Logistic regression

Maarten Speekenbrink

Experimental Psychology
University College London

Statistics lecture 10

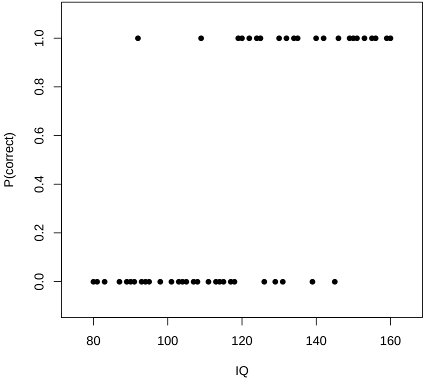
Outline

1. Why logistic regression?
2. The logistic transformation
3. Example: know your participants...
4. Interpretation
5. Model evaluation and selection
   - Individual predictors
   - Model error
   - Model comparison
   - Model fit
6. In practice
7. Multinomial logits

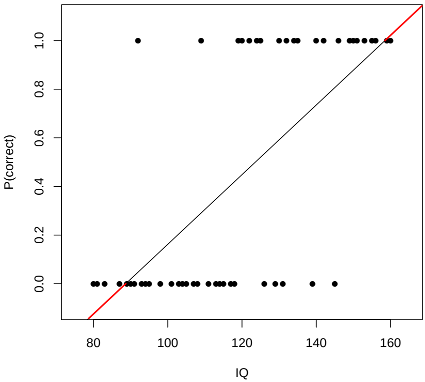
Why logistic regression?

- Dichotomous (binary) variables
  - yes, no
  - correct, incorrect
  - mentally ill, not mentally ill
  - etc.
- Distribution not normal but binomial
- No homogeneity of variance
  - variance of a binomial variable depends on mean

Linear regression

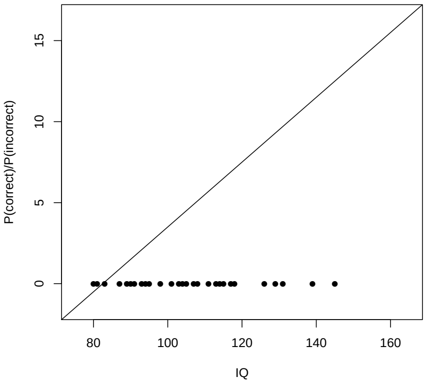


Linear regression



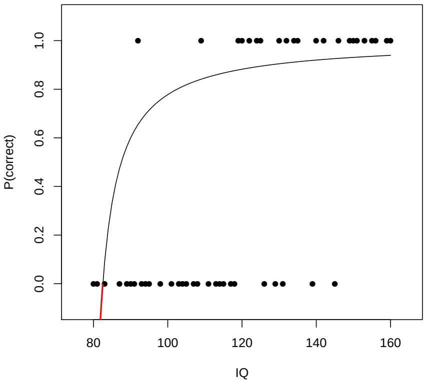
$$P(\texttt{correct}_i) = \beta_0 + \beta_1 \texttt{IQ}_i + \epsilon_i$$

Regression of the odds



$$\frac{P(\texttt{correct})}{1 - P(\texttt{correct}_i)} = \beta_0 + \beta_1 \texttt{IQ}_i + \epsilon_i$$

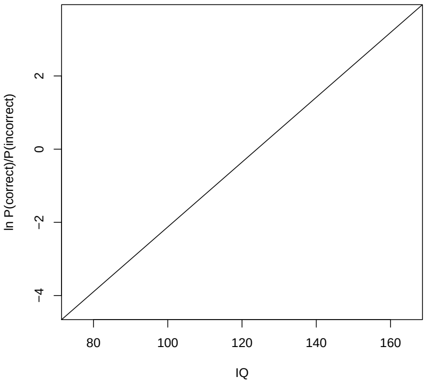
Regression of the odds



$$\frac{P(\texttt{correct})}{1 - P(\texttt{correct}_i)} = \beta_0 + \beta_1 \texttt{IQ}_i + \epsilon_i$$

$$\hat{P}(\texttt{correct}_i) = \frac{\beta_0 + \beta_1 \texttt{IQ}_i}{1 + \beta_0 + \beta_1 \texttt{IQ}_i}$$

