win or loose.
  - Present or absent.

Logistic regression is an extension of regression for categorical (typically binary) response variables.

Logistic regression: some examples

- survival after a surgery depending on age, length of surgery, ...
- whether purchase occurs on an online shop depending on age, income, website characteristics, ...
- whether speech errors occur depending on alcohol level
- when certain linguistic variation is observed depending on speed of utterance, stress, social group, ...
- whether one votes to a political party (or not) depending on age, income, ethnicity, ...

Simple regression: a refresher

$$y_i = a + bx_i + e_i$$

y is the *response* (or outcome, or dependent) variable. The index i represent each unit observation/measurement (sometimes called a 'case').

x is the *predictor* (or explanatory, or independent) variable.

a is the intercept.

b is the slope of the regression line.

$a + bx$ is the *deterministic* part of the model (we sometimes use $\hat{y}$).

e is the residual, error, or the variation that is not accounted for by the model. Assumed to be (approximately) normally distributed with 0 mean (e_i are assumed to be i.i.d).

Regression assumptions

independence cases (observations) should be independent.

linearity the relation between 'y vs. x' should be linear.

normality residuals should be normally distributed with 0 mean.

constant variance residual variance should be constant.

Multiple regression

$$y_i = \underbrace{a + b_1 x_{i,1} + b_2 x_{2,i} + \ldots + b_k x_{k,i}}_{\hat{y}} + e_i$$

- Multiple regression is a generalization of the simple regression, where we predict the outcome using multiple predictors.
- Multicollinearity causes problems in estimation and interpretation of multiple-regression models.
- Model selection (finding a model that fits the data well, but not more complex than necessary) is important.

Multiple regression: example

We will try to determine reaction time differences in a lexical decision task in two experimental conditions 'C_1' and 'C_2', and also investigate the effect of age. Here is how our data looks like,

Reaction time (ms)	condition	age
563	C_1	38
562	C_2	35
550	C_1	20
573	C_2	31
⋮	⋮	⋮

Multiple regression: example (2)

Running multiple regression (R output):

```
Call:
lm(formula = log(reaction.time) ~ age + condition)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.3246836  0.0035730 1770.13  < 2e-16 ***
age         0.0001977  0.0001024    1.93   0.0613 .
conditionC2 0.0167303  0.0014068   11.89 3.33e-14 ***
---
Residual standard error: 0.004374 on 37 DF
Multiple R-squared: 0.7927,    Adj. R-squared: 0.7815
F-statistic: 70.74 on 2 and 37 DF,  p-value: 2.276e-13
```

Note: you should do model diagnostics before interpreting the model.

Logistic Regression

- Logistic regression is a special case of regression, where a binary outcome is predicted based on a number of predictors.
- The main trick is to predict probability of an event, rather than the outcome directly. This allows converting a categorical variable to a numeric variable (probabilities).
- One needs to overcome difficulties due to regression assumptions.
  - Probabilities are strictly bounded between 0 and 1.
  - Error (residuals) with binary (or proportion) response is not normally distributed.

An example for logistic regression

We go back to our lexical decision task example. In similar experiments typically we also whether the reaction was correct or incorrect.
Our data actually looks like this:

Reaction time (ms)	condition	age	correct
563	C_1	38	1
562	C_2	35	1
550	C_1	20	1
573	C_2	31	0
⋮	⋮	⋮	⋮

Our aim is to predict probability of correct decision.

Trying ordinary regression

Model the outcome directly:

```
lm(formula = correct ~ age + condition)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.461957   0.243223   1.899   0.0653 .
age          0.016037   0.006973   2.300   0.0272 *
conditionC2 -0.209907   0.095766  -2.192   0.0348 *
...
```

$$P(\mathrm{correct}) = 0.462 + 0.016 \times \mathrm{age} - 0.210 \times \mathrm{condition}$$

Trying ordinary regression

Model the outcome directly:

```
lm(formula = correct ~ age + condition)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.461957   0.243223   1.899   0.0653 .
age          0.016037   0.006973   2.300   0.0272 *
conditionC2 -0.209907   0.095766  -2.192   0.0348 *
...
```

$$\mathrm{P}(\mathrm{correct}) = 0.462 + 0.016 \times \mathrm{age} - 0.210 \times \mathrm{condition}$$

Estimated probability of correct response to condition 1 from a 40-year old is: $0.462 + 0.016 \times 40 - 0.210 \times 0 = 1.102$.

Response variable (probability) is not bounded between 0 and 1.

