# Sta 111 - Summer II 2017
# Probability and Statistical Inference

## 24. Logistic regression

Lu Wang

Duke University, Department of Statistical Science

August 7, 2017

## Outline

At this point we have covered:

- Simple linear regression: one predictor - $y$ and $x$

- Multiple linear regression: multiple predictors - $y$ and $x_1$, $x_2, \ldots$
  - Relationship between numerical response and multiple numerical and/or categorical predictors

What we haven't seen is what to do when the predictors are weird (nonlinear, complicated dependence structure, etc.) or when the response is weird (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.

From *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

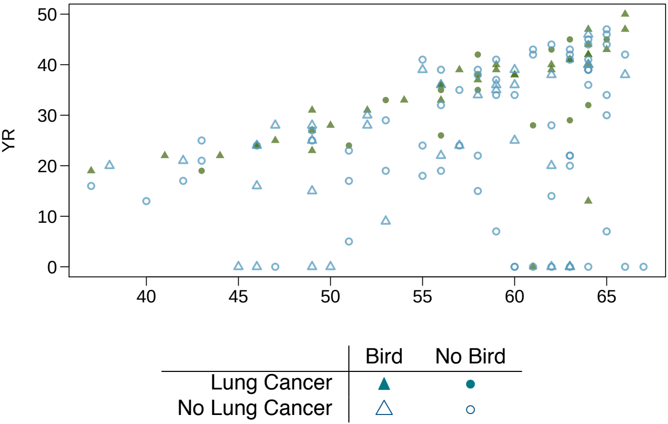
| | LC | FM | SS | BK | AG | YR | CD |
|---|---|---|---|---|---|---|---|
| 1 | LungCancer | Male | Low | Bird | 37.00 | 19.00 | 12.00 |
| 2 | LungCancer | Male | Low | Bird | 41.00 | 22.00 | 15.00 |
| 3 | LungCancer | Male | High | NoBird | 43.00 | 19.00 | 15.00 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 147 | NoCancer | Female | Low | NoBird | 65.00 | 7.00 | 2.00 |

| | |
|---|---|
| LC | Whether subject has lung cancer |
| FM | Gender 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) |

---

*Note: NoCancer is the reference response 0, LungCancer is the non-reference response 1.*

# Birdkeeping and Lung Cancer - EDA



- How do we come up with a model that will let us explore this relationship?

- Even if we set `NoLungCancer` to 0 and `LungCancer` to 1, this isn't something we can transform our way out of - cannot apply linear regression directly.

- One way to think about the problem - we can treat `NoLungCancer` and `LungCancer` as successes and failures arising from a binomial distribution where the probability of a success is given by a transformation of a linear model of the predictors.

It turns out that this is a very general way of addressing this type of problem in regression, and the resulting models are called *generalized linear models (GLMs)*. Logistic regression is just one example of this type of model.

All *generalized linear models* have the following three characteristics:

1. A probability distribution describing the outcome variable
2. A linear model
   - $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$
3. A link function that relates the linear model to the parameter of the outcome distribution
   - $g(p) = \eta$ or $p = g^{-1}(\eta)$

- Logistic regression is a GLM used to model a *binary categorical outcome* using numerical and categorical predictors.

- We assume the outcome variable follows a binomial distribution and therefore want to model the probability $p$ of success for a given set of predictors.

- To finish specifying the Logistic model we just need to establish a reasonable link function that connects $(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)$ to $p$. There are a variety of options but the most commonly used is the logit function.

### Logit function

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right), \text{ for } 0 \le p \le 1$$

- The *logit function* takes a value between 0 and 1 and maps it to a value between $-\infty$ and $+\infty$.
- The *inverse logit function* takes a value between $-\infty$ and $+\infty$ and maps it to a value between 0 and 1.

**Inverse logit (logistic) function**

$$g^{-1}(x) = \frac{\exp(x)}{1+\exp(x)} = \frac{1}{1+\exp(-x)}$$

The logit formulation is very useful when it comes to interpreting the model since logit can be interpreted as the log *odds* of a success.

# Odds

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

## Odds

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

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

The three GLM criteria give us:

$$
\begin{aligned}
y_i &\sim \text{Bernoulli}(p_i) \\
\eta &= \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \\
\text{logit}(p) &= \eta
\end{aligned}
$$

From which we arrive at,

$$
p_i = \frac{\exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i})}{1 + \exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i})}
$$

In R we fit a GLM using `glm` and we must also specify the type of GLM by the `family` argument.

# Example - Birdkeeping and Lung Cancer - Model

```
summary(glm(LC ~ FM + SS + BK + AG + YR + CD, data=bird, family=binomial))
```

```
## Call:
## glm(formula = LC ~ FM + SS + BK + AG + YR + CD, family = binomial,
##     data = bird)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.93736    1.80425  -1.074 0.282924
## FMFemale     0.56127    0.53116   1.057 0.290653
## SSHigh       0.10545    0.46885   0.225 0.822050
## BKBird       1.36259    0.41128   3.313 0.000923 ***
## AG          -0.03976    0.03548  -1.120 0.262503
## YR           0.07287    0.02649   2.751 0.005940 **
## CD           0.02602    0.02552   1.019 0.308055
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 187.14  on 146  degrees of freedom
## Residual deviance: 154.20  on 140  degrees of freedom
## AIC: 168.2
##
## Number of Fisher Scoring iterations: 5
```

