Multiple Regression 5 :Logistic Regression and Intro to Generalized Linear Models

Outline:

1. Logistic Regression with a single continuous predictor
2. Logistic Regression with multiple predictors
3. Generalized Linear models ~ post poned

Examples:

1. Beetle: Logistic Regression with a single continuous predictor
2. Birdkeeping: Logistic Regression with multiple predictors
3. Elephants: Poisson Regression

1. Logistic Regression

In many research studies, the response variable may be represented as one of two possible values (0 or 1, yes or no, dead or alive). In other words, the response variable is binary.

1 not 0 categorical

When the response variable is binary, we are interested in probability of an "event" occurring = p = P(Y=1). We want to relate p to a linear combination of predictor variables. However, p varies between 0 and 1.

The model often used to study the association between a binary response and a set of predictor variables is **Logistic Regression**.

MR

1. normally dist'd => not w/ binary response
2. confined to 0 and 1

★ Simple Logistic Regression Model:

P(X) = P(Y = 1) = P(success)

Let p(x) be the probability that y equals 1 when the predictor variable equals x.

logit = log odds scale

intercept

"slope"

R returns in log odds

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$$

odds scale

$$\left(\frac{p(x)}{1-p(x)}\right) = e^{(\beta_0+\beta_1 x)}$$

most common

probability scale

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

report in probability

beetles dying

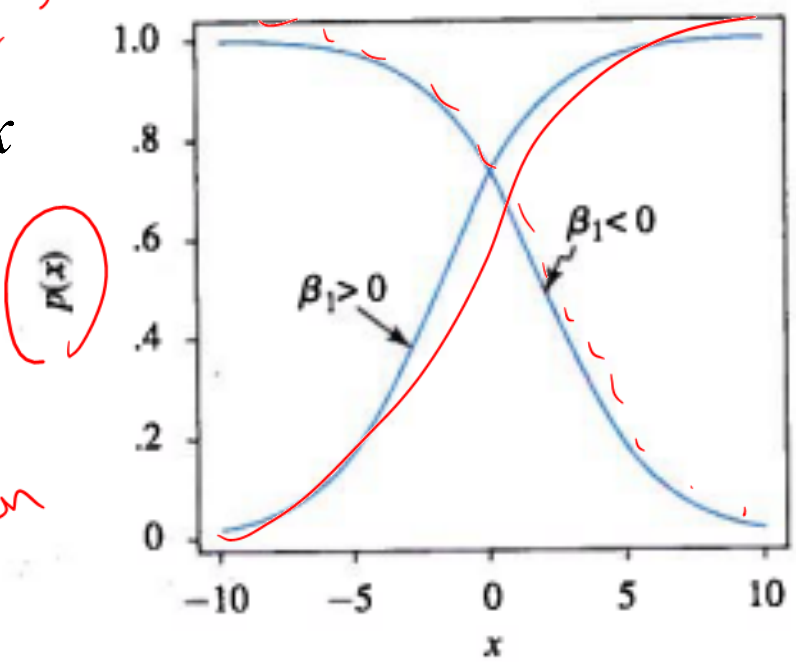

FIGURE 12.5
Logistic regression functions

Parameter Interpretation in Logistic Regression:

$$Odds = \frac{p(x)}{1-p(x)} = \underline{e^{(\beta_0+\beta_1 x)}}$$

Intercept (β_0):

When x=0: $Odds = e^{(\beta_0)}$

So, when x=0 the odds of the event is a function of just β_0.

Slope (β_1):

A 1 unit increase in x gives:

one unit increase

$$Odds = e^{\beta_0+\beta_1(x+1)} = e^{(\beta_0+\beta_1 x)}e^{\beta_1}$$

So, a one unit increase in x multiples the odds by e^{β_1}.

60

Beetle Example: In a pesticide study, approximately sixty beetles were tested at 8 doses of a pesticide. A particular beetle is found to be either dead (success) or alive (failure).