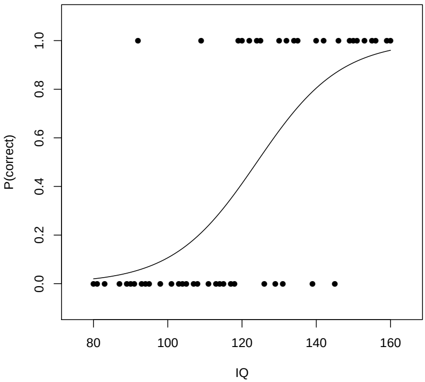
Regression of the log odds (logit)



$$\ln\left(\frac{P(\texttt{correct}_i)}{1-P(\texttt{correct}_i)}\right) = \beta_0 + \beta_1 \texttt{IQ}_i + \epsilon_i$$

$$\ln(a) = b \quad \leftrightarrow \quad \exp(b) = a$$
$$\log_c(a) = b \quad \leftrightarrow \quad c^b = a$$

Regression of the log odds (logit)

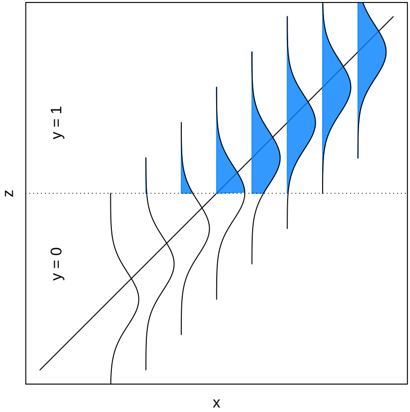


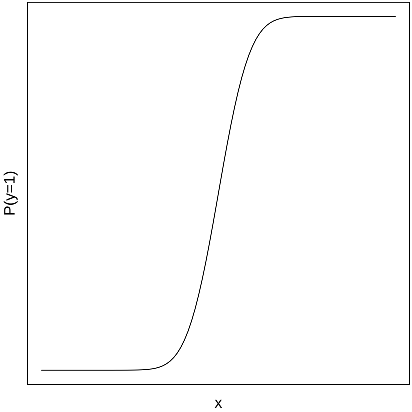
$$\ln\left(\frac{P(\texttt{correct}_i)}{1 - P(\texttt{correct}_i)}\right) = \beta_0 + \beta_1 \texttt{IQ}_i + \epsilon_i$$

$$\widehat{P}(\texttt{correct}_i) = \frac{\exp(\beta_0 + \beta_1 \texttt{IQ}_i)}{1 + \exp(\beta_0 + \beta_1 \texttt{IQ}_i)}$$

$$\ln(a) = b \quad \leftrightarrow \quad \exp(b) = a$$
$$\log_c(a) = b \quad \leftrightarrow \quad c^b = a$$

A latent variable justification. . .





Assumptions

Three main assumptions:

- The (conditional) probabilities are a logistic function of the predictors
- Dependent variable follows binomial distribution
- The observations are independent

Example: Know your participants. . .

Cowles and Davis (1987) performed a study on the characteristics of volunteer participants in psychology studies. They obtained scores on the Eysenck Personality Inventory from 1421 undergraduate students. Respondents were also asked to indicate their willingness to participate in further research. The scores on the EPI did not depart greatly from those norms for American college students. The question is whether volunteers are different from non-volunteers.

Cowles, M., & Davis, C. (1987). The subject matter of psychology: Volunteers. *British Journal of Social Psychology*, *26*, 97–102.

Results

Neuroticism and Extraversion were centered.
Contrast code for gender: $\texttt{Gender}_i = -1$ (male), $\texttt{Gender}_i = 1$ (female)

effect	b	$SE(b)$	Wald	p
intercept	-0.366	0.056	42.129	$< .001$
Neuroticism (N)	0.005	0.012	0.190	.663
Extraversion (E)	0.065	0.015	19.333	$< .001$
Gender (G)	0.134	0.056	5.609	.018
N × E	-0.009	0.003	8.848	.003
N × G	0.003	0.012	0.047	.828
E × G	0.006	0.015	0.165	.685
N × E × G	0.004	0.003	1.932	.164

Interpreting coefficients

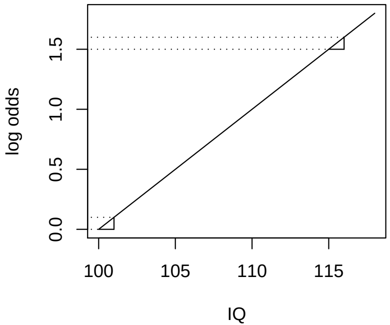
- intercept (b_0)
  - log odds when all $X_j = 0$
- slope (b_j)
  - change in log odds with one unit increase of predictor X_j
- exponentiated slope ($\exp(b_j)$ or e^{b_j})
  - multiplicative effect on odds
- E.g. $\ln\left(\frac{P(Y_i=1)}{P(Y_i=0)}\right) = -10 + 0.1 \times \texttt{IQ}_i$

