Statistiek II

John Nerbonne

Dept of Information Science
j.nerbonne@rug.nl
Wtih important improvements by Hartmut Fitz!

April 14, 2015

university of groningen

Today: logistic regression

Idea: predict **categorical** variable using regression

Examples:

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

Logistic regression very popular, especially in sociolinguistics

Regression technique attractive

Logistic regression attractive technique because

- allows prediction of one variable value based on one or more others
- allows prediction of the probability of the occurrence of an event
- allows 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—transform categorical to numeric variable
- note: independent variables may be numeric or categorical—as in regression in general, simple or multiple

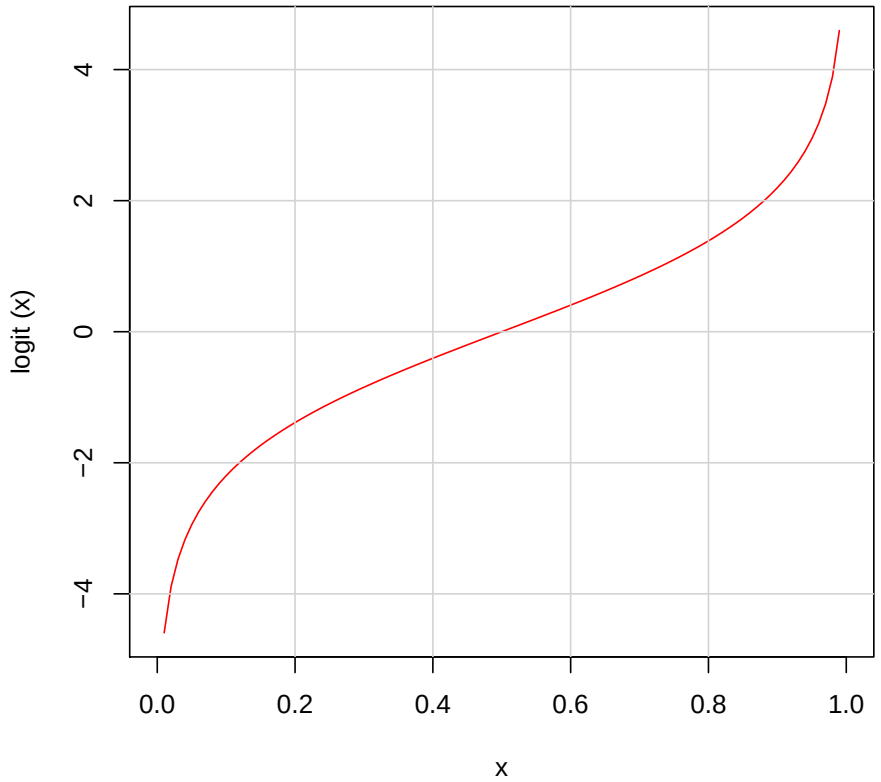
Chance as dependent variable

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

The `logit` function

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$



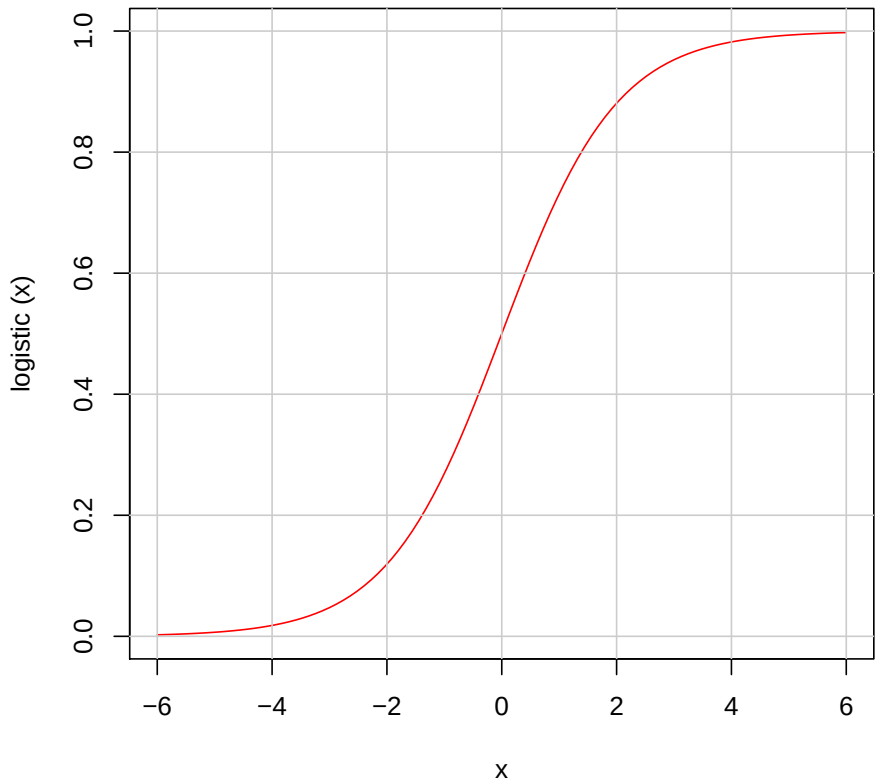
p	0.01	0.05	0.10	0.30	0.5	0.7	0.9	0.95	0.99
logit(p)	-4.6	-2.9	-2.2	-0.8	0.0	0.8	2.2	2.9	4.6

Logit 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 Barcelona against Man United are 4 to 1
- probability of Barcelona winning is 0.8, $\frac{p}{(1-p)} = \frac{0.8}{0.2} = 4:1$
- why the name logistic?

Why 'logistic' regression?

Logistic function: $f(x) = \frac{1}{1+e^{-x}}$



Note that values of logistic function are bounded: $0 < f(x) < 1$

Logit vs logistic

`logit` function is the **inverse** of the `logistic` function, i.e., $\texttt{logistic}(\texttt{logit}(p)) = p$:

$$
\begin{aligned}
\ln\left(\frac{p}{1-p}\right) &= \text{logit}(p) &&\Leftrightarrow\\
\frac{p}{1-p} &= e^{\text{logit}(p)} &&\Leftrightarrow\\
p &= e^{\text{logit}(p)} \cdot (1-p) &&\Leftrightarrow\\
p &= e^{\text{logit}(p)} - p \cdot e^{\text{logit}(p)} &&\Leftrightarrow\\
p + p \cdot e^{\text{logit}(p)} &= e^{\text{logit}(p)} &&\Leftrightarrow\\
p(1 + e^{\text{logit}(p)}) &= e^{\text{logit}(p)} &&\Leftrightarrow\\
p &= \frac{e^{\text{logit}(p)}}{(1+e^{\text{logit}(p)})} &&\Leftrightarrow\\
p &= \frac{e^{\text{logit}(p)}}{(1+e^{\text{logit}(p)})} \cdot \frac{e^{-\text{logit}(p)}}{e^{-\text{logit}(p)}} &&\Leftrightarrow\\
p &= \frac{e^{0}}{e^{-\text{logit}(p)}+e^{0}} &&\Leftrightarrow\\
p &= \frac{1}{1+e^{-\text{logit}(p)}}
\end{aligned}
$$

Strategy: predict `logit` values

$\texttt{logit}(p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$, where $x_1, \ldots, x_p$ are the independent variables

- try to find **optimal** $\beta_0, \ldots, \beta_p$ for given data
- 'optimal' here does **not** mean minimizing the sum of squared deviations
- logistic regression uses a **maximum likelihood** method; maximizes probability of obtaining the observed results given the fitted regression coefficients
- hence, significance statistics in logistic regression **different** from linear regression
- note that we are seeking a **non-linear** relationship

Example 1: predicting the dative alternation

Question: how does an English speaker determine which of the alternative dative structures to choose to convey a given message about a transfer event?

