R06 - Logistic Regression

HCI/PSYCH 522
Iowa State University

March 31, 2022

Overview

- Individual data
  - Bernoulli distribution
  - Logistic regression model
  - Admission as a function of GRE
- Grouped data
  - Binomial distribution
  - Logistic regression model
  - Probability of staying healthy as a function of Vitamin C intake
- Other examples
  - Probability of extinction as a function of island size
  - Cancer occurrence as a function of breast-feeding
  - Admission as a function of GRE, GPA, and school rank

Bernoulli Distribution

Let Y be a random variable that indicates "success". For example,

- Winning a game
- Having fewer than 3 errors on a task
- Clicking on an ad

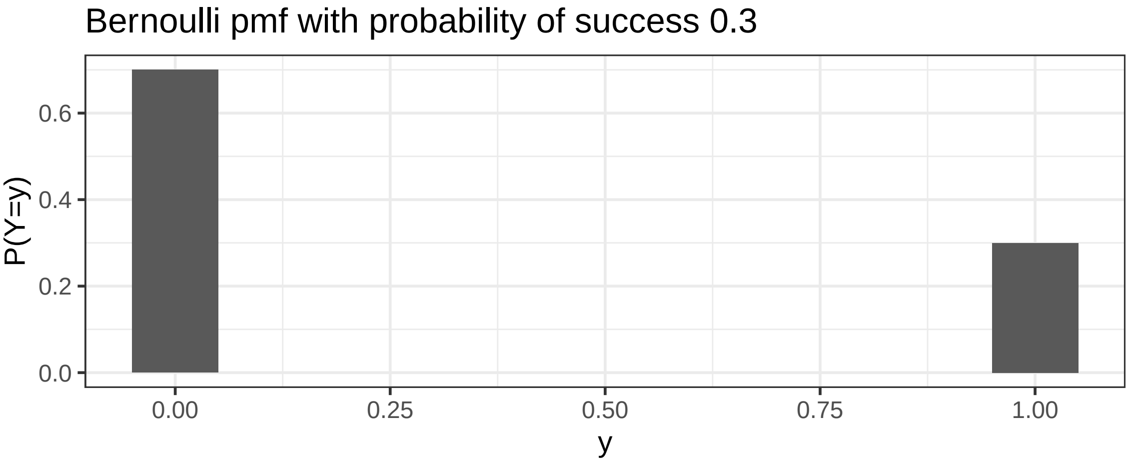
Then Y has a Bernoulli distribution with probability of success $0 < \theta < 1$ and we write $Y \sim Ber(\theta)$. The probability mass function is

$$P(Y = y) = \theta^y (1 - \theta)^{1-y} \quad \text{for } y = 0, 1$$

and we can find that

$$E[Y] = \theta \quad \text{and} \quad Var[Y] = \theta(1 - \theta).$$

Bernoulli pmf



Bernoulli probability of success

Suppose the Bernoulli probability of success changes due to some other variable. For example,

- Time of day
- Sex/gender
- Length of a game

A logistic regression model allows the probability of success to change according to these independent variables.

Logistic regression model

For observation i, let

- Y_i be the indicator of success and
- X_i be the value of an independent variable.

