R05 - Poisson and Logistic Regression

HCI/PSYCH 522
Iowa State University

March 24, 2022

Overview

- Poisson Regression
  - Poisson Distribution
  - Poisson Regression Model
  - O-ring incidents as a function of temperature
- Logistic Regression
  - Bernoulli Distribution
  - Logistic Regression Model
  - Probability of staying with as a function of Vitamin C

Poisson Distribution

Let Y be a random variable that is a count over some amount of time or space where the count has no obvious upper maximum. For example,

- Number of visitors to a website in the next hour
- Number of chatbot uses for an individual during registration
- Number of clicks in a certain region of the screen during a game

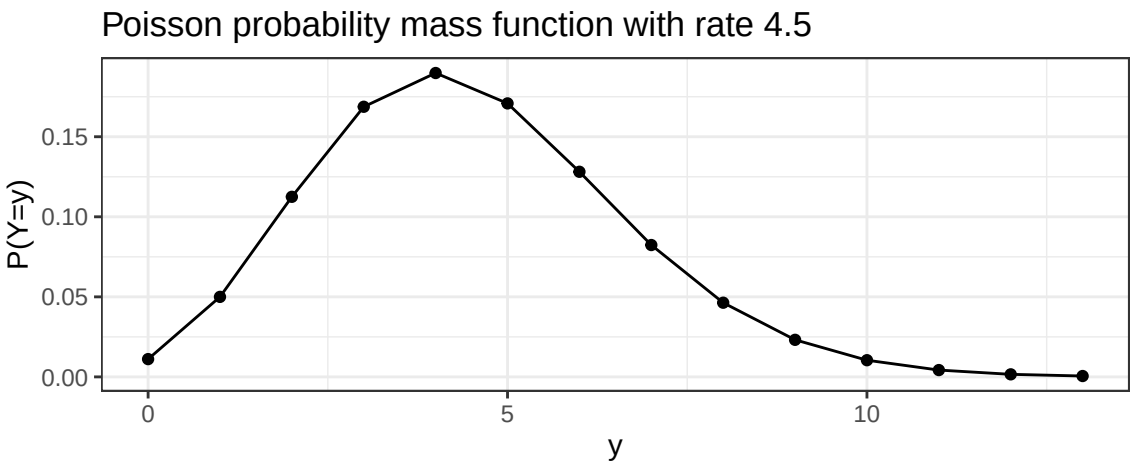
Then Y has a Poisson distribution with rate parameter $\lambda > 0$ and we write $Y \sim Po(\lambda)$. The probability mass function (pmf) is

$$P(Y = y) = \frac{\lambda^y e^{\lambda}}{y!} \quad \text{for } y = 0, 1, 2, \ldots$$

and we can find that

$$E[Y] = \lambda \quad \text{and} \quad Var[Y] = \lambda.$$

Poisson pmf


Poisson probability mass function with rate 4.5
P(Y=y)
0.15
0.10
0.05
0.00
0
5
10
y

Poisson rate changes according to some independent variable

Suppose the Poisson rate parameter changes due to some other variable. For example,

- Time of day
- Sex/gender
- Length of a game

A Poisson regression model allows the rate to change according to these independent variables.

