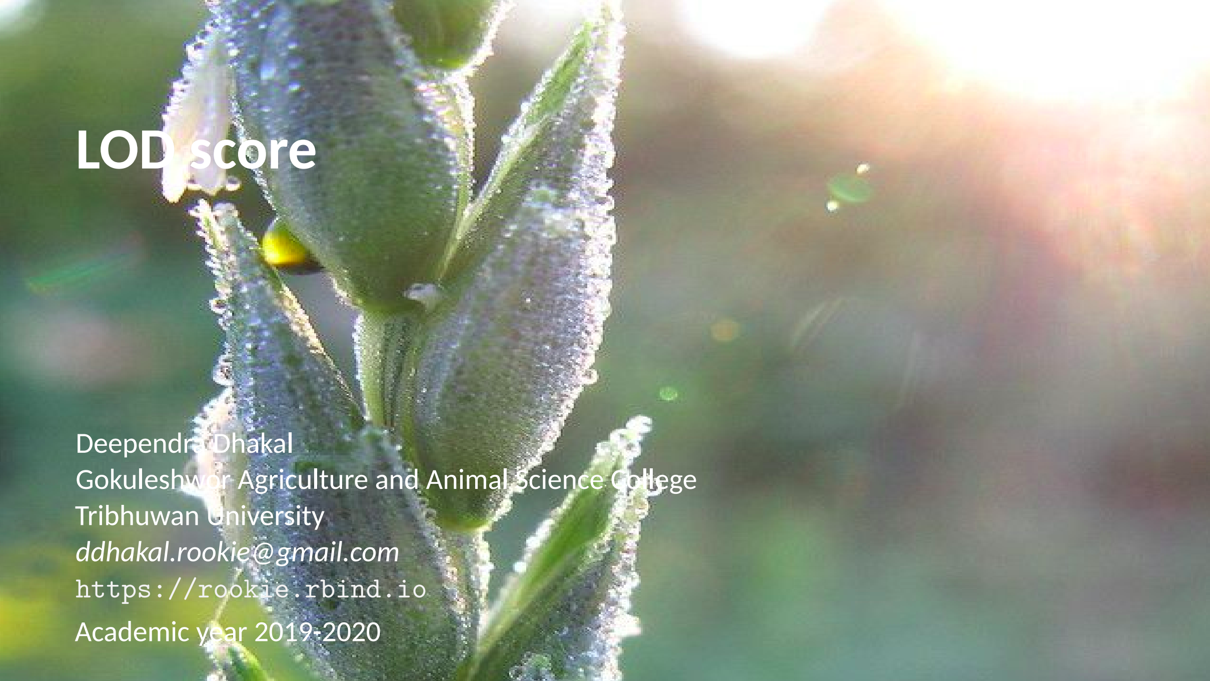


LOD score

Deependra Dhakal
Gokuleshwor Agriculture and Animal Science College
Tribhuwan University
ddhakal.rookie@gmail.com
`https://rookie.rbind.io`

Academic year 2019-2020

Outline

1. LOD score
2. Bibliography

Outline

1 LOD score

2 Bibliography

LOD score

- Genetic markers located on the same chromosome tend to remain together during sexual reproduction (linkage groups).
- That is, they do not exhibit independent assortment. Consequently, there are as many linkage groups as there are homologous pairs of chromosomes.
- Manual linkage analysis is feasible if only a few markers are being studied.
- Modern linkage map construction is a computerized operation, feasible through mapping software packages. - These computer software packages use the coded information from the segregating population to determine recombination frequencies.
- The basic calculation is a ratio (odds ratio) of linkage versus no linkage, expressed as the log of the ratio (logarithm of odds or LOD).