The (simple) logistic regression model is

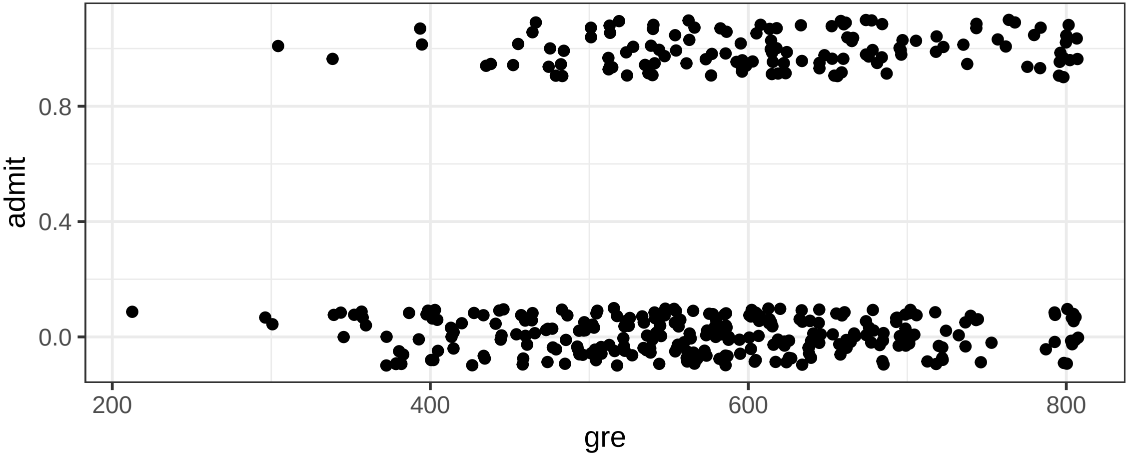
$$Y_i \stackrel{ind}{\sim} Ber(\theta_i) \quad \text{where} \quad \log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_1 X_i$$

In this model, $100 * (e^{\beta_1} - 1)$ is the percent change in the odds $\left(\frac{\theta}{1-\theta}\right)$ of success when the independent variable increases by 1.

Admission as a function of GRE

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

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

summary(admission)

##      admit             gre             gpa             rank      
##  Min.   :0.0000   Min.   :220.0   Min.   :2.260   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:520.0   1st Qu.:3.130   1st Qu.:2.000  
##  Median :0.0000   Median :580.0   Median :3.395   Median :2.000  
##  Mean   :0.3175   Mean   :587.7   Mean   :3.390   Mean   :2.485  
##  3rd Qu.:1.0000   3rd Qu.:660.0   3rd Qu.:3.670   3rd Qu.:3.000  
##  Max.   :1.0000   Max.   :800.0   Max.   :4.000   Max.   :4.000  
```

Admission as a function of GRE



Admission as a function of GRE

```
m <- glm(admit ~ gre, data = admission)
summary(m)

##
## Call:
## glm(formula = admit ~ gre, data = admission)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4755  -0.3415  -0.2522   0.5989   0.8966
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.1198407  0.1190510  -1.007 0.314722
## gre          0.0007442  0.0001988   3.744 0.000208 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.2103746)
##
```

Admission as a function of GRE

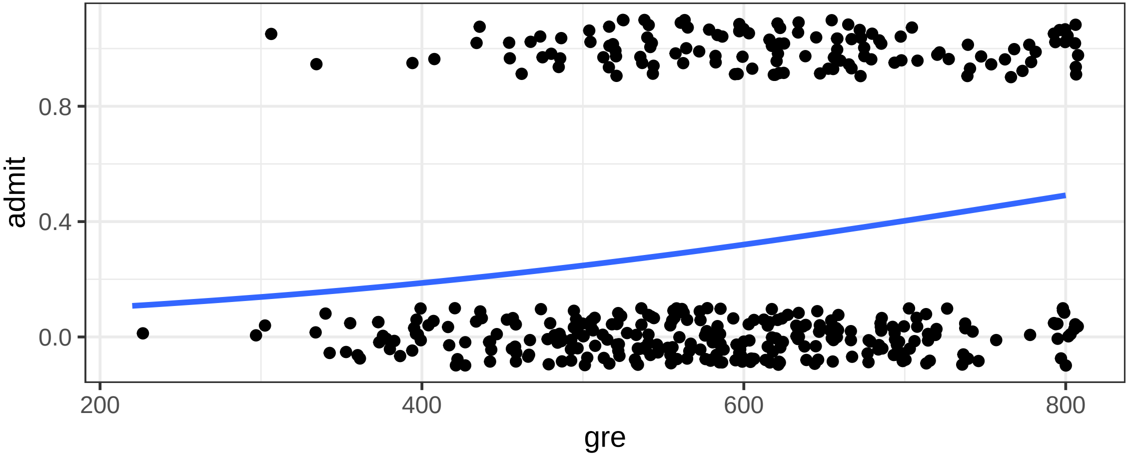
```
ci <- 100*(exp(confint(m)[2,])-1)
ci

##       2.5 %     97.5 %
## 0.03546236 0.11343950
```

For each 1 point increase in GRE score, the percent change in odds of admission is (0.035, 0.113)%.

