

Advanced Statistical Methods for Research

Solution to exam questions on logistic regression

Suggested solutions appear in bold

INSTRUCTIONS: Plan to spend about 1 hour on the problems on logistic regression. Answer each question and show work in the space provided on the exam. Turn in the **entire** exam when you are done or when time is up.

1. An experiment is conducted on the toxicity of doses of an insecticide on the tobacco budworm moth. In the experiment batches of 20 male moths were exposed for 3 days to the insecticide and the number in each batch that were dead or knocked down was recorded. The data are given below.

Dose Level	1	2	4	8	16	32
$\log_2(\text{Dose Level})$	0	1	2	3	4	5
Dead or down	1	4	9	13	18	20

- (a) In general, describe the relationship between the dose level and the proportion of male moths dead or down.

As the dose increases, the proportion dead or down also increases. Only 5% are dead at dose level 1, but 100% are dead at dose level 32.

- (b) What is the observed proportion of dead or down at dose level 16?

$$p = \frac{18}{20} = \mathbf{0.90}$$

- (c) What are the observed odds of dead or down at dose level 16?

$$\frac{p}{1-p} = \frac{0.90}{0.10} = \mathbf{9 \text{ or } 9 \text{ to } 1 \text{ odds.}}$$

A logistic regression of the proportion of dead or down on the $\log_2(\text{Dose Level})$ is run. Below is the summary from S+.

Coefficients:

	Value	Std. Error	t value
(Intercept)	-2.818555	0.5479524	-5.143796
log2dose	1.258949	0.2120484	5.937086

Null Deviance: 71.13758 on 5 degrees of freedom

Residual Deviance: 1.88097 on 4 degrees of freedom

(d) Give the logistic regression equation.

$$\hat{\pi}'_i = \text{Predicted log} \left(\frac{\pi_i}{1-\pi_i} \right) = -2.81856 + 1.25895 \log_2(\text{Dose})$$

(e) Use the equation in (d) to predict the log-odds for a dose of 16. What are the predicted odds? What is the predicted proportion of dead and down male moths?

$$\text{Dose} = 16 \quad \log_2(\text{Dose}) = 4$$

$$\hat{\pi}'_i = -2.81856 + 1.25895(4) = 2.21724$$

$$\frac{\hat{\pi}_i}{1-\hat{\pi}_i} = e^{2.21724} = 9.18195$$

$$\hat{\pi}_i = \frac{9.18195}{10.18195} = 0.90179 \text{ or } 90\%$$

(f) If we increase the dose from 16 to 32, by what multiple will the predicted odds increase?

$$e^{1.25895} = 3.52 \text{ the predicted odds will be multiplied by } 3.52.$$

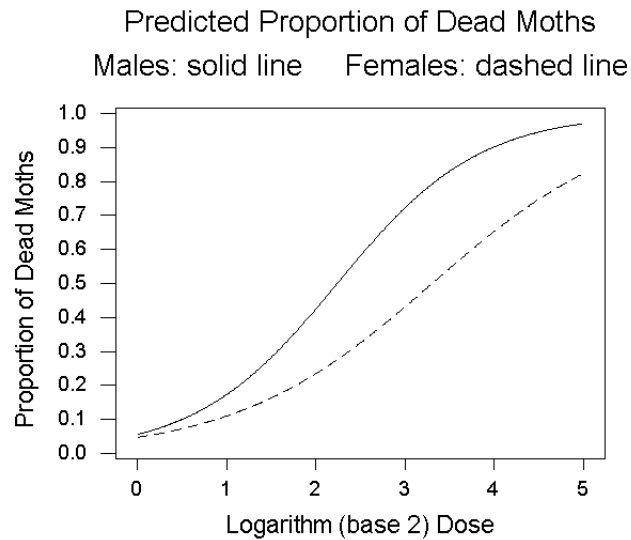
$$\text{Dose} = 32 \quad \log_2(\text{Dose}) = 5$$

$$\hat{\pi}'_i = -2.81856 + 1.25895(5) = 3.47619$$

$$\frac{\hat{\pi}_i}{1-\hat{\pi}_i} = e^{3.47619} = 32.33629$$

$$\frac{32.33629}{9.18195} = 3.52172$$

The experiment is run again with batches of 20 female moths. The number of dead and down moths is again recorded and a logistic regression is run. Below is a plot of the fitted logistic regressions for males and females.



- (g) According to the plot, is there much of a difference between male and female moths in terms of the proportion of dead or down? Explain briefly.

