Generalized Linear Models and Logistic Regression

Author: Nicholas G Reich, OpenIntro

This material is part of the **statsTeachR** *project*

Today's Lecture

- Generalized linear models (GLMs)
- Logistic regression

[Note: more on logistic regression can be found in the OpenIntro textbook, Chapter 8. These slides are based, in part, on the slides from OpenIntro.]

Regression so far ...

At this point we have covered:

- Simple linear regression
  - Relationship between numerical response and a numerical or categorical predictor
- Multiple regression
  - Relationship between numerical response and multiple numerical and/or categorical predictors
  - What to do when the relationships with the predictors are complex (nonlinear, skewed distribution, interactions, confounding, etc.)

What we haven't covered is what to do when the response is not continuous (i.e. categorical, count data, etc.)

Example - Birdkeeping and Lung Cancer

A 1972 - 1981 health survey in The Hague, Netherlands, discovered an association between keeping pet birds and increased risk of lung cancer. To investigate birdkeeping as a risk factor, researchers conducted a case-control study of patients in 1985 at four hospitals in The Hague (population 450,000). They identified 49 cases of lung cancer among the patients who were registered with a general practice, who were age 65 or younger and who had resided in the city since 1965. They also selected 98 controls from a population of residents having the same general age structure.

Ramsey, F.L. and Schafer, D.W. (2002). The Statistical Sleuth: A Course in Methods of Data Analysis (2nd ed)

Example - Birdkeeping and Lung Cancer - Data

```
library(Sleuth3)
birds = case2002
head(birds)
```

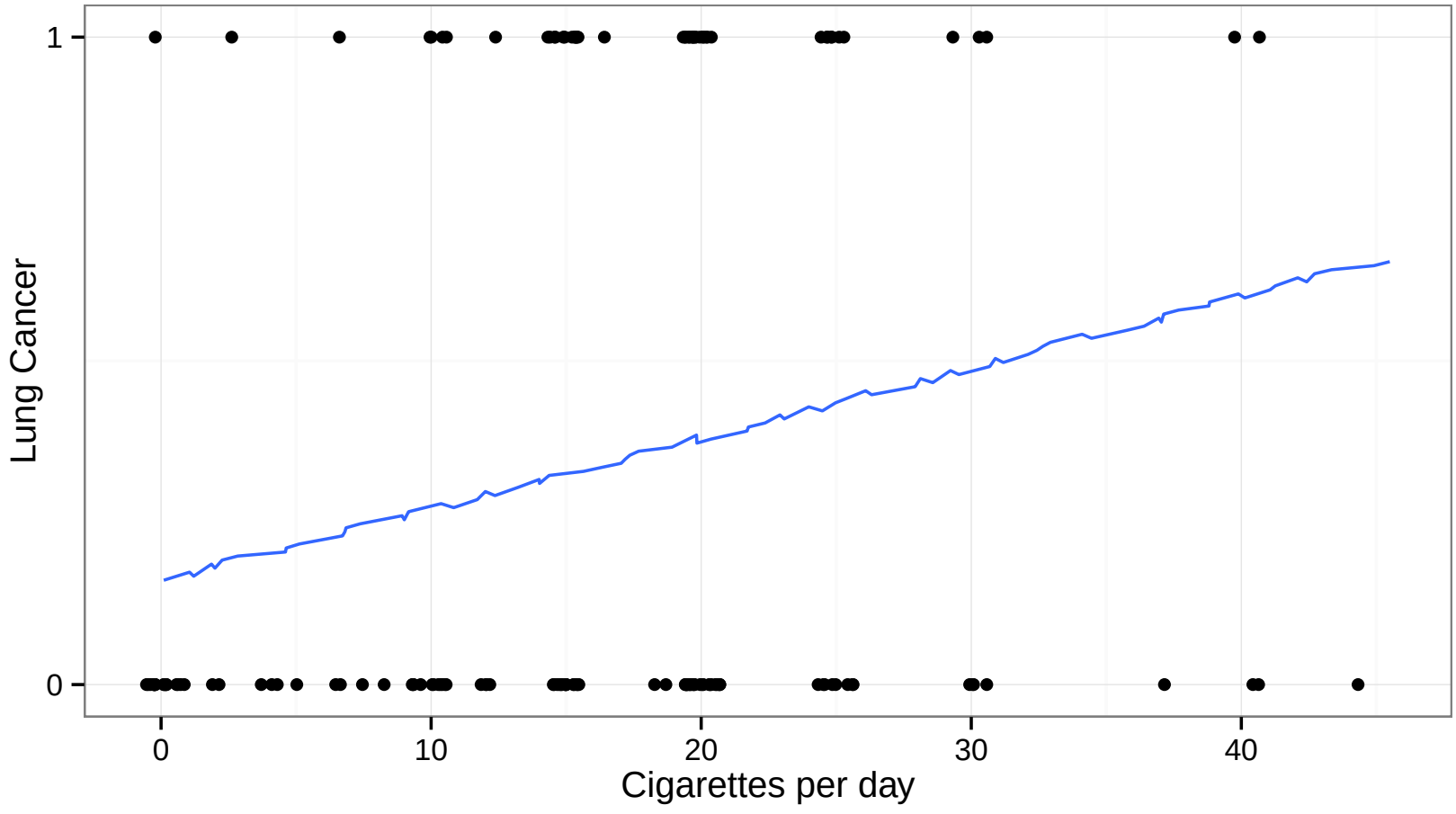
```
##           LC   FM   SS     BK AG YR CD
## 1 LungCancer Male  Low   Bird 37 19 12
## 2 LungCancer Male  Low   Bird 41 22 15
## 3 LungCancer Male High NoBird 43 19 15
## 4 LungCancer Male  Low   Bird 46 24 15
## 5 LungCancer Male  Low   Bird 49 31 20
## 6 LungCancer Male High NoBird 51 24 15
```

