Lecture 7

Logistic and Poisson Regression

Reading: Faraway 2016 Chapters 2.1-2.5; 5.1; 8.1; ISLR 2021 Chapter 4.2; 4.3.1-4.3.4; 4.6

STAT 8020 Statistical Methods II

Whitney Huang
Clemson University

Agenda

1. Logistic Regression: Binary response modeling
2. Poisson Regression: Count data modeling
3. Generalized Linear Models (GLMs)

Horseshoe Crab Mating [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`

This dataset provides a motivating example for logistic regression

The response is binary:

$$y = \begin{cases} 1 & \text{satellite males present} \\ 0 & \text{otherwise} \end{cases}$$

Why Not SLR? Why Logistic Regression?

Let $\pi(x) = P(Y = 1 \mid x)$, where Y is a binary response.

Simple Linear Regression (SLR): $\pi(x) = \beta_0 + \beta_1 x$

Problems with SLR for binary data:

- Predicted probabilities may be < 0 or > 1
- Variability changes with the mean
- Linear relationship is often unrealistic for probabilities

Logistic Regression:
Instead of modeling probability directly, logistic regression models the log-odds:

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

Applying the inverse transformation gives

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

Key idea: The logistic model guarantees valid probability estimates.

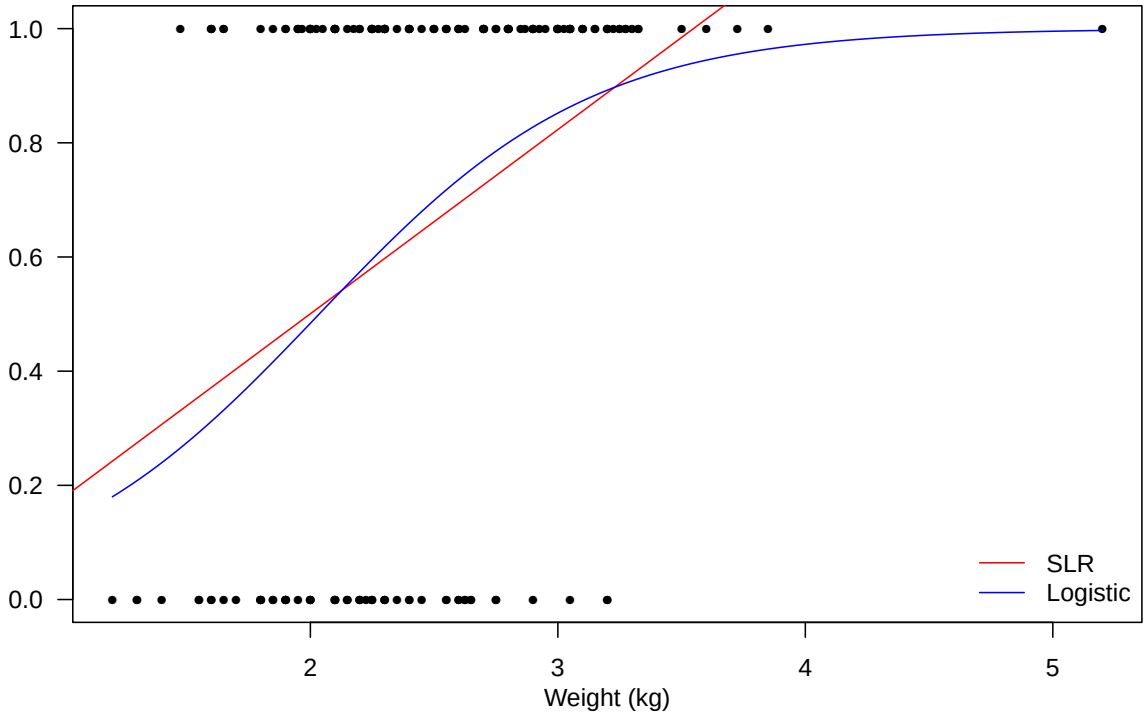
SLR vs Logistic Regression: Horseshoe Crab Mating Data

