Introduction to Logistic Regression

Author: Nicholas G Reich

Made available under the Creative Commons Attribution-ShareAlike 3.0 Unported License: http://creativecommons.org/licenses/by-sa/3.0/deed.en_US

Today's Lecture

- Logistic regression

[Note: more on logistic regression can be found in Kaplan, Chapter 16 and the OpenIntro Statistics textbook, Chapter 8. These slides are based, in part, on the slides from OpenIntro.]

Regression so far ...

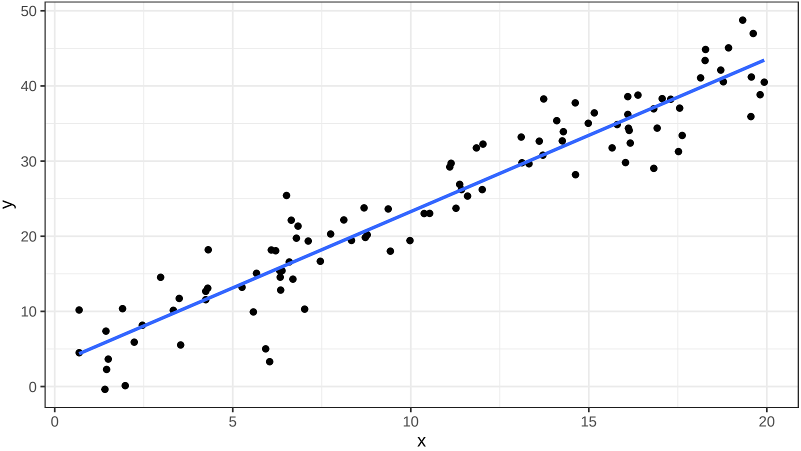
At this point we have covered:

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

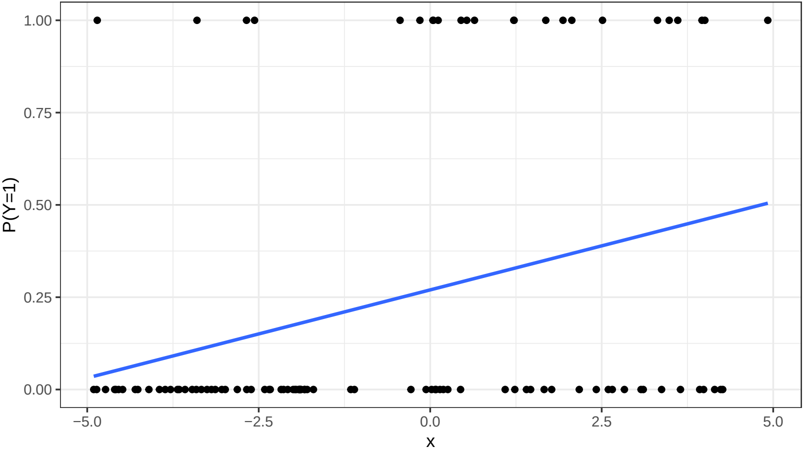
Refresher: regression estimates the average response

In linear regression, when y is a continuous variable, our model estimates for us the average y, $\hat{y}$, value for a particular value of x.



Refresher: regression estimates the average response

When y is a binary outcome variable, regression models can still estimate for us the average y, $\hat{y}$, but now it represents something different: the probability that $y = 1$, or $P(Y = 1)$.



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

```r
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.

Example - Birdkeeping and Lung Cancer - Data

What types of associations do you expect to see between the predictors below and lung cancer? Might you expect any interactions to be present?

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)

Interpreting linear regressions of binary data

We can use linear regression for binary data, and for *very simple models* it gives reasonable and interpretable output.

$$\hat{L}C_i = \beta_0$$

```
birds$LCnum <- as.numeric(birds$LC=="LungCancer")
sum(birds$LCnum)/nrow(birds)

