# STAT 401A - Statistical Methods for Research Workers

## Logistic and Poisson regression

Jarad Niemi (Dr. J)

Iowa State University

last updated: December 8, 2014

# Linear regression

The linear regression model

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

where

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

What if $Y_i$ is a binary or a count? Use

- logistic regression or
- Poisson regression.

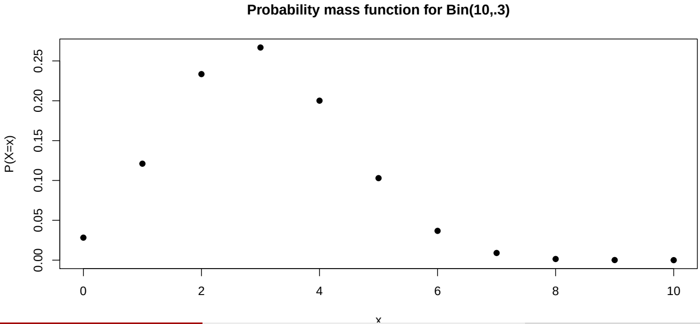
# Binomial distribution

The probability mass function of the binomial distribution is

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

Properties:

- $E[Y] = np$
- $V[Y] = np(1-p)$



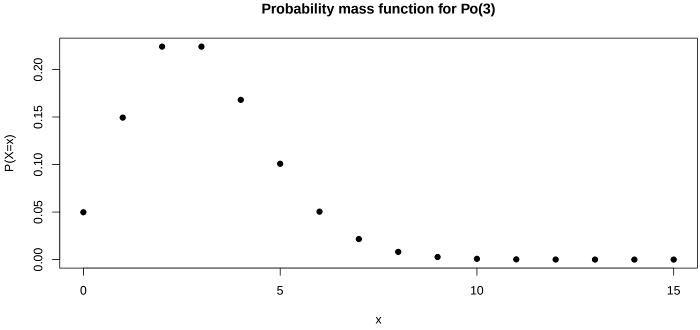
# Poisson distribution

The probability mass function of the Poisson distribution is

$$P(Y = y) = \frac{\mu^y e^{-\mu}}{y!} \qquad \mu > 0, \; y = 0, 1, 2, \ldots$$

Properties:

- $E[Y] = V[Y] = \mu$



# Is Poisson or binomial more appropriate?

- Use Poisson when there is no technical upper limit to how high the count could be.
- Use binomial when you know a technical upper limit, this becomes $n$.

Examples

- Binomial
  - Number of head coin flips out of 10 trials
  - Whether or not somebody has lung cancer
  - Number of species that went extinct since last census
- Poisson
  - Number of cars through an intersection in 10 minutes
  - Number of successful matings for African elephants
  - Number of salamanders found in a 49 $m^2$ area

# Logistic regression

The model

$$Y_i \stackrel{ind}{\sim} Bin(n_i, p_i)$$

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}$$

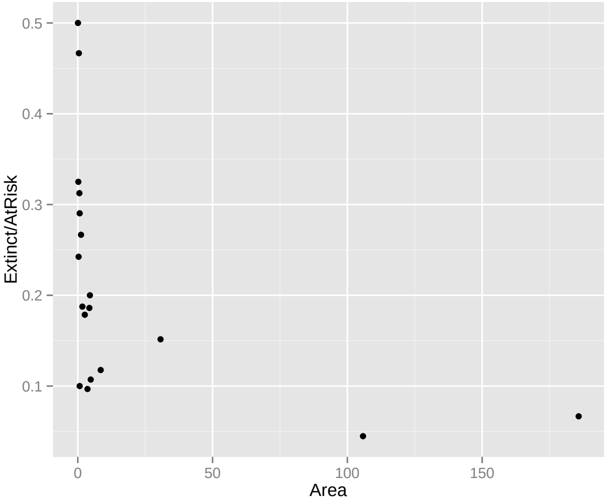
where

- $Y_i$ is an integer from 0 to $n_i$
- Bin refers to the binomial distribution
- Note: if $\text{logit}(p) = \eta$ then $p = \frac{e^\eta}{1+e^\eta}$