LOD score

- The LOD score compares the likelihood of obtaining the test data if the two loci are indeed linked, to the likelihood of observing the same data purely by chance.
- A **LOD value** or score measures the likelihood of linkage between two markers, a score of more than three usually being the cut-off minimum for mapping.
- Positive LOD scores favour the presence of linkage, whereas negative LOD scores indicate that linkage is less likely.
- A LOD of three indicates a 1000:1 odds in favor of genetic linkage (linkage between the two markers is a thousand times more likely than no linkage).
- The researcher may vary the stringency of mapping by, for example, lowering the LOD score to detect a greater level of linkage.

Linear regression models are defined by the equation

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_q x_q + \epsilon; \epsilon \sim N(0, \sigma^2)$$

which calculates the response Y directly. Logisitc regression is defined in a similar manner,

Consider a model with two predictors, x_1 and x_2, and one binary (Bernoulli) response variable Y, which we denote $p = P(Y = 1)$. We assume a linear relationship between the predictor variables, and the log-odds of the event that $Y = 1$.

This linear relationship can be written in the following mathematical form (where ℓ is the log-odds, b is the base of the logarithm, and β_i are parameters of the model):

$$\ell = \log_b \left(\frac{p(\mathbf{x})}{1 - p(\mathbf{x})} \right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}$$

- In binary context, the odds are probability for a ‘positive’ event ($Y = 1$) divided by the probability of ‘negative’ event ($Y = 0$).

$$\frac{p(\mathbf{x})}{1 - p(\mathbf{x})} = \frac{P[Y = 1|\mathbf{X} = \mathbf{x}]}{P[Y = 0|\mathbf{X} = \mathbf{x}]}$$

- The logistic regression equation guarantees that a value between 0 and 1 is calculated. This is evident the when the inverse logit transformation is applied, which results in a “direct” probability prediction.

$$p(\mathbf{x_i}) = P[Y = 1|\mathbf{X} = \mathbf{x}] = \frac{\exp^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{i(p-1)}}}{1 + \exp^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{i(p-1)}}}$$

- Note that this is prediction of “probability”, not a numerical value. This probability value must be translated to a categorical prediction.

We can recover the odds by exponentiating the log-odds:

$$\frac{p}{1-p} = b^{\beta_0+\beta_1 x_1+\beta_2 x_2}.$$

By simple algebraic manipulation, the probability that $Y = 1$ is

$$p = \frac{b^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{b^{\beta_0+\beta_1 x_1+\beta_2 x_2} + 1} = \frac{1}{1 + b^{-(\beta_0+\beta_1 x_1+\beta_2 x_2)}}$$

The above formula shows that once β_i are fixed, we can easily compute either the log-odds that $Y = 1$ for a given observation, or the probability that $Y = 1$ for a given observation. The main use-case of a logistic model is to be given an observation (x_1, x_2), and estimate the probability p that $Y = 1$. In most applications, the base b of the logarithm is usually taken to be e.

Example: Inerpreting logarithm of odds

We consider an example with $b = 10$, and coefficients $\beta_0 = -3$ $\beta_1 = 1$ $\beta_2 = 2$. To be concrete, the model is

$$\log_{10} \frac{p}{1-p} = \ell = -3 + x_1 + 2x_2$$

where p is the probability of the event that $Y = 1$.

- This can be interpreted as follows:
  - $\beta_0 = -3$ is the y-intercept. It is the log-odds of the event that $Y = 1$, when the predictors $x_1 = x_2 = 0$. By exponentiating, we can see that when $x_1 = x_2 = 0$ the odds of the event that $Y = 1$ are 1-to-1000, or 10^{-3} . Similarly, the probability of the event that $Y = 1$ when $x_1 = x_2 = 0$ can be computed as $1/(1000 + 1) = 1/1001$.
  - $\beta_1 = 1$ means that increasing x_1 by 1 increases the log-odds by 1. So if x_1 increases by 1, the odds that $Y = 1$ increase by a factor of 10^1.
  - $\beta_2 = 2$ means that increasing x_2 by 1 increases the log-odds by 2. So if x_2 increases by 1, the odds that $Y = 1$ increase by a factor of 10^2. Note how the effect of x_2 on the log-odds is twice as great as the effect of x_1, but the effect on the odds is 10 times greater.
- In order to estimate the parameters β_i from data, one must do logistic regression.