```
ci <- 100*(exp(10*confint(m)[2,])-1)
ci

##      2.5 %    97.5 %
## 0.3551901 1.1402034
```

For each 10 point increase in GRE score, the percent change in odds of admission is (0.355, 1.14)%.

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.

```
##   Dose Number WithoutIllness ProportionWithout
## 1 0.00   1158            267             0.231
## 2 0.25    331             74             0.224
## 3 1.00    552            130             0.236
## 4 2.00    308             65             0.211
```

Binomial Distribution

Let Y be a random variable the count of the number of "successes" in a group. For example,

- Number of games won
- Number of individuals having 3 or fewer errors on a task
- Number of visitors clicking on an ad

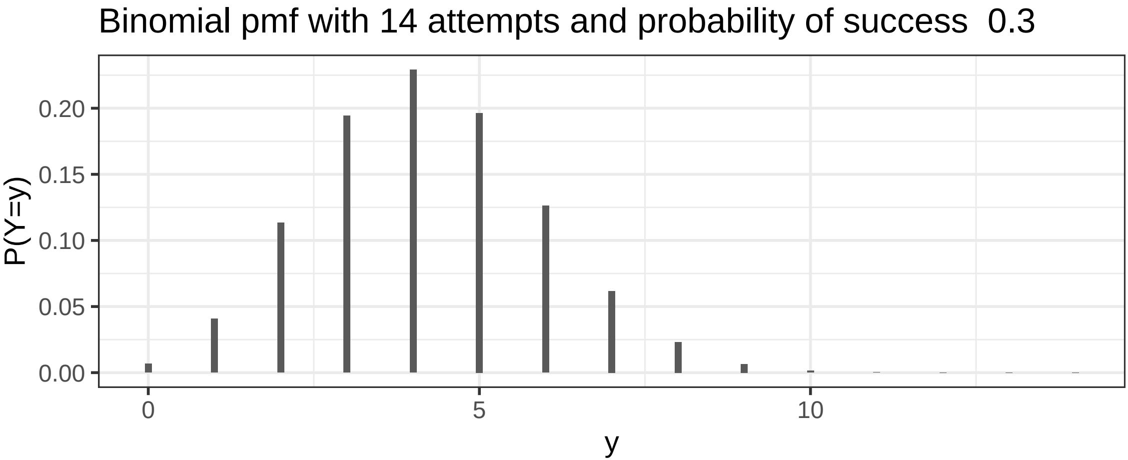
Then Y has a Binomial distribution with number of attempts n and probability of success $0 < \theta < 1$ and we write $Y \sim Bin(n, \theta)$. The probability mass function is

$$P(Y = y) = \binom{n}{y} \theta^y (1-\theta)^{1-y} \quad \text{for } y = 0, 1, \ldots, n$$

and we can find that

$$E[Y] = n\theta \quad \text{and} \quad Var[Y] = n\theta(1-\theta).$$

Binomial pmf



Binomial probability of success

Suppose the probability of success changes due to some other variable: For example,

- Time of day
- Sex/gender
- Length of a game

A logistic regression model allows the probability of success to change according to these independent variables.

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

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

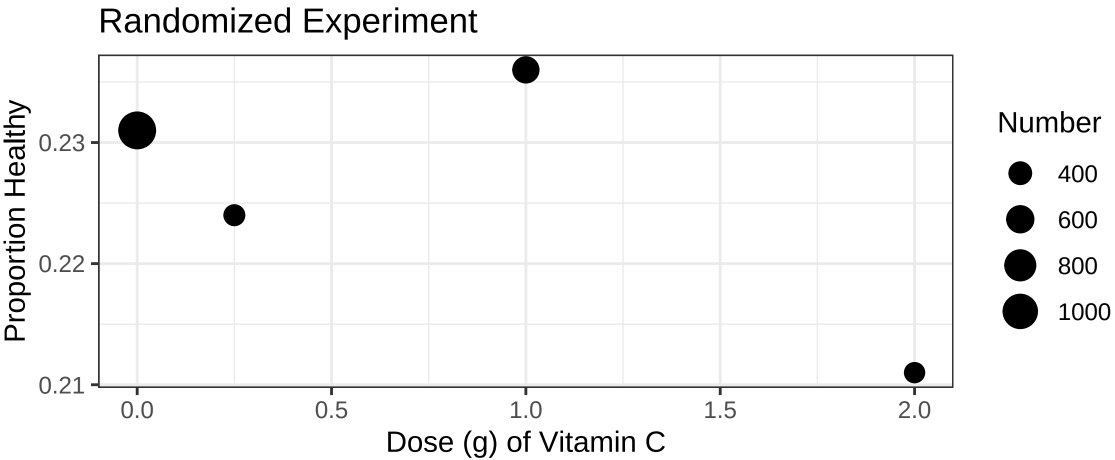
In this model, $100 * (e^{\beta_1} - 1)$ is the percent change in the odds $\left(\frac{\theta}{1-\theta}\right)$ of success when the independent variable increases by 1.

Vitamin C effect on incidence of colds

