Set S01 - Logistic Regression

STAT 401 (Engineering) - Iowa State University

April 17, 2017

Linear regression

The linear regression model

$$Y_i \overset{ind}{\sim} N(\mu_i, \sigma^2)$$
$$\mu_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}$$

where

- Y_i is continuous
- X_i is continuous or categorical (indicator variables)

What if Y_i is

- binary or
- a count of the number of success out of some total?

Logistic regression

Let

$$Y_i = \begin{cases} 1 & \text{if observation } i \text{ is a ``success'} \\ 0 & \text{otherwise.} \end{cases}$$

and X_i be an explanatory variable that affects the probability of success θ_i for observation i.

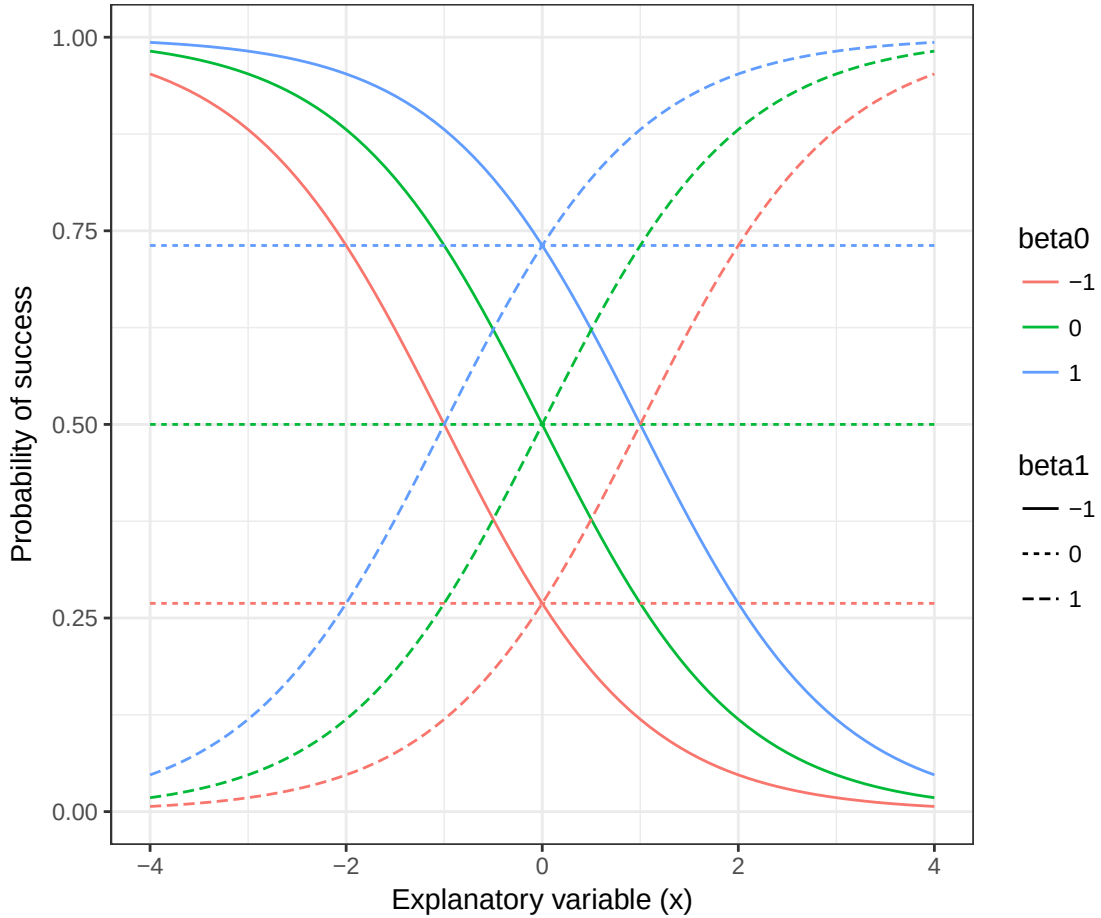
Then a logistic regression model is

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

where the logistic function of X_i is

$$\theta_i = f(X_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_i)}}.$$

```
d <- expand.grid(b0 = c(-1,0,1), b1 = c(-1,0,1), x = seq(-4,4,by=0.1)) %>%
  mutate(theta = 1/(1+exp(-(b0+b1*x))),
         beta0 = as.factor(b0),
         beta1 = as.factor(b1))

ggplot(d, aes(x,theta,color=beta0,linetype=beta1,group=interaction(beta0,beta1))) +
  geom_line() +
  theme_bw() +
  labs(x="Explanatory variable (x)", y="Probability of success")
```



Interpretation

When $X_i = 0$, then

$$E[Y_i|X_i = 0] = \theta_i = \frac{1}{1 + e^{-\beta_0}}$$

thus β_0 determines the probability of success when the explanatory variable is zero.