Transforming the response variable

- Instead of predicting the probability, p, we can predicts odds.
- Odds in favor of an event (e.g., FC Groningen winning against Ajax) is defined based on p, the probability of the event occurring, as

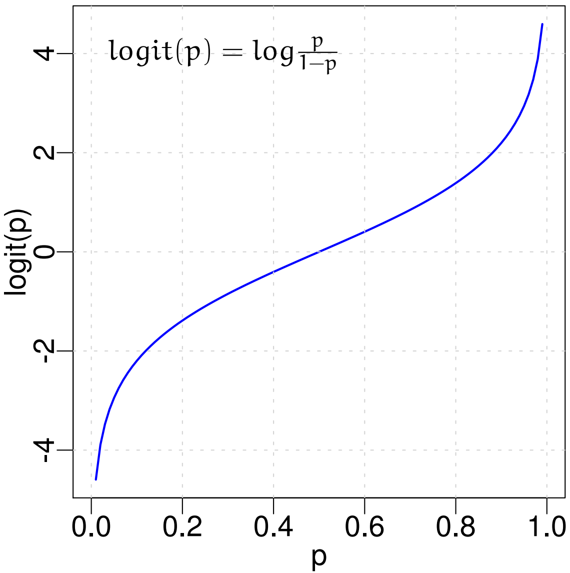
$$\mathrm{odds} = \frac{p}{1-p}$$

- odds are bounded between 0 and infinity, $0 \leq \mathrm{odds} \leq \infty$.
- further if we take the logarithm of odds,

$$\log \frac{p}{1-p} = \mathrm{logit}(p)$$

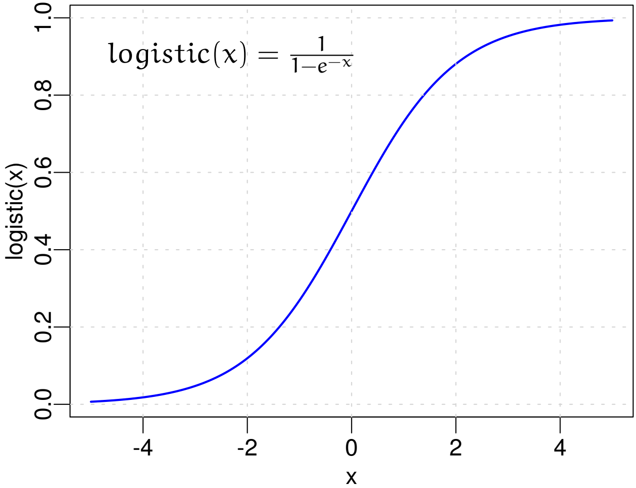
Bounds of logit is $-\infty \leq \mathrm{logit}(p) \leq +\infty$

The logit function



But, what about the name ‘logistic’?

Logistic function is the inverse of the logit function:



Trying ordinary regression with logit transform

```
lm(formula = logit(correct) ~ age + condition)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.27875    1.78213  -0.156   0.8766
age          0.11751    0.05109   2.300   0.0272 *
conditionC2 -1.53802    0.70169  -2.192   0.0348 *
--
Residual standard error: 2.182 on DF
Multiple R-squared:  0.2501,    Adj. R-squared:  0.2095
F-statistic: 6.169 on 2 and 37 DF,  p-value: 0.004873
```

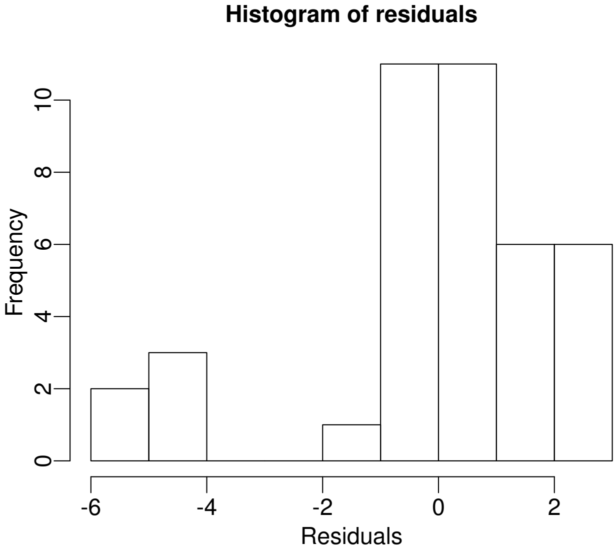
Now our estimate is:

$$\mathrm{logit}(\mathrm{P}(\mathrm{correct})) = -0.279 + 0.118 \times \mathrm{age} - 1.538 \times \mathrm{condition}$$

The model is more correct, but interpretation is difficult.