```
##   Dose Number WithoutIllness ProportionWithout
## 1 0.00   1158            267             0.231
## 2 0.25    331             74             0.224
## 3 1.00    552            130             0.236
## 4 2.00    308             65             0.211
```

Z

Vitamin C effect on incidence of colds



Logistic regression model for proportion healthy

```
m <- glm(cbind(WithoutIllness, Number - WithoutIllness) ~ Dose,
         data = ex2113, family = binomial)
summary(m)

##
## Call:
## glm(formula = cbind(WithoutIllness, Number - WithoutIllness) ~
##     Dose, family = binomial, data = ex2113)
##
## Deviance Residuals:
##        1         2         3         4
## -0.06857  -0.27405   0.57021  -0.35303
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.20031    0.06167 -19.464   <2e-16 ***
## Dose        -0.03465    0.07113  -0.487    0.626
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

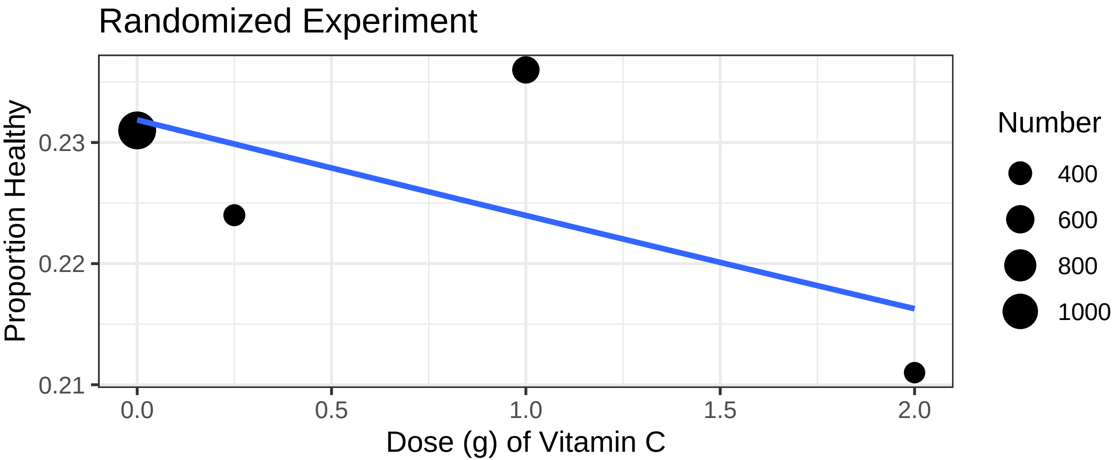
Logistic regression model for proportion healthy

```
ci <- 100*(exp(confint(m)[2,])-1)
ci

##      2.5 %    97.5 %
## -16.09864  10.89977
```

Manuscript statement:

Each gram increase in Vitamin C causes the odds of staying healthy to change by (-16, 11)%.

