Lecture 21 - Logistic Regression

Sta102

June 15, 2016

Colin Rundel & Mine Çetinkaya-Rundel

GLMs

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

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

Example - Donner Party

In 1846 the Donner and Reed families left Springfield, Illinois, for California by covered wagon. In July, the Donner Party, as it became known, reached Fort Bridger, Wyoming. There its leaders decided to attempt a new and untested route to the Sacramento Valley. Having reached its full size of 87 people and 20 wagons, the party was delayed by a difficult crossing of the Wasatch Range and again in the crossing of the desert west of the Great Salt Lake. The group became stranded in the eastern Sierra Nevada mountains when the region was hit by heavy snows in late October. By the time the last survivor was rescued on April 21, 1847, 40 of the 87 members had died from famine and exposure to extreme cold.

From *Ramsey, Schafer (2002). The Statistical Sleuth*

Example - Donner Party - Data

	Age	Sex	Status
1	23.00	Male	Died
2	40.00	Female	Survived
3	40.00	Male	Survived
4	30.00	Male	Died
5	28.00	Male	Died
⋮	⋮	⋮	⋮
43	23.00	Male	Survived
44	24.00	Male	Died
45	25.00	Female	Survived

Example - Donner Party - EDA

Status vs. Gender:

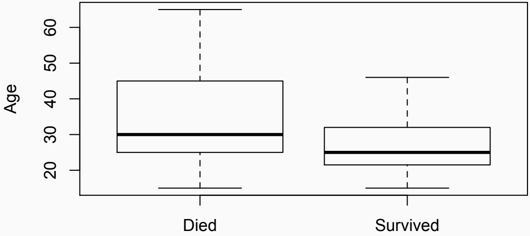
	Male	Female
Died	20	5
Survived	10	10

Example - Donner Party - EDA

Status vs. Gender:

	Male	Female
Died	20	5
Survived	10	10

Status vs. Age:



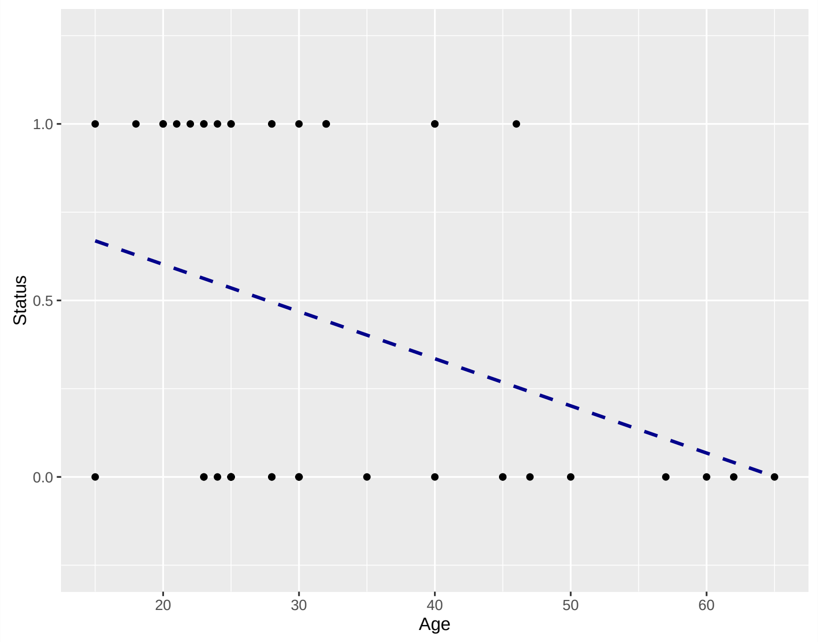
Example - Donner Party - ???

It seems clear that both age and gender have an effect on someone's survival, how do we come up with a model that will let us explore this relationship?

Example - Donner Party - ???

It seems clear that both age and gender have an effect on someone's survival, how do we come up with a model that will let us explore this relationship?

Even if we set Died to 0 and Survived to 1, this isn't something we can reasonably fit a linear model to - we need something more.



1.0
0.5
0.0
Status
20
30
40
50
60
Age

Bernoulli Data?

Another way to think about the problem:

- We can treat each outcome (Survived and Died) as successes and failures arising from separate Bernoulli trials
- Each Bernoulli trial can have a separate probability of success

$$y_i \sim \text{Bern}(p_i)$$

- We can then use the predictor variables to model that probability of success (p_i)
- We can't just use a linear model for p_i (since p_i must be between 0 and 1) but we can transform the linear model to have the appropriate range.

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

Generalized linear models

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

All generalized linear models have the following three characteristics:

Generalized linear models

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

All generalized linear models have the following three characteristics:

1. A probability distribution describing a generative model for the outcome variable

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

All generalized linear models have the following three characteristics:

1. A probability distribution describing a generative model for the outcome variable
2. A linear model

$$\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

Generalized linear models

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

All generalized linear models have the following three characteristics:

1. A probability distribution describing a generative model for the outcome variable
2. A linear model
$$\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$
3. A link function that relates the linear model to the parameter of the outcome distribution

Logistic Regression

Logistic Regression

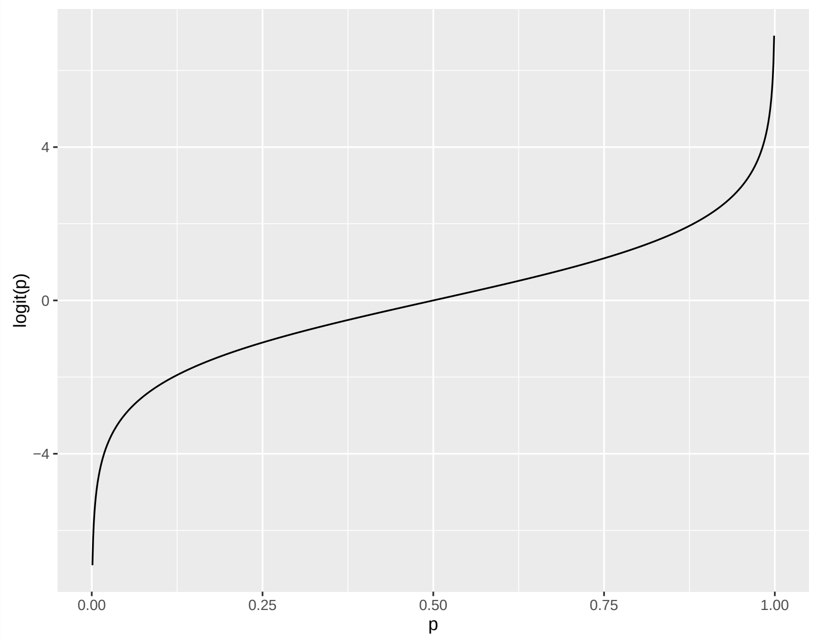
Logistic regression is a GLM used to model a binary categorical outcome using numerical and categorical predictors.

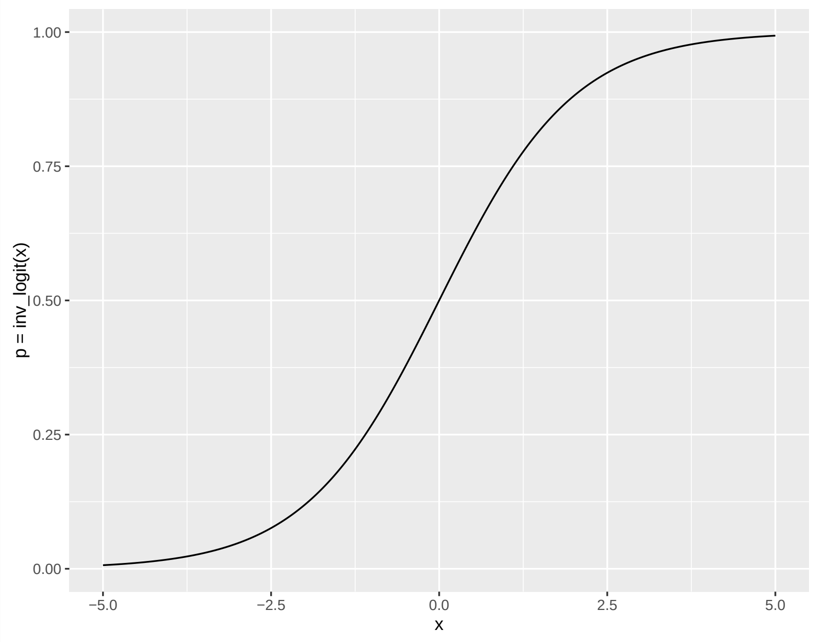
Logistic regression is a GLM used to model a binary categorical outcome using numerical and categorical predictors.

To finish specifying the Logistic model we just need to define a reasonable link function that connects η_i to p_i. There are a variety of options but the most commonly used is the logit function.

Logit function:

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


logit(p)
4
0
−4
0.00
0.25
0.50
0.75
1.00
p


1.00
0.75
0.50
0.25
0.00
p = inv_logit(x)
−5.0
−2.5
0.0
2.5
5.0
x

Properties of the Logit

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

Inverse logit (logistic) function:

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

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

This formulation is also useful for interpreting the model, since the logit can be interpreted as the log odds of a success - more on this later.

The logistic regression model

The three GLM criteria give us:

$$y_i \sim \text{Bern}(p_i)$$

$$\eta_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i}$$

$$\text{logit}(p_i) = \eta_i$$

From which we get,

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

Example - Donner Party - Model

In R we fit a GLM in the same way as a linear model except we use `glm` instead of `lm`. (We specify the type of GLM to fit using the `family` argument)

```
summary(glm(Status ~ Age, data=donner, family=binomial))
```

```
##
## Call:
## glm(formula = Status ~ Age, family = binomial, data = donner)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5401  -1.1594  -0.4651   1.0842   1.7283
##
## Coefficients:
##             Estimate Std. Error z value Pr(>♦z♦)
## (Intercept)  1.81852    0.99937   1.820   0.0688 .
## Age         -0.06647    0.03222  -2.063   0.0391 *
...
```

Example - Donner Party - Prediction

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Example - Donner Party - Prediction

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a newborn (Age=0):

Example - Donner Party - Prediction

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a newborn (Age=0):

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 0$$

$$\frac{p}{1-p} = \exp(1.8185) = 6.16$$

$$p = 6.16/7.16 = 0.86$$

Example - Donner Party - Prediction (cont.)

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a 25 year old:

Example - Donner Party - Prediction (cont.)

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a 25 year old:

$$\begin{aligned}
\log\left(\frac{p}{1-p}\right) &= 1.8185 - 0.0665 \times 25 \\
\frac{p}{1-p} &= \exp(0.156) = 1.17 \\
p &= 1.17/2.17 = 0.539
\end{aligned}$$

Example - Donner Party - Prediction (cont.)

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a 25 year old:

$$\begin{aligned}\log\left(\frac{p}{1-p}\right) &= 1.8185 - 0.0665 \times 25 \\ \frac{p}{1-p} &= \exp(0.156) = 1.17 \\ p &= 1.17/2.17 = 0.539\end{aligned}$$

Odds / Probability of survival for a 50 year old:

Example - Donner Party - Prediction (cont.)

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a 25 year old:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 25$$

$$\frac{p}{1-p} = \exp(0.156) = 1.17$$

$$p = 1.17/2.17 = 0.539$$

Odds / Probability of survival for a 50 year old:

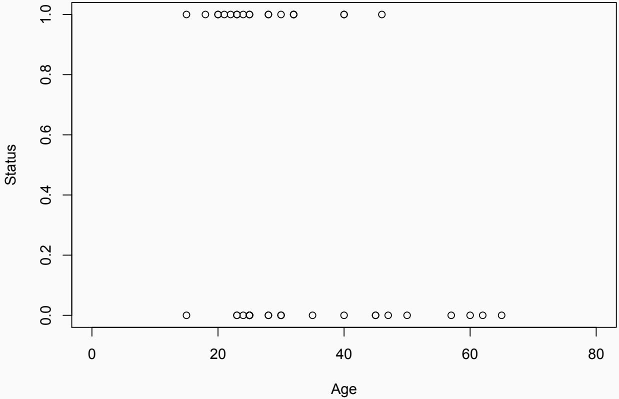
$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 0$$

$$\frac{p}{1-p} = \exp(-1.5065) = 0.222$$

$$p = 0.222/1.222 = 0.181$$

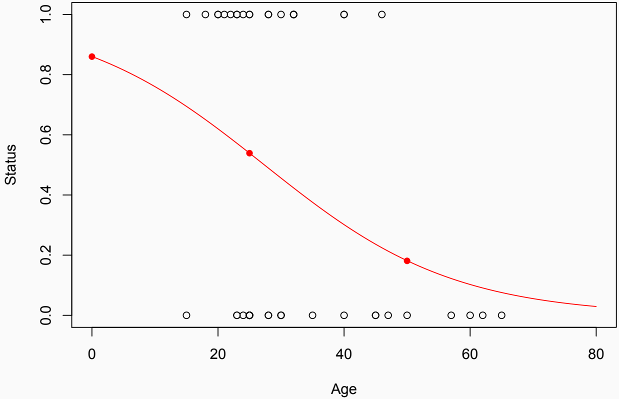
Example - Donner Party - Prediction (cont.)

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$



Example - Donner Party - Prediction (cont.)

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$



Example - Donner Party - Interpretation

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

Simple interpretation is only possible in terms of *log odds* and *log odds ratios* for intercept and slope terms.

Intercept: The *log odds* of survival for a party member with an age of 0. From this we can calculate the odds or probability, but additional calculations are necessary.

Slope: For a unit increase in age (being 1 year older) how much will the *log odds ratio* change, not particularly intuitive. More often than not we care only about sign and relative magnitude.

Example - Donner Party - Interpretation - Intercept

Value of η when all predictors are 0:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665(0) = 1.8185$$

$$\frac{p}{1-p} = \exp(1.8185) = 6.16$$

$$p = 6.16/(6.16+1) = 0.86$$

Example - Donner Party - Interpretation - Slope

$$\begin{aligned}
\log\left(\frac{p_1}{1-p_1}\right) &= 1.8185 - 0.0665(x+1) \\
&= 1.8185 - 0.0665x - 0.0665 \\
\log\left(\frac{p_2}{1-p_2}\right) &= 1.8185 - 0.0665x
\end{aligned}$$

$$\begin{aligned}
\log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_2}{1-p_2}\right) &= -0.0665 \\
\log\left(\frac{p_1}{1-p_1} \bigg/ \frac{p_2}{1-p_2}\right) &= -0.0665 \\
\frac{p_1}{1-p_1} \bigg/ \frac{p_2}{1-p_2} &= \exp(-0.0665) = 0.94
\end{aligned}$$

Example - Donner Party - Age and Gender

```
summary(glm(Status ~ Age + Sex, data=donner, family=binomial))
```

```
##
## Call:
## glm(formula = Status ~ Age + Sex, family = binomial, data = donner)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7445  -1.0441  -0.3029   0.8877   2.0472
##
## Coefficients:
##             Estimate Std. Error z value Pr(>♦z♦)
## (Intercept)  1.63312    1.11018   1.471   0.1413
## Age         -0.07820    0.03728  -2.097   0.0359 *
## SexFemale    1.59729    0.75547   2.114   0.0345 *
...
```

Gender slope: When the other predictors are held constant this is the log odds ratio between the contrast (Female) and the reference level (Male).

Example - Donner Party - Gender Models

Just like MLR we can plug in gender to arrive at two status vs age models for men and women respectively.

General model:

$$\log\left(\frac{p_1}{1-p_1}\right) = 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times \text{Sex}$$

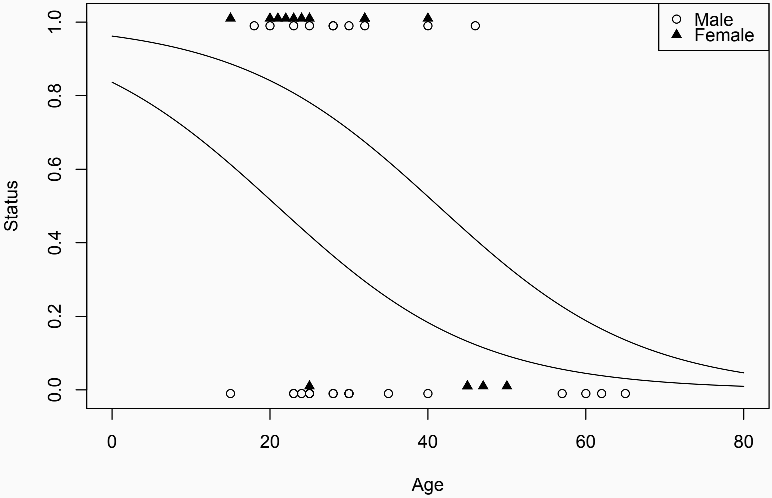
Male model:

$$\begin{aligned}\log\left(\frac{p_1}{1-p_1}\right) &= 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times \mathit{0} \\ &= 1.63312 + -0.07820 \times \text{Age}\end{aligned}$$

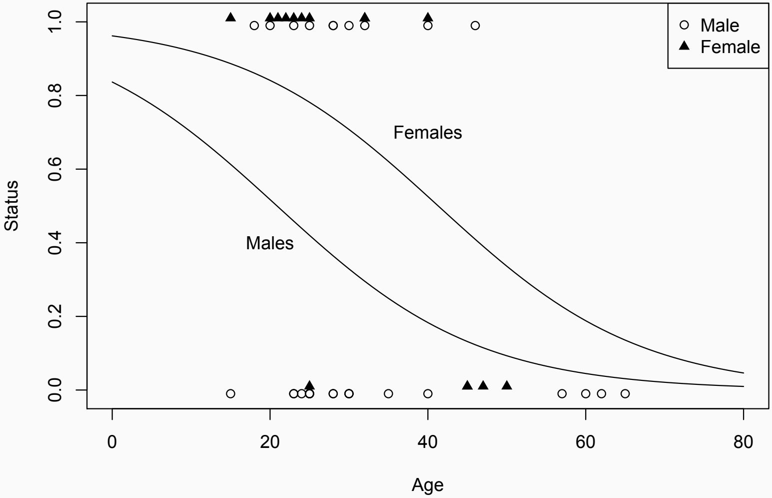
Female model:

$$\begin{aligned}\log\left(\frac{p_1}{1-p_1}\right) &= 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times \mathit{1} \\ &= 3.23041 + -0.07820 \times \text{Age}\end{aligned}$$

Example - Donner Party - Gender Models (cont.)



Example - Donner Party - Gender Models (cont.)



Hypothesis test for the model

```
summary(glm(Status ~ Age + Sex, data=donner, family=binomial))
```

```
##
## Call:
## glm(formula = Status ~ Age + Sex, family = binomial, data = donner)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7445  -1.0441  -0.3029   0.8877   2.0472
##
## Coefficients:
##             Estimate Std. Error z value Pr(>♦z♦)
## (Intercept)  1.63312    1.11018   1.471   0.1413
## Age         -0.07820    0.03728  -2.097   0.0359 *
## SexFemale    1.59729    0.75547   2.114   0.0345 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 51.256  on 42  degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

Hypothesis tests for a coefficient

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

We can still perform inference for individual coefficients, the basic framework is the same as SLR/MLR except we use a Z test instead of a t test.

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

Testing for the slope of Age

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	*-0.0782*	*0.0373*	*-2.10*	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

Testing for the slope of Age

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	*-0.0782*	*0.0373*	*-2.10*	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

$$H_0 : \beta_{age} = 0$$

$$H_A : \beta_{age} \neq 0$$

Testing for the slope of Age

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

$$H_0 : \beta_{age} = 0$$

$$H_A : \beta_{age} \neq 0$$

$$Z = \frac{\hat{\beta_{age}} - \beta_{age}}{SE_{age}} = \frac{-0.0782 - 0}{0.0373} = -2.10$$

$$\text{p-value} = P(|Z| > 2.10) = P(Z > 2.10) + P(Z < -2.10)$$

$$= 2 \times 0.0178 = 0.0359$$

Confidence interval for age slope coefficient

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

Remember, the interpretation for a slope is the change in log odds ratio per unit change in the predictor.

Confidence interval for age slope coefficient

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

Remember, the interpretation for a slope is the change in log odds ratio per unit change in the predictor.

Log odds ratio:

$$CI = PE \pm CV \times SE = -0.0782 \pm 1.96 \times 0.0373 = (-0.1513, -0.0051)$$

Confidence interval for age slope coefficient

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

Remember, the interpretation for a slope is the change in log odds ratio per unit change in the predictor.

Log odds ratio:

$$CI = PE \pm CV \times SE = -0.0782 \pm 1.96 \times 0.0373 = (-0.1513, -0.0051)$$

Odds ratio:

$$\exp(CI) = (\exp(-0.1513), \exp(-0.0051) = (0.8596, 0.9949)$$

Bird Keeping and Lung Cancer?

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.

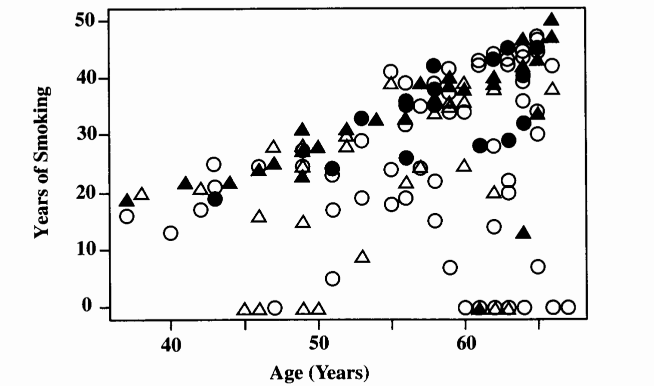
From *Ramsey, F.L. and Schafer, D.W. (2002). The Statistical Sleuth: A Course in Methods of Data Analysis (2nd ed)*

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	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)

Note - NoCancer is the reference response (0 or failure), LungCancer is the contrast response (1 or success).



	Bird	No Bird
Lung Cancer	▲	●
No Lung Cancer	△	○

Model

```
summary({g=glm(LC ~ ., data=bird, family=binomial)})
```

```
##
## Call:
## glm(formula = LC ~ ., family = binomial, data = bird)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5642  -0.8333  -0.4605   0.9808   2.2460
##
## 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
## ---
...
```

Model Selection

```r
library(MASS)
g2 = stepAIC(g)
```

```
## Start:  AIC=168.2
## LC ~ FM + SS + BK + AG + YR + CD
##
##          Df Deviance    AIC
## - SS     1   154.25 166.25
## - CD     1   155.24 167.24
## - FM     1   155.32 167.32
## - AG     1   155.49 167.49
## <none>       154.20 168.20
## - YR     1   163.93 175.93
## - BK     1   165.87 177.87
##
## Step:  AIC=166.25
## LC ~ FM + BK + AG + YR + CD
##
##          Df Deviance    AIC
## - FM     1   155.32 165.32
```

Model Selection - Results

```
summary(g2)


##
## Call:
## glm(formula = LC ~ BK + YR, family = binomial, data = bird)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6093  -0.8644  -0.5283   0.9479   2.0937
##
## Coefficients:
##             Estimate Std. Error z value Pr(>♦z♦)
## (Intercept) -3.18016    0.63640  -4.997 5.82e-07 ***
## BKBird       1.47555    0.39588   3.727 0.000194 ***
## YR           0.05825    0.01685   3.458 0.000544 ***
...
```

Interpretation

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.1802	0.6364	-5.00	0.0000
BKBird	1.4756	0.3959	3.73	0.0002
YR	0.0582	0.0168	3.46	0.0005

Interpretation

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.1802	0.6364	-5.00	0.0000
BKBird	1.4756	0.3959	3.73	0.0002
YR	0.0582	0.0168	3.46	0.0005

Keeping all other predictors constant then,

Interpretation

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.1802	0.6364	-5.00	0.0000
BKBird	1.4756	0.3959	3.73	0.0002
YR	0.0582	0.0168	3.46	0.0005

Keeping all other predictors constant then,

- The odds ratio of getting lung cancer for bird keepers vs non-bird keepers is $\exp(1.4756) = 4.37$.

Interpretation

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.1802	0.6364	-5.00	0.0000
BKBird	1.4756	0.3959	3.73	0.0002
YR	0.0582	0.0168	3.46	0.0005

Keeping all other predictors constant then,

- The odds ratio of getting lung cancer for bird keepers vs non-bird keepers is $\exp(1.4756) = 4.37$.
- The odds ratio of getting lung cancer for an additional year of smoking is $\exp(0.0582) = 1.06$.

Interpretation

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.1802	0.6364	-5.00	0.0000
BKBird	1.4756	0.3959	3.73	0.0002
YR	0.0582	0.0168	3.46	0.0005

Keeping all other predictors constant then,

- The odds ratio of getting lung cancer for bird keepers vs non-bird keepers is $\exp(1.4756) = 4.37$.
- The odds ratio of getting lung cancer for an additional year of smoking is $\exp(0.0582) = 1.06$.

What do these numbers mean in practice?

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.

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.

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*.

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

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

Back to the birds - Low Incidence

What is the probability of lung cancer in a bird keeper if we knew that $P(\text{lung cancer}|\text{no birds}) = 0.05$?

$$OR = \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{P(\text{lung cancer}|\text{no birds})/[1 - P(\text{lung cancer}|\text{no birds})]}$$

$$= \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{0.05/[1 - 0.05]} = 4.37$$

Back to the birds - Low Incidence

What is the probability of lung cancer in a bird keeper if we knew that $P(\text{lung cancer}|\text{no birds}) = 0.05$?

$$OR = \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{P(\text{lung cancer}|\text{no birds})/[1 - P(\text{lung cancer}|\text{no birds})]}$$

$$= \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{0.05/[1 - 0.05]} = 4.37$$

$$P(\text{lung cancer}|\text{birds}) = \frac{4.37 \times \frac{0.05}{0.95}}{1 + 4.37 \times \frac{0.05}{0.95}} = 0.187$$

Back to the birds - Low Incidence

What is the probability of lung cancer in a bird keeper if we knew that $P(\text{lung cancer}|\text{no birds}) = 0.05$?

$$OR = \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{P(\text{lung cancer}|\text{no birds})/[1 - P(\text{lung cancer}|\text{no birds})]}$$

$$= \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{0.05/[1 - 0.05]} = 4.37$$

$$P(\text{lung cancer}|\text{birds}) = \frac{4.37 \times \frac{0.05}{0.95}}{1 + 4.37 \times \frac{0.05}{0.95}} = 0.187$$

$$RR = P(\text{lung cancer}|\text{birds})/P(\text{lung cancer}|\text{no birds}) = 0.187/0.05 = 3.74$$

Back to the birds - High Incidence

What is the probability of lung cancer in a bird keeper if we knew that $P(\text{lung cancer}|\text{no birds}) = 0.25$?

$$OR = \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{P(\text{lung cancer}|\text{no birds})/[1 - P(\text{lung cancer}|\text{no birds})]}$$

$$= \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{0.25/[1 - 0.25]} = 4.37$$

Back to the birds - High Incidence

What is the probability of lung cancer in a bird keeper if we knew that $P(\text{lung cancer}|\text{no birds}) = 0.25$?

$$OR = \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{P(\text{lung cancer}|\text{no birds})/[1 - P(\text{lung cancer}|\text{no birds})]}$$

$$= \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{0.25/[1 - 0.25]} = 4.37$$

$$P(\text{lung cancer}|\text{birds}) = \frac{4.37 \times \frac{0.25}{0.75}}{1 + 4.37 \times \frac{0.25}{0.75}} = 0.593$$

Back to the birds - High Incidence

What is the probability of lung cancer in a bird keeper if we knew that $P(\text{lung cancer}|\text{no birds}) = 0.25$?

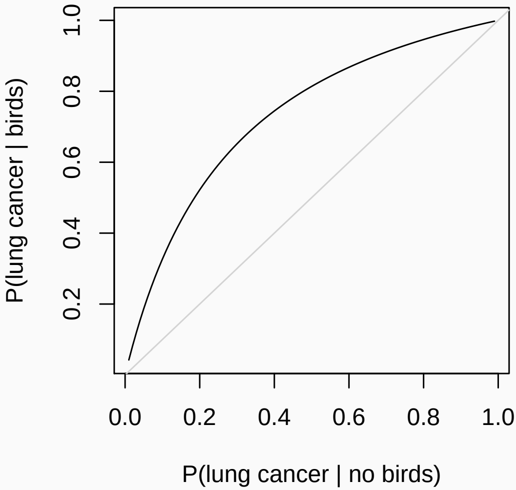
$$OR = \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{P(\text{lung cancer}|\text{no birds})/[1 - P(\text{lung cancer}|\text{no birds})]}$$

$$= \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{0.25/[1 - 0.25]} = 4.37$$

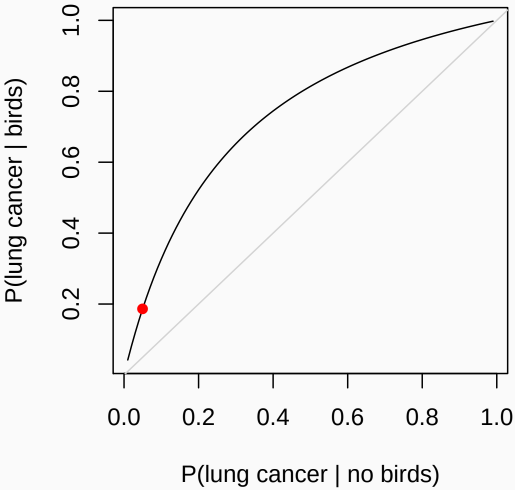
$$P(\text{lung cancer}|\text{birds}) = \frac{4.37 \times \frac{0.25}{0.75}}{1 + 4.37 \times \frac{0.25}{0.75}} = 0.593$$

$RR = P(\text{lung cancer}|\text{birds})/P(\text{lung cancer}|\text{no birds}) = 0.593/0.25 = 2.37$

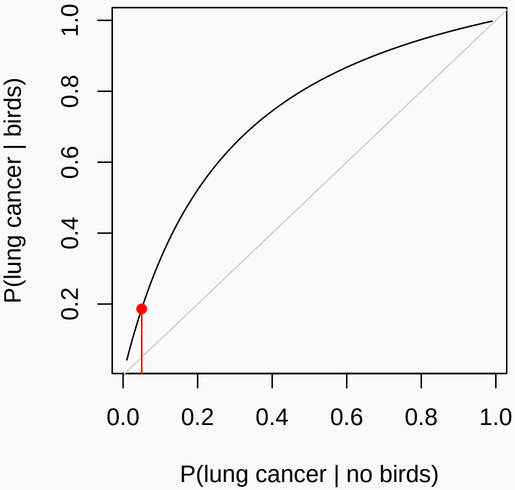
Bird OR Curve



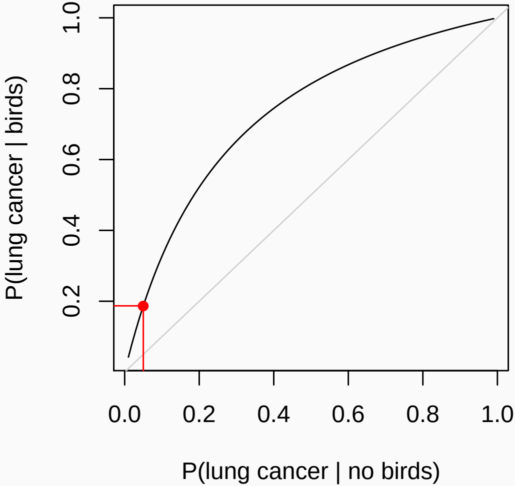
Bird OR Curve



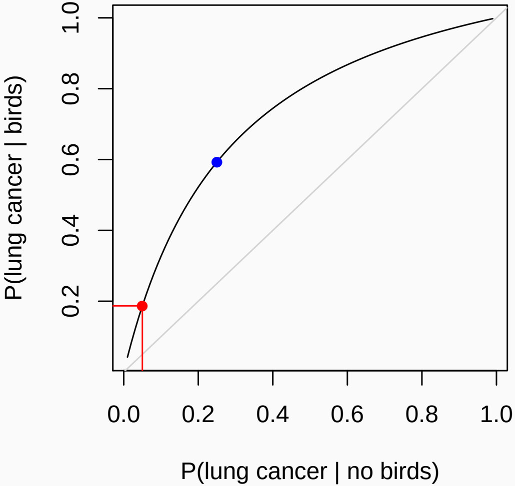
Bird OR Curve



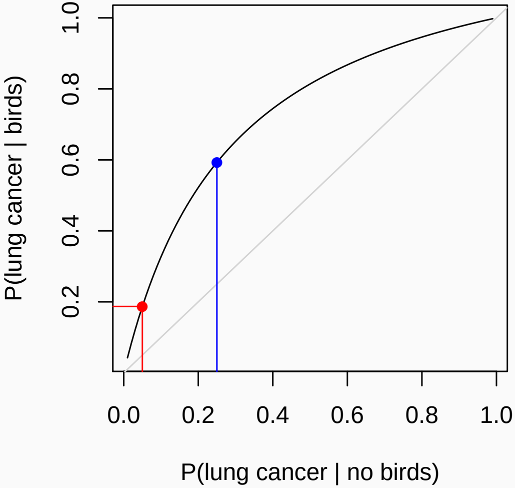
Bird OR Curve



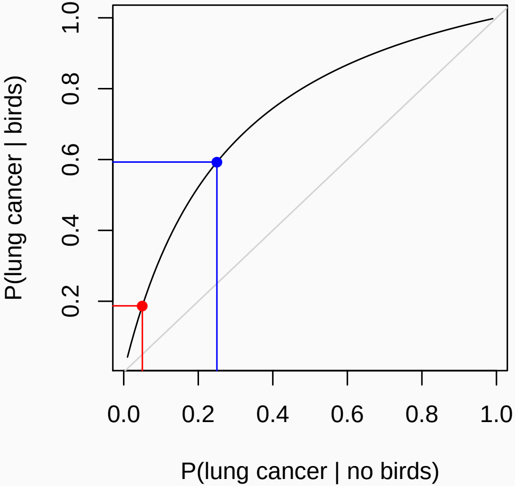
Bird OR Curve



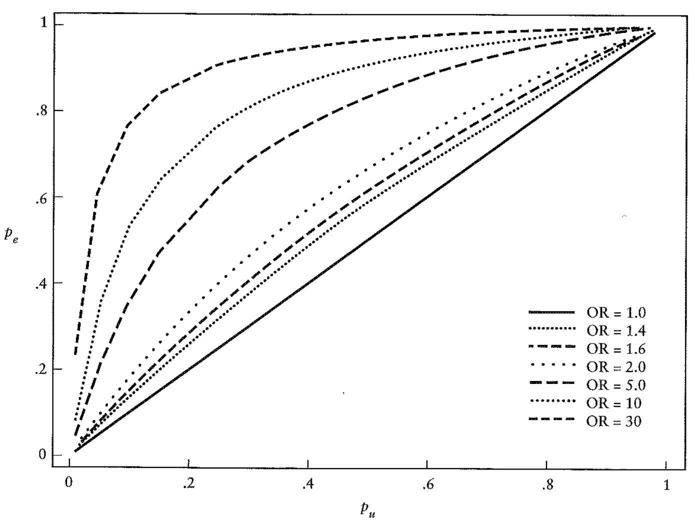
Bird OR Curve



Bird OR Curve



OR Curves



Residuals

Using the logistic regression model we can predict probabilities,

$$\hat{p}_i = \text{logit}^{-1}\left(b_0 + b_1 x_1 + \ldots + b_k x_k\right)$$

Residuals

Using the logistic regression model we can predict probabilities,

$$\hat{p}_i = \text{logit}^{-1}\left(b_0 + b_1 x_1 + \ldots + b_k x_k\right)$$

MLR-like Residual:

$$r_i = y_i - \hat{p}_i$$

Residuals

Using the logistic regression model we can predict probabilities,

$$\hat{p}_i = \text{logit}^{-1}\left(b_0 + b_1 x_1 + \ldots + b_k x_k\right)$$

MLR-like Residual:

$$r_i = y_i - \hat{p}_i$$

Deviance Residual:

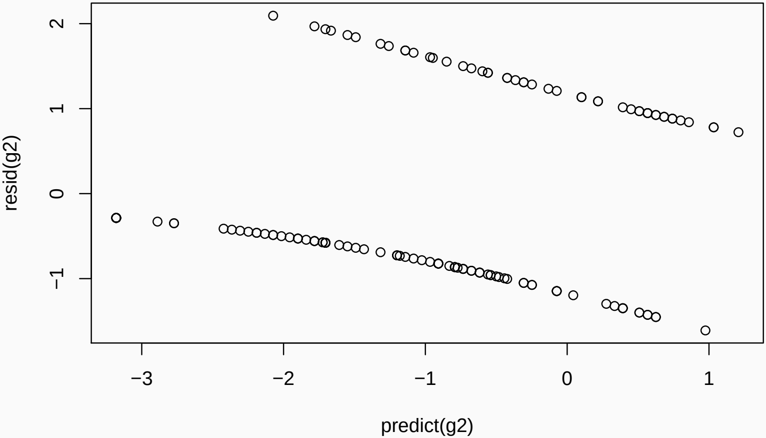
$$r_i = -s_i \sqrt{-2\left(y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)\right)}$$

where

$$s_i = \begin{cases} 1 & \text{if } y_i = 1 \\ -1 & \text{if } y_i = 0 \end{cases}$$

Diagnostics?

```
plot(predict(g2),resid(g2))
```



Diagnostics - Binning

```r
library(arm)
binnedplot(predict(g2),resid(g2))
```