IQ	$\ln\left(\frac{P(Y=1)}{P(Y=0)}\right)$	$\frac{P(Y=1)}{P(Y=0)}$	$P(Y=1)$
100	0	1	0.500
101	0.1	1.105	0.525
115	1.5	4.482	0.818
116	1.6	4.953	0.832

Interpreting coefficients

- intercept (b_0)
  - log odds when all $X_j = 0$
- slope (b_j)
  - change in log odds with one unit increase of predictor X_j
- exponentiated slope ($\exp(b_j)$ or e^{b_j})
  - multiplicative effect on odds
- E.g. $\ln\left(\frac{P(Y_i=1)}{P(Y_i=0)}\right) = -10 + 0.1 \times \texttt{IQ}_i$

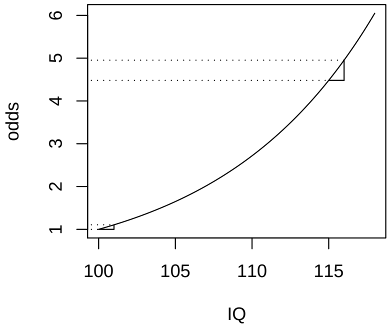
IQ	$\ln\left(\frac{P(Y=1)}{P(Y=0)}\right)$	$\frac{P(Y=1)}{P(Y=0)}$	$P(Y=1)$
100	0	1	0.500
101	0.1	1.105	0.525
115	1.5	4.482	0.818
116	1.6	4.953	0.832



Interpreting coefficients

- intercept (b_0)
  - log odds when all $X_j = 0$
- slope (b_j)
  - change in log odds with one unit increase of predictor X_j
- exponentiated slope ($\exp(b_j)$ or e^{b_j})
  - multiplicative effect on odds
- E.g. $\ln\left(\frac{P(Y_i=1)}{P(Y_i=0)}\right) = -10 + 0.1 \times \mathtt{IQ}_i$

IQ	$\ln\left(\frac{P(Y=1)}{P(Y=0)}\right)$	$\frac{P(Y=1)}{P(Y=0)}$	$P(Y=1)$
100	0	1	0.500
101	0.1	1.105	0.525
115	1.5	4.482	0.818
116	1.6	4.953	0.832

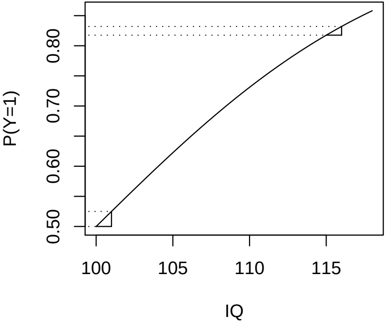


$$\frac{1.105}{1} = \frac{4.953}{4.482} = 1.105$$

Interpreting coefficients

- intercept (b_0)
  - log odds when all $X_j = 0$
- slope (b_j)
  - change in log odds with one unit increase of predictor X_j
- exponentiated slope ($\exp(b_j)$ or e^{b_j})
  - multiplicative effect on odds
- E.g. $\ln\left(\frac{P(Y_i=1)}{P(Y_i=0)}\right) = -10 + 0.1 \times \mathtt{IQ}_i$

IQ	$\ln\left(\frac{P(Y=1)}{P(Y=0)}\right)$	$\frac{P(Y=1)}{P(Y=0)}$	$P(Y=1)$
100	0	1	0.500
101	0.1	1.105	0.525
115	1.5	4.482	0.818
116	1.6	4.953	0.832



Know your participants. . .

Neuroticism and Extraversion were centered.

Contrast code for gender: $\texttt{Gender}_i = -1$ (male), $\texttt{Gender}_i = 1$ (female)

effect	b	$SE(b)$	Wald	p	$\exp b$
intercept	-0.366	0.056	42.129	$< .001$	0.693
Neuroticism (N)	0.005	0.012	0.190	.663	1.005
Extraversion (E)	0.065	0.015	19.333	$< .001$	1.067
Gender (G)	0.134	0.056	5.609	.018	1.143
N × E	-0.009	0.003	8.848	.003	0.991
N × G	0.003	0.012	0.047	.828	1.003
E × G	0.006	0.015	0.165	.685	1.006
N × E × G	0.004	0.003	1.932	.164	1.004

Know your participants...