Linear regression:

$$\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x, \hat{\beta}_0 = -0.1449(0.1472), \hat{\beta}_1 = 0.3227(0.0588)$$

Logistic regression:

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



The logistic model respects the probabilistic structure of binary data

Properties of Logistic Regression

- β_1 controls the relationship between x and the probability:

$$\beta_1 > 0 \Rightarrow \pi(x) \uparrow \text{ as } x \uparrow$$

$$\beta_1 < 0 \Rightarrow \pi(x) \downarrow \text{ as } x \uparrow$$

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

  is constant, so the probability does not depend on x

- Properties of the logistic curve:
  - The curve is locally approximately linear:

$$\frac{d\pi(x)}{dx} = \beta_1 \pi(x)\big(1 - \pi(x)\big)$$

  - The midpoint occurs at

$$\pi\left(-\frac{\beta_0}{\beta_1}\right) = 0.5$$

  - Larger $|\beta_1|$ produces a steeper transition in probability

Odds Ratio Interpretation

From the logistic regression model,

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

the odds can be written as

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

If x increases by 1 unit,

$$\text{Odds at } (x+1) = \exp(\beta_1) \times \text{Odds at } x$$

Therefore,

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

Horseshoe crab example

$$\hat{\beta}_1 = 1.8151 \quad \Rightarrow \quad e^{1.8151} = 6.14$$

A 1 kg increase in weight multiplies the odds of attracting satellite males by about 6.1

Parameter Estimation

Logistic regression parameters are estimated using the method of maximum likelihood (MLE).

Statistical model

$$Y_i \sim \text{Bernoulli}(\pi_i)$$

where

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

Likelihood function

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

Maximum likelihood estimates: $\hat{\beta}_0, \hat{\beta}_1$
are the parameter values that maximize the likelihood.