Model checking: is the transformation enough?

Let's look at the residual distribution:



Regression for binary response: problems so far

- Binary data is strictly bounded, probabilities smaller than 0 and larger than 1 are meaningless.
  - We solve this by transforming the response variable using logit function (other transformations are also possible).
- When response is binary, the residuals are not normally distributed and residual variance is not constant.
- We estimate the coefficients without assuming normally distributed error.
- This requires giving up the familiar (and uniquely determined) least-squares regression. Estimation is done with more complicated procedures (typically maximum likelihood estimation).
- More precisely, logistic regression is an instance of generalized linear models (GLMs).

Logistic regression

Logistic regression is similar to (multiple) regression, differences are:

- Categorical response variable.
- Non-normal errors.
- Relationship is non-linear (linear after the transformation).
- Estimation is (typically) done using maximum likelihood estimation (MLE). Least-squares regression is not applicable.

Maximum likelihood estimation

$$\mathrm{logit}(p_i) = a + b_1 x_{1,i} + \ldots + b_k x_{k,i} + e_i$$

- Maximum likelihood estimation tries to find the set of model parameters, or coefficients, a, b_1, ...b_k, which make the data most likely (or minimizes the error).
- MLE is an iterative search for the optimum parameter values. There is no exact solution.
- In some cases, MLE may fail to find a solution.
- If errors are normally distributed, MLE is equivalent to least-squares estimation.

Logistic regression: example

We return to the same example, this time fitting a correct model:

```
glm(formula = correct ~ age + condition, family =
    binomial)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    14.7647  3620.9079   0.004   0.9967
age             0.1904     0.1073   1.775   0.0759 .
conditionC2   -19.1351  3620.9069  -0.005   0.9958
--
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 30.142  on 39  degrees of freedom
Residual deviance: 17.708  on 37  degrees of freedom
AIC: 23.708
Number of Fisher Scoring iterations: 19
```

Estimated equation is:

$$\mathrm{logit}(P(\mathrm{correct})) = 14.7647 + 0.1904 \times \mathrm{age} - 19.1351 \times \mathrm{condition}$$

Logistic regression: interpreting the coefficients

$$\log \frac{P(\text{correct})}{1 - P(\text{correct})} = 14.7647 + 0.1904 \times \text{age} - 19.1351 \times \text{condition}$$

Slope of a variable indicates the change in log odds for unit-change in the predictor (while other predictors kept constant). For example, one unit change in 'age', say from 40 to 41, means

$$\begin{aligned}
\log \frac{P(\text{correct}_{41})}{1-P(\text{correct}_{41})} - \log \frac{P(\text{correct}_{40})}{1-P(\text{correct}_{40})} &= 0.1904 \\
e^{\log \frac{P(\text{correct}_{41})}{1-P(\text{correct}_{41})} - \log \frac{P(\text{correct}_{40})}{1-P(\text{correct}_{40})}} &= \underbrace{e^{0.1904}}_{\exp(b)} \\
\frac{\text{odds}(\text{correct}_{41})}{\text{odds}(\text{correct}_{40})} &= 1.209733
\end{aligned}$$

Logistic regression: interpreting the coefficients

$$\log \frac{\mathrm{P(correct)}}{1 - \mathrm{P(correct)}} = 14.7647 + 0.1904 \times \mathrm{age} - 19.1351 \times \mathrm{condition}$$

Slope of a variable indicates the change in log odds for unit-change in the predictor (while other predictors kept constant). For example, one unit change in 'age', say from 40 to 41, means

$$\begin{aligned}
\log \frac{\mathrm{P(correct_{41})}}{1-\mathrm{P(correct_{41})}} - \log \frac{\mathrm{P(correct_{40})}}{1-\mathrm{P(correct_{40})}} &= 0.1904 \\
e^{\log \frac{\mathrm{P(correct_{41})}}{1-\mathrm{P(correct_{41})}} - \log \frac{\mathrm{P(correct_{40})}}{1-\mathrm{P(correct_{40})}}} &= \underbrace{e^{0.1904}}_{\exp(b)} \\
\frac{\mathrm{odds(correct_{41})}}{\mathrm{odds(correct_{40})}} &= 1.209733
\end{aligned}$$

When age is increased form 40 to 41, odds of correct response increase 1.2 times. Note: the change is non-linear.