Poisson Regression Model

For observation i, let

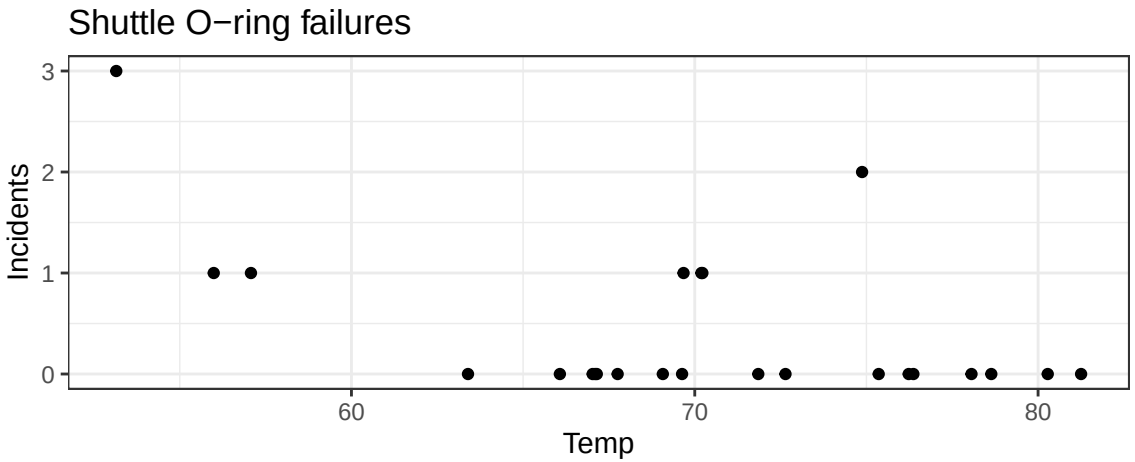
- Y_i be the count and
- X_i be the value of an independent variable.

The (simple) Poisson regression model is

$$Y_i \overset{ind}{\sim} Po(\lambda_i) \quad \text{where} \quad \log(\lambda_i) = \beta_0 + \beta_1 X_i$$

In this model, $100(e^{\beta_1} - 1)$ will be the percent change in mean salary per unit increase in X.

Number of O-ring problems



Poisson regression for O-rings

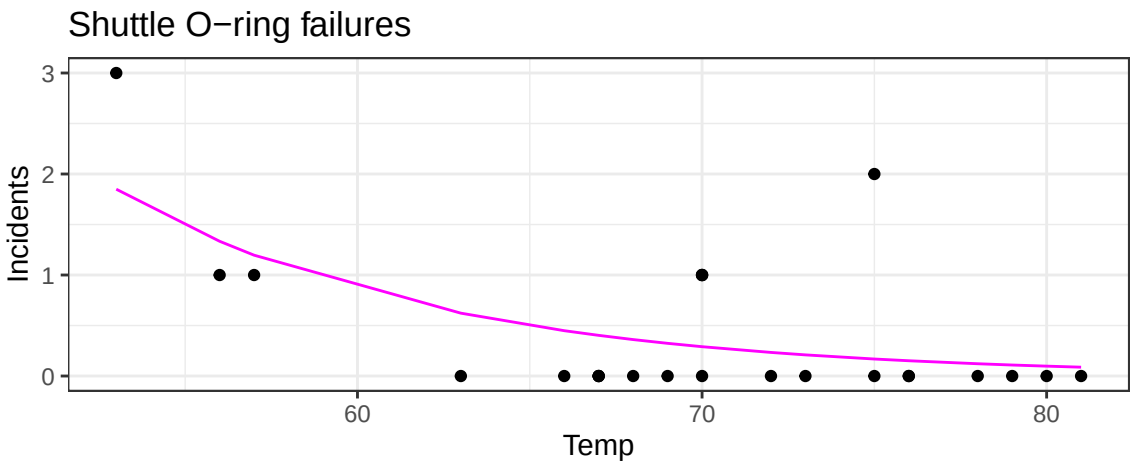
```
m <- glm(Incidents ~ Temp, data = ex2223, family = poisson)
summary(m)

##
## Call:
## glm(formula = Incidents ~ Temp, family = poisson, data = ex2223)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1155  -0.8158  -0.5495  -0.2731   2.4972
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.38844    2.50849   2.547  0.01087 *
## Temp        -0.10894    0.03937  -2.767  0.00566 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
```

Poisson regression for O-rings

```
ci <- 100*(exp(confint(m)[2,])-1)
ci

##      2.5 %     97.5 %
## -17.259598  -3.152303
```

Manuscript statement:
Each one degree Fahrenheit increase in temperature is associated with the mean number of O-ring incidents decreasing by (3, 17)%.