The odds of success when $X_1 = x$ is

$$\frac{\theta_1}{1 - \theta_1} = e^{\beta_0 + \beta_1 x}.$$

The probability of success when $X_2 = x + 1$ is

$$\frac{\theta_2}{1 - \theta_2} = e^{\beta_0 + \beta_1 (x+1)} = e^{\beta_0 + \beta_1 x + \beta_1}.$$

Thus, the multiplicative change in the odds for a 1 unit increase in x is

$$\frac{\frac{\theta_2}{1-\theta_2}}{\frac{\theta_1}{1-\theta_1}} = \frac{e^{\beta_0 + \beta_1 x + \beta_1}}{e^{\beta_0 + \beta_1 x}} = e^{\beta_1}$$

This is also referred to as an odds ratio.

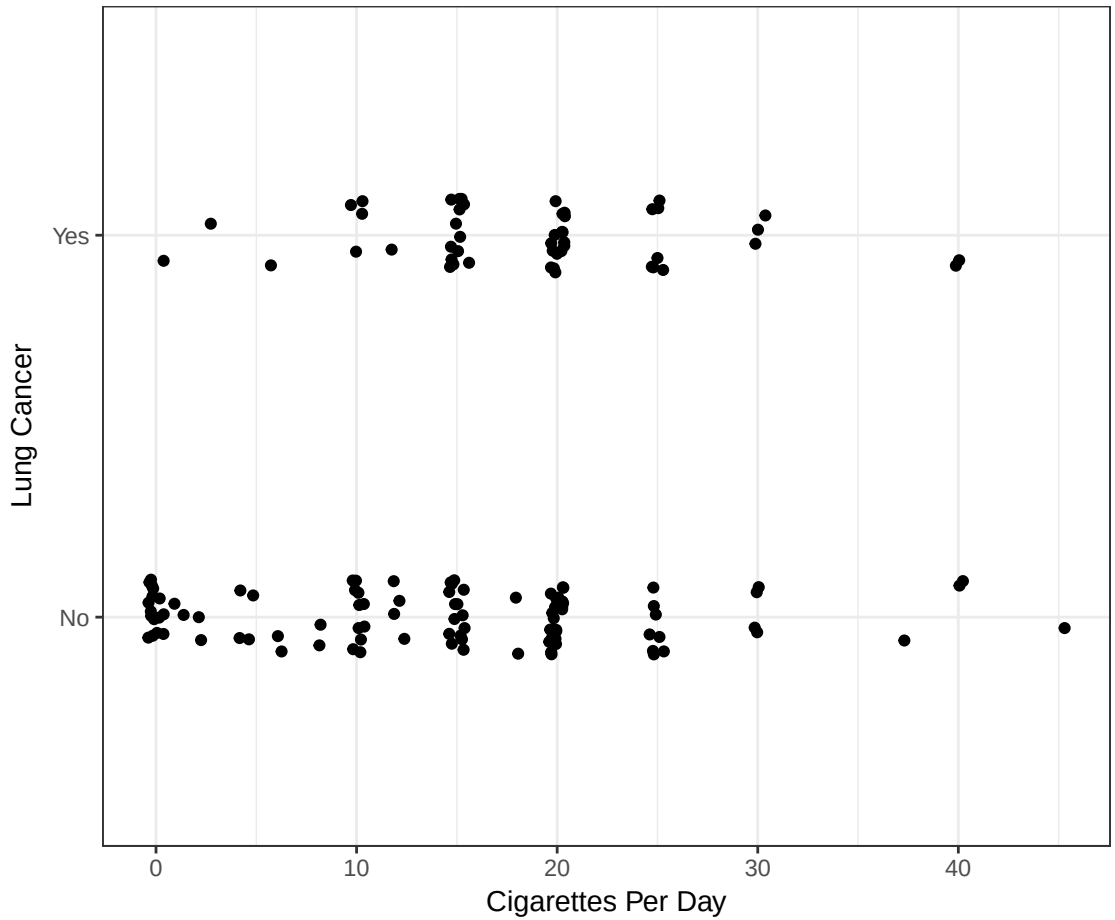
Lung cancer due to smoking

To *prove* a causal relationship between lung cancer and smoking, there should be clear evidence that there is a dose response between lung cancer and smoking.