1

X = log (dose) log dose not required
N = number tested at each dose
Y = number that died at each dose (out of N)
p = true probability that an individual beetle will die.

Then Y is binomial (N, p) at each dose and X is a continuous predictor variable.

This is an example of designed experiment with **grouped** data. *

In R, we will use generalized linear model

```
glm(Y ~ X, family=binomial(link="logit"))
```

to fit the logistic regression model.

Beetle Example:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -59.282      4.995  -11.87   <2e-16 ***
lgdose        33.519      2.814   11.91   <2e-16 ***
```

- $\hat{\beta}_0$

- $\hat{\beta}_1$

Estimated slope is 33.519.

*Interpretation: A one unit increase in lgdose (x) multiplies the predicted odds of death by $e^{33.519} = 3.6072\text{x}10^{14}$!

Note that the range of lgdose is only 1.69 – 1.88 for this study!

range < .2

Consider a smaller increase (0.1): $e^{33.519(0.1)} = 28.5$

Another smaller increase (0.01): $e^{33.519(0.01)} = 1.4$

A confidence interval for the odds ratio is found by exponentiating the confidence interval for the slope.

95% CI for the “slope”: (28.27950, 39.34206)

default

95% CI for the odds ratio: (1.9126x10^{12}, 1.2190x10^{17})

$$\left(e^{28.28}, e^{39.34}\right)$$

Notes about logistic regression:

1. Note that the responses may or may not be grouped.* The glm() function can handle both formats, but be careful about formatting!

2. Parameter estimation is based on the “maximum likelihood” (ML) method, in which the parameters are estimated by the values of β_0 and β_1 that maximize the probability of observing the data that you actually did observe. This probability is calculated based on the distribution of Y (which in this case is binomial).

3. Finding the ML estimates is an iterative procedure, but is better behaved than nonlinear regression, and doesn’t require that you provide starting values.

4. The estimated response can be plotted by putting the estimates into the equation. (Works here because we have single predictor and grouped data.)

5. Approximate tests of parameters are given in the output.

6. Approximate CIs can be found using the `confint()` function. Exponentiate bounds to report results on odds ratio scale.

7. The `glm()` function accepts categorical and continuous predictor variables.

8. The `step()` function can be used to perform stepwise logistic regression based on AIC criteria. We can also use `dredge()` from `MuMIn` for best subsets selection.

9. We can calculate "pseudo" R^2 value for logistic regression models. See Beetles example for calculation of McFadden's pseudo R2. Be aware that there are several different definitions of the pseudo R^2 for logistic regression and they can arrive at different values

10. The `dose.p()` function from the `MASS` library, can be used to find the log dose that is required to achieve a given percentage mortality and construct confidence intervals. Example: LD_{50} is the log dose required to achieve 50% mortality.

2. Logistic Regression with Multiple Predictor Variables

Birdkeeping Example (Ramsey and Shafer): A retrospective case-control study of the relationship between lung cancer and birdkeeping (The Hague, Netherlands, 1973-81). 49 lung cancer patients, age < 65 identified and 98 controls obtained from the same area and general age range.

LC = lung cancer (LungCancer, NoCancer)
Sex = patient sex (M, F)
SS = socioeconomic status (High, Low)
BK = birdkeeping (Bird, NoBird)
Age = age (in years)
YR = Years smoked
CD = cigarettes/day

Birdkeeping Model Selection:

1. step() to run stepwise selection based on AIC.
2. dredge() from MuMIn for best subsets selection based on AIC.

For this example, the same model is selected by either approach. The "final" model includes BK and YR.

Why is AGE not in this model? Since people usually start smoking in their teens, "years smoking" contains information about age among those that smoke.

Some researchers like to find the best model based on background variables, without the variable of interest (BK), then add the variable of interest to the best model and test for significance.

Birdkeeping Results (Model1):

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.70460    0.56267  -3.030 0.002450
BKNoBird    -1.47555    0.39588  -3.727 0.000194
YR           0.05825    0.01685   3.458 0.000544
```

Recall that when a categorical predictor (in this case BK) is included in the model, R creates indicator variables. By default the first level (in this case Bird) is used as a baseline or reference.

