

Lecture 7

Logistic Regression and Poisson Regression

Reading: Faraway, 2016 Chapters 2, 5

DSA 8020 Statistical Methods II

February 15-19, 2021

Whitney Huang
Clemson University

1 Logistic Regression

2 Poisson Regression

3 Generalized Linear Model

A Motivating Example: Horseshoe Crab Malting [Brockmann, 1996, Agresti, 2013]



sat	y	weight	width
8	1	3.05	28.3
0	0	1.55	22.5
9	1	2.30	26.0
0	0	2.10	24.8
4	1	2.60	26.0
0	0	2.10	23.8
0	0	2.35	26.5
0	0	1.90	24.7
0	0	1.95	23.7
0	0	2.15	25.6

Source: <https://www.britannica.com/story/horseshoe-crab-a-key-player-in-ecology-medicine-and-more>

We are going to use this data set to illustrate **logistic regression**. The response variable is y : whether there are males clustering around the female

Let $P(Y = 1) = \pi \in [0, 1]$, and x be the predictor (weight in the previous example). In SLR we have

$$\pi(x) = \beta_0 + \beta_1 x,$$

which will lead to invalid estimate of π (i.e., > 1 or < 0).

Logistic Regression

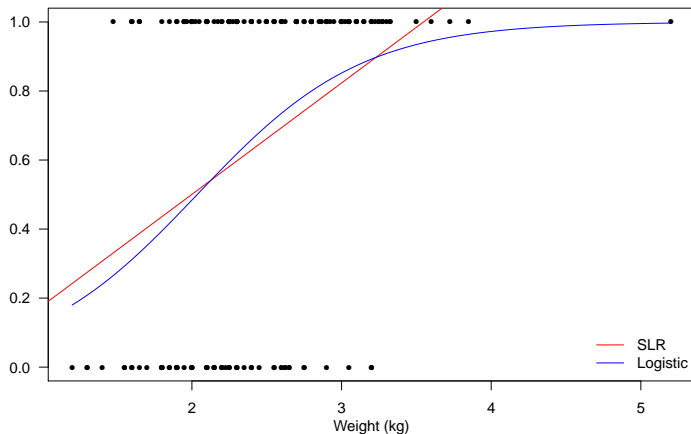
$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x.$$

- $\log\left(\frac{\pi}{1-\pi}\right)$: the log-odds or the logit

- $\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \in (0, 1)$

Logistic Regression Fit

$$\hat{\pi}(x) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}}, \hat{\beta}_0 = -3.6947(0.8802), \hat{\beta}_1 = 1.8151(0.3767).$$



Logistic Regression

Poisson Regression

Generalized Linear
Model

- Similar to SLR, Sign of β_1 indicates whether $\pi(x) \uparrow$ or \downarrow as $x \uparrow$
- If $\beta_1 = 0$, then $\pi(x) = e^{\beta_0} / (1 + e^{\beta_0})$ is a constant w.r.t x (i.e., π does not depend on x)
- Curve can be approximated at fixed x by straight line to describe rate of change: $\frac{d\pi(x)}{dx} = \beta_1 \pi(x)(1 - \pi(x))$
- $\pi(-\beta_0/\beta_1) = 0.5$, and $1/\beta_1 \approx$ the distance of x values with $\pi(x) = 0.5$ and $\pi(x) = 0.75$ (or $\pi(x) = 0.25$)

Recall $\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x$, we have the odds

$$\frac{\pi(x)}{1 - \pi(x)} = \exp(\beta_0 + \beta_1 x).$$

If we increase x by 1 unit, the the odds becomes

$$\exp(\beta_0 + \beta_1(x + 1)) = \exp(\beta_1) \times \exp(\beta_0 + \beta_1 x).$$

$$\Rightarrow \frac{\text{Odds at } x+1}{\text{Odds at } x} = \exp(\beta_1), \forall x$$

Example: In the horseshoe crab example, we have $\hat{\beta}_1 = 1.8151 \Rightarrow e^{1.8151} = 6.14 \Rightarrow$ **Estimated odds of satellite multiply by 6.1 for 1 kg increase in weight.**

In logistic regression we use **method of maximum likelihood** to estimate the parameters:

- **Statistical model:** $Y_i \sim \text{Bernoulli}(\pi(x_i))$ where

$$\pi(x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}.$$

- **Likelihood function:** We can write the joint probability density of the data $\{x_i, y_i\}_{i=1}^n$ as

$$\prod_{i=1}^n \left[\frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right]^{y_i} \left[\frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)} \right]^{(1-y_i)}.$$

We treat this as a function of parameters (β_0, β_1) given data.