Example: Probability of passing an exam versus hours of study

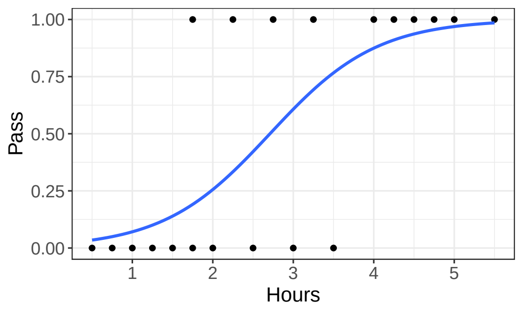
- A group of 20 students spends between 0 and 6 hours studying for an exam. How does the number of hours spent studying affect the probability of the student passing the exam?
- The reason for using logistic regression for this problem is that the values of the dependent variable, pass and fail, while represented by "1" and "0", are not cardinal numbers. If the problem was changed so that pass/fail was replaced with the grade 0–100 (cardinal numbers), then simple regression analysis could be used.

LOD score

- The table shows the number of hours each student spent studying, and whether they passed (1) or failed (0).

Hours	Pass	Hours1	Pass1
0.50	0	2.8	1
0.75	0	3.0	0
1.00	0	3.2	1
1.25	0	3.5	0
1.50	0	4.0	1
1.75	0	4.2	1
1.75	1	4.5	1
2.00	0	4.8	1
2.25	1	5.0	1
2.50	0	5.5	1

The graph shows the probability of passing the exam versus the number of hours studying, with the logistic regression curve fitted to the data.



The logistic regression gives the following output:

term	estimate	std.error	statistic	p.value
(Intercept)	-4.1	1.76	-2.3	0.02
Hours	1.5	0.63	2.4	0.02

The output indicates that hours studying is significantly associated with the probability of passing the exam (p = 0.0167, Wald test). The output also provides the coefficients for Intercept = -4.0777 and Hours = 1.5046. The coefficient for Hours values indicate change in the log odds of passing the exam due to one unit change in Hours (i.e. β_1 coefficient). These coefficients are entered in the logistic regression equation to estimate the odds (probability) of passing the exam:

$$\begin{aligned} \text{Log-odds of passing exam} &= 1.5046 \cdot \text{Hours} - 4.0777 = 1.5046 \cdot (\text{Hours} - 2.71) \\ \text{Odds of passing exam} &= \exp(1.5046 \cdot \text{Hours} - 4.0777) = \exp(1.5046 \cdot (\text{Hours} - 2.71)) \\ \text{Probability of passing exam} &= \frac{1}{1 + \exp(-(1.5046 \cdot \text{Hours} - 4.0777))} \end{aligned}$$

- One additional hour of study is estimated to increase log-odds of passing by 1.5, so multiplying odds of passing by exp(1.5046) $\approx$ 4.5. The form with the x-intercept (2.71) shows that this estimates even odds (log-odds 0, odds 1, probability 1/2) for a student who studies 2.71 hours.
- For example, for a student who studies 2 hours, entering the value Hours = 2 in the equation gives the estimated probability of passing the exam of 0.26:

$$\text{Probability of passing exam} = \frac{1}{1 + \exp\left(-\left(1.5046 \cdot 2 - 4.0777\right)\right)} = 0.26$$

- Similarly, for a student who studies 4 hours, the estimated probability of passing the exam is 0.87:

$$\text{Probability of passing exam} = \frac{1}{1 + \exp\left(-\left(1.5046 \cdot 4 - 4.0777\right)\right)} = 0.87$$

This table shows the probability of passing the exam for several values of hours studying.