Neuroticism and Extraversion were centered.

Contrast code for gender: $\texttt{Gender}_i = -1$ (male), $\texttt{Gender}_i = 1$ (female)

effect	b	$SE(b)$	Wald	p	$\exp b$
intercept	-0.366	0.056	42.129	$< .001$	0.693
Neuroticism (N)	0.005	0.012	0.190	.663	1.005
Extraversion (E)	0.065	0.015	19.333	$< .001$	1.067
Gender (G)	0.134	0.056	5.609	.018	1.143
N × E	-0.009	0.003	8.848	.003	0.991
N × G	0.003	0.012	0.047	.828	1.003
E × G	0.006	0.015	0.165	.685	1.006
N × E × G	0.004	0.003	1.932	.164	1.004

Average female:

$$P(\texttt{volunteer}) = \frac{\exp(-0.366 + 0.134)}{1 + \exp(-0.366 + 0.134)} = 0.442$$

Know your participants...

Neuroticism and Extraversion were centered.
Contrast code for gender: $\texttt{Gender}_i = -1$ (male), $\texttt{Gender}_i = 1$ (female)

effect	b	$SE(b)$	Wald	p	$\exp b$
intercept	-0.366	0.056	42.129	$< .001$	0.693
Neuroticism (N)	0.005	0.012	0.190	.663	1.005
Extraversion (E)	0.065	0.015	19.333	$< .001$	1.067
Gender (G)	0.134	0.056	5.609	.018	1.143
N × E	-0.009	0.003	8.848	.003	0.991
N × G	0.003	0.012	0.047	.828	1.003
E × G	0.006	0.015	0.165	.685	1.006
N × E × G	0.004	0.003	1.932	.164	1.004

Average female (odds):

$$\frac{P(\texttt{volunteer})}{1 - P(\texttt{volunteer})} = \frac{0.442}{0.558} = 0.792$$

Know your participants...

Neuroticism and Extraversion were centered.

Contrast code for gender: $\texttt{Gender}_i = -1$ (male), $\texttt{Gender}_i = 1$ (female)

effect	b	$SE(b)$	Wald	p	$\exp b$
intercept	-0.366	0.056	42.129	$< .001$	0.693
Neuroticism (N)	0.005	0.012	0.190	.663	1.005
Extraversion (E)	0.065	0.015	19.333	$< .001$	1.067
Gender (G)	0.134	0.056	5.609	.018	1.143
N × E	-0.009	0.003	8.848	.003	0.991
N × G	0.003	0.012	0.047	.828	1.003
E × G	0.006	0.015	0.165	.685	1.006
N × E × G	0.004	0.003	1.932	.164	1.004

Average male:

$$P(\texttt{volunteer}) = \frac{\exp(-0.366 - 0.134)}{1 + \exp(-0.366 - 0.134)} = 0.377$$

Know your participants...

Neuroticism and Extraversion were centered.

Contrast code for gender: $\texttt{Gender}_i = -1$ (male), $\texttt{Gender}_i = 1$ (female)

effect	b	$SE(b)$	Wald	p	$\exp b$
intercept	-0.366	0.056	42.129	$< .001$	0.693
Neuroticism (N)	0.005	0.012	0.190	.663	1.005
Extraversion (E)	0.065	0.015	19.333	$< .001$	1.067
Gender (G)	0.134	0.056	5.609	.018	1.143
N × E	-0.009	0.003	8.848	.003	0.991
N × G	0.003	0.012	0.047	.828	1.003
E × G	0.006	0.015	0.165	.685	1.006
N × E × G	0.004	0.003	1.932	.164	1.004

Average male (odds):

$$\frac{P(\texttt{volunteer})}{1 - P(\texttt{volunteer})} = \frac{0.377}{0.623} = 0.607$$

Know your participants. . .

Neuroticism and Extraversion were centered.