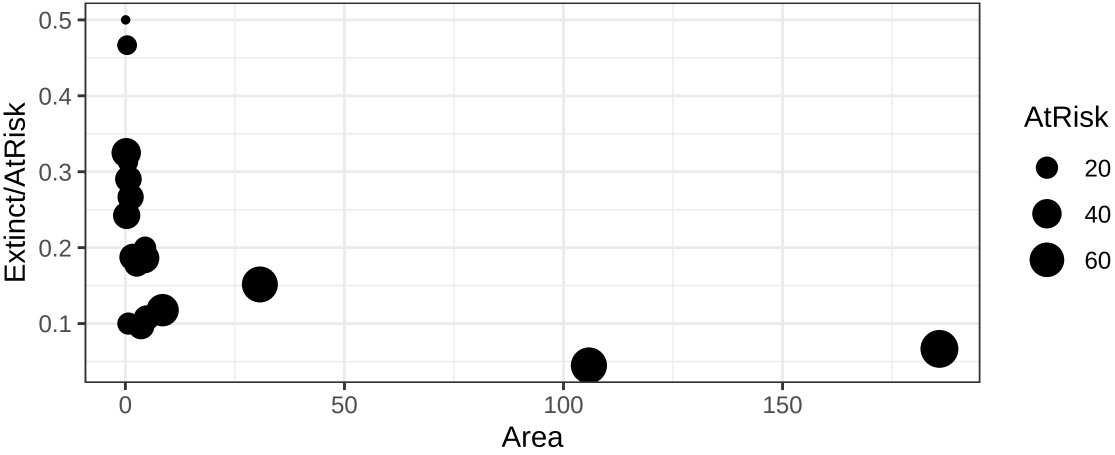
Vitamin C effect on incidence of colds



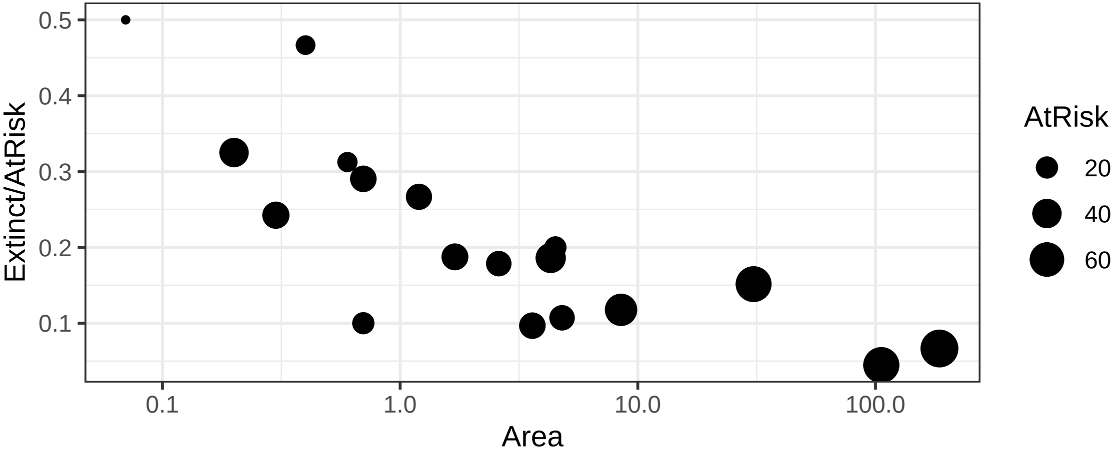
Probability of extinction as a function of island size

```
Sleuth3::case2101

##               Island   Area AtRisk Extinct
## 1         Ulkokrunni 185.80     75       5
## 2          Maakrunni 105.80     67       3
## 3          Ristikari  30.70     66      10
## 4     Isonkivenletto   8.50     51       6
## 5     Hietakraasukka   4.80     28       3
## 6           Kraasukka   4.50     20       4
## 7          Lansiletto   4.30     43       8
## 8         Pihlajakari   3.60     31       3
## 9                Tyni   2.60     28       5
## 10       Tasasenletto   1.70     32       6
## 11             Raiska   1.20     30       8
## 12        Pohjanletto   0.70     20       2
## 13               Toro   0.70     31       9
## 14         Luusiletto   0.60     16       5
## 15       Vatunginletto  0.40     15       7
## 16       Vatunginnokka  0.30     33       8
## 17           Tiirakari  0.20     40      13
```

