Exercises logodds and probabilities

Stéphanie M. van den Berg

March 14, 2020

1 Exercises

1.1 From probability to logodds:

Given: In the Netherlands, 51% of the inhabitants is female.

1. If we randomly pick someone from this Dutch population, what is the probability that that person is female?
2. If we randomly pick someone from this Dutch population, what are the odds that that person is female over being male? (:)
3. If we randomly pick someone from this Dutch population, what are the odds that that person is male over being female? (:)
4. What is the odds of randomly picking an inhabitant that is female, expressed as one number?
5. What is the odds of randomly picking an inhabitant that is male, expressed as one number?
6. What is the logodds of randomly picking an inhabitant that is female?
7. What is the logodds of randomly picking an inhabitant that is male?

Answers:

1. 0.51
2. 51 to 49 (51:49).
3. 49:51.
4. 51/49=1.04
5. 49/51=0.96
6. ln(51/49) = ln(1.04) = 0.04
7. ln(49/51) = ln(0.96) = -0.04

1.2 From logoddss to probabilities:

Given: In the Netherlands, 51% of the inhabitants are female. Females tend to get older than males, so if we predict sex by age, we should expect a higher probability of a female for older ages. Suppose we have the following linear model for the relationship between age (in years) and the logodds of being female:

$$logodds_{female} = -0.01 + 0.01 \times age,$$

1. What is the predicted logodds of being female for a person of age 20?
2. What is the predicted logodds of being female for a person of age 90?
3. What is the predicted odds of being female for a person of age 20?
4. What is the predicted odds of being female for a person of age 90?
5. What are the predicted odds of being female for a person of age 20?
6. What are the predicted odds of being female against being male for a person of age 90?
7. What is the predicted probability of being female against being male for a person of age 20?
8. What is the predicted probability of being female for a person of age 90?
9. What is the predicted probability of being MALE for a person of age 90?

Answers:

1. $-0.01 + 0.01 \times 20 = 0.19$
2. $-0.01 + 0.01 \times 90 = 0.89$
3. $exp(0.19) = 1.21$
4. $exp(0.89) = 2.44$
5. 1.21 to 1, or 1.21:1
6. 2.44 to 1, or 2.44:1
7. 1.21/ (1.21 + 1)= 0.55
8. 2.44 / (2.44 + 1)= 0.71
9. 1 - 0.71 = 0.29

2 Some extra questions

A big data analyst constructs a model that predicts whether an account on Twitter belongs to either a real person or organisation, or to a bot.

1. For one account, a user of this model finds a logodds of 4.5 that the account belongs to a bot. What is the corresponding probability that the twitter account belongs to a bot? Give the calculation.

2. For a short tweet with only a hyperlink, the probability that it comes from a bot is only 10%. What is the logodds that corresponds to this probability? Give the calculation.

Answers:

1. The logodds is 4.5, so the oddsratio is exp(4.5)=90.0. The odds of being a bot is then 90:1. The probability of being a bot is 90/ (90+1)= 0.99

2. Out of 100 tweets with only a hyperlink, 10 are by bots and 90 are by real persons or organisations. So the odds of coming from a bot are 10:90. The odds is therefore 10/90 = 0.11. When we take the natural logarithm of this odds, we get the logodds: ln(0.11) = -2.21.

3 Logistic regression

Table 1: Taking the train to Paris data.

train	age	sex_male	income	business
1	35.12	1	7544.00	1
1	66.66	1	7096.00	0
0	42.77	1	29261.00	1
0	72.63	0	24977.00	0
1	76.25	0	876.00	1
0	19.87	1	126943.00	1

Using the train data in Table 1, we try to predict whether people take the train or not by their purpose of their trip: business or not.

1. What does the SPSS syntax look like? Note the data in Table 1.

2. Suppose the results look like those in Figure 1. What is the predicted probability of taking the train for people that travel for business? Provide the calculations.

3. Suppose the results look like those in Figure 1. What is the predicted probability of taking the train for people that travel NOT for business? Provide the calculations.

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-1.155	.1196	-1.389	-.921	93.321	1	.000
business	-.050	.1531	-.351	.250	.108	1	.742
(Scale)	1[a]						

Dependent Variable: train
Model: (Intercept), business

a. Fixed at the displayed value.