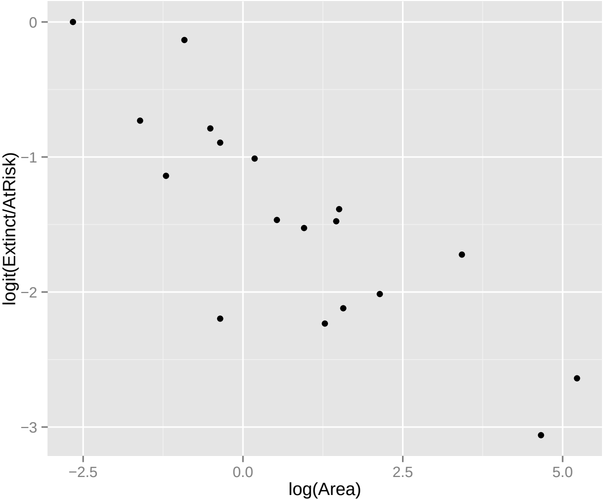
# Number of species that have gone extinct

```
             Island   Area AtRisk Extinct
1        Ulkokrunni 185.80     75       5
2         Maakrunni 105.80     67       3
3         Ristikari  30.70     66      10
4    Isonkivenletto   8.50     51       6
5    Hietakraasukka   4.80     28       3
6          Kraasukka   4.50     20       4
7        Lansiletto   4.30     43       8
8       Pihlajakari   3.60     31       3
9              Tyni   2.60     28       5
10     Tasasenletto   1.70     32       6
11           Raiska   1.20     30       8
12      Pohjanletto   0.70     20       2
13             Toro   0.70     31       9
14       Luusiletto   0.60     16       5
15    Vatunginletto   0.40     15       7
16    Vatunginnokka   0.30     33       8
17        Tiirakari   0.20     40      13
18 Ristikarenletto   0.07      6       3
```

# Is there a relationship between the probability of extinction and island size?



# Is there a relationship between the probability of extinction and island size?



# Parameter estimation

Fit the model

$$Y_i \stackrel{ind}{\sim} \text{Bin}(n_i, p_i) \qquad \text{logit}(p_i) = \beta_0 + \beta_1 X_i$$

where

- $Y_i$ is the number of extinctions on island $i$
- $n_i$ is the total extinctions possible (the number at risk) on island $i$
- $X_{i,1}$ is the logarithm of the area for island $i$

and

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

# Logistic regression in R

```
Call:
glm(formula = cbind(Extinct, AtRisk - Extinct) ~ log(Area), family = "binomial",
    data = case2101)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.71726  -0.67722   0.09726   0.48365   1.49545

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.19620    0.11845 -10.099  < 2e-16 ***
log(Area)   -0.29710    0.05485  -5.416 6.08e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 45.338  on 17  degrees of freedom
Residual deviance: 12.062  on 16  degrees of freedom
AIC: 75.394

Number of Fisher Scoring iterations: 4
                 2.5 %     97.5 %
(Intercept) -1.4330322 -0.9680656
log(Area)   -0.4077542 -0.1922731
```

# Logistic regression parameter interpretation

- At an area size of 1 [log(area)=0], the probability of extinction is estimated to be 23% with a 95% confidence interval of (19%, 38%).

$$\frac{e^{-1.1962}}{1+e^{-1.1962}} = 0.23 \qquad \frac{e^{-1.4283}}{1+e^{-1.4283}} = 0.19 \qquad \frac{e^{-0.9640}}{1+e^{-0.9640}} = 0.38$$

- With all other variables held constant, a unit increase in log(area) is associated with a 0.74 [$= e^{-0.2971}$] multiplicative change in the odds, e.g. from log(area)=0 to log(area)=1

$$\begin{aligned} 0.74\,\text{odds}_0 = \text{odds}_1 &\implies 0.74\tfrac{p_0}{1-p_0} = \tfrac{p_1}{1-p_1} \\ 0.74\tfrac{0.23}{1-0.23} = \tfrac{p_1}{1-p_1} &\implies 0.17 = \tfrac{p_1}{1-p_1} \implies p_1 = 0.15 \end{aligned}$$

- Since we used the logarithm of area, each doubling of area is associated with a multiplicative change in the odds of 0.81 [$= 2^{-0.2971}$] and each 10-fold increase in area is associated with a multiplicative change in the odds of 0.50 [$= 10^{-0.2971}$].