The maximization has no closed-form solution and must 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
```

New terms in `glm()` output:

- `Null deviance`: fit with intercept only
- `Residual deviance`: fit after adding predictors
- Smaller deviance indicates better model fit

Inference: Confidence Intervals

A 95% confidence interval for β_i is

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

Horseshoe Crab Example

For β_1,

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

Exponentiating the interval gives a confidence interval for the odds ratio:

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

A 1 kg increase in weight is estimated to multiply the odds by between 2.94 and 12.85.

Inference: Hypothesis Test

Null and Alternative Hypotheses:

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

$H_a : \beta_1 \neq 0$

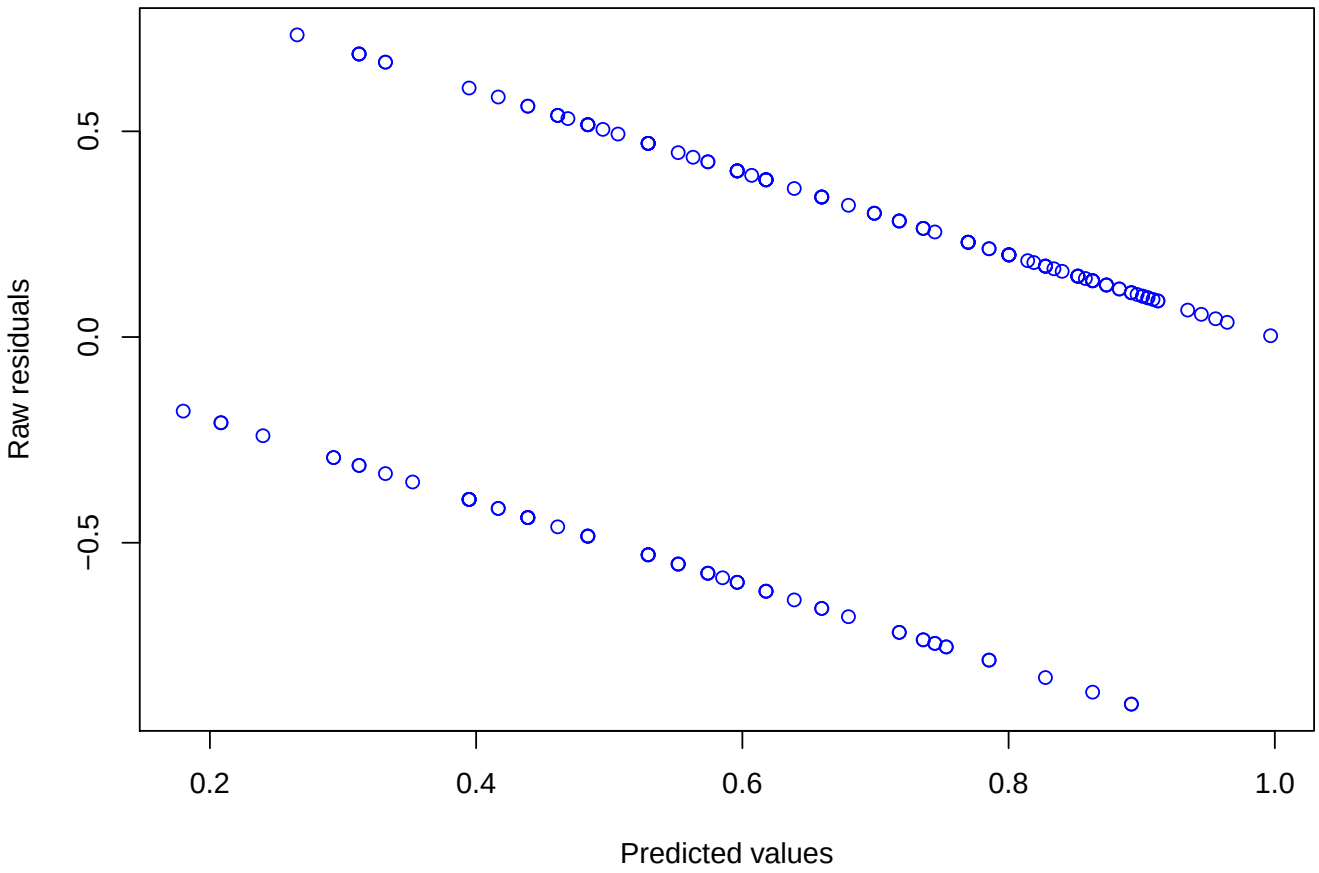
Test Statistics:

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

$\Rightarrow p$-value $= 1.45 \times 10^{-6}$

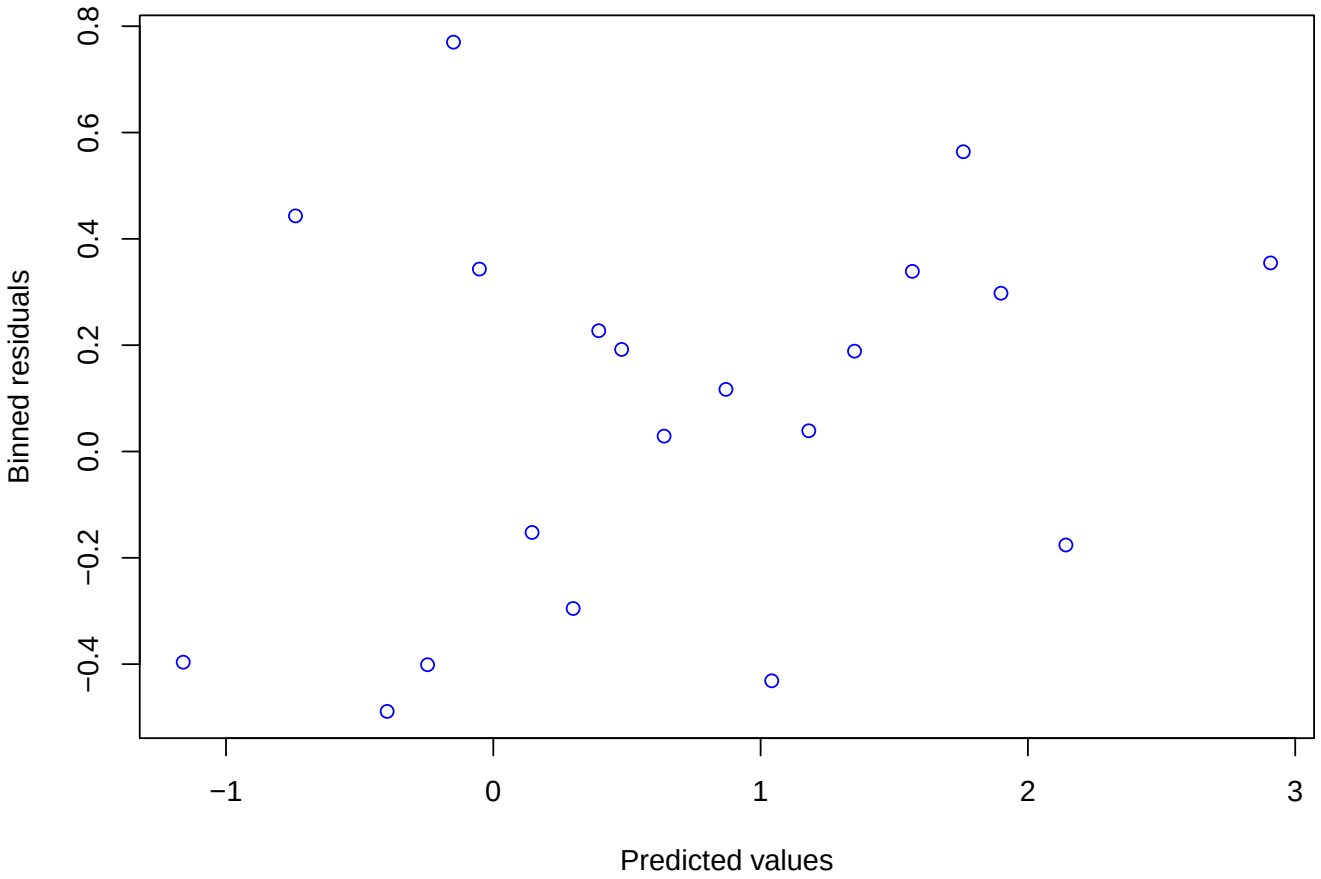
We have sufficient evidence that `weight` has positive effect on π, the probability of having satellite male horse-shoe crabs

Diagnostic: Raw Residual Plot



The raw residual plot is not very informative because the response variable, y, only takes two possible values

Diagnostic: Binned Residual Plot



- Grouping the residuals into bins and calculating the average for each bin
- $\log\left(\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}\right)$ is plotted on the horizontal axis (rather than the $\hat{\pi}(x)$) to provide better spacing

Model Selection with `step()`

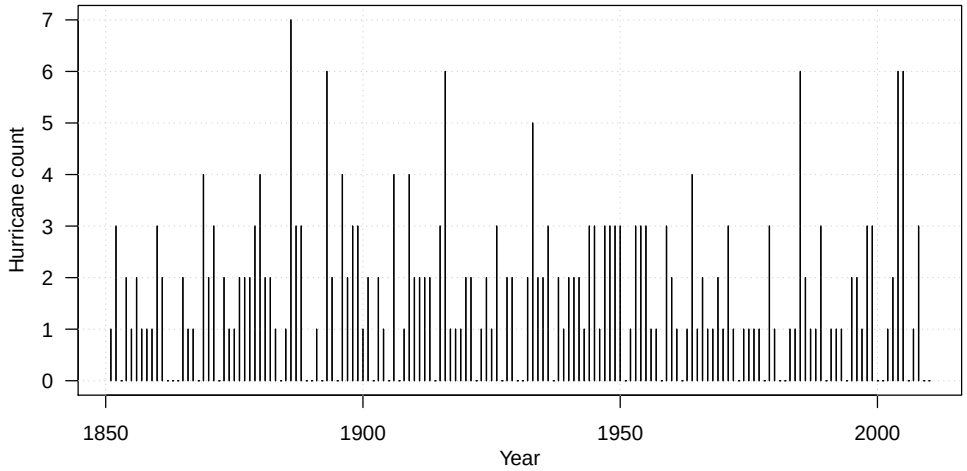
```
> logitFit2 <- glm(y ~ weight + width, data = crab, family = "binomial")
> step(logitFit2)
Start:  AIC=198.89
y ~ weight + width