Contrast code for gender: $\texttt{Gender}_i = -1$ (male), $\texttt{Gender}_i = 1$ (female)

effect	b	$SE(b)$	Wald	p	$\exp b$
intercept	-0.366	0.056	42.129	$< .001$	0.693
Neuroticism (N)	0.005	0.012	0.190	.663	1.005
Extraversion (E)	0.065	0.015	19.333	$< .001$	1.067
Gender (G)	0.134	0.056	5.609	.018	1.143
N × E	-0.009	0.003	8.848	.003	0.991
N × G	0.003	0.012	0.047	.828	1.003
E × G	0.006	0.015	0.165	.685	1.006
N × E × G	0.004	0.003	1.932	.164	1.004

Odds-ratio (female-male):

$$\frac{\text{Odds female}}{\text{Odds male}} = \frac{0.792}{0.607} = 1.307 = \exp(2 \times 0.134)$$

Tests for individual predictors

- Similar to GLM:
  - $H_0 : \beta_j = 0$
  - $H_a : \beta_j \neq 0$
- Wald statistic:
  - like t tests for parameters in linear regression
  - $W = \left(\frac{\hat{b}}{SE(\hat{b})}\right)^2$
  - approximately Chi-Square distributed with $df = 1$
  - not so good for small samples
- Model comparison (likelihood-ratio statistic):
  - compare MODEL A with all predictors to MODEL C without predictor of interest (slope fixed to 0)
  - better with small samples

Model likelihood

- Probability of data according to the model

$$\begin{aligned} L &= P(Y_1, Y_2, \ldots, Y_n | \text{model}) \\ &= P(Y_1|\text{model}) \times P(Y_2|\text{model}) \times \ldots \times P(Y_n|\text{model}) \\ &= \prod_{i=1}^{n} P(Y_i|\text{model}) \end{aligned}$$

- Taking log turns this into sum (easier to handle)

$$\log L = \log \left(\prod_{i=1}^{n} P(Y_i|\text{model}) \right) = \sum_{i=1}^{n} \log P(Y_i|\text{model})$$

- Use the *negative* log likelihood as a measure of error

Model comparison

Compare two models by difference in likelihood

- To compare MODEL A with PA parameters to MODEL C with PC parameters, use the statistic

$$\chi^2 = -2 \log L(C) - (-2 \log L(A))$$

  - Chi-Square distributed with $df = \text{PA} - \text{PC}$
  - equivalent to -2 times the *log likelihood ratio*, i.e.

$$-2 \log \frac{L(C)}{L(A)}$$

- Similar idea as F test for GLM

Pseudo R-Squared

In GLM, R^2 is proportion of variance accounted for. In logistic regression, define "pseudo-R^2" as proportion of likelihood accounted for

- Three versions, always compare MODEL A to MODEL C with intercept only
  - McFadden's pseudo R-Squared

$$\text{pseudo-R}^2 = \frac{\log L(C) - \log L(A)}{\log L(C)} = 1 - \frac{\log L(A)}{\log L(C)}$$

  - Cox-Snell R-Squared

$$\text{pseudo-R}^2 = 1 - \exp\left(-\frac{2}{N}(\log L(A) - \log L(C))\right)$$

  - Nagelkerke's R Squared:

$$\text{pseudo-R}^2 = \frac{1 - \exp\left(-\frac{2}{N}(\log L(A) - \log L(C))\right)}{1 - \exp\left(\frac{2}{N}\log L(C)\right)}$$

Test for model fit

A common test to assess whether model fits the data:

- Compare model to the most general model
- Most general model is the *saturated* model with a parameter for each unique predictor pattern
- Test statistic: $-2\log L(A) - (-2\log L(\text{saturated}))$
- approximately Chi-Square distributed with $df = n - \text{PA}$ (n is total number of unique predictor patterns)
- If observations come from different individuals, each forms his/her own group
  - $\log L(\text{saturated}) = 0$
  - Test statistic simplified to $-2\log L(A)$, with $df = n - \text{PA}$

Know your participants. . .

effect	b	$SE(b)$	Wald	p	$\exp b$
intercept	-0.366	0.056	42.129	$< .001$	0.693
Neuroticism (N)	0.005	0.012	0.190	.663	1.005
Extraversion (E)	0.065	0.015	19.333	$< .001$	1.067
Gender (G)	0.134	0.056	5.609	.018	1.143
N × E	-0.009	0.003	8.848	.003	0.991
N × G	0.003	0.012	0.047	.828	1.003
E × G	0.006	0.015	0.165	.685	1.006
N × E × G	0.004	0.003	1.932	.164	1.004

$-2\log L(A) = 1895.243$ $\quad$ $-2\log L(C) = 1933.506$

Omnibus test:

$$\chi^2 = 1933.506 - 1895.243 = 38.263$$

with df = PA – PC = $8 - 1 = 7$ and $P(\chi^2_7 \geq 38.263) < .001$

Know your participants. . .