This is not necessarily a problem, but it might be more convenient to have the "NoBird" group act as baseline. Two options:

1. Stay with default, but invert the odds ratio and corresponding CI.
2. Reorder the factor levels and refit the model.

Birdkeeping Results (Model2 after reordering factor levels):

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.18016    0.63640  -4.997 5.82e-07
BKBird       1.47555    0.39588   3.727 0.000194
YR           0.05825    0.01685   3.458 0.000544
```

Comparing the output for the 2 models (before and after reordering factor levels for BK). While the Intercept and BK estimates are different, the AIC and tests corresponding BK and YR are the same. They are the same model!

For parameter interpretation on next slide, we will work with Model2.

Birdkeeping Parameter Interpretation:

1. For each additional year of smoking, the odds of getting lung cancer are multiplied by:

$$e^{0.0582} = 1.06 \qquad (1.027,\ 1.098)\ 95\%\ \text{C.I.}$$

within each birdkeeping group (other variables, including age, which are not in the model are not held constant).

2. The odds of birdkeepers getting lung cancer are multiplied by:

$$e^{1.4756} = 4.37 \qquad (2.05,\ 9.75)\ \ 95\%\ \text{C.I.}$$

compared to non-birdkeepers, number of years smoking held constant.

Pairwise Comparisons using emmeans

In many cases, we might be interested in calculating emmeans and/or running pairwise comparisons on a categorical predictor.

This can be done using `emmeans()` from `emmeans` package. By default, the emmeans will be returned on the logit scale (difficult to interpret). But with the `type = “response”` option, we get probabilities and odds ratios (easier to interpret).

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$$

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

Hosmer-Lemeshow Lack of Fit test for Logistic Regression:

H_0: The model fits. vs H_A: The model does not fit.

In R, this test can be done with `hoslem.test()` from the `ResourceSelection` package. R versions do not seem to work correctly with grouped data.

The observations are divided into a number of groups of roughly the same size based on the percentiles of the estimated probabilities. The user needs to specify the number of groups.

The lack of fit statistic is compared to the chi-squared distribution.

For Birdkeeping (p-value = 0.4792) the models appear to fit.

What does it mean if we conclude the model doesn’t fit the data?
One possible explanation is that logistic regression isn’t the right model form for the data.

Another possibility is that an “**overdispersed**” model (or a model allowing different variance) may be more appropriate. Recall that for a Binomial RV Y, Var(Y) = Np(1-p).
If Var(Y) = kNp(1-p) for k>1, then it is “overdispersed”, possibly indicating dependent events due to clustering or batch effects.

In R the `summary(, dispersion = )` option provides ways to estimate and correct for the overdispersion factor. If this option is used, the estimated parameters stay the same, but the SE (and CIs) will change.

3. Generalized Linear Models

The normal and logistic regression models we have considered are special cases of a more general structure: the "Generalized Linear Model". Such models are "generalized" in two *senses:*

1) **Distribution of Y:** So far, we have assumed that Y is Normal or Binomial. These are special cases in what is called the "Exponential Family" of distributions that includes:

a) Poisson
b) Negative Binomial
c) Gamma

Generalized Linear Models allow Y to have a distribution in any one of these groups (and more).

2) **Relationship between the mean of Y and the prediction equation (Link function):**

In linear regression the mean of Y is related to the prediction equation by the "identity" function:

$$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{ik} x_{ik}$$

In logistic regression the mean of Y is related to the prediction equation by the "logit" function (logit=log(odds)):

$E(Y_i) = np_i$ (ignore the *n*, because it is a const.)

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{ik} x_{ik}$$

Generalized Linear Models allow us to pick any of the exponential family **distributions** and relate the mean of Y to the prediction equation via a **link** function (monotonic function).

