R10 - Logistic Regression

HCI/PSYCH 522
Iowa State University

April 14, 2022

Overview

- Individual data
  - Admission as a function of GRE and GPA
- Grouped data
  - Effect of moth color and distance on predation
  - $+$ interaction between color and distance

Logistic regression model

For observation i, let

- Y_i be the indicator of success and
- $X_{i,p}$ be the value of the pth independent variable.

The (simple) logistic regression model is

$$Y_i \stackrel{ind}{\sim} Ber(\theta_i) \quad \text{where} \quad = \log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}$$

In this model,

- e^{β_0} is the odds when all independent variables are zero and
- $100(e^{\beta_p} - 1)$ is the percent increase in the odds $\left(\frac{\theta}{1-\theta}\right)$ of success when the pth independent variable increases by 1 holding other independent variables constant.

Admission

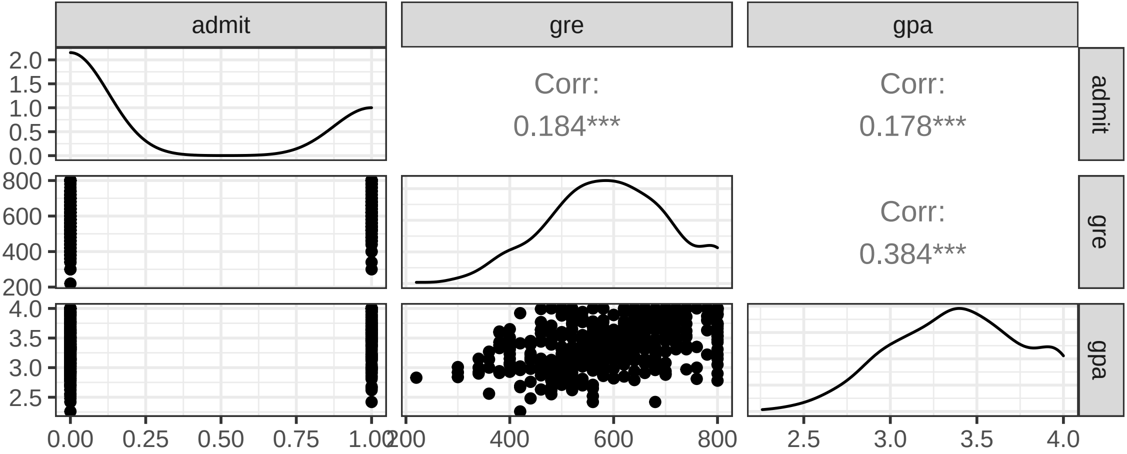
```
admission <- read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv") %>% select(-rank)
head(admission)

##   admit gre  gpa
## 1     0 380 3.61
## 2     1 660 3.67
## 3     1 800 4.00
## 4     1 640 3.19
## 5     0 520 2.93
## 6     1 760 3.00

summary(admission)

##      admit             gre             gpa
##  Min.   :0.0000   Min.   :220.0   Min.   :2.260
##  1st Qu.:0.0000   1st Qu.:520.0   1st Qu.:3.130
##  Median :0.0000   Median :580.0   Median :3.395
##  Mean   :0.3175   Mean   :587.7   Mean   :3.390
##  3rd Qu.:1.0000   3rd Qu.:660.0   3rd Qu.:3.670
##  Max.   :1.0000   Max.   :800.0   Max.   :4.000
```

Admission



admit
gre
gpa
Corr:
0.184***
Corr:
0.178***
Corr:
0.384***
admit
gre
gpa
2.0
1.5
1.0
0.5
0.0
800
600
400
200
4.0
3.5
3.0
2.5
0.00
0.25
0.50
0.75
1.00
200
400
600
800
2.5
3.0
3.5
4.0

Admission

Here's code for a 3d interactive graphic. Unfortunately I can't figure out how to include it in the pdf.

```
plot_ly(admission, x = ~gre, y = ~gpa, z = ~admit, color = ~rank)
```

Admission

```
m <- glm(admit ~ I(gre-580) + I(gpa-3.4), data = admission, family = binomial)
summary(m)

##
## Call:
## glm(formula = admit ~ I(gre - 580) + I(gpa - 3.4), family = binomial,
##     data = admission)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2730  -0.8988  -0.7206   1.3013   2.0620
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.822846   0.112926  -7.287 3.18e-13 ***
## I(gre - 580)  0.002691   0.001057   2.544   0.0109 *
## I(gpa - 3.4)  0.754687   0.319586   2.361   0.0182 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Admission as a function of GRE