- **Maximum likelihood estimate:** The maximizer $\hat{\beta}_0, \hat{\beta}_1$ is the maximum likelihood estimate (MLE). This maximization can only be solved numerically.

Horseshoe Crab Logistic Regression Fit

```
> logitFit <- glm(y ~ weight, data = crab, family = "binomial")  
> summary(logitFit)
```

Call:

```
glm(formula = y ~ weight, family = "binomial", data = crab)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.1108	-1.0749	0.5426	0.9122	1.6285

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.6947	0.8802	-4.198	2.70e-05 ***
weight	1.8151	0.3767	4.819	1.45e-06 ***

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 225.76 on 172 degrees of freedom
Residual deviance: 195.74 on 171 degrees of freedom
AIC: 199.74

Number of Fisher Scoring iterations: 4

A 95% confidence interval of the parameter β_i is

$$\hat{\beta}_i \pm z_{0.025} \times \text{SE}(\hat{\beta}_i), \quad i = 0, 1$$

Horseshoe Crab Example

A 95% (Wald) confidence interval of β_1 is

$$1.8151 \pm 1.96 \times 0.3767 = [1.077, 2.553]$$

Therefore a 95% CI of e^{β_1} , the **multiplicative effect** on odds of 1-unit increase in x , is

$$[e^{1.077}, e^{2.553}] = [2.94, 12.85]$$

Null and Alternative Hypotheses:

$H_0 : \beta_1 = 0 \Rightarrow Y$ is independent of $X \Rightarrow \pi(x)$ is a constant

$H_a : \beta_1 \neq 0$

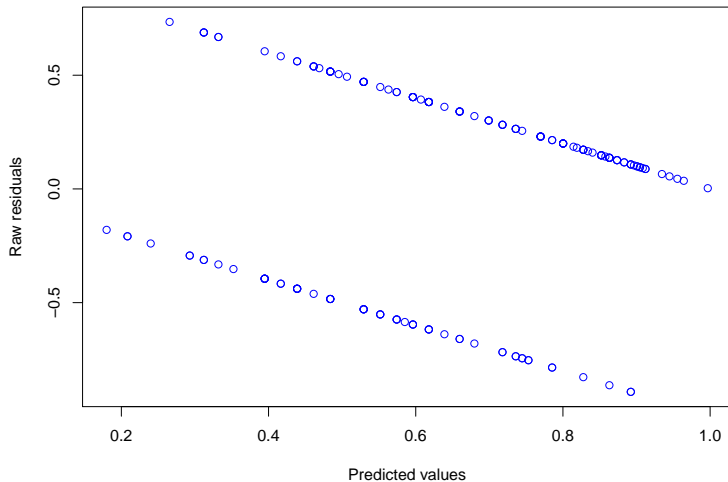
Test Statistics:

$$z_{obs} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{1.8151}{0.3767} = 4.819.$$

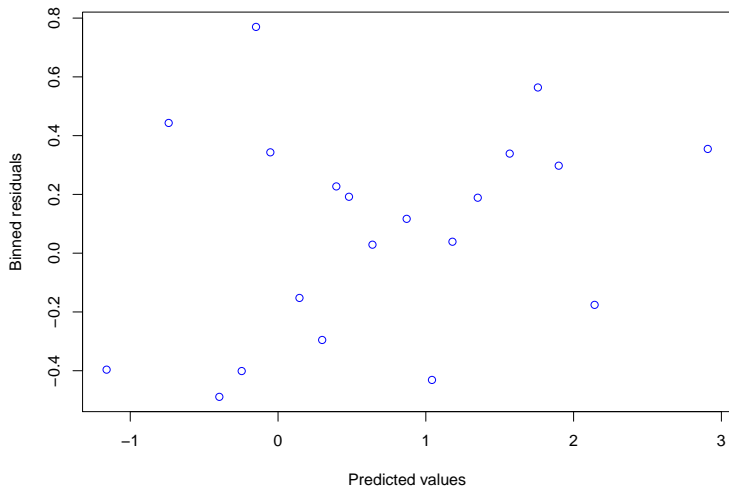
P-value = 1.45×10^{-6}

We have sufficient evidence that `weight` has positive effect on π , the probability of having satellite male horse-shoe crabs

Diagnostic: Raw Residual Plot



Diagnostic: Binned Residual Plot



Model Selection

