

## Set S02 - Poisson Regression

STAT 401 (Engineering) - Iowa State University

April 19, 2017

# Linear regression

For continuous  $Y_i$ , we have linear regression

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

For binary or count with an upper maximum  $Y_i$ , we have logistic regression

$$Y_i \stackrel{ind}{\sim} \text{Bin}(n_i, \theta_i)$$
$$\log \left( \frac{\theta_i}{1-\theta_i} \right) = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}$$

What if  $Y_i$  is a count without a maximum?

# Poisson regression

Let  $Y_i \in \{0, 1, 2, \dots\}$  be a count (typically over some amount of time or some amount of space) with associated explanatory variables  $X_{i,1}, X_{i,2}, \dots$

Then a Poisson regression model is

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

and

$$\log(\lambda_i) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \dots + \beta_p X_{i,p}$$

# Interpretation

When all explanatory variables are zero, then

$$E[Y_i | X_{i,1} = 0, \dots, X_{i,p} = 0] = \lambda_i = e^{\beta_0}$$

thus  $\beta_0$  determines the **expected response when all explanatory variables are zero**.

More generally,

$$E[Y_i | X_{i,1} = x_1, \dots, X_{i,p} = x_p] = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}.$$

If  $X_{i,1}$  increases by one unit, we have

$$E[Y_i | X_{i,1} = x_1 + 1, \dots, X_{i,p} = x_p] = e^{\beta_0 + \beta_1(x_1 + 1) + \dots + \beta_p x_p} = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p} e^{\beta_1}$$

Thus

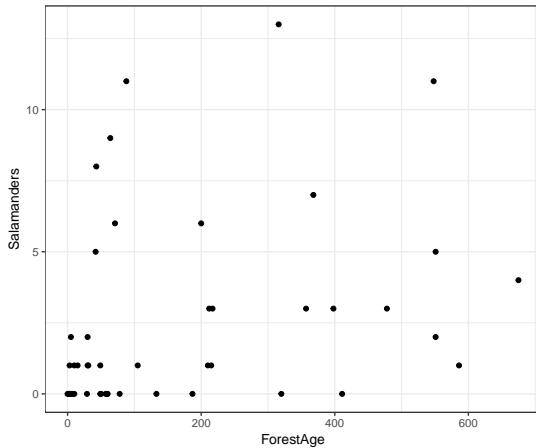
$$\frac{E[Y_i | X_{i,1} = x_1 + 1, \dots, X_{i,p} = x_p]}{E[Y_i | X_{i,1} = x_1, \dots, X_{i,p} = x_p]} = e^{\beta_1}.$$

Thus  $e^{\beta_p}$  is the **multiplicative effect on the mean response for a one unit increase in the associated explanatory variable when holding all other explanatory variables constant**.

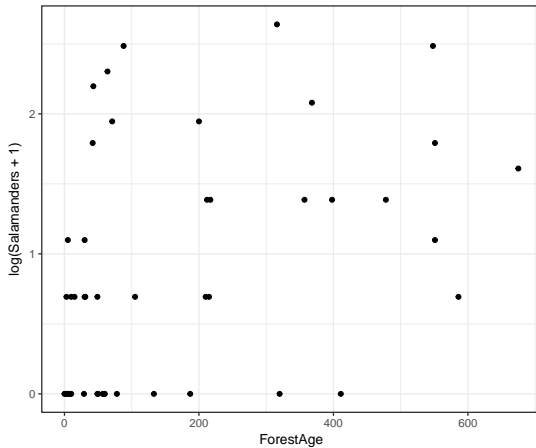
# Salamander habitat

*The Del Norte Salamander (plethodon elongates) is a small (57 cm) salamander found among rock rubble, rock outcrops and moss-covered talus in a narrow range of northwest California. To study the habitat characteristics of the species and particularly the tendency of these salamanders to reside in dwindling old-growth forests, researchers selected 47 sites from plausible salamander habitat in national forest and parkland. Randomly chosen grid points were searched for the presence of a site with suitable rocky habitat. At each suitable site, a 7 metre by 7 metre search area was examined for the number of salamanders it contained.*

```
ggplot(Sleuth3::case2202, aes(ForestAge, Salamanders)) +  
  geom_point() +  
  theme_bw()
```



```
ggplot(Sleuth3::case2202, aes(ForestAge, log(Salamanders+1))) +  
  geom_point() +  
  theme_bw()
```



# Analysis

```
m <- glm(Salamanders ~ ForestAge,
          data = Sleuth3::case2202,
          family = "poisson")

summary(m)
```

Call:  
 glm(formula = Salamanders ~ ForestAge, family = "poisson", data = Sleuth3::case2202)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.6970	-1.8539	-0.7987	0.2144	4.4582

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.5040207	0.1401385	3.597	0.000322 ***
ForestAge	0.0019151	0.0004155	4.609	4.05e-06 ***

---  
 Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

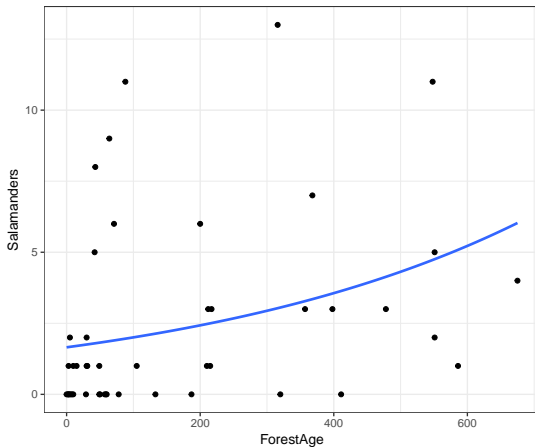
(Dispersion parameter for poisson family taken to be 1)

Null deviance: 190.22 on 46 degrees of freedom  
 Residual deviance: 170.65 on 45 degrees of freedom  
 AIC: 259.7

Number of Fisher Scoring iterations: 6



```
ggplot(Sleuth3::case2202, aes(ForestAge, Salamanders)) +  
  geom_point() +  
  stat_smooth(method="glm",  
             se=FALSE,  
             method.args = list(family="poisson")) +  
  theme_bw()
```



# Salamander habitat (cont.)

```
m <- glm(Salamanders ~ ForestAge * PctCover,
  data = Sleuth3::case2202,
  family = "poisson")

summary(m)
```

Call:

```
glm(formula = Salamanders ~ ForestAge * PctCover, family = "poisson",
  data = Sleuth3::case2202)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.9710	-1.3237	-0.7378	0.6114	3.9136

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.388e+00	5.038e-01	-2.754	0.00588 **
ForestAge	-2.812e-03	6.799e-03	-0.414	0.67918
PctCover	3.147e-02	6.145e-03	5.121	3.04e-07 ***
ForestAge:PctCover	3.141e-05	7.625e-05	0.412	0.68033

---

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: 121.13 on 43 degrees of freedom  
 AIC: 214.19

Number of Fisher Scoring iterations: 6

# Offset

If not all counts are based on the same amount of time or space, we need to account for the amount of time or space used. To do this, we can include an **offset**.

Let  $T_i$  represent the amount of time or space, then a Poisson regression model with an offset is

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

and

$$\log(\lambda_i) = \log(T_i) + \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p}.$$

The offset is  $\log(T_i)$  and can be thought of as an explanatory variable with a known coefficient of 1. Note that

$$\log E[Y_i/T_i] = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p}$$

so we are effectively modeling the **rate**.

# Airline crash data

When considering airline crash data, we need to account for the fact that airlines are (typically) flying more miles year over year.

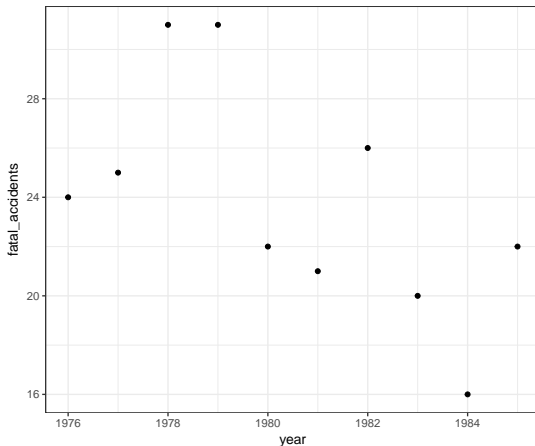
```
airline = data.frame(year=1976:1985,
                     fatal_accidents = c(24,25,31,31,22,21,26,20,16,22),
                     passenger_deaths = c(734,516,754,877,814,362,764,809,223,1066),
                     death_rate = c(0.19,0.12,0.15,0.16,0.14,0.06,0.13,0.13,0.03,0.15)) %>%
  mutate(miles_flown = passenger_deaths / death_rate)
```

```
airline
```

	year	fatal_accidents	passenger_deaths	death_rate	miles_flown
1	1976	24	734	0.19	3863.158
2	1977	25	516	0.12	4300.000
3	1978	31	754	0.15	5026.667
4	1979	31	877	0.16	5481.250
5	1980	22	814	0.14	5814.286
6	1981	21	362	0.06	6033.333
7	1982	26	764	0.13	5876.923
8	1983	20	809	0.13	6223.077
9	1984	16	223	0.03	7433.333
10	1985	22	1066	0.15	7106.667

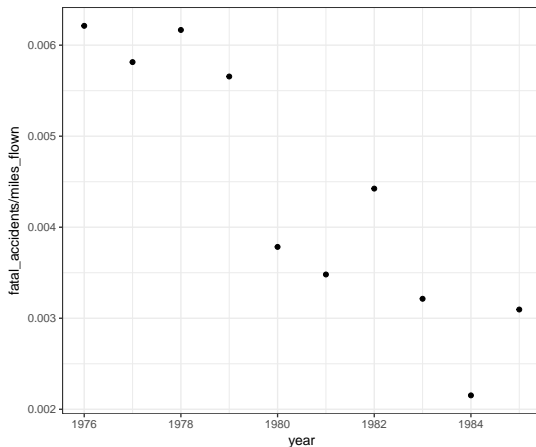
# Visualize airline crash data

```
ggplot(airline, aes(year, fatal_accidents)) +  
  geom_point() +  
  scale_x_continuous(breaks= scales::pretty_breaks()) +  
  theme_bw()
```



# Visualize airline crash data

```
ggplot(airline, aes(year, fatal_accidents/miles_flown)) +  
  geom_point() +  
  scale_x_continuous(breaks= scales::pretty_breaks()) +  
  theme_bw()
```



# Offset in R

```
m <- glm(fatal_accidents ~ year + offset(log(miles_flown)),
        data = airline,
        family = "poisson")

summary(m)
```

Call:

```
glm(formula = fatal_accidents ~ year + offset(log(miles_flown)),
    family = "poisson", data = airline)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.2829	-0.5813	-0.1230	0.7254	1.0211

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	201.32854	45.62354	4.413	1.02e-05 ***
year	-0.10442	0.02304	-4.532	5.84e-06 ***

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 26.133 on 9 degrees of freedom  
Residual deviance: 5.457 on 8 degrees of freedom  
AIC: 59.426

Number of Fisher Scoring iterations: 4

# Offset in R

```
m <- glm(fatal_accidents ~ year + log(miles_flown),
         data = airline,
         family = "poisson")

confint(m) # No evidence coefficient for log(miles_flown) is incompatible with 1
```

	2.5 %	97.5 %
(Intercept)	-134.5369352	415.57465599
year	-0.2192575	0.07628503
log(miles_flown)	-1.6508503	2.64154996



# Likelihood ratio tests

To compare nested generalized linear models, we use likelihood ratio tests.

Suppose we have a model  $p(y|\theta)$  for our data and two hypotheses

- $H_0 : \theta = \theta_0$  and
- $H_A : \theta \neq \theta_0$ .

Then the likelihood is  $L(\theta) = p(y|\theta)$  and the likelihood ratio statistics is

$$\lambda = \frac{L(\theta_0)}{L(\hat{\theta}_{MLE})} = \frac{p(y|\theta_0)}{p(y|\hat{\theta}_{MLE})}.$$

Asymptotically (as we have more data) under  $H_0$ ,

$$\text{deviance} = -2 \log(\lambda) \xrightarrow{d} \chi_v^2$$

where  $\chi_v^2$  is a chi-squared distribution with  $v$  degrees of freedom and  $v$  is the number of parameters in  $\theta$ , i.e. the number of parameters set to a known value.

The pvalue is

$$pvalue = P(\chi_v^2 > -2 \log(\lambda)).$$

# $\chi^2$ -distributions

If  $X \sim \chi_v^2$ , then  $X$  has a chi-squared distribution with  $v$  degrees of freedom.

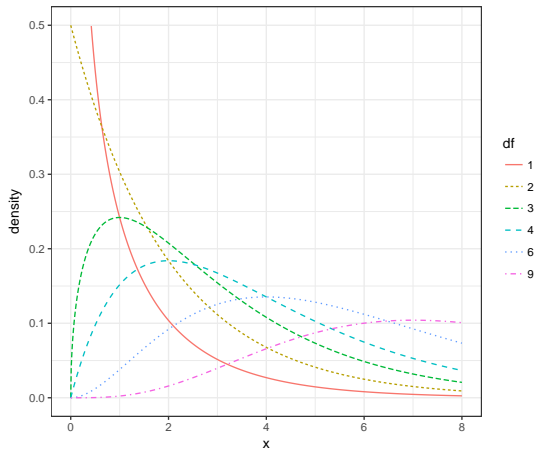
The probability density function is

$$p(x) = \frac{1}{2^{\frac{v}{2}} \Gamma\left(\frac{v}{2}\right)} x^{\frac{v}{2}-1} e^{-\frac{x}{2}}$$

with support  $x \in [0, \infty)$ . We have

$$\begin{aligned} E[X] &= v \\ \text{Var}[X] &= 2v. \end{aligned}$$

# $\chi^2$ -distribution visualization



# Likelihood ratio tests in R

```
m <- glm(Salamanders ~ ForestAge * PctCover,
  data = Sleuth3::case2202,
  family = "poisson")

anova(m, test="Chi")
```

Analysis of Deviance Table

Model: poisson, link: log

Response: Salamanders

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			46	190.22	
ForestAge	1	19.573	45	170.65	9.681e-06 ***
PctCover	1	49.342	44	121.30	2.150e-12 ***
ForestAge:PctCover	1	0.170	43	121.13	0.6797

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1