## [1] 0.3333333

summary(lm(LCnum ~ 1, data=birds))$coef

##              Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 0.3333333 0.03901372 8.544004 1.574829e-14
```

Interpreting linear regressions of binary data

We can use linear regression for binary data, and for *very simple models* it gives reasonable and interpretable output.

$$\hat{L}C_i = \beta_0 + \beta_1 * FM_i$$

```
mod1 <- lm(LCnum ~ FM, data=birds)
round(summary(mod1)$coef, 3)

##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.333      0.079   4.214        0
## FMMale         0.000      0.091   0.000        1
```

What is this model's estimated probability of lung cancer for men? for women?

Interpreting linear regressions of binary data

But if the model gets too complicated, then it can produce some tricky results.

$$\hat{L}C_i = \beta_0 + \beta_1 \cdot FM_i + \beta_2 \cdot YR_i$$

```
mod2 <- lm(LCnum ~ FM + YR, data=birds)
round(summary(mod2)$coef, 3)

##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.119      0.092   1.294    0.198
## FMMale        -0.150      0.094  -1.592    0.114
## YR             0.012      0.003   4.044    0.000
```

What is this model's estimated probability of lung cancer for men who have never smoked? for women who never smoked?

Brief detour: what are the “odds”?

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

For some event E,

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

and

$$\mathrm{P}(E) = \frac{odds(E)}{1 + odds(E)}$$

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

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

Odds Ratios

Odds Ratios compare the odds of an event in two different groups. For some outcome of interest (say, disease) in two groups, (e.g. exposed and unexposed),

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

Facts about Odds Ratios

- ORs have range of $(0, \infty)$.
- OR $= 1$ means no difference between the groups.
- They have a multiplicative scale: e.g. OR $= 0.5$ and OR $= 2$ both indicate that one group has twice the odds of another.
- This means that the log OR is on an additive scale of odds (This is important for logistic regression!).
- OR is not a ratio of probabilities.

Unadjusted association btw lung cancer and sex

```
library(epitools)
birds$LC <- relevel(birds$LC, ref="NoCancer")
(tmp <- with(birds, table(FM, LC)))

##         LC
## FM        NoCancer LungCancer
##   Female       24         12
##   Male         74         37

oddsratio(tmp)$measure

##         odds ratio with 95% C.I.
## FM        estimate     lower    upper
##   Female 1.0000000        NA       NA
##   Male   0.9954866 0.4516538 2.280673
```

Do men have different odds of lung cancer compared to women, without adjustment for possible confounders?

Unadjusted association btw lung cancer and birdkeeping

```
birds$BK <- relevel(birds$BK, ref="NoBird")
(tmp <- with(birds, table(BK, LC)))

##         LC
## BK       NoCancer LungCancer
##   NoBird       64         16
##   Bird         34         33

oddsratio(tmp)$measure

##         odds ratio with 95% C.I.
## BK       estimate   lower    upper
##   NoBird 1.000000      NA       NA
##   Bird   3.827991 1.86773 8.124253
```

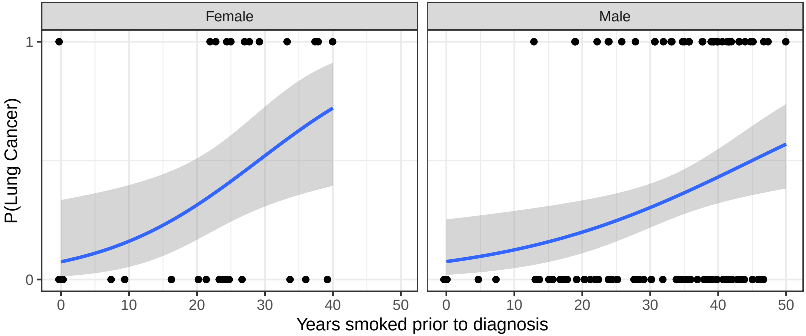
Do birdkeepers have different odds of lung cancer compared to non-birdkeepers, without adjustment for possible confounders?

Lung cancer as a function of years smoked

Modeling the log-odds is one solution to the problem of linearity.

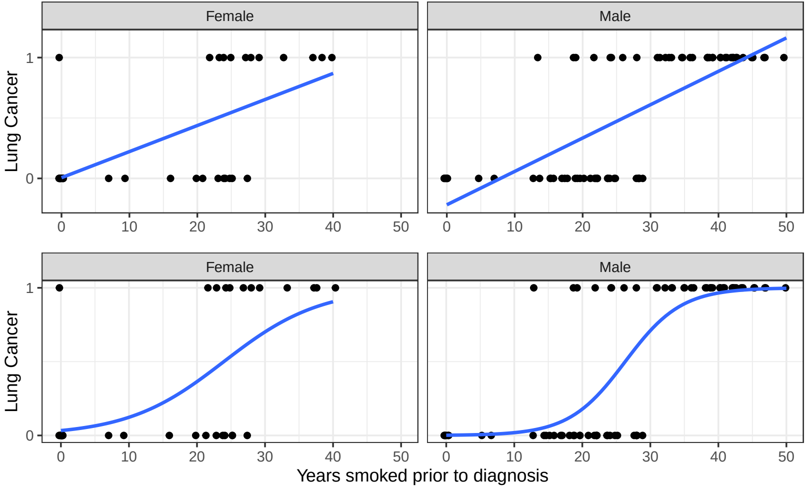
$$\log odds(LC) \sim FM + YR$$