```
> logitFit2 <- glm(y ~ weight + width, data = crab, family = "binomial")  
> step(logitFit2)
```

Start: AIC=198.89

y ~ weight + width

	Df	Deviance	AIC
- weight	1	194.45	198.45
<none>		192.89	198.89
- width	1	195.74	199.74

Step: AIC=198.45

y ~ width

	Df	Deviance	AIC
<none>		194.45	198.45
- width	1	225.76	227.76

Call: glm(formula = y ~ width, family = "binomial", data = crab)

Coefficients:

(Intercept)	width
-12.3508	0.4972

Degrees of Freedom: 172 Total (i.e. Null); 171 Residual

Null Deviance: 225.8

Residual Deviance: 194.5 AIC: 198.5

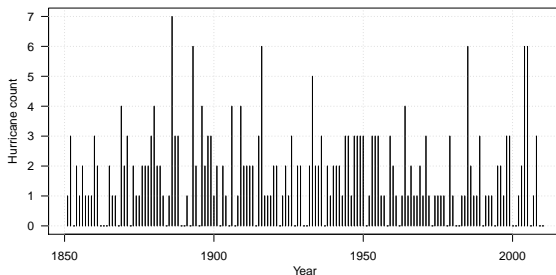
Daily COVID-19 Cases in South Carolina



Each day shows new cases reported since the previous day · Updated less than 19 hours ago ·

Source: [The New York Times](#) · [About this data](#)

Number of landfalling hurricanes per hurricane season



So far we have talked about:

- Linear regression: $Y = \beta_0 + \beta_1 x + \varepsilon, \varepsilon \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$
- Logistic Regression: $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x, \quad \pi = P(Y = 1)$

Count data

- Counts typically have a right skewed distribution
- Counts are not necessarily binary

We could use [Poisson Regression](#) to model count data

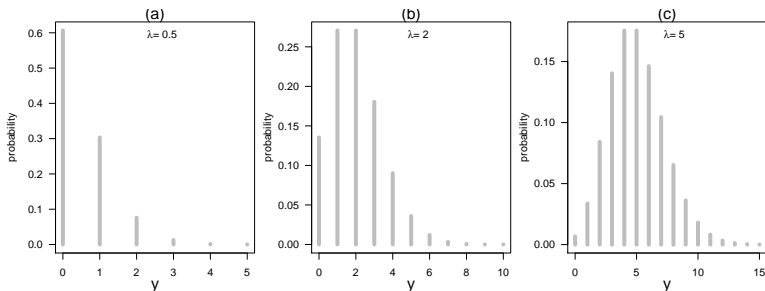
- If Y follow a Poisson distribution, then we have

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

where λ is the rate parameter that describe the event occurrence frequency

- $E(Y) = \text{Var}(Y) = \lambda$ if $Y \sim \text{Pois}(\lambda)$, $\lambda > 0$
- A useful model to describe the probability of a given number of events occurring in a fixed interval of time or space

Poisson Probability Mass Function



- (a), $\lambda = 0.5$: distribution gives highest probability to $y = 0$ and falls rapidly as $y \uparrow$
- (b), $\lambda = 2$: a skew distribution with longer tail on the right
- (c), $\lambda = 5$: distribution become more normally shaped

Flying-Bomb Hits on London During World War II [Clarke, 1946; Feller, 1950]

The City of London was divided into 576 small areas of one-quarter square kilometers each, and the number of areas hit exactly k times was counted. There were a total of 537 hits, so the average number of hits per area was $\frac{537}{576} = 0.9323$. The observed frequencies in the table below are remarkably close to a **Poisson distribution** with rate $\lambda = 0.9323$

Hits	0	1	2	3	4	5+
Observed	229	211	93	35	7	1
Expected	226.7	211.4	98.5	30.6	7.1	1.6

US Landfalling Hurricanes

Logistic Regression
and Poisson
Regression

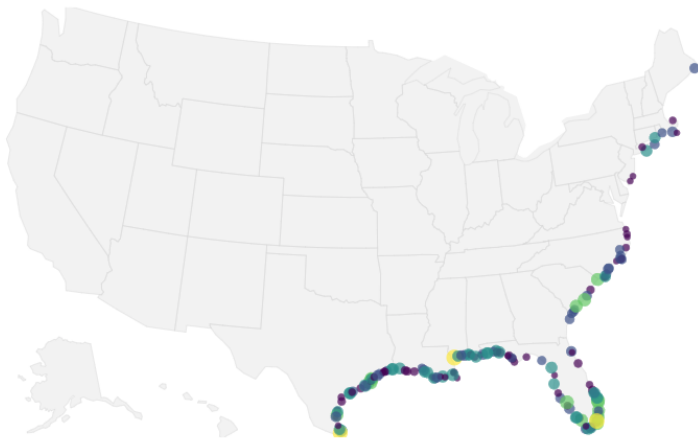
CLEMSON
UNIVERSITY

Logistic Regression