LC	Whether subject has lung cancer
FM	Sex of subject
SS	Socioeconomic status
BK	Indicator for birdkeeping
AG	Age of subject (years)
YR	Years of smoking prior to diagnosis or examination
CD	Average rate of smoking (cigarettes per day)

NoCancer is the reference response (0 or failure), LungCancer is the non-reference response (1 or success) - this matters for interpretation.

Lung cancer as a function of cigarettes per day

```
(p <- qplot(CD, as.numeric(LC=="LungCancer")*1, data=birds, geom=c("point", "smooth"),
            position=position_jitter(w=.7, h=0), method="lm", se=FALSE) +
     ylab("Lung Cancer") + xlab("Cigarettes per day") + scale_y_continuous(breaks=c(0,1)))
```

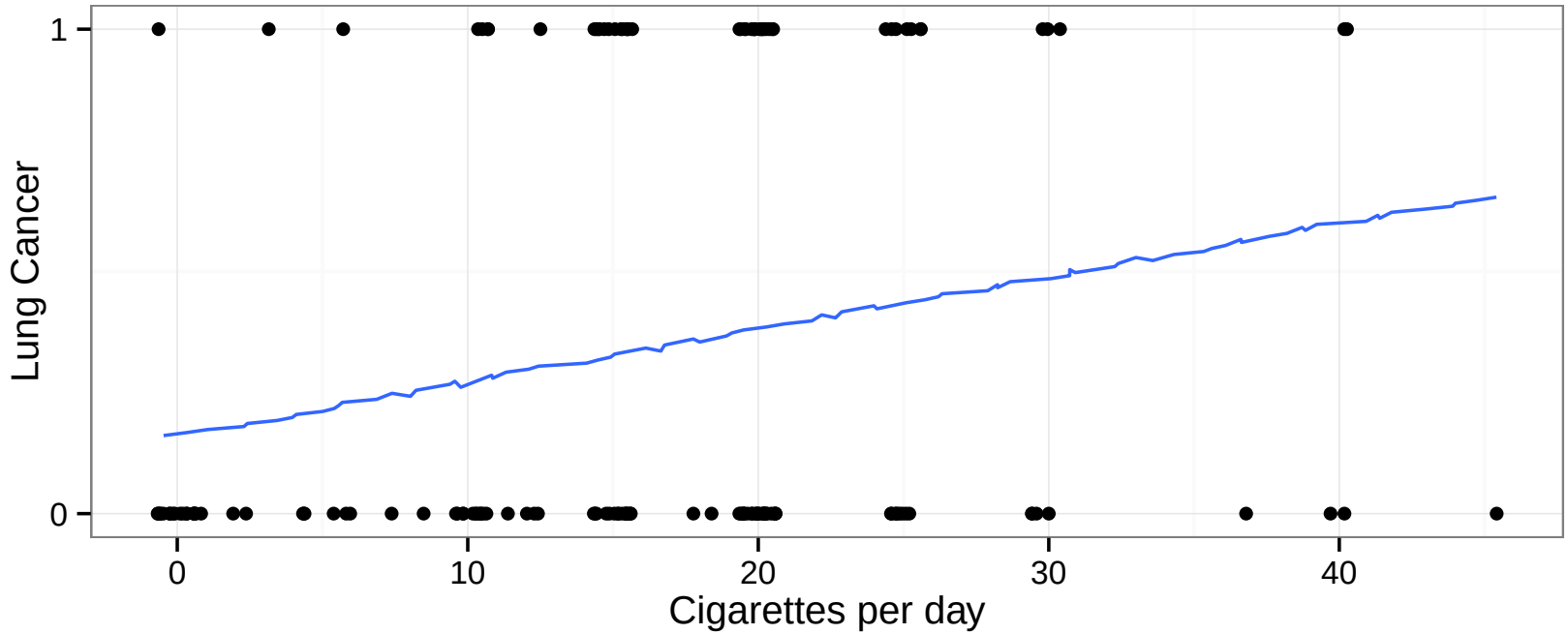


Generalized linear models

$$\mathbb{E}[Y|x] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Why not just use MLR when outcomes not continuous?

- Linearity assumption may be more unreasonable than usual.

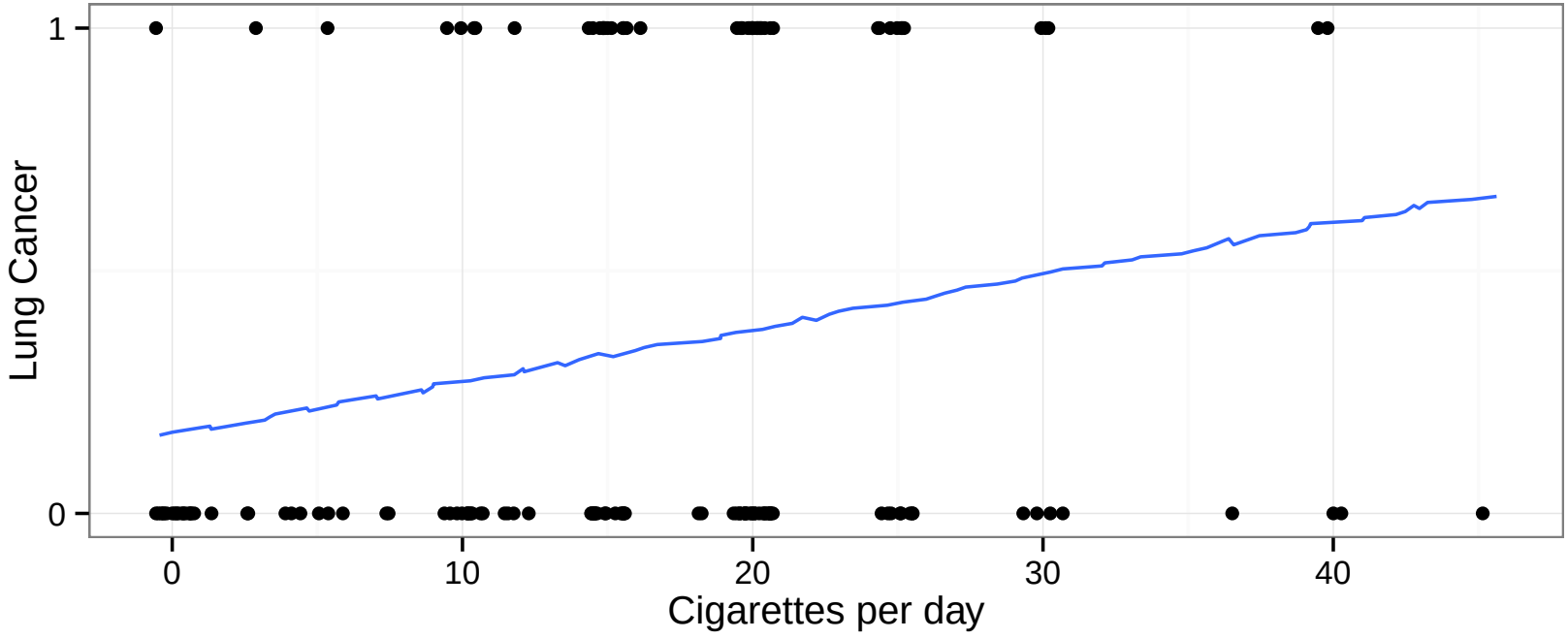


Generalized linear models

$$\mathbb{E}[Y|x] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Why not just use MLR when outcomes not continuous?