But since lung cancer is binary, we need to compare the proportion of individuals who have lung cancer to those who don't amongst the individuals who smoke about the same amount.

To investigate the causes of lung cancer, researchers conducted a case-control study where the 49 cases of individuals with lung cancer where *matched* with 98 controls from a population of residents having the same general age structure. (In case-control studies, the intercept does not have our standard interpretation because it is determined by our sampling.)

```
lung_cancer <- Sleuth3::case2002 %>%
  mutate(`Lung Cancer` = ifelse(LC=="NoCancer", "No","Yes"),
         `Cigarettes Per Day` = CD)

ggplot(lung_cancer, aes(`Cigarettes Per Day`, `Lung Cancer`)) +
  geom_jitter(height=0.1) +
  theme_bw()
```

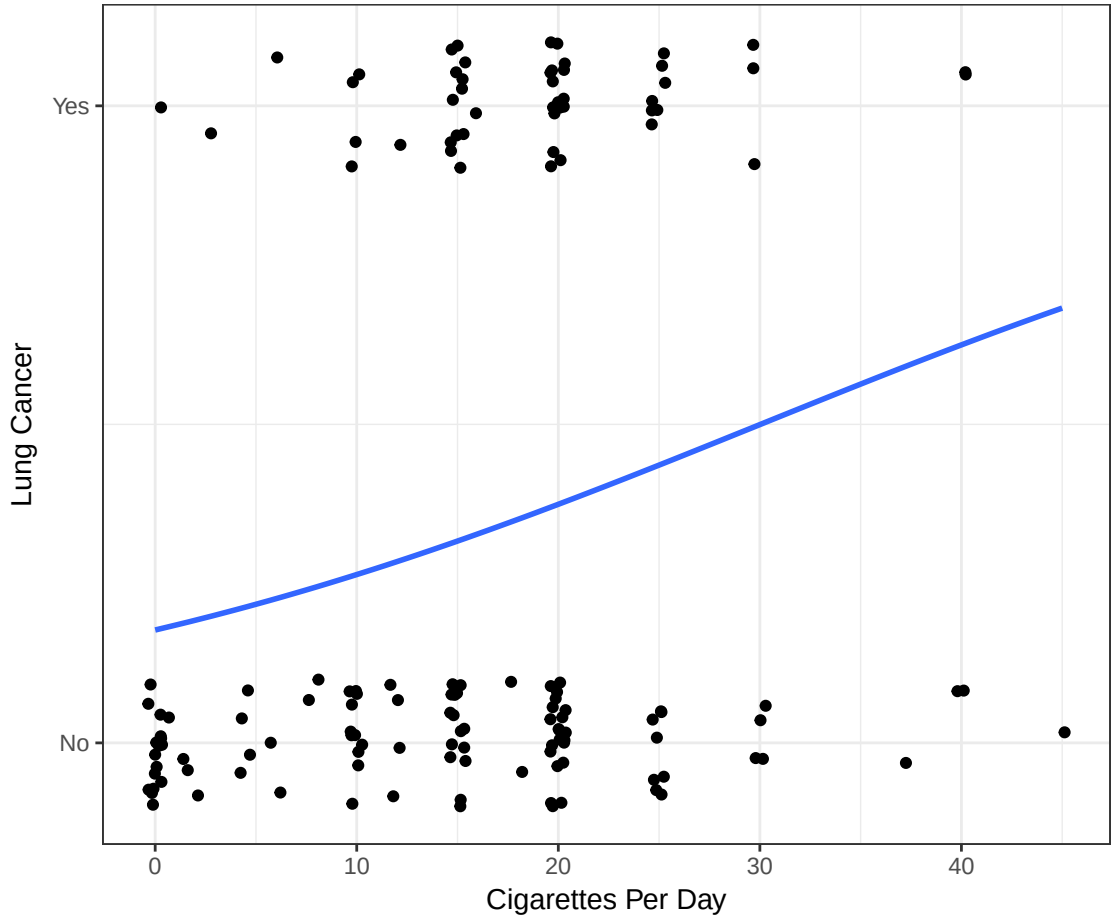


Analysis

```
m <- glm(`Lung Cancer`=="Yes" ~ `Cigarettes Per Day`,
         data = lung_cancer,
         family = "binomial")

summary(m)
```

```
Call:
glm(formula = `Lung Cancer` == "Yes" ~ `Cigarettes Per Day`,
    family = "binomial", data = lung_cancer)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5148  -0.9688  -0.7166   1.3449   1.8603

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.53541    0.37707  -4.072 4.66e-05 ***
`Cigarettes Per Day`  0.05113    0.01939   2.637  0.00836 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 187.14  on 146  degrees of freedom
Residual deviance: 179.62  on 145  degrees of freedom
AIC: 183.62

Number of Fisher Scoring iterations: 4
```

```
ggplot(lung_cancer, aes(`Cigarettes Per Day`, 1*(`Lung Cancer` == "Yes"))) +
  geom_jitter(height=0.1) +
  stat_smooth(method="glm",
              se=FALSE,
              method.args = list(family="binomial"))  +
  theme_bw() +
  scale_y_continuous(breaks=c(0,1), labels=c("No","Yes")) +
  labs(y = "Lung Cancer")
```



Grouping

Often data are grouped:

```
lung_cancer_grouped <- lung_cancer %>%
  group_by(`Cigarettes Per Day`) %>%
  summarize(`Number of individuals` = n(),
            `Number with lung cancer` = sum(`Lung Cancer` == "Yes"),
            `Number without lung cancer` = sum(`Lung Cancer` == "No"),
            `Proportion with lung cancer` = `Number with lung cancer`/`Number of individuals`)