Probability of extinction as a function of island size



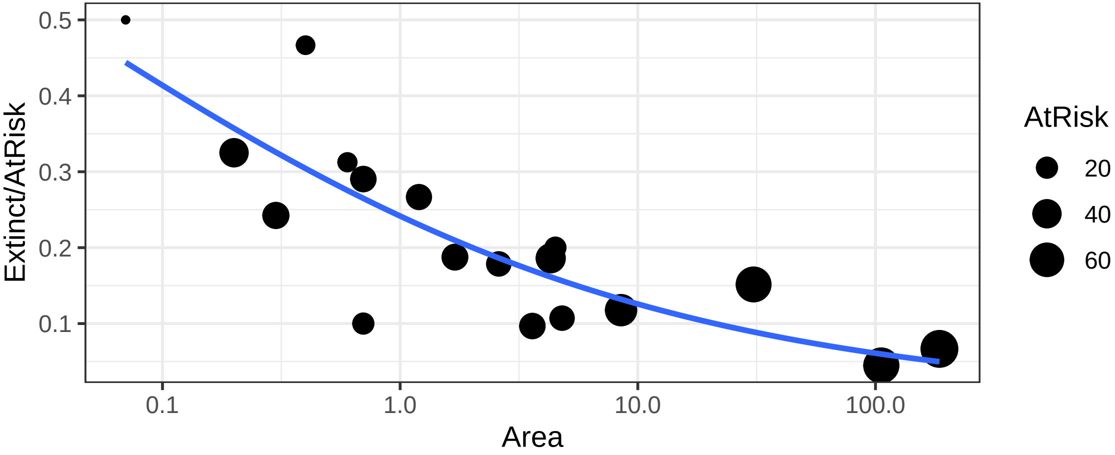
Probability of extinction as a function of island size



Probability of extinction as a function of island size

```
m <- glm(cbind(Extinct, AtRisk - Extinct) ~ Area,
         data = Sleuth3::case2101, family = binomial)
summary(m)

##
## Call:
## glm(formula = cbind(Extinct, AtRisk - Extinct) ~ Area, family = binomial,
##     data = Sleuth3::case2101)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6526  -1.0661  -0.1877   1.0038   2.1860
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.305957   0.117339 -11.130  < 2e-16 ***
## Area        -0.010121   0.002684  -3.771 0.000163 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Probability of extinction as a function of island size



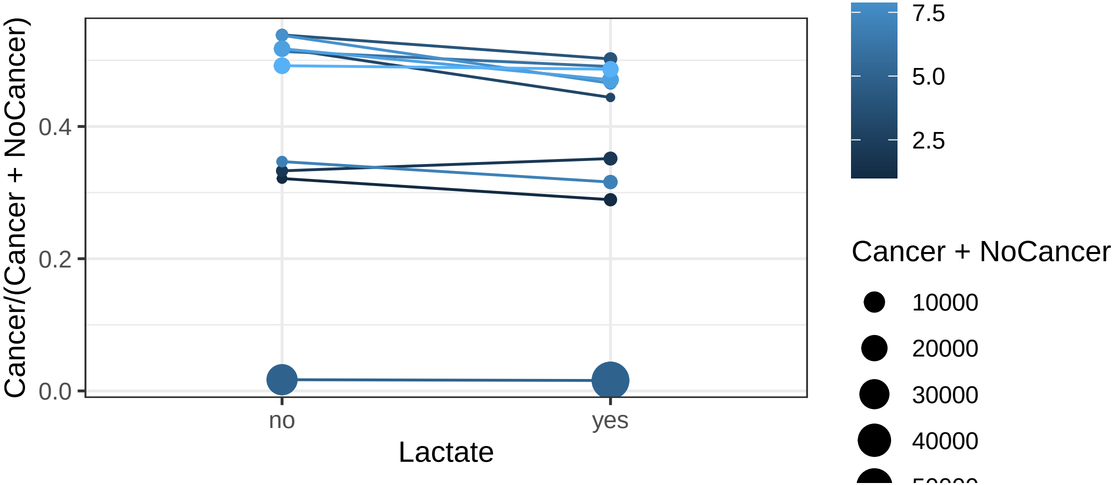
Cancer occurrence as a function of breast-feeding