```
1/(1+exp(-coef(m)[1]))      # probability of acceptance with GRE 580 and GPA 3.4

## (Intercept)
##   0.3051598

1/(1+exp(-confint(m)[1,]))

##     2.5 %    97.5 %
## 0.2595473 0.3531931

100*(exp(coef(m)[-1])-1)

## I(gre - 580) I(gpa - 3.4)
##    0.2694307  112.6945379

100*(exp(confint(m)[-1,])-1)

##                   2.5 %      97.5 %
## I(gre - 580)  0.0637599   0.4803211
## I(gpa - 3.4) 14.4251749 301.5376560
```

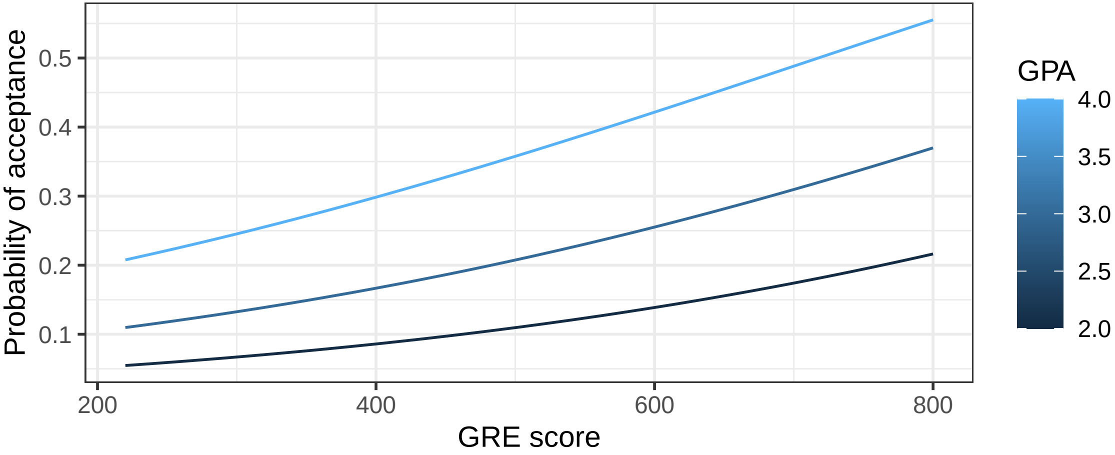
Interpretation

- With a GRE of 580 and GPA of 3.4, the probability of acceptance is 31% (26, 35).
- After adjusting for GPA, each 1 point increase in GRE score is associated with a 0.27% (0.06, 0.48) increase in the odds of acceptance.
- After adjusting for GPA, each 100 point increase in GRE score is associated with a 31% (7, 61) increase in the odds of acceptance.
- After adjusting for GRE, each 1 point increase in GPA score is associated with a 113% (14, 302) increase in the odds of acceptance.

Admission as a function of GRE

```
nd <- expand.grid(gre = seq(220,800,length=101), gpa = 2:4)
nd$p <- predict(m, newdata = nd, type="response")
ggplot(nd, aes(x = gre, y = p, color = gpa, group = gpa)) +
  geom_line() +
  labs(x = "GRE score", y = "Probability of acceptance", color = "GPA")
```

Admission as a function of GRE



Grouped data

If the data are grouped, then the analysis is basically the same, but the mathematics and code look a bit different.