Figure 1: SPSS output of a generalized linear model for predicting taking the train from purpose of the trip.

4. Suppose the results look like those in Figure 2. What is the predicted probability of taking the train for people that travel for business? Provide the calculations.

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-1.205	.0957	-1.393	-1.018	158.757	1	.000
[business=.00]	.050	.1531	-.250	.351	.108	1	.742
[business=1.00]	0[a]	.	.	.	.	.	.
(Scale)	1[b]						

Dependent Variable: train
Model: (Intercept), business

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

Figure 2: SPSS output of a generalized linear model for predicting taking the train from purpose of the trip.

5. Suppose the results look like those in Figure 2. What is the predicted probability of taking the train for people that travel NOT for business? Provide the calculations.

6. On the basis of this SPSS output, do business travellers tend to take the train more or less often than non-business travellers? Motivate your answer.

7. Suppose in SPSS output for logistic regression, you find an intercept value of 0.5 with a standard error of 0.1. There is a corresponding Wald chi-square value of 25. Explain where this Wald chi-square value comes from.

8. Suppose we have data on coin flips as in Table **??**:

ID	Heads	weight	type
1	0	2.7831226	5cents
2	1	0.8058492	10cents
3	1	3.1401581	1Euro
4	1	1.0156831	10cents
5	1	4.4503490	1Euro

If we want to predict the outcome of the coin flip, on the basis of the type of coin, should we use a linear model, a linear mixed model, or a generalized linear model? Motivate your answer.

If we want to predict the weight of the coin, on the basis of the type of the coin, should we use a linear model, a linear mixed model, or a generalized linear model? Motivate your answer.

Answers:

1. It could look like this (using WITH, treating the independent variable as quantitative):

```
GENLIN train (REFERENCE=FIRST) WITH business
  /MODEL business
 DISTRIBUTION=BINOMIAL LINK=LOGIT
  /PRINT CPS DESCRIPTIVES   SOLUTION.
```

or like this (using BY, treating the independent variable as qualitative)

```
GENLIN train (REFERENCE=FIRST) BY business
  /MODEL business
 DISTRIBUTION=BINOMIAL LINK=LOGIT
  /PRINT CPS DESCRIPTIVES   SOLUTION.
```

2. People that travel for business score 1 on the business variable. So the predicted logodds for those people is $-1.155 - 0.050 \times 1 = -1.205$. The odds is the $exp(-1.205) = 0.299692$. So the odds of going by train are 0.30 to 1. This is equivalent to 3 to 10. So suppose we have 13 trips, 3 are by train and 10 are not by train. So the probability of a trip being by train equals $3/13 = 0.23$. Or logit(-1.205)=exp(-1.205)/(1+exp(-1.205))=0.3/1.3=0.23

3. People that travel NOT for business score 0 on the business variable. So the predicted logodds for those people is $-1.155 - 0.050 \times 0 = -1.155$. The odds is the $exp(-1.155) = 0.3150575$. So the odds of going by train are 0.32 to 1. This is equivalent to 32 to 100. So suppose we have 132 trips, 32 are by train and 100 are not by train. So the probability of a trip being by train equals $32/132 = 0.24$.

4. p=exp(-1.205)/(1+exp(-1.205))=0.3/1.3=0.23

5. p=exp(-1.205+0.050)/(1+exp(-1.205+0.050))= exp(-1.155)/(1+exp(-1.155))=0.32/1.32=0.24

6. Less often, since the slope for the dummy variable [**business=0**] is positive. Business trips are the reference category, and relative to that non-business trips get an extra slope of 0.050, so a higher logodds and therefore a higher probability. So if non-business trips have a higher probability of being by train, then business trips have a lower probablity of being by train. We also see that for the answers to the previous questions: the probability of taking the train for business trips is 0.23 and for non-business trips it is 0.24.

7. See the formula in the text: $X^2 = \frac{B^2}{SE^2} = \frac{0.5^2}{0.1^2} = \frac{0.25}{0.01} = 25$.

8. If we want to predict the outcome of the coin flip, on the basis of the type of coin, we should use a generalized linear model, because the dependent variable is dichotomous (has only 2 values), so the residuals can never have a normal distribution.

   If we want to predict the weight of the coin, on the basis of the type of the coin, we should use a linear model, because the dependent variable is continuous.