```
Sleuth3::ex2119 %>%
  filter(Study == 5) %>%
  mutate(p <- Cancer / (Cancer + NoCancer))

##   Study Lactate Cancer NoCancer p <- Cancer/(Cancer + NoCancer)
## 1     5      no    565    32693                      0.01698839
## 2     5     yes    894    55735                      0.01578696
```

Cancer occurrence as a function of breast-feeding

```
m <- glm(cbind(Cancer, NoCancer) ~ Lactate,
         data = Sleuth3::ex2119 %>% filter(Study == 5),
         family = binomial)
summary(m)

##
## Call:
## glm(formula = cbind(Cancer, NoCancer) ~ Lactate, family = binomial,
##     data = Sleuth3::ex2119 %>% filter(Study == 5))
##
## Deviance Residuals:
## [1]  0  0
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.05809    0.04243 -95.637   <2e-16 ***
## Lactateyes  -0.07457    0.05419  -1.376    0.169
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Cancer occurrence as a function of breast-feeding



Cancer occurrence as a function of breast-feeding

```
m <- glm(cbind(Cancer, NoCancer) ~ Lactate + factor(Study),
         data = Sleuth3::ex2119,
         family = binomial)
summary(m)

##
## Call:
## glm(formula = cbind(Cancer, NoCancer) ~ Lactate + factor(Study),
##     family = binomial, data = Sleuth3::ex2119)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.70217  -0.57823  -0.00853   0.47100   1.43668
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.78029    0.05846 -13.348  < 2e-16 ***
## Lactateyes     -0.10943    0.02303  -4.751 2.02e-06 ***
## factor(Study)2  0.21757    0.07050   3.086  0.00203 **
## factor(Study)3  0.77526    0.10433   7.431 1.08e-13 ***
```

Cancer occurrence as a function of breast-feeding

```
library("lme4")
m <- glmer(cbind(Cancer, NoCancer) ~ Lactate + (1|Study),
         data = Sleuth3::ex2119,
         family = binomial)
summary(m)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
##  Family: binomial  ( logit )
## Formula: cbind(Cancer, NoCancer) ~ Lactate + (1 | Study)
##    Data: Sleuth3::ex2119
##
##      AIC      BIC   logLik deviance df.resid
##    248.5    251.5   -121.2    242.5       17
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.69254 -0.60223  0.01219  0.44225  1.44836
##
## Random effects:
##  Groups Name        Variance Std.Dev.
```

Admission as a function of GRE, GPA, and school rank

```
m <- glm(admit ~ gre + gpa + factor(rank),
       data = admission, family = binomial)
summary(m)

##
## Call:
## glm(formula = admit ~ gre + gpa + factor(rank), family = binomial,
##     data = admission)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6268  -0.8662  -0.6388   1.1490   2.0790
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.989979   1.139951  -3.500 0.000465 ***
## gre             0.002264   0.001094   2.070 0.038465 *
## gpa             0.804038   0.331819   2.423 0.015388 *
## factor(rank)2 -0.675443   0.316490  -2.134 0.032829 *
## factor(rank)3 -1.340204   0.345306  -3.881 0.000104 ***
```

Summary

- Logistic regression
  - Dependent variable is a count with clear upper maximum
  - Interpret $100(e^{\beta_1} - 1)$ as the percent change in odds