Number of O-ring problems



Bernoulli Distribution

Let Y be a random variable that indicates "success". For example,

- Winning a game
- Having fewer than 3 errors on a task
- Clicking on an ad

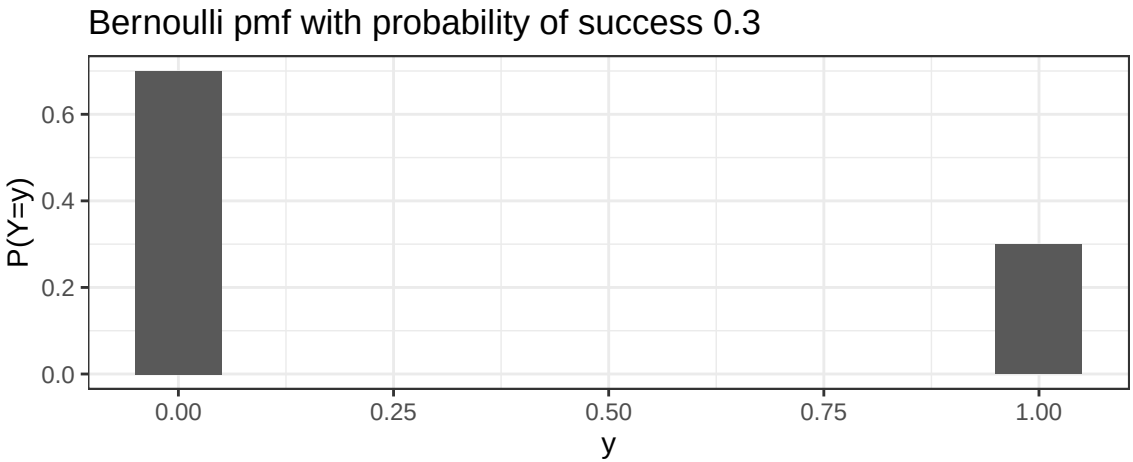
Then Y has a Bernoulli distribution with probability of success $0 < \theta < 1$ and we write $Y \sim Ber(\theta)$. The probability mass function is

$$P(Y = y) = \theta^y (1 - \theta)^{1-y} \quad \text{for } y = 0, 1$$

and we can find that

$$E[Y] = \theta \quad \text{and} \quad Var[Y] = \theta(1 - \theta).$$

Bernoulli pmf



Bernoulli probability of success

Suppose the Bernoulli probability of success changes due to some other variable. For example,

- Time of day
- Sex/gender
- Length of a game

A logistic regression model allows the probability of success to change according to these independent variables.

Logistic regression model

For observation i, let

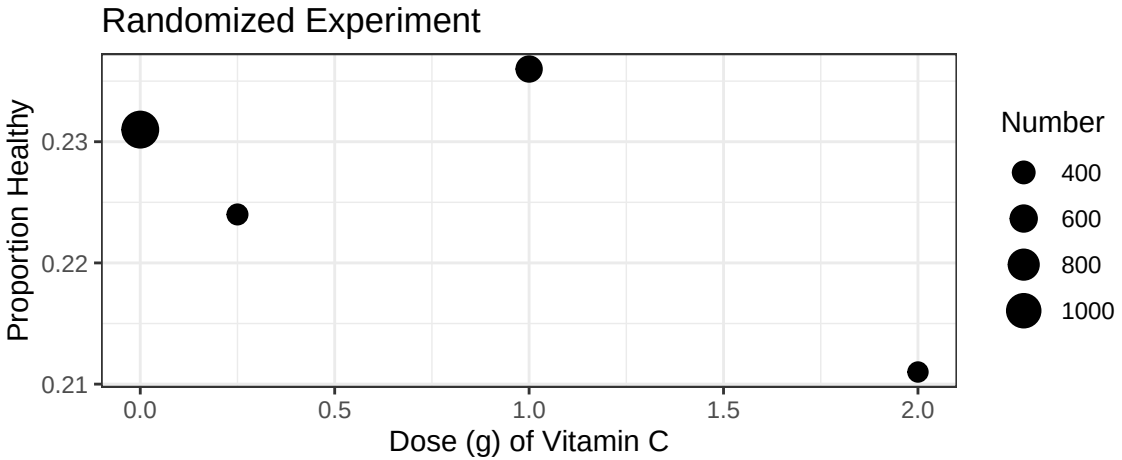
- Y_i be the indicator of success and
- X_i be the value of an independent variable.