effect	b	$SE(b)$	Wald	p	$\exp b$
intercept	-0.366	0.056	42.129	$< .001$	0.693
Neuroticism (N)	0.005	0.012	0.190	.663	1.005
Extraversion (E)	0.065	0.015	19.333	$< .001$	1.067
Gender (G)	0.134	0.056	5.609	.018	1.143
N × E	-0.009	0.003	8.848	.003	0.991
N × G	0.003	0.012	0.047	.828	1.003
E × G	0.006	0.015	0.165	.685	1.006
N × E × G	0.004	0.003	1.932	.164	1.004

$-2\log L(A) = 1895.243 \qquad -2\log L(C) = 1933.506$

McFadden pseudo-$R^2 = \frac{1933.506 - 1895.243}{1933.506} = 0.0197$

Know your participants. . .

effect	b	$SE(b)$	Wald	p	$\exp b$
intercept	-0.366	0.056	42.129	$< .001$	0.693
Neuroticism (N)	0.005	0.012	0.190	.663	1.005
Extraversion (E)	0.065	0.015	19.333	$< .001$	1.067
Gender (G)	0.134	0.056	5.609	.018	1.143
N × E	-0.009	0.003	8.848	.003	0.991
N × G	0.003	0.012	0.047	.828	1.003
E × G	0.006	0.015	0.165	.685	1.006
N × E × G	0.004	0.003	1.932	.164	1.004

$-2 \log L(A) = 1895.243$ $\quad$ $-2 \log L(C) = 1933.506$

Cox-Snell pseudo-$R^2 = 1 - \exp(-\frac{2}{1421} \times (-947.6215 - -966.753)) = 0.0266$

Know your participants. . .

effect	b	$SE(b)$	Wald	p	$\exp b$
intercept	-0.366	0.056	42.129	$< .001$	0.693
Neuroticism (N)	0.005	0.012	0.190	.663	1.005
Extraversion (E)	0.065	0.015	19.333	$< .001$	1.067
Gender (G)	0.134	0.056	5.609	.018	1.143
N × E	-0.009	0.003	8.848	.003	0.991
N × G	0.003	0.012	0.047	.828	1.003
E × G	0.006	0.015	0.165	.685	1.006
N × E × G	0.004	0.003	1.932	.164	1.004

$-2 \log L(A) = 1895.243$ $\qquad$ $-2 \log L(C) = 1933.506$

Nagelkerke pseudo-$R^2 = \frac{1 - \exp(-\frac{2}{1421} \times (-947.6215 - -966.753))}{1 - \exp(-\frac{2}{1421} \times (-966.753))} = 0.0357$

Know your participants...

effect	b	$SE(b)$	Wald	p	$\exp b$
intercept	-0.366	0.056	42.129	$< .001$	0.693
Neuroticism (N)	0.005	0.012	0.190	.663	1.005
Extraversion (E)	0.065	0.015	19.333	$< .001$	1.067
Gender (G)	0.134	0.056	5.609	.018	1.143
N × E	-0.009	0.003	8.848	.003	0.991
N × G	0.003	0.012	0.047	.828	1.003
E × G	0.006	0.015	0.165	.685	1.006
N × E × G	0.004	0.003	1.932	.164	1.004

$-2\log L(A) = 1895.243$ $\quad$ $-2\log L(C) = 1933.506$

Model fit test (against saturated model):

$$\chi^2 = 1895.243$$

with df $= n -$ PA $= 1421 - 8 = 1413$ and $P(\chi^2_{1413} \geq 1895.243) < .001$

Hosmer-Lemeshow Chi-Square

- Continuous predictors: often only one observation with particular predicted probability
  - expected frequency < 5
  - problem for Chi-Square
- Hosmer-Lemeshow test
  - divide sample into g (roughly) equal groups (e.g., deciles: $g = 10$)
  - groups form cells for Chi-Square
  - test with $df = g - 2$

In practice

For categorical predictors, use contrast codes as in ANOVA

Apart from assumptions, similar concerns as linear regression:

- Sample size
- Outliers
- Multicollinearity

More than two categories

Logistic regression can also be used when dependent variable is polytomous (with $m > 2$ categories)

- Nominal categories:
  - Labour, conservative, lib dem
  - etc ...
- Ordered categories:
  - Yes, maybe, no
  - Fundamentalist, moderate, liberal
  - etc ...
- General idea:
  - For k categories, fit $k-1$ simultaneous logistic regression models

Further reading

- DeMaris, A. (1995). A tutorial in logistic regression. *Journal of Marriage and Family*, *57*, 956-968. (on Moodle)
- Speekenbrink, M. (2012). *Logistic regression* (lecture notes)