         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
```

Key takeaway: `step()` compares models using AIC. Lower AIC suggests a better trade-off between fit and complexity.

Count Data

Daily COVID-19 Cases in South Carolina

Number of landfalling hurricanes per hurricane season



Key takeaway: Many real-world responses are counts, which are discrete, nonnegative, and often right-skewed

Modeling Count Data

So far we have talked about:

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

Count data

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

We can use Poisson Regression to model count data

Poisson Distribution

- A useful model to describe the probability of a given number of events occurring in a fixed interval of time or space

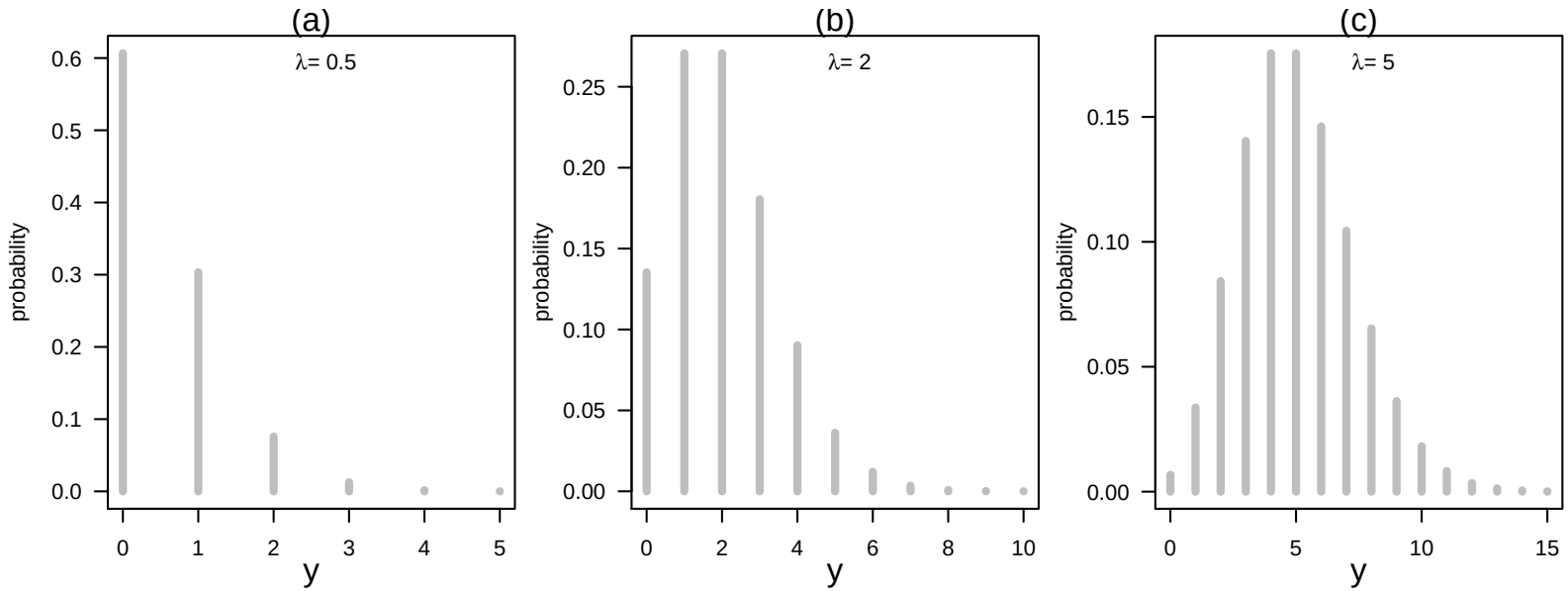
- If Y follow a Poisson distribution, then we have

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

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

- $\mathrm{E}(Y) = \mathrm{Var}(Y) = \lambda$ if $Y \sim \mathrm{Pois}(\lambda), \quad \lambda > 0$

Poisson Probability Mass Function



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

As λ increases, the distribution becomes less skewed and more symmetric

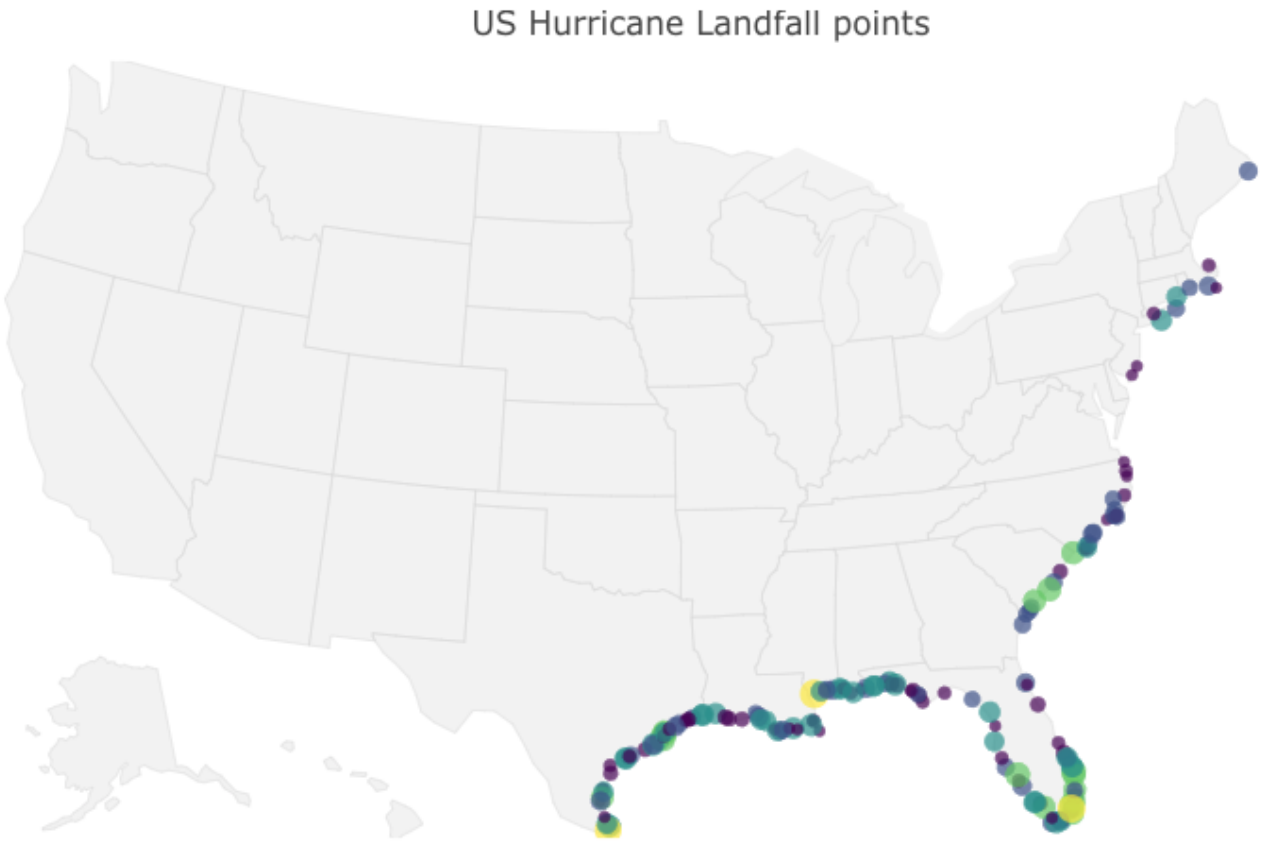
Flying-Bomb Hits on London During World War II [Clarke, 1946; Feller, 1950]

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

Observed frequencies closely match the theoretical Poisson distribution $\Rightarrow$ the random occurrence assumption is plausible