```
Sleuth3::case2102
```

```
##    Morph Distance Placed Removed
## 1  light      0.0     56      17
## 2   dark      0.0     56      14
## 3  light      7.2     80      28
## 4   dark      7.2     80      20
## 5  light     24.1     52      18
## 6   dark     24.1     52      22
## 7  light     30.2     60       9
## 8   dark     30.2     60      16
## 9  light     36.4     60      16
## 10  dark     36.4     60      23
## 11 light     41.5     84      20
## 12  dark     41.5     84      40
## 13 light     51.2     92      24
## 14  dark     51.2     92      39
```

Logistic regression model

For group g, let

- n_g be the number of individuals in the group ,
- Y_g be the indicator of success, and
- X_g be the value of an independent variable associated with group g.

The (simple) logistic regression model is

$$Y_g \stackrel{ind}{\sim} Bin(n_g, \theta_g) \quad \text{where} \quad \log\left(\frac{\theta_g}{1-\theta_g}\right) = \beta_0 + \beta_1 X_{g,1} + \cdots + \beta_p X_{g,p}$$

In this model,

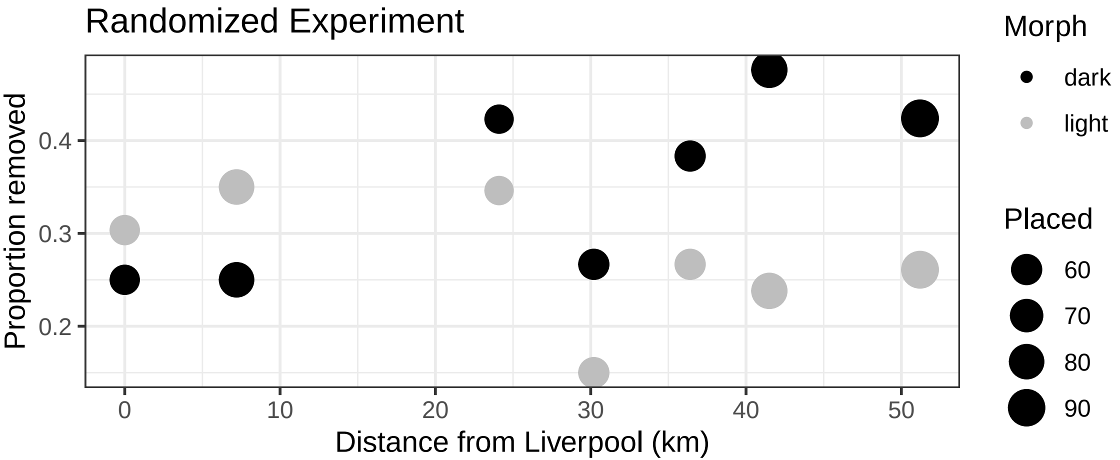
- e^{β_0} is the odds when all independent variables are zero and
- $100(e^{\beta_p} - 1)$ is the percent increase in the odds $\left(\frac{\theta}{1-\theta}\right)$ of success when the pth independent variable increases by 1 holding other independent variables constant.

Natural selection

```
Sleuth3::case2102

##    Morph Distance Placed Removed
## 1  light      0.0     56      17
## 2   dark      0.0     56      14
## 3  light      7.2     80      28
## 4   dark      7.2     80      20
## 5  light     24.1     52      18
## 6   dark     24.1     52      22
## 7  light     30.2     60       9
## 8   dark     30.2     60      16
## 9  light     36.4     60      16
## 10  dark     36.4     60      23
## 11 light     41.5     84      20
## 12  dark     41.5     84      40
## 13 light     51.2     92      24
## 14  dark     51.2     92      39
```

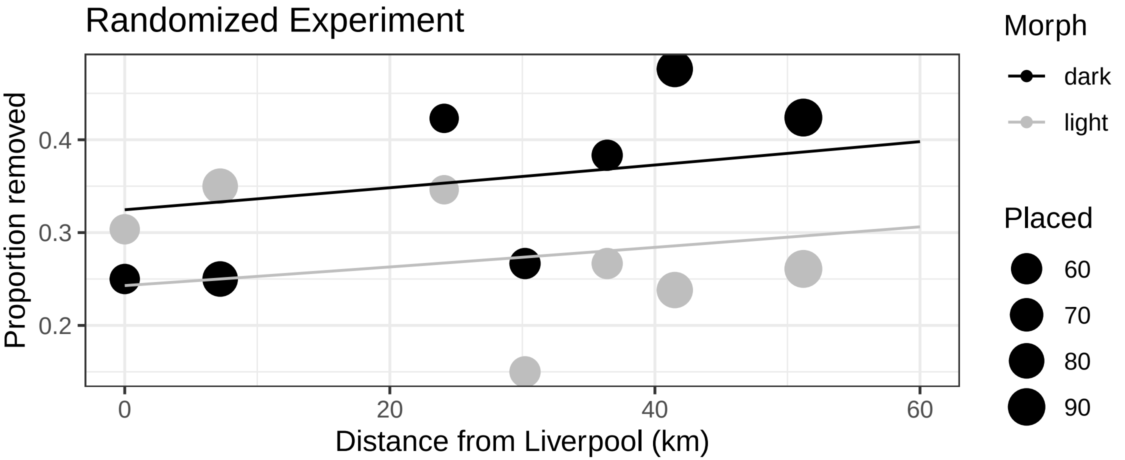
Natural selection



Logistic regression model for proportion removed