At lower doses there is not much difference between the proportions dead and down for male and female moths. However, at higher doses the difference becomes marked. For example, at Dose=1 ($\log_2(\text{Dose}) = 0$) both males and females have a proportion dead or down of about 5%. At Dose=8 ($\log_2(\text{Dose}) = 3$), males have over 70% dead or down while females have less than 50%.

- (h) According to the plot, if you wish to kill or knock down 50% of the males what dose should you use?

$$\log_2(\text{Dose}) = 2.2 \quad \text{or} \quad \text{Dose} = 4.6$$

- (i) According to the plot, if you wish to kill or knock down 50% of the females what dose should you use?

$$\log_2(\text{Dose}) = 3.2 \quad \text{or} \quad \text{Dose} = 9.2$$

2. Short answer essay questions.

- (a) When modeling the relationship between a numerical predictor variable and a binary response, why is the logit transformation a good idea?

Often the relationship between a binary response and an explanatory (predictor) variable is non-linear (*i.e.* curved). The logit transformation takes the curved relationship and straightens it out so that we can fit a linear model to the logits.

- (b) The usual assumptions placed on the error terms in ordinary least squares regression are:

- independently distributed
- identically distributed (equal variance)
- normally distributed

Which of these assumptions are violated when dealing with binary response data? Explain briefly how each is violated.

Binary response violates the equal variance assumption. Proportions at the extremes (close to 0 and 1) tend to have smaller variances than proportions near the middle (close to 0.5).

Binary response violates the normality assumption. Since binary responses are 0's and 1's, the errors will be binary as well. Therefore, the errors will not be distributed according to a normal distribution.

Independent observations (often accomplished by random sampling) is an important assumption in all regression. We still need the independence of observed responses with binary response data.

3. Each autumn, individuals (especially older persons or the chronically ill) are encouraged to get a flu shot. Fifty persons are selected at random from a health clinic client list and asked if they actually went to get a flu shot. A client who got a flu shot has a response of $Y = 1$, if no flu shot, the response is $Y = 0$. Other data collected were age (Age) and health awareness (Aware), for which higher values indicate greater awareness.

Simple logistic regressions were run on Age and Aware separately.

Coefficients:

	Value	Std. Error
(Intercept)	-6.57492	2.12560
Age	0.13302	0.04439

Null Deviance: 68.03 on 49 degrees of freedom

Residual Deviance: 56.08 on 48 degrees of freedom

Coefficients:

	Value	Std. Error
(Intercept)	-7.39019	2.09332
Aware	0.13486	0.03884

Null Deviance: 68.03 on 49 degrees of freedom

Residual Deviance: 49.28 on 48 degrees of freedom

- (a) In the model with Age alone, is there a significant lack of fit? Is the variable Age significant? Use tests based on deviances to support your answers.

Lack of fit: The residual deviance is 56.08 on 48 df. With 50 df a $\chi^2 = 56.33$ has a P-value of 0.25. This large P-value indicates that there is no significant lack of fit.

Age: The change in deviance between the null and residual is 68.03 minus 56.08 or 11.95 on 1 df. With 1 df, a $\chi^2 = 11.95$ has a P-value between 0.0005 and 0.001. Such a small P-value indicates that such a large change in deviance is not attributable to chance alone. That is, Age is a significant explanatory variable in the prediction of whether people get or do not get a flu shot.

- (b) If you were to pick only one variable, Age or Aware, to model the binary response of whether a client received a flu shot, which one would you choose? Support your choice statistically.

Given the choice between using either Age or Aware, one should choose Aware since the residual deviance is smaller for the model with Aware than it is for the model with Age. This smaller residual deviance indicates a better fit.

A multiple logistic regression was run with both Age and Aware in the model.

Coefficients:

	Value	Std. Error
(Intercept)	-21.58213	6.33966
Age	0.22175	0.07360
Aware	0.20348	0.06206

Null Deviance: 68.03 on 49 degrees of freedom

Residual Deviance: 32.42 on 47 degrees of freedom

- (c) What is the z-test statistic for the variable Age in this multiple logistic regression analysis? Is it statistically significant?

The z-test statistic for Age is $z = \frac{0.22175}{0.07360} = 3.013$. This has a P-value of 0.0013. The small P-value indicates that such a large slope coefficient is not likely to have happened by chance. Age is therefore a significant explanatory variable in the multiple logistic regression model

- (d) Use the change in deviance to test the significance of adding the variable Aware to the simple model using Age.

The residual deviance for the model with both Age and Aware included is 32.42. The residual deviance for the model with only Age included is 56.08. The change in deviance between these two models is 56.08 minus 32.42 or 23.66 on 1 df. This is highly significant with a P-value of virtually zero. Adding Aware to the model that already includes Age significantly improves the fit of the model.