- Equal variance assumption often violated ($Var[Y|x] = \sigma^2$).

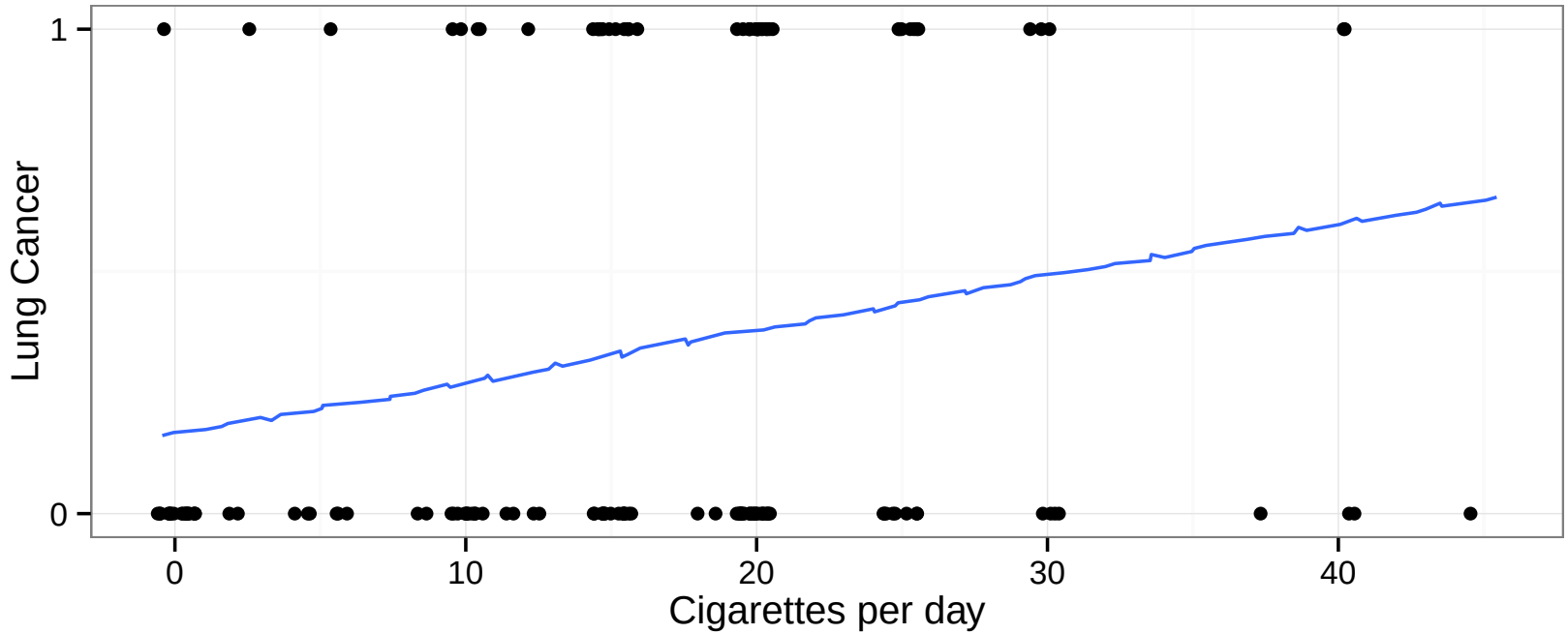


Generalized linear models

$$\mathbb{E}[Y|x] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Why not just use MLR when outcomes not continuous?

- Assumption of normal errors explicitly violated.