# Logistic regression with multiple explanatory variables

```
Call:
glm(formula = LC ~ FM + SS + BK + AG + YR + CD, family = "binomial",
    data = case2002)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2460  -0.9808   0.4605   0.8333   1.5642

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.09196    1.75465  -0.052 0.958204
FMMale       0.56127    0.53116   1.057 0.290653
SSLow        0.10545    0.46885   0.225 0.822050
BKNoBird     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
---
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: 154.20  on 140  degrees of freedom
AIC: 168.2

Number of Fisher Scoring iterations: 5
```

# Poisson regression

$$Y_i \stackrel{ind}{\sim} Po(\mu_i)$$

$$\log(\mu_i) = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}$$

where

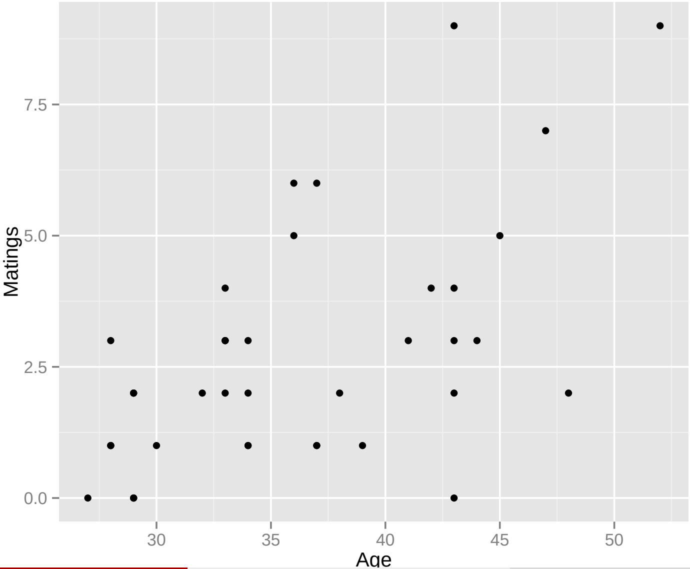
- $Y_i$ is a non-negative integer
- Po refers to the Poisson distribution

# African elephant mating

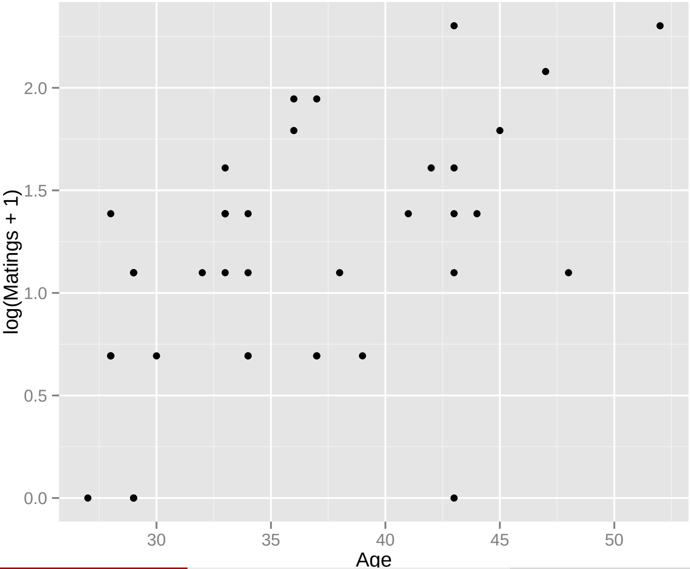
```
head(case2201,10)
```

```
   Age Matings
1   27       0
2   28       1
3   28       1
4   28       1
5   28       3
6   29       0
7   29       0
8   29       0
9   29       2
10  29       2
```

# Is there a relationship between Matings and Age?



# Is there a relationship between Matings and Age?



# Poisson regression

```
m = glm(Matings~Age, data=case2201, family="poisson")
summary(m)



Call:
glm(formula = Matings ~ Age, family = "poisson", data = case2201)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.80798  -0.86137  -0.08629   0.60087   2.17777

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.58201    0.54462  -2.905  0.00368 **
Age          0.06869    0.01375   4.997 5.81e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 75.372  on 40  degrees of freedom
Residual deviance: 51.012  on 39  degrees of freedom
AIC: 156.46

Number of Fisher Scoring iterations: 5
```

