IE 330 Logistic Regression

(C) Seifu Chonde

Binary Logistic Regression is a special type of regression where binary response variable is related to a set of explanatory variables, which can be discrete and/or continuous.

The important point here to note is that in linear regression, the expected values of the response variable are modeled based on combination of values taken by the predictors. In logistic regression **Probability** or **Odds** of the response taking a particular value is modeled based on combination of values taken by the predictors. Like regression , we make an explicit distinction between a response variable and one or more predictor (explanatory) variables.

Logistic regression is applicable, for example, if:

we want to model the probabilities of a response variable as a function of some explanatory variables, e.g. "success" of admission as a function of gender.

we want to perform descriptive discriminate analyses such as describing the differences between individuals in separate groups as a function of explanatory variables, e.g. student admitted and rejected as a function of gender

we want to predict probabilities that individuals fall into two categories of the binary response as a function of some explanatory variables, e.g. what is the probability that a student is admitted given she is a female

we want to classify individuals into two categories based on explanatory variables, e.g. classify new students into "admitted" or "rejected" group depending on their gender.

These ideas boil down to when the response variable is nominal, ordinal or binary we use Logistic Regression. Examples of these variables are grouped variables like Age or Yes/No responses from the course project.

Recall simple linear regression:

Objective: model the expected value of a continuous variable, Y, as a linear function of the continuous predictor, X, $E(Yi) = \beta_0 + \beta_1 x_i$

Model structure: $Y_i = \beta_0 + \beta_1 x_i + e_i$

Model assumptions: Y is iid as normal, errors are normally distributed, $e_i \sim N(0, \sigma^2)$, and independent, and X is fixed, and constant variance ??2.

Parameter estimates and interpretation: $\hat{\beta}_0$ is estimate of β_0 or the intercept, and $\hat{\beta}_1$ is estimate of the slope, etc... Do you recall, what is the interpretation of the intercept and the slope?

Model fit: R^2, residual analysis, F-statistic

Now lets consider binary logistic regression: * Objective: model the log odds of probability of "success" as a fuction of explanatory variable(s) * Model structure: $log(\frac{\pi}{(1-\pi)}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$ * Model assumptions: Y is iid as exponential (binomial, Poisson, multinomial, normal, etc), errors are independent but NOT necessarily normal, uses Maximum Likelihood Estimation vice OLS (large-sample size required), heuristic rule that not more than 20% of counts can have less than 5 (see example) * Parameter estimates and interpretation: $\hat{\beta}_0$ is estimate of probability of a success when all variables are at their "zero"" level, $\hat{\beta}_i$ is estimate of association between probability of a success and variable x_i at a "non-zero"" level. * Model fit: not covered in this class

Now let's consider our first example of logistic regression. Below are data on the relationship between the proportion of male turtles and incubation temperature for eggs from New Mexico. We are interested in determining if there exists a significant relationship between the sex of the turtle and the incubation temperature.

```
# A simple example of logistic regression

# Our first step is to input the data into columns
temp <- c(27.2, 28.3, 29.9)
male <- c(0, 8, 8)
female <- c(10, 4, 2)

# Now we make a dataframe of our columns
example.data <- as.data.frame(cbind(temp, male, female))

# This can be visualized as
example.data
```

```
##   temp male female
## 1 27.2    0     10
## 2 28.3    8      4
## 3 29.9    8      2
```

```
summary(example.data)
```

```
##       temp           male          female
##  Min.   :27.2   Min.   :0.00   Min.   : 2.00
##  1st Qu.:27.8   1st Qu.:4.00   1st Qu.: 3.00
##  Median :28.3   Median :8.00   Median : 4.00
##  Mean   :28.5   Mean   :5.33   Mean   : 5.33
##  3rd Qu.:29.1   3rd Qu.:8.00   3rd Qu.: 7.00
##  Max.   :29.9   Max.   :8.00   Max.   :10.00
```