```
(p <- ggplot(birds, aes(x=YR, y=LCnum)) +
   geom_jitter(height=0) + facet_grid(.~FM) +
   stat_smooth(method='glm', method.args=list(family='binomial')) +
   ylab("P(Lung Cancer)") + xlab("Years smoked prior to diagnosis") +
   scale_y_continuous(breaks=c(0,1)))
```



A more drastic example: why we use the log-odds

I dropped individuals who smoked >30 years prior to diagnosis who did not have LC.



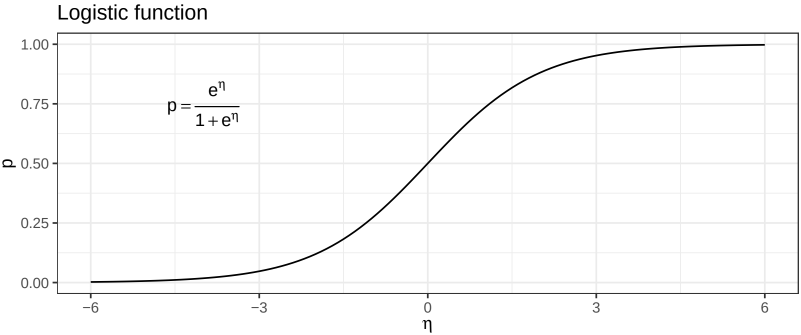
Logistic regression has log(odds) as the link

A logistic regression model can be defined as follows:

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

$$Pr(Y_i = 1|\mathbf{x}_i) = p_i$$

$$logit(p_i) = \log \frac{p_i}{1 - p_i} = \eta = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$



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")
fm1 <- glm(LCnum ~ BK + FM + SS +  AG + YR + CD,
           data=birds, family=binomial)
```

Example - Birdkeeping and Lung Cancer - Interpretation

```
summary(fm1)$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,

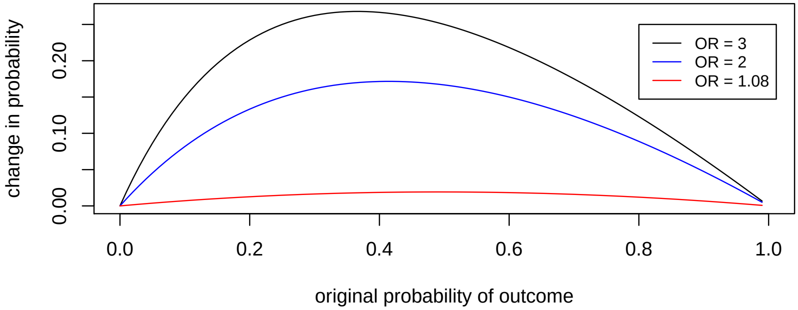
- The odds ratio of getting lung cancer for bird keepers vs non-bird keepers is $\exp(1.3626) = 3.91$.
- The odds ratio of getting lung cancer for an additional year of smoking is $\exp(0.0729) = 1.08$. I.e. for every year an individual smokes, the odds of lung cancer increase by 8%.
- The odds ratio of getting lung cancer for an additional 10 years of smoking is $\exp(0.0729 * 10) = 2.07$.

Building Intuition: How do ORs modify absolute risk?

If you have a 1% risk of lung cancer, what does a 8% increase in odds mean? How about a 100% or 200% increase in odds?