Hours	log_odds	p_value	odds
0.0	-4.08	0.02	0.02
1.0	-2.57	0.07	0.08
2.0	-1.07	0.26	0.34
3.0	0.44	0.61	1.55
4.0	1.94	0.87	6.96
5.0	3.45	0.97	31.36
6.0	4.95	0.99	141.20
7.0	6.45	1.00	635.75
8.0	7.96	1.00	2862.50
9.0	9.46	1.00	12888.56
10.0	10.97	1.00	58031.48
2.7	0.00	0.50	1.00

The output from the logistic regression analysis gives a p-value of p = 0.0167, which is based on the Wald z-score. Rather than the Wald method, the recommended method to calculate the p-value for logistic regression is the likelihood-ratio test (LRT), which for this data gives p = 0.0006.

Example: Proportions of female children at various ages during adolescence who have reached menarche

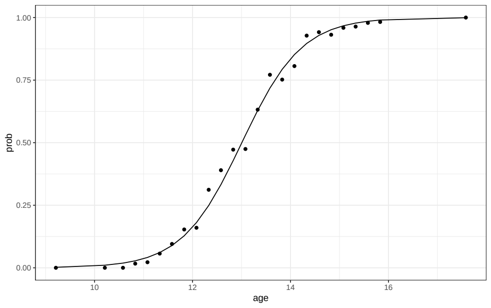
The coefficient returned by a logistic regression in r is a logit, or the log of the odds. To convert logits to odds ratio, you can exponentiate it, as you've done above. To convert logits to probabilities, you can use the function `exp(logit)/(1+exp(logit))`. However, there are some things to note about this procedure.

First, lets generate some reproducible data to illustrate. We fit a generalized linear model to menarche dataset with response variable as function of Age. The coefficients displayed are for logits, just as in your example. The `predict()` gives the predicted value in terms of logits.

LOD score

term	estimate	std.error	statistic	p.value
(Intercept)	-21.2	0.77	-28	0
Age	1.6	0.06	28	0

If we plot these data and this model, we see the sigmoidal function that is characteristic of a logistic model fit to binomial data.



Note that the change in probabilities is not constant - the curve rises slowly at first, then more quickly in the middle, then levels out at the end. The difference in probabilities between 10 and 12 is far less than the difference in probabilities between 12 and 14. This means that it’s impossible to summarise the relationship of age and probabilities with one number without transforming probabilities.

To answer the specific questions: How do you interpret odds ratios?

The odds ratio for the value of the intercept is the odds of a “success” (in your data, this is the odds of taking the product) when x = 0 (i.e. zero thoughts). The odds ratio for your coefficient is the increase in odds above this value of the intercept when you add one whole x value (i.e. x=1; one thought). Using the menarche data:

names	x
(Intercept)	0.0
Age	5.1

We could interpret this as the odds of menarche occurring at age = 0 is .00000000006. Or, basically impossible. Exponentiating the age coefficient tells us the expected increase in the odds of menarche for each unit of age. In this case, it’s just over a quintupling. An odds ratio of 1 indicates no change, whereas an odds ratio of 2 indicates a doubling, etc.

Your odds ratio of 2.07 implies that a 1 unit increase in ‘Thoughts’ increases the odds of taking the product by a factor of 2.07.

How do you convert odds ratios of thoughts to an estimated probability of decision?

You need to do this for selected values of thoughts, because, as you can see in the plot above, the change is not constant across the range of x values. If you want the probability of some value for thoughts, get the answer as follows:

```
exp(intercept +
coef*THOUGHT_Value)/(1+(exp(intercept+coef*THOUGHT_Value)))
```

For another example refer to: Mapping QTL in populations with known pedigrees in Griffiths et al. (2015) (pp 742).

Outline

1 LOD score

2 Bibliography

References

Griffiths, Anthony JF, Susan R Wessler, Sean B Carroll, Doebley John, and others. 2015. *An Introduction to Genetic Analysis*. Macmillan.