lung_cancer_grouped
```

```
# A tibble: 19  5
   `Cigarettes Per Day` `Number of individuals` `Number with lung cancer` `Number without lung cancer`
                  <int>                   <int>                     <int>                       <int>
1                     0                      17                         1                          16
2                     1                       2                         0                           2
3                     2                       2                         0                           2
4                     3                       1                         1                           0
5                     4                       2                         0                           2
6                     5                       2                         0                           2
7                     6                       3                         1                           2
8                     8                       2                         0                           2
9                    10                      15                         4                          11
10                   12                       5                         1                           4
11                   15                      27                        12                          15
12                   16                       1                         1                           0
13                   18                       2                         0                           2
14                   20                      38                        16                          22
15                   25                      15                         7                           8
16                   30                       7                         3                           4
17                   37                       1                         0                           1
```

Binomial distribution

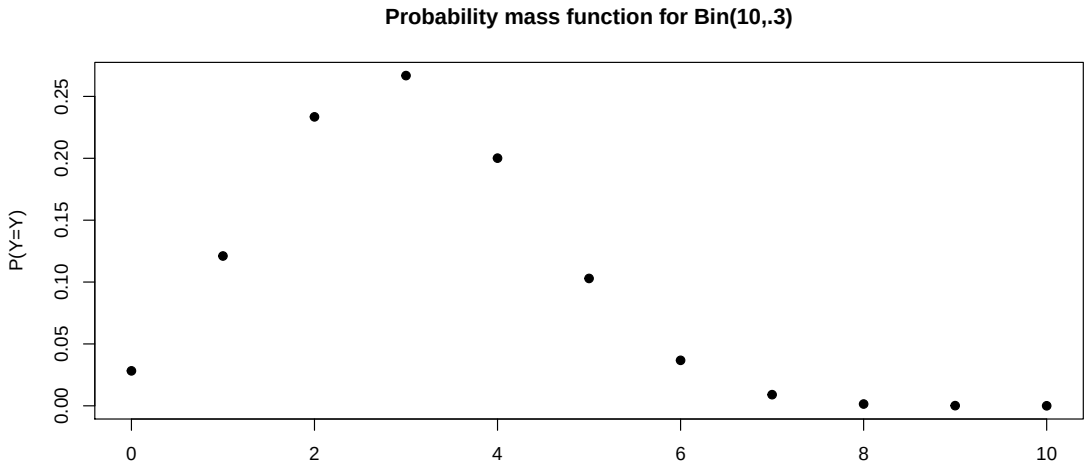
The probability mass function of the binomial distribution is

$$P(Y = y) = \binom{n}{y} \theta^y (1-\theta)^{n-y} \qquad y = 0, 1, 2, \ldots, n$$