# Example - Birdkeeping and Lung Cancer - Model

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -1.9374 | 1.8043 | -1.07 | 0.2829 |
| FMFemale | 0.5613 | 0.5312 | 1.06 | 0.2907 |
| SSHigh | 0.1054 | 0.4688 | 0.22 | 0.8221 |
| BKBird | 1.3626 | 0.4113 | 3.31 | 0.0009 |
| AG | -0.0398 | 0.0355 | -1.12 | 0.2625 |
| YR | 0.0729 | 0.0265 | 2.75 | 0.0059 |
| CD | 0.0260 | 0.0255 | 1.02 | 0.3081 |

Model:

$$\log\left(\frac{p}{1-p}\right) = -1.9374 + 0.5613\text{FMFemale} + 0.1054\text{SSHigh} + 1.3626\text{BKBird} - 0.0398\text{AG} + 0.0729\text{YR} + 0.0260\text{CD}$$

## Slope Interpretation - Categorical Variable

Just like MLR we can plug in `BK` to arrive at two status for `Bird` and `NoBird` respectively, *while the other predictors are held constant.*

`Bird` model: $\log\left(\frac{p_1}{1-p_1}\right) = \ldots + 1.3626 \times 1 + \ldots$

`NoBird` model: $\log\left(\frac{p_0}{1-p_0}\right) = \ldots + 1.3626 \times 0 + \ldots$

$$\textit{change in log odds} \rightarrow \log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_0}{1-p_0}\right) = 1.3626$$

$$\textit{log odds ratio} \rightarrow \log\left(\frac{p_1}{1-p_1} \Big/ \frac{p_0}{1-p_0}\right) = 1.3626$$

$$\textit{odds ratio} \rightarrow \frac{p_1}{1-p_1} \Big/ \frac{p_0}{1-p_0} = \exp(1.3626) = 3.9063$$

*`BK` slope*: Keeping all other predictors constant, this is the *log odds ratio* of getting lung cancer for bird keepers (given level) vs non-bird keepers (reference level).

## Slope Interpretation - Numerical Variable

- When the other predictors are held constant, for a unit increase in `YR` (additional year of smoking), how much will the log odds change?

$$\log\left(\frac{p}{1-p}\right) = \ldots + 0.0729x + \ldots$$

$$\log\left(\frac{p'}{1-p'}\right) = \ldots + 0.0729(x+1) + \ldots$$

$$\log\left(\frac{p'}{1-p'}\right) - \log\left(\frac{p}{1-p}\right) = 0.0729$$

$$\log\left(\frac{p'}{1-p'}\Bigg/\frac{p}{1-p}\right) = 0.0729$$

$$\frac{p'}{1-p'}\Bigg/\frac{p}{1-p} = \exp(0.0729) = 1.0756$$

*`YR` slope:* Keeping all other predictors constant, this is the *change in log odds* of getting lung cancer for an additional year of smoking (per unit change in the predictor).

## Common mistake: odds ratio vs relative risk

- 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 most common mistake made when interpreting logistic regression is to treat an odds ratio as a ratio of probabilities.
- Bird keepers are <u>not</u> 3.91x more likely to develop lung cancer than non-bird keepers.

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

$$RR = \frac{P(\text{Cancer}|\text{Bird})}{P(\text{Cancer}|\text{NoBird})} = \frac{p_1}{p_0}$$

$$OR = \frac{P(\text{Cancer}|\text{Bird})/[1 - P(\text{Cancer}|\text{Bird})]}{P(\text{Cancer}|\text{NoBird})/[1 - P(\text{Cancer}|\text{NoBird})]} = \frac{p_1/(1-p_1)}{p_0/(1-p_0)}$$

## Testing for the slope of `BKBird`

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -1.9374 | 1.8043 | -1.07 | 0.2829 |
| FMFemale | 0.5613 | 0.5312 | 1.06 | 0.2907 |
| SSHigh | 0.1054 | 0.4688 | 0.22 | 0.8221 |
| BKBird | *1.3626* | *0.4113* | *3.31* | *0.0009* |
| AG | -0.0398 | 0.0355 | -1.12 | 0.2625 |
| YR | 0.0729 | 0.0265 | 2.75 | 0.0059 |
| CD | 0.0260 | 0.0255 | 1.02 | 0.3081 |

The basic setup is exactly the same as what we've seen before except that we use a Z test.

$H_0 : \beta_j = 0$ when other explanatory variables are included in the model.

$H_A : \beta_j \neq 0$ when other explanatory variables are included in the model.

$$Z = \frac{\hat{\beta}_j - \beta_j}{SE} = \frac{1.3620 - 0}{0.4113} = 3.31$$

$$\text{p-value} = P(|Z| > 3.31) = P(Z > 3.31) + P(Z < -3.31)$$
$$= 0.0009$$

---

*Note: The only tricky bit, which is way beyond the scope of this course, is how the standard error is calculated.*

## Confidence interval for `BKBird` slope

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -1.9374 | 1.8043 | -1.07 | 0.2829 |
| FMFemale | 0.5613 | 0.5312 | 1.06 | 0.2907 |
| SSHigh | 0.1054 | 0.4688 | 0.22 | 0.8221 |
| BKBird | *1.3626* | *0.4113* | 3.31 | 0.0009 |
| AG | -0.0398 | 0.0355 | -1.12 | 0.2625 |
| YR | 0.0729 | 0.0265 | 2.75 | 0.0059 |
| CD | 0.0260 | 0.0255 | 1.02 | 0.3081 |

Recall that the interpretation for the `BKBird` slope is the *log odds ratio* of getting lung cancer for bird keepers vs non-bird keepers.

Log odds ratio:

$$CI = PE \pm CV \times SE = 1.3626 \pm 1.96 \times 0.4113 = (0.5565, 2.1687)$$

Odds ratio:

$$\exp(CI) = (\exp(0.5565), \exp(2.1687)) = (1.7446, 8.7469)$$