```
m <- glm(cbind(Removed, Placed - Removed) ~ Distance + Morph,
         data = case2102, family = binomial)
summary(m)

##
## Call:
## glm(formula = cbind(Removed, Placed - Removed) ~ Distance + Morph,
##     family = binomial, data = case2102)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.28292  -1.16122   0.00237   1.03757   1.98945
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.732690   0.151221  -4.845 1.27e-06 ***
## Distance     0.005314   0.004002   1.328  0.18422
## Morphlight  -0.404052   0.139377  -2.899  0.00374 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

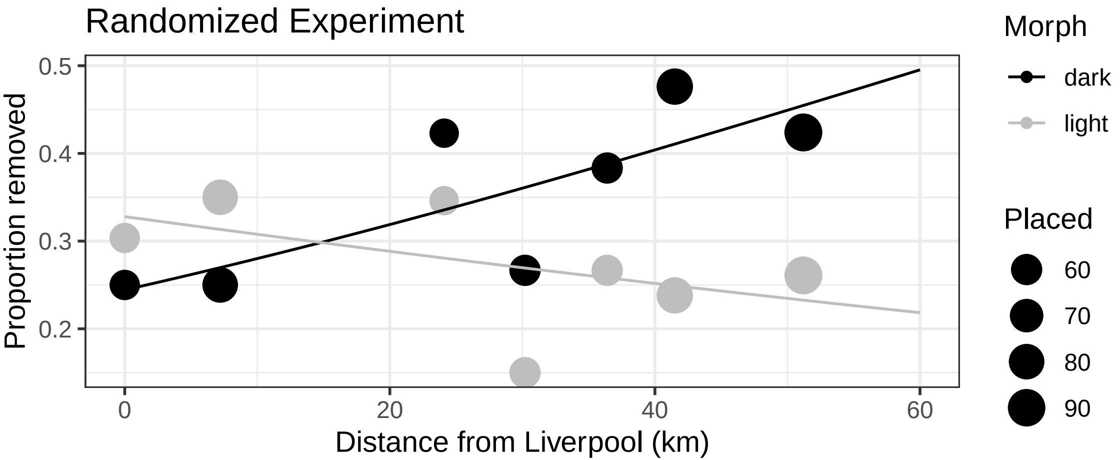

Randomized Experiment
Proportion removed
0.2
0.3
0.4
0
20
40
60
Distance from Liverpool (km)
Morph
dark
light
Placed
60
70
80
90

Logistic regression model for proportion removed

```
m <- glm(cbind(Removed, Placed - Removed) ~ Distance + Morph + Distance:Morph,
         data = case2102, family = binomial)
summary(m)

##
## Call:
## glm(formula = cbind(Removed, Placed - Removed) ~ Distance + Morph +
##     Distance:Morph, family = binomial, data = case2102)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.21183  -0.39883   0.01155   0.68292   1.31242
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -1.128987   0.197906  -5.705 1.17e-08 ***
## Distance            0.018502   0.005645   3.277 0.001048 **
## Morphlight          0.411257   0.274490   1.498 0.134066
## Distance:Morphlight -0.027789   0.008085  -3.437 0.000588 ***
## ---
```

Plot with fitted lines



emtrends

```
em <- emmeans(m, ~ Morph, at = list(Distance = 15))
em_ci <- confint(em, type = "response")
em_ci

##  Morph  prob     SE  df asymp.LCL asymp.UCL
##  dark  0.299 0.0273 Inf     0.248     0.355
##  light 0.298 0.0264 Inf     0.249     0.352
##
## Confidence level used: 0.95
## Intervals are back-transformed from the logit scale
```

```
et <- emtrends(m, ~ Morph, var = "Distance")
et_ci <- confint(et)
et_ci

##  Morph Distance.trend      SE  df asymp.LCL asymp.UCL
##  dark         0.01850 0.00565 Inf   0.00744   0.02957
##  light       -0.00929 0.00579 Inf  -0.02063   0.00206
##
## Confidence level used: 0.95
```

Manuscript statements

- At 15 km from Liverpool, both light and dark morphology had 30% (25, 36) removed.
- For dark morphology, each additional km away from Liverpool resulted in a 1.9% (0.7, 3) percent increase in odds.
- For light morphology, each additional km away from Liverpool resulted in a 0.9% (-0.2, 2.1) percent decrease in odds.

Summary

For binary data or counts with a clear upper maximum, logistic regression is an appropriate model.