Properties:

- $E[Y] = n\theta$
- $V[Y] = n\theta(1-\theta)$

```
xx = 0:10
plot(xx, dbinom(xx, 10, .3), main="Probability mass function for Bin(10,.3)",
     xlab="y", ylab="P(Y=Y)", pch=19)
```



Logistic regression for grouped data

Let Y_i be the number of success out of n_i attempts in group i. Then a logistic regression model is

$$Y_i \overset{ind}{\sim} Bin(n_i, \theta_i)$$

$$\text{logit}(\theta_i) = \log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_1 X_i$$

where

- Y_i is an integer from 0 to n_i
- Bin refers to the binomial distribution

Logistic regression in R

```
m = glm(cbind(‘Number with lung cancer‘, ‘Number without lung cancer‘) ~ ‘Cigarettes Per Day‘,
        data = lung_cancer_grouped,
        family="binomial")
summary(m)
```

```
Call:
glm(formula = cbind(‘Number with lung cancer‘, ‘Number without lung cancer‘) ~
    ‘Cigarettes Per Day‘, family = "binomial", data = lung_cancer_grouped)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5148  -1.0253  -0.5070   0.3305   1.7922

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.53541    0.37707  -4.072 4.66e-05 ***
‘Cigarettes Per Day‘  0.05113    0.01939   2.637  0.00836 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28.651  on 18  degrees of freedom
Residual deviance: 21.141  on 17  degrees of freedom
AIC: 48.879

Number of Fisher Scoring iterations: 4
```

```
confint(m)
```

Effect of birdkeeping on lung cancer

The data set we have been analyzing was actually constructed to investigate the relationship between birdkeeping and lung cancer. But, since we know smoking increase the probability of developing lung cancer, we want to control for the effect of smoking when assessing the effect of bird keeping. Thus, we will run a logistic regression with both smoking and bird-keeping to determine the effect of bird-keeping on lung cancer.