Logistic regression: inference and model fit

- In logistic regression the significance testing for coefficients are carried out using z-statistics (also known as 'Wald statistic' or 'Wald's z').
- We do not have r^2 as measure of fit (approximations exists, but none have all the properties of r^2.
- $-2\mathrm{LL}$ (-2 times Log Likelihood, smaller better) is the measure of fit in maximum likelihood estimation.
- Log likelihood is not bounded like r^2, comparisons are valid only on the same data set.
- $-2\mathrm{LL}$ is approximately χ^2 distributed, a model can be tested against null model (similar to F-test for least-squares regression).

Logistic regression: where things may go wrong

- Overdispersion is the case when variance in the data is higher than expected.
  - Logistic regression requires response to follow binomial distribution.
  - Variance of binomial distribution is predictable from it's mean.
- In case of complete separation, i.e., when predictors perfectly separate cases of success and failure, logistic regression parameters cannot be estimated.
- Logistic regression may also fail to find a solution, especially when data is not evenly distributed.
- Logistic regression is multiple regression, other problems, such as collinearity is also problems for logistic regression.

Logistic regression: another example

We would like to guess whether a child would develop dyslexia or not based on a test applied to pre-verbal children. Here is a simplified problem:

- We test children when they are less than 2 years of age.
- The hypothesis is that the test may be relevant in diagnosing dyslexia.
- We observe the same children when they are in the school age, and note whether they are diagnosed with dyslexia or not.
- Our data looks like the following:

Test score	Dyslexia
8.2	0
2.2	1
6.2	1
⋮	⋮

* The research question is from the longitudinal study by Ben Maasen and his colleagues. Data is fake as usual.

Example: the analysis

```
glm(formula = dys ~ score, family = binomial, data = d)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6558  -0.5927  -0.2519   0.3590   1.8567
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.4373     1.5758   2.816  0.00486 **
score        -0.9419     0.2920  -3.225  0.00126 **
---
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 54.548  on 39  degrees of freedom
Residual deviance: 30.337  on 38  degrees of freedom
AIC: 34.337
Number of Fisher Scoring iterations: 5
```

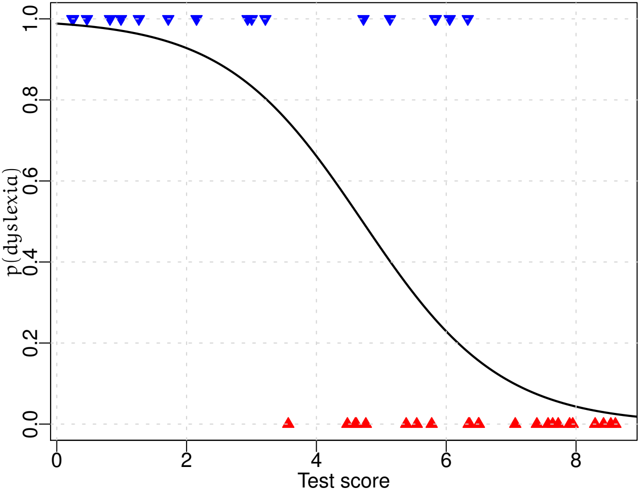
$$\mathrm{logit}(p(\mathrm{dyslexia})) = 4.4373 - 0.9419 \times \mathrm{score}$$

Example: interpretation

$$\mathrm{logit}(p(\mathrm{dyslexia})) = 4.4373 - 0.9419 \times \mathrm{score}$$

score	$\mathrm{logit} = \log \frac{p}{1-p}$	$\mathrm{odds} = \frac{p}{1-p}$	p
1	3.50	32.96	0.97
3	1.61	5.01	0.83
5	−0.27	0.76	0.43
7	−2.16	0.12	0.10
9	−3.10	0.05	0.04

Visualizing the logistic regression



Summary

- Logistic regression is applicable when responseis binomial.
- Two major differences from ordinary least-squares regression:
  - Response variable is strictly bounded. logit transformation make sure regression output is mapped to range $[0, 1]$.
  - Errors are not normal. GLM framework allows non-normal errors. However, we use maximum likelihood estimation instead of least squares.
- Problems to watch out for:
  - overdisperion
  - complete separation, or insufficient data for estimation
  - like in multiple regression: multicollinearity
- Extensions
  - Multinomial logistic regression, when there are more than two categories of the response variable.
  - Genralized linear mixed-effect models when observations are not independent.

Next time

- A quick summary.
- Time for your questions.