Choice between prepositional dative and the double object structure.

prepositional dative:	... gave [toys] [to the children]	V NP PP
double object dative:	... gave [the children] [toys]	V NP NP

Bresnan, Cueni, Nikitina & Baayen (2005) try to predict the use of the dative alternation with logistic regression model

Example 1: predicting the dative alternation

Data: 2360 dative observations from the Switchboard collection of recorded telephone conversations

All these datives were annotated for the explanatory variables

Regression model postulates 14 explanatory variables that might influence the choice of alternative dative structures

Some variables were:

- accessibility of theme/recipient
- definiteness of theme/recipient
- animacy/person of recipient
- number/concreteness/definiteness of theme

Example 1: predicting the dative alternation

After fitting the model to the data, the model fit was evaluated:

Classification Table for Model A (1 = PP; cut value = 0.50)				
		Predicted:		% Correct
		0	1	
Observed:	0	1796	63	97%
	1	115	386	77%
			Overall:	92%

Hence, the logistic regression model correctly classifies 92% of the data overall

Example 2: 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.

Method:

- Labov walked into 3 NYC department stores (Saks, Macy's and S. Klein)
- stores cater to distinct social classes (high, middle and low, respectively)
- asked shop assistants for departments which were on the fourth floor
- seeking repetition of 'fourth floor' by pretending not to understand

Example 2: Labov's NYC /r/ study

Data from Labov's study:

Percentage of r-use in three NYC department stores

	Saks (%)	Macy's (%)	S. Klein (%)
Cons. [r]	30	20	4
Vocalic [ə]	6	74	50
Mixed /r/	32	31	17
Number	68	125	71

Saks: high social class
Macy's: middle class
S. Klein: low social class

Mixed /r/: mixed allophones [r,ə]

Analyzing social influence on /r/

What statistical test is needed to ask whether social status influences pronunciation of /r/?

- χ^2 test of independence:
  - is one categorical variable dependent on another (or are they 'randomly related')?
- We employ logistic regression here for two reasons:
  - to measure the degree of dependence
  - to combine analysis 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 are 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

- 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

Variable	B	S.E.	Wald	df	Sig	Exp(B)
SOC_STAT			43.90	2	.000	
SOC_STAT(1)	4.13	.69	36.38	1	.000	62.49
SOC_STAT(2)	1.22	.58	4.44	1	.035	3.38
Constant	-2.53	.52	23.63	1	.000	

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

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

Interpreting SPSS output

$$
\begin{aligned}
\texttt{logit}(p) &= -2.5 + 4.1 s_1 && \text{Saks, } s_1 = 1 \\
&= -2.5 + 1.2 s_2 && \text{Macy's, } s_2 = 1 \\
&= -2.5 && \text{S. Klein, } s_1 = s_2 = 0 \\
\\
&= -2.5 + 4.1 = 1.6 && \text{Saks} \\
&= -2.5 + 12. = -1.3 && \text{Macy's} \\
&= -2.5 && \text{S. Klein}
\end{aligned}
$$

Checking interpretation of output

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

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

These indeed match the data to be predicted

SPSS output

Variable	B	S.E.	Wald	df	Sig	Exp(B)
SOC_STAT			43.90	2	.000	
SOC_STAT(1)	4.13	.69	36.38	1	.000	62.49
SOC_STAT(2)	1.22	.58	4.44	1	.035	3.38
Constant	-2.53	.52	23.63	1	.000	

- Note that all variables are significant
- Exp(B) $= e^{\beta}$: $e^{4.13} = 62.18$
  $e^{1.22} = 3.38$

Hypothesis testing

We test each model parameter for significance, e.g.,

Null hypothesis H_0: $\beta_1 = 0$
Alternative hypothesis H_a: $\beta_1 \neq 0$

Compute the **test statistic**

$$z = \frac{b_1}{\mathrm{SE}_{b_1}}$$

Under H_0, z^2 has approximately χ^2 distribution with 1 degree of freedom.

Sometimes (e.g., in SPSS) called **Wald** statistics

Confidence interval for the slope

A level C **confidence interval** for the slope β_1 is

$$b_1 \pm z^* \mathrm{SE}_{b_1}$$

z^* is the value for the standard Normal density curve with area C between $-z^*$ and z^*.