To run the binary logistic regression here we make use of the glm command from R. The glm command takes a matrix of success/failures and regresses it on a set of variables. Additionally, one may specify the data and the type of glm (logistic regression is in the binomial family).

```
# General Linear Model -- Logistic Regression Syntax glm (
# cbind(successColumn, failureColumn) ~ variable1 + ... + variableN, data =
# data.source, family = binomial)

# Using our data and saving the model to the variable my.log.regression:
my.log.regression <- glm(cbind(male, female) ~ temp, data = example.data, family = binomial)

# print the regression model
my.log.regression
```

```
##
## Call:  glm(formula = cbind(male, female) ~ temp, family = binomial,
##     data = example.data)
##
## Coefficients:
## (Intercept)         temp
##      -41.74         1.47
##
## Degrees of Freedom: 2 Total (i.e. Null);  1 Residual
## Null Deviance:       19.1
## Residual Deviance: 6.08  AIC: 15.3
```

```
# print a summary of the regression model indicating our statistically
# significant predictors
summary(my.log.regression)
```

```
##
## Call:
## glm(formula = cbind(male, female) ~ temp, family = binomial,
##     data = example.data)
##
## Deviance Residuals:
##      1       2       3
## -1.774   1.428  -0.944
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -41.742     14.520   -2.87   0.0040 **
## temp           1.470      0.513    2.87   0.0041 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 19.0770  on 2  degrees of freedom
## Residual deviance:  6.0787  on 1  degrees of freedom
## AIC: 15.34
##
## Number of Fisher Scoring iterations: 5
```

This tells us that our regression model is: $log(\frac{P(male)}{Prob(female)}) = -41.74 + 1.47 * temp$
As we see from the double stars (or from the p-values) our intercept and temperature are statistically significant predictors of the probability (odds) of a turtle being a male.

Now let's plot what this is looking like:

```
# plot temeprature vs. the probability of having a male
plot(temp, male/(male + female), xlab = "Temperature", ylab = "Probability of Male",
    ylim = c(0, 1))

# add the logistic regression curve to previous plot
curve(predict(my.log.regression, data.frame(temp = x), type = "resp"), add = TRUE)

# add points of data to logistic regression to previous plot
points(temp, fitted(my.log.regression), pch = 20)
```

Finally, let's do some predictions

1. At what temperature do you expect a 50:50 split of males and females?
2. What is the probability that a male hatches from an incubated egg at 27?

```
# Answer to Question 1 log(0.5/0.5) = -41.74 + 1.47*temp temp = 28.4
(log(0.5/0.5) + 41.74)/1.47
```

```
## [1] 28.39
```

```
# Answer to Question 2 log(P(male)/(1-P(male))) = -41.74 + 1.47*(27) this is
# equivalent to the following with some rounding errors P(male) = exp(-41.74
# + 1.47*27) / ( 1 + exp(-41.74 + 1.47*27) )
exp(-41.74 + 1.47 * 27)/(1 + exp(-41.74 + 1.47 * 27))
```

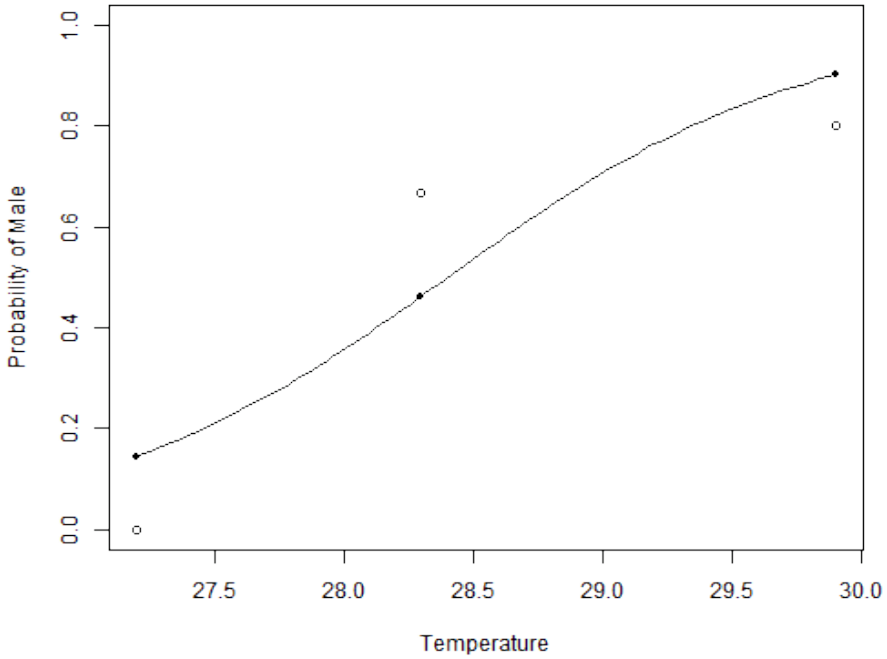

Figure 1: plot of chunk unnamed-chunk-3

```
## [1] 0.1141
```

```
# also this can be more accurately calculated from
odds <- exp(predict(my.log.regression, data.frame(temp = 27)))
odds/(1 + odds)
```

```
##      1
## 0.1127
```