Poisson Regression

Generalized Linear
Model

US Hurricane Landfall points



Source: <https://www.kaggle.com/gi0vanni/analysis-on-us-hurricane-landfalls>

Number of US Landfalling Hurricanes Per Hurricane Season

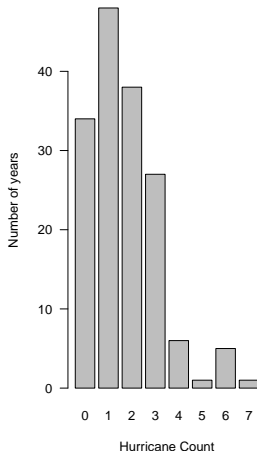
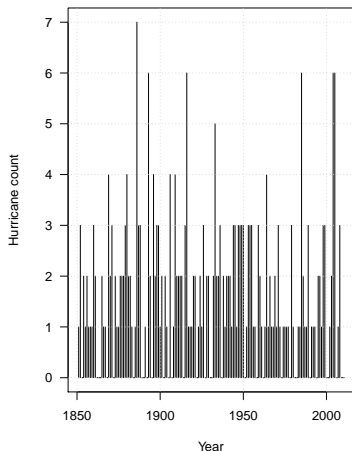
Logistic Regression
and Poisson
Regression



Logistic Regression

Poisson Regression

Generalized Linear
Model



Research question: Can the variation of the annual counts be explained by some environmental variable, e.g., Southern Oscillation Index (SOI)?

Some Potentially Relevant Predictors

- Southern Oscillation Index (SOI): an indicator of wind shear
- Sea Surface Temperature (SST): an indicator of oceanic heat content
- North Atlantic Oscillation (NAO): an indicator of steering flow
- Sunspot Number (SSN): an indicator of upper air temperature

Hurricane Count vs. Environmental Variables

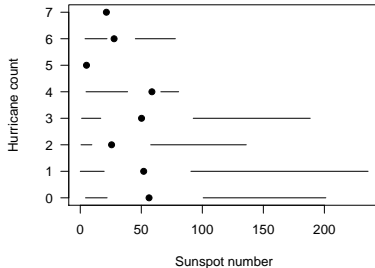
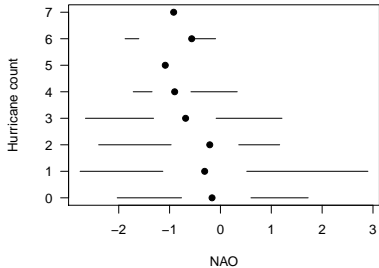
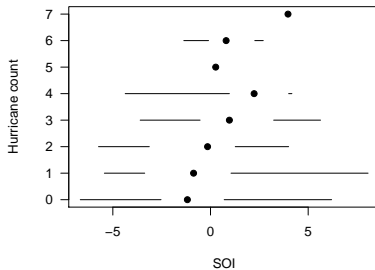
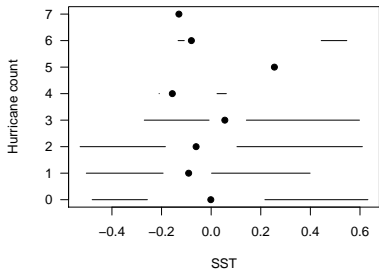
Logistic Regression
and Poisson
Regression



Logistic Regression

Poisson Regression

Generalized Linear
Model



$$\log(\lambda) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$$

$$\Rightarrow Y \sim \text{Pois}(\lambda = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}))$$

- Model **the logarithm of the mean response** as a linear combination of the predictors
- Parameter estimation is carry out using **maximum likelihood method**
- Interpretation of β' s: every one unit increase in x_j , given that the other predictors are held constant, the **λ increases by a factor of $\exp(\beta_j)$**

Poisson Regression Model:

$$\log(\lambda_{\text{Count}}) \sim \text{SOI} + \text{NAO} + \text{SST} + \text{SSN}$$

Table: Coefficients of the Poisson regression model.

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.5953	0.1033	5.76	0.0000
SOI	0.0619	0.0213	2.90	0.0037
NAO	-0.1666	0.0644	-2.59	0.0097
SST	0.2290	0.2553	0.90	0.3698
SSN	-0.0023	0.0014	-1.68	0.0928

⇒ every one unit increase in SOI, the hurricane rate increases by a factor of $\exp(0.0619) = 1.0639$ or 6.39%.

Issue with Linear Regression Fit

Linear Regression Model:

$$E(\text{Count}) \sim \text{SOI} + \text{NAO} + \text{SST} + \text{SSN}$$

Table: Coefficients of the linear regression model.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.8869	0.1876	10.06	0.0000
SOI	0.1139	0.0402	2.83	0.0053
NAO	-0.2929	0.1173	-2.50	0.0137
SST	0.4314	0.4930	0.88	0.3830
SSN	-0.0039	0.0024	-1.66	0.1000

If we use this fitted model to predict the mean hurricane count,
say SOI = -3, NAO=3, SST = 0, SSN=250