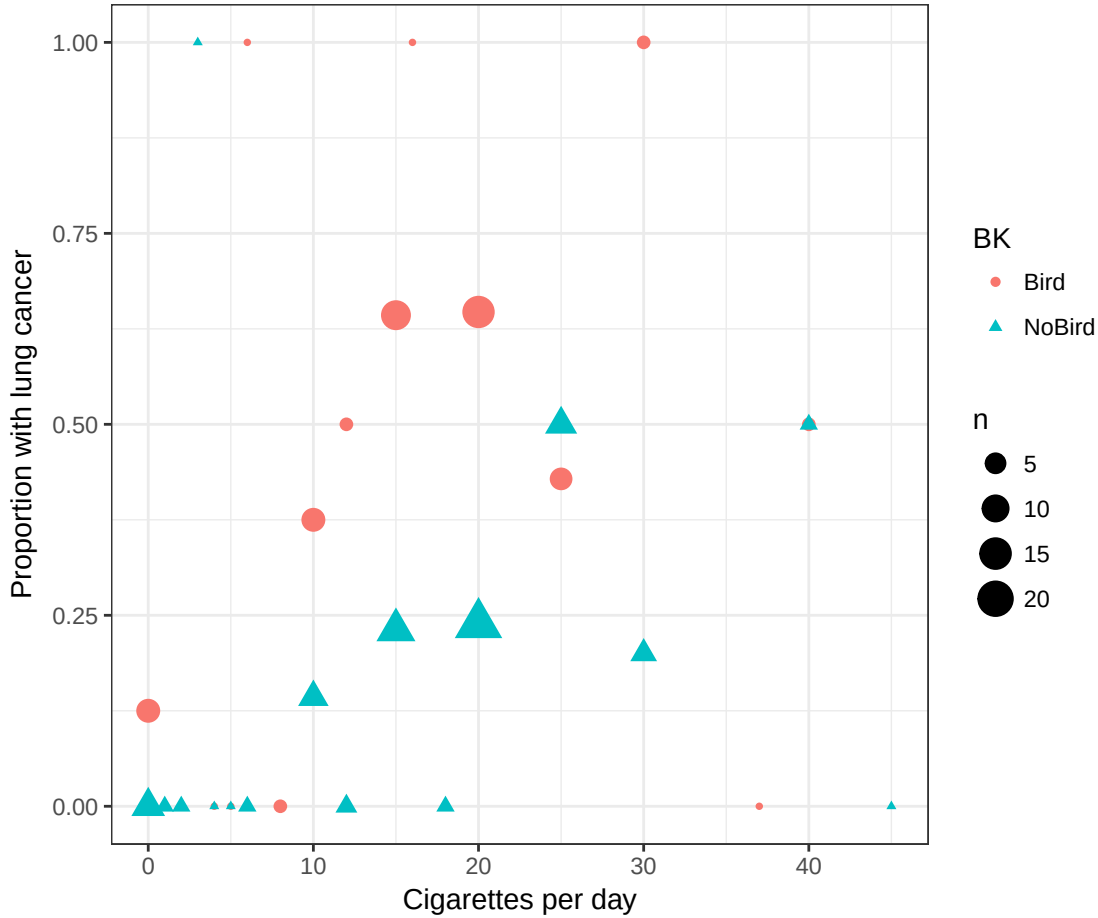
Summarize data

```
lung_cancer_bird <- Sleuth3::case2002 %>%
  group_by(CD, BK) %>%
  summarize(y = sum(LC == "LungCancer"),
            n = n(),
            p = y/n)

lung_cancer_bird
```

```
Source: local data frame [30 x 5]
Groups: CD [?]

      CD     BK     y     n     p
   <int> <fctr> <int> <int> <dbl>
1      0   Bird     1     8 0.125
2      0 NoBird     0     9 0.000
3      1 NoBird     0     2 0.000
4      2 NoBird     0     2 0.000
5      3 NoBird     1     1 1.000
6      4   Bird     0     1 0.000
7      4 NoBird     0     1 0.000
8      5   Bird     0     1 0.000
9      5 NoBird     0     1 0.000
10     6   Bird     1     1 1.000
# ... with 20 more rows
```

Visualize data

```
ggplot(lung_cancer_bird, aes(CD, p, size=n, color=BK, shape=BK)) +
  geom_point() +
  theme_bw() +
  labs(x="Cigarettes per day", y="Proportion with lung cancer")
```



Model

Let Y_i be the number of success out of n_i attempts in group i with explanatory variables $X_{i,1}$ and $X_{i,2}$. Then a logistic regression model is

$$Y_i \overset{ind}{\sim} Bin(n_i, \theta_i)$$

$$\text{logit}(\theta_i) = \log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2}$$

The interpretation is

- The probability of success is $\frac{1}{1+e^{-\beta_0}}$ when all explanatory variables are zero. (Except in a case-control study.)
- The odds ratio for a one unit increase in $X_{i,1}$ is e^{β_1} when holding all other explanatory variables constant.
- The odds ratio for a one unit increase in $X_{i,2}$ is e^{β_2} when holding all other explanatory variables constant.

Logistic regression with multiple explanatory variables

```
# LC is binary
summary(m <- glm(cbind(y,n-y) ~ CD + BK, data=lung_cancer_bird, family="binomial"))


Call:
glm(formula = cbind(y, n - y) ~ CD + BK, family = "binomial",
    data = lung_cancer_bird)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7167  -0.9555  -0.5413   0.4025   2.1594

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.94683    0.41319  -2.291 0.021935 *
CD           0.05838    0.02087   2.797 0.005157 **
BKNoBird    -1.45760    0.38856  -3.751 0.000176 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 53.386  on 29  degrees of freedom
Residual deviance: 30.612  on 27  degrees of freedom
AIC: 66.07

Number of Fisher Scoring iterations: 4
```

```
nd <- expand.grid(CD = 0:45, BK=c("Bird","NoBird"))
pd <- cbind(nd, data.frame(p=predict(m, newdata = nd, type = "response")))

ggplot() +
  geom_point(data = lung_cancer_bird, aes(CD, p, size=n, color=BK, shape=BK)) +
  geom_line(data = pd, aes(CD, p, color=BK, linetype=BK)) +
  theme_bw() +
  labs(x="Cigarettes per day", y="Proportion with lung cancer")
```