Generalized linear models: defined

All generalized linear models have the following three characteristics:

1. **A probability distribution** describing the outcome variable
   - e.g. $Y \sim \text{Bernoulli}(p) \longrightarrow \mathbb{E}[Y|p] = p$.
2. **A linear model**
   - $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$
3. **A link function** that relates the linear model to the parameter of the outcome distribution
   - $g(\mathbb{E}[Y]) = g(p) = \eta$ or $\mathbb{E}[Y] = p = g^{-1}(\eta)$

Gaussian MLR is a special case of a GLM

For continuous outcome, we often do this

1. **A probability distribution** describing the outcome variable
   - $Y|X \sim \text{Normal}(\mu, \sigma^2) \longrightarrow \mathbb{E}[Y|X] = \mu$.
2. **A linear model**
   - $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$
3. **A link function** that relates the linear model to the parameter of the outcome distribution
   - $g(\mathbb{E}[Y|X]) = g(\mu) = \mu = \eta$

$$g(\mathbb{E}[Y|X]) = E[Y|X] = \mu = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Logistic regression: a common GLM for 0/1 outcome data

1. **A probability distribution** describing the outcome variable
   - $Y|X \sim \text{Bernoulli}(p) \longrightarrow \mathbb{E}[Y|X] = Pr(Y = 1|X) = p$.
2. **A linear model**
   - $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$
3. **A link function** that relates the linear model to the parameter of the outcome distribution
   - $g(\mathbb{E}[Y|X]) = g(p) = logit(p) = \log \frac{p}{1-p} = \eta$

$$g(\mathbb{E}[Y|X]) = logit[Pr(Y = 1|X)] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Odds

Odds are another way of quantifying the probability of an event, commonly used in gambling (and logistic regression).

For some event E,

$$\text{odds}(E) = \frac{P(E)}{P(E^c)} = \frac{P(E)}{1 - P(E)}$$

Similarly, if we are told the odds of E are x to y then

$$\text{odds}(E) = \frac{x}{y} = \frac{x/(x+y)}{y/(x+y)}$$

which implies

$$P(E) = x/(x+y), \quad P(E^c) = y/(x+y)$$

Logistic regression has log(odds) as the link

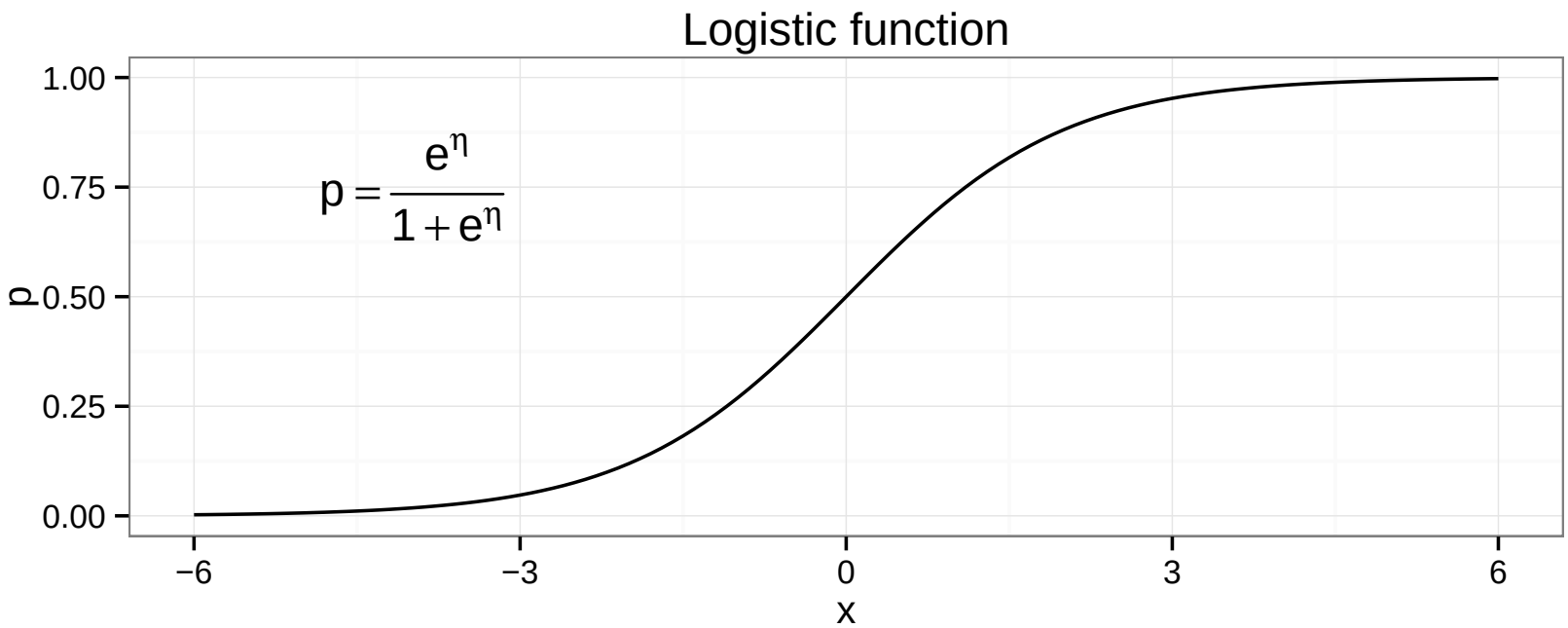
A logistic regression model can be defined as follows:

$$Y_i|\mathbf{x}_i \sim \text{Bernoulli}(p_i)$$

$$\mathbb{E}[Y_i|\mathbf{x}_i] = Pr(Y_i = 1|\mathbf{x}_i) = p_i$$

$$g(p_i) = logit(p_i) = \log \frac{p_i}{1 - p_i}$$

$$logit(\mathbb{E}[Y_i|\mathbf{x}_i]) = logit(p_i) = \eta = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$



Example - Birdkeeping and Lung Cancer - Model

$$logitPr(LC = 1|\mathbf{x}) = \beta_0 + \beta_1 BK + \beta_2 FM + \beta_3 SS + \beta_4 AG + \beta_5 YR + \beta_6 CD$$

```
birds$LCnum <- as.numeric(birds$LC=="LungCancer")
birds$BK <- relevel(birds$BK, ref="NoBird")
lm1 <- glm(LCnum ~ BK + FM + SS +  AG + YR + CD,
           data=birds, family=binomial)
```

Example - Birdkeeping and Lung Cancer - Interpretation

```
summary(lm1)$coef
```

```
##                Estimate Std. Error    z value     Pr(>|z|)
## (Intercept) -1.27063830 1.82530568 -0.6961236 0.4863514508
## BKBird       1.36259456 0.41127585  3.3130916 0.0009227076
## FMMale      -0.56127270 0.53116056 -1.0566912 0.2906525319
## SSLow       -0.10544761 0.46884614 -0.2249088 0.8220502474
## AG          -0.03975542 0.03548022 -1.1204952 0.2625027758
## YR           0.07286848 0.02648741  2.7510612 0.0059402544
## CD           0.02601689 0.02552400  1.0193110 0.3080553359
```

Keeping all other predictors constant then,

- The odds ratio of getting lung cancer for bird keepers vs non-bird keepers is $\exp(1.3626) = 3.91$.
- The odds ratio of getting lung cancer for an additional year of smoking is $\exp(0.0729) = 1.08$.

What do the numbers not mean ...

The most common mistake made when interpreting logistic regression is to treat an odds ratio as a ratio of probabilities.

Bird keepers are *not* 4x more likely to develop lung cancer than non-bird keepers.

This is the difference between relative risk and an odds ratio.

$$RR = \frac{P(\text{disease}|\text{exposed})}{P(\text{disease}|\text{unexposed})}$$

$$OR = \frac{P(\text{disease}|\text{exposed})/[1 - P(\text{disease}|\text{exposed})]}{P(\text{disease}|\text{unexposed})/[1 - P(\text{disease}|\text{unexposed})]}$$

To match or not to match

Case-control studies are common for (rare) binary outcomes

- Randomly selected controls $\longrightarrow$ vanilla logistic regression
- Matched controls $\longrightarrow$ conditional logistic regression

Conditional logistic regression

- Accounts for the fact that you have "adjusted" for some variables in the design.
- Calculates an OR for each matched-set/pair, then "averages" across sets
- Forfeits ability to estimate effects of matched variables, but design can substantially improve power.
- Implemented in R with `clogit()`.

Important notes about GLMs

On logistic regression in particular...

- There are other link functions for binary data (e.g. probit, cloglog).
- Other, less parameteric methods may be appropriate here too – e.g. CART, random forests, classification algorithms.

Beyond the scope of this course, but interesting topics...

- How are logistic models (and other GLMS) fitted?
- Can we perform the same kind of model diagnostics to determine whether a model provides a good fit to data?
- ROC curves and classification rules