In R, generalized linear models are fit using the `glm( )` function. The form of the `glm` function is

glm(*formula***, family=***familytype***(link=***linkfunction***), data=)**

Default links for each family are shown below but other links are allowed. See `help(family)` for other allowable link functions for each family.

Family	**Default Link Function**
binomial	(link = "logit")
gaussian	(link = "identity")
Gamma	(link = "inverse")
inverse.gaussian	(link = "1/mu^2")
poisson	(link = "log")

Elephant Example: Poisson Regression

Age and Mating Success of Male Elephants (Ramsey and Schafer)

41 male elephants in Amboseli National Park, Kenya

Y_i = number of successful matings (ith elephant)

X_i = Age

The original analysis suggested the following generalized linear model: Y_i has a Poisson distribution with mean μ_i , where

$$\log(\mu_i) = \beta_0 + \beta_1 x_i$$

Here we consider the log link used by the study authors, and also consider the identity link as an alternative.
The function that transforms the mean of Y into a linear regression model is called the “link” function, because it links the mean of Y to the linear prediction model.

Elephant Analysis Strategy

1. Simple linear regression model.
2. Simple linear regression after square root transformation on matings.
3. Generalized linear model with Poisson Y’s and an identity link.
4. Generalized linear model with Poisson Y’s and a log link.
5. We will compare models 3 and 4 (Poisson regressions identity and log links) using AIC.

Elephant Analysis Results:

1. Simple linear regression shows approximate linearity, but a clear trend toward variance increasing with predicted value.
2. A sqrt(Y) transformation solves the problem of increasing variance. However, there is now an outlier and perhaps some evidence of curvature.
3. The Poisson regression with identity link has AIC= 155.50.
4. The Poisson regression with log link has AIC = 156.46.

Conclusion: The model with the identity link has a slight edge, but the choice is far from clear. See plots.

Parameter Interpretation for Elephants Analysis:

Identity link:

The slope parameter (0.2018) is interpreted in the usual way: The average number of successful matings increases by 0.2018 for each additional year of age.

Log link:

The slope parameter (0.0687) is interpreted multiplicatively:

$$\log(\mu_i) = \beta_0 + \beta_1(x_{i1} + 1)$$
$$\mu_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_1) = \exp(\beta_0)\exp(\beta_1 x_{i1})\exp(\beta_1)$$

The average number of successful matings is multiplied by $e^{0.0687}$ = 1.071, for each additional year of age.

Comments about Likelihood, AIC, and "Deviance":

$$AIC = -2\log(likelihood) + 2p$$

1. "Likelihood" is the probability of observing the data that was actually observed. The data are considered fixed, and the likelihood is thought of as a function of the parameters.

2. Since likelihood is a probability, it depends on which distribution is assumed for the data. Normal, Binomial, Poisson, etc. distributions will all have different formulas for the likelihood and AIC.

3. Constants that do not involve the parameters are usually dropped.

4. The "deviance" is defined relative to a "saturated" model.

$$Deviance = -2\log(L_1) - [-2\log(L_0)]$$

Goodness of Fit and Overdispersion:

For the Poisson regression (without random effects), the (Pearson chi-square)/ DF value measures the goodness of the model fit. This info is not part of the default output.

This value should be close to 1. A value greater than 1 might indicate overdispersion.

For the elephant example, with either the identity or log link the value is 1.13 or 1.16.

Recall that for a Poisson RV Y, Var(Y) = E(Y) = μ.
If Var(Y) = kμ for $k>1$, then it is overdispersed, possibly indicating clustering of events.

In R the `summary(, dispersion = )` option provides ways to estimate and correct for the overdispersion factor. If this option is used, the estimated parameters stay the same, but the SE (and CIs) will change.

Another (related) alternative when we are modeling count data would be to use the Negative Binomial distribution instead of Poisson. The negative binomial has a more flexible variance.

Negative binomial distribution is not available directly from `glm()` but can be found in the `MASS` package.