```
> predict(lmFull, newdata = data.frame(SOI = -3, NAO = 3, SST = 0, SSN = 250))  
1  
-0.318065
```

This number does not make sense

Model Selection

```
> step(PoiFull)
```

```
Start: AIC=479.64
```

```
All ~ SOI + NAO + SST + SSN
```

		Df	Deviance	AIC
- SST	1	175.61	478.44	
<none>		174.81	479.64	
- SSN	1	177.75	480.59	
- NAO	1	181.58	484.41	
- SOI	1	183.19	486.02	

```
Step: AIC=478.44
```

```
All ~ SOI + NAO + SSN
```

		Df	Deviance	AIC
<none>		175.61	478.44	
- SSN	1	178.29	479.12	
- NAO	1	183.57	484.41	
- SOI	1	183.91	484.74	

```
Call: glm(formula = All ~ SOI + NAO + SSN, family = "poisson", data = df)
```

```
Coefficients:
```

(Intercept)	SOI	NAO	SSN
0.584957	0.061533	-0.177439	-0.002201

```
Degrees of Freedom: 144 Total (i.e. Null); 141 Residual
```

```
Null Deviance: 197.9
```

```
Residual Deviance: 175.6 AIC: 478.4
```

- **Gaussian Linear Model:**

$$Y \sim N(\mu, \sigma^2), \quad \mu = \mathbf{X}^T \boldsymbol{\beta}$$

- **Bernoulli Linear Model:**

$$Y \sim \text{Bernoulli}(\pi), \quad \log\left(\frac{\pi}{1-\pi}\right) = \mathbf{X}^T \boldsymbol{\beta}$$

- **Poisson Linear Regression:**

$$Y \sim \text{Poisson}(\lambda), \quad \log \lambda = \mathbf{X}^T \boldsymbol{\beta}$$