```
change_in_prob <- function(orig_prob, OR){
    new_odds <- orig_prob / (1-orig_prob) * OR
    return( new_odds/(1+new_odds) - orig_prob)
}
change_in_prob(orig_prob=.01, OR=1.08)

## [1] 0.0007913669
```



What the numbers do 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.

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

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

Measuring accuracy of a model for binary outcomes

A common metric for evaluating models for binary outcomes is simply called the “accuracy”.
It requires specifying a threshold (c) for the predicted probability ($\hat{p}_i$) that an individual has the outcome of interest. For a given threshold, we say that

$$\hat{y}_i(c) = \begin{cases} 1, & \text{if } p_i \geq c \\ 0, & \text{if } p_i < c \end{cases}$$

And then the accuracy, for a given c is defined

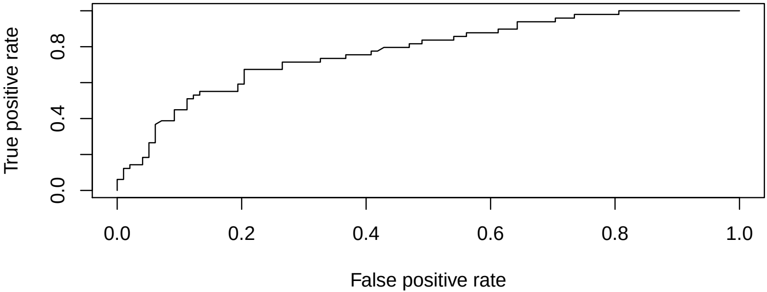
$$accuracy(c) = \sum_{i=1}^{N} \frac{1 - |\hat{y}_i(c) - y_i|}{N}$$

.

Measuring accuracy across all thresholds

To gain a full picture of the accuracy of your model, you want to compute the accuracy for all possible values of c. This is what the ROC curve does:

```
library(ROCR)
birds$phat <- predict(fm1, type="response")
pred <- prediction(birds$phat, birds$LCnum)
perf <- performance(pred,"tpr","fpr")
plot(perf)
```

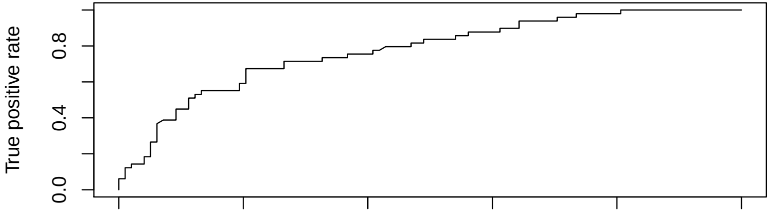


Measuring accuracy across all thresholds

The "area under the curve" (AUC) is a common single measure of predictive accuracy of a model. It can be interpreted as *the probability, if one random individual with the outcome and one without were drawn from your dataset, that your model would accurately assign a higher p_i to the one with the outcome.*

```
## this prints out the computed area under the curve
performance(pred,"auc")@y.values


## [[1]]
## [1] 0.7746772
```



Important notes about GLMs

On logistic regression in particular...

- There are other link functions for binary data (e.g. probit, cloglog).
- Other, less parametric 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?
- Sensitivity and specificity

Mathematical details

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 $\mathbb{E}[Y] = p = g^{-1}(\eta)$

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
   - $g(\mathbb{E}[Y|X]) = g(p) = logit(p) = \log \frac{p}{1-p} = \eta$

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