In the example (for C=95%):

$$4.13 \pm (1.96) \cdot 0.69 = 4.13 \pm 0.14$$

We are 95% confident that the slope is between 3.99 and 4.27

Predictions and correctness

	Predicted		
	[ə] Macy's/ Klein	[r] Saks	Perc correct
Observed			
[ə]	124	6	95.38%
[r]	24	30	55.56%
			83.70%

Table shows the prediction of the variable coded for status.

Note that we are predicting that Saks' pronunciations should be all [r] and the others all [ə] (schwa).

Log likelihood

Suppose there are n observations, where the positive value [r] was seen k times and the null value [ə] was seen $(n-k)$ times.

Let p be the probability of observing [r].

Try to estimate p which makes observed data most likely.

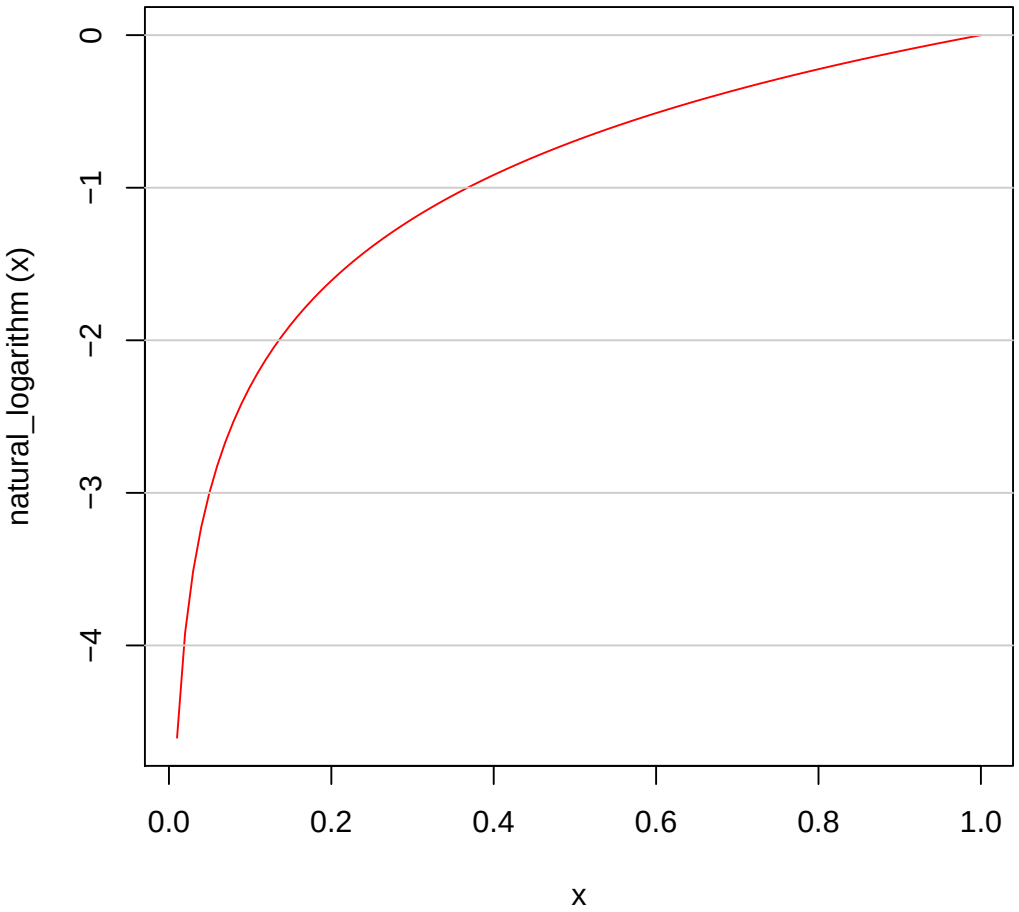
Log likelihood L is given by

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

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

$-2L$ follows 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 quality of the model using log likelihood and estimate parameters to obtain optimal value.

We obtain **optimal** value by using the overall frequencies as a best guess:

	Pronunciation of /r/	
Social status	cons [r]	vocalic [ə]
high	30	6
medium	20	74
low	4	50
total	54	130
best guess	0.293	0.707

Simplest model—no social class

Simplest model without social class:

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

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:

	Pronunciation of /r/		
Social status	cons [r]	vocalic [ə]	prob. [r]
high	30	6	0.833
non-high	24	124	0.162

$-2L$ in new (two-class) model

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

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)

Alternative: Interpret POINT BI-SERIAL CORRELATION coeff. like Pearson's corr. coeff. (r)

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 as new variable
- Labov's data is not available except in the tabular form used, so we cannot examine residuals here

Example 3: predicting admission to Grad school

Question: How do variables such as

- GRE (Graduate Record Exam scores)
- GPA (grade point average)
- prestige of the program

affect admission into graduate school?

The response variable—admit/don't admit (1/0)—is a binary variable.

We use logistic regression to predict the odds of being admitted based on three predictors: GRE score, GPA, and quality of program (high $= 1$, low $= 0$)

Example 3: predicting admission to Grad school

Admission data:

Obs	admit	gre	topnotch	gpa
1	0	380	0	3.61
2	1	660	1	3.67
3	1	800	1	4.00
4	1	640	0	3.19
5	0	520	0	2.93
6	1	760	0	3.00
⋮	⋮	⋮	⋮	⋮
400	0	600	0	3.89

Note that two predictors are numerical, one categorical

Check for empty cells

Check if any cells (created by cross-tabulation of categorical and response variables) are empty.

If this occurs, there may be difficulties running the logit model.

	admit	
topnotch	0	1
0	238.00	97.00
1	35.00	30.00

None of the cells are too small or empty (has no cases), so we can use logistic regression

R output

We regress admission on GRE, GPA and quality of program

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	−4.6008	1.0964	−4.20	0.0000	***
gre	0.0025	0.0011	2.31	0.0207	*
gpa	0.6676	0.3253	2.05	0.0401	*
topnotch	0.4372	0.2919	1.50	0.1341	
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	

Table of coefficients shows that both GRE and GPA are statistically significant while topnotch is not.

Odds-ratios

To interpret the coefficients as odds-ratios we exponentiate them

(Intercept)	gre	gpa	topnotch
0.01004366	1.00247990	1.94946627	1.54840220

and determine the 95% confidence intervals:

	2.5 %	97.5 %
(Intercept)	0.00	0.08
gre	1.00	1.00
gpa	1.04	3.72
topnotch	0.87	2.74

For a one unit increase in GPA, the odds of being admitted to graduate school increase by a factor of 1.94 (but note that GPA confidence interval is large)

Log-likelihood test

Want to test whether difference between current model (three predictors) and null model (only intercept) is statistically significant.

We compare the log-likelihoods ($-2L$) of the null model and the current model with χ^2 test for 3 degrees of freedom ($= \#$ predictors)

Obtain a χ^2 of 21.85, with a p-value of less than 0.00004

Indicates that our model as a whole fits significantly better than an 'empty' model

Predict probabilities to interpret logistic regression results

Q: What is the probability of getting accepted into grad school based on the quality of the undergrad program only?

Fix GRE and GPA at mean, and predict probabilities based on 'topnotch' variable only (using our model)

	gre	gpa	topnotch	topnotchP
1	587.70	3.39	0.00	0.29
2	587.70	3.39	1.00	0.39

Table shows: predicted probability of being accepted into the graduate program is 0.29 if the undergraduate institution was not "top notch" (topnotch $= 0$) and 0.39 if it was (topnotch $= 1$).

Predict probabilities to interpret logistic regression results

Q: What is the probability of getting accepted into grad school based on the GRE scores only?

Fix GPA and topnotch at mean, and predict probabilities based on GRE only (using our model)

	gre	greP
1	200.00	0.15
2	300.00	0.18
3	400.00	0.22
4	500.00	0.26
5	600.00	0.31
6	700.00	0.37
7	800.00	0.43

The probability of getting admitted into grad school is 0.15 with a GRE of 200 and increases to 0.43 with a GRE of 800.

Logistic regression wrapped up

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

$$\texttt{logit}(p) = \ln(\tfrac{p}{1-p})$$

- independent variable may be numeric or categorical

End of course material

Good luck with preparing for the exam!

Come see me if you have questions
(make appt. at `j.nerbonne@rug.nl`)