The (simple) logistic regression model is

$$Y_i \overset{ind}{\sim} Ber(\theta_i) \quad \text{where} \quad \log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_1 X_i$$

In this model, $100 * (e^{\beta_1} - 1)$ is the percent change in the odds $\left(\frac{\theta}{1-\theta}\right)$ of success.

Vitamin C effect on incidence of colds



Logistic regression model for proportion healthy

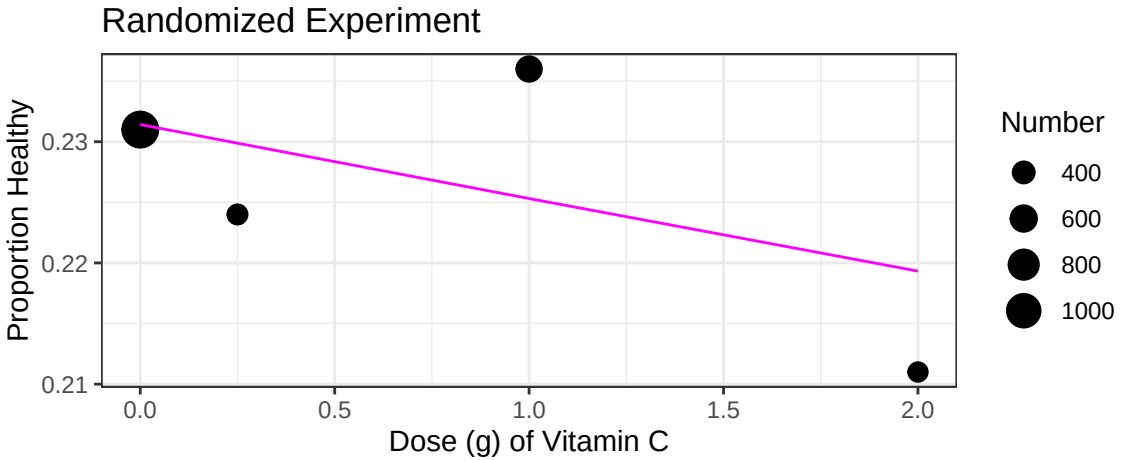
```
m <- glm(cbind(WithoutIllness, Number-WithoutIllness) ~ Dose,
         data = ex2113, family = binomial)
summary(m)

##
## Call:
## glm(formula = cbind(WithoutIllness, Number - WithoutIllness) ~
##     Dose, family = binomial, data = ex2113)
##
## Deviance Residuals:
##        1         2         3         4
## -0.06857  -0.27405   0.57021  -0.35303
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.20031    0.06167 -19.464   <2e-16 ***
## Dose        -0.03465    0.07113  -0.487    0.626
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Logistic regression model for proportion healthy

```
##     2.5 %    97.5 %
## -16.09864  10.89977
```

Manuscript statement:
Each gram increase in Vitamin C causes the odds of staying healthy to change by (-16, 11)%.

Vitamin C effect on incidence of colds



Summary

- Poisson regression
  - Dependent variable is a count with no clear upper maximum
  - Interpret $100(e^{\beta_1} - 1)$ as the percent change in rate
- Logistic regression
  - Dependent variable is a count with clear upper maximum
  - Interpret $100(e^{\beta_1} - 1)$ as the percent change in odds
- Causal inference
  - Observational study $\rightarrow$ association
  - Randomized experiment $\rightarrow$ cause