# Shifting the intercept

```
mAge = median(case2201$Age)
m = glm(Matings~I(Age-mAge), data=case2201, family="poisson")
summary(m)



Call:
glm(formula = Matings ~ I(Age - mAge), family = "poisson", data = case2201)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.80798  -0.86137  -0.08629   0.60087   2.17777

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.75355    0.11761   6.407 1.48e-10 ***
I(Age - mAge)  0.06869    0.01375   4.997 5.81e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 75.372  on 40  degrees of freedom
Residual deviance: 51.012  on 39  degrees of freedom
AIC: 156.46

Number of Fisher Scoring iterations: 5
```

# Shifting the intercept

```
confint(m)
```

```
                   2.5 %     97.5 %
(Intercept)   0.51288577 0.97468553
I(Age - mAge) 0.04167776 0.09563762
```

# Poisson regression parameter interpretation

- At the median age of 34, the expected number of matings is 2.1 $[= e^{0.7535}]$ with a 95% confidence interval of (1.7,2.7).
- With all other variables held constant, for each year increase in age there is a multiplicative effect on the expected number of matings of 1.07 $[= e^{0.0687}]$ with a 95% confidence interval of (1.04,1.10), e.g.

$$\begin{array}{llll}
\mu(\text{age} = 35) & = \mu(\text{age} = 34) \cdot 1.07 & = 2.1 \cdot 1.07 & = 2.28 \\
\mu(\text{age} = 44) & = \mu(\text{age} = 34) \cdot 1.07^{10} & = 2.1 \cdot 1.07^{10} & = 4.2
\end{array}$$

# Drop-in-deviance test

To test whether a set of explanatory variables should be in the model, a drop-in-deviance test should be used. This is analogous to the extra-sums-of-squares F-test for normally distributed data.

The deviance is $-2\log L(\hat{\theta}_{MLE})$. The drop-in-deviance test statistic is

$$Deviance_{reduced} - Deviance_{full}$$

which, if the null hypothesis is true, has a $\chi^2_v$ where $v$ is the difference in the number of parameters between the full and reduced models.

# Drop-in deviance test for age squared

Fit the model with only age (reduced model):

| Criterion | DF | Value | Value/DF |
|---|---|---|---|
| Deviance | 39 | 51.0116 | 1.3080 |

Fit the model with age and age squared (full model):

| Criterion | DF | Value | Value/DF |
|---|---|---|---|
| Deviance | 38 | 50.8262 | 1.3375 |

Drop-in-deviance test:

$$\text{Dev}_{red} - \text{Dev}_{full} = 51.0116 - 50.8262 = 0.1854$$

compare this to a $\chi^2_1$, i.e.

$$P(\chi^2_1 > 0.1854) = 0.67$$

# Drop-in-deviance test

```
anova(glm(Matings~Age, data=case2201, family="poisson"),
      glm(Matings~Age + I(Age^2), data=case2201, family="poisson"),
      test="Chi")


Analysis of Deviance Table

Model 1: Matings ~ Age
Model 2: Matings ~ Age + I(Age^2)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        39     51.012
2        38     50.826  1  0.18544   0.6667
```

# Poisson regression with multiple explanatory variables

```
summary(m <- glm(Salamanders~PctCover+ForestAge, data=case2202, family="poisson"))


Call:
glm(formula = Salamanders ~ PctCover + ForestAge, family = "poisson",
    data = case2202)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9484  -1.3649  -0.7072   0.6243   3.8417

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.483e+00  4.573e-01  -3.244  0.00118 **
PctCover     3.249e-02  5.735e-03   5.666 1.46e-08 ***
ForestAge   -2.111e-05  4.981e-04  -0.042  0.96620
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 190.22  on 46  degrees of freedom
Residual deviance: 121.30  on 44  degrees of freedom
AIC: 212.36

Number of Fisher Scoring iterations: 5
```

# Drop-in-deviance tests

```
# Perform all the drop-in-deviance tests
drop1(m, test="Chi")


Single term deletions

Model:
Salamanders ~ PctCover + ForestAge
          Df Deviance    AIC    LRT Pr(>Chi)
<none>         121.30 212.36
PctCover   1   170.65 259.70 49.342 2.15e-12 ***
ForestAge  1   121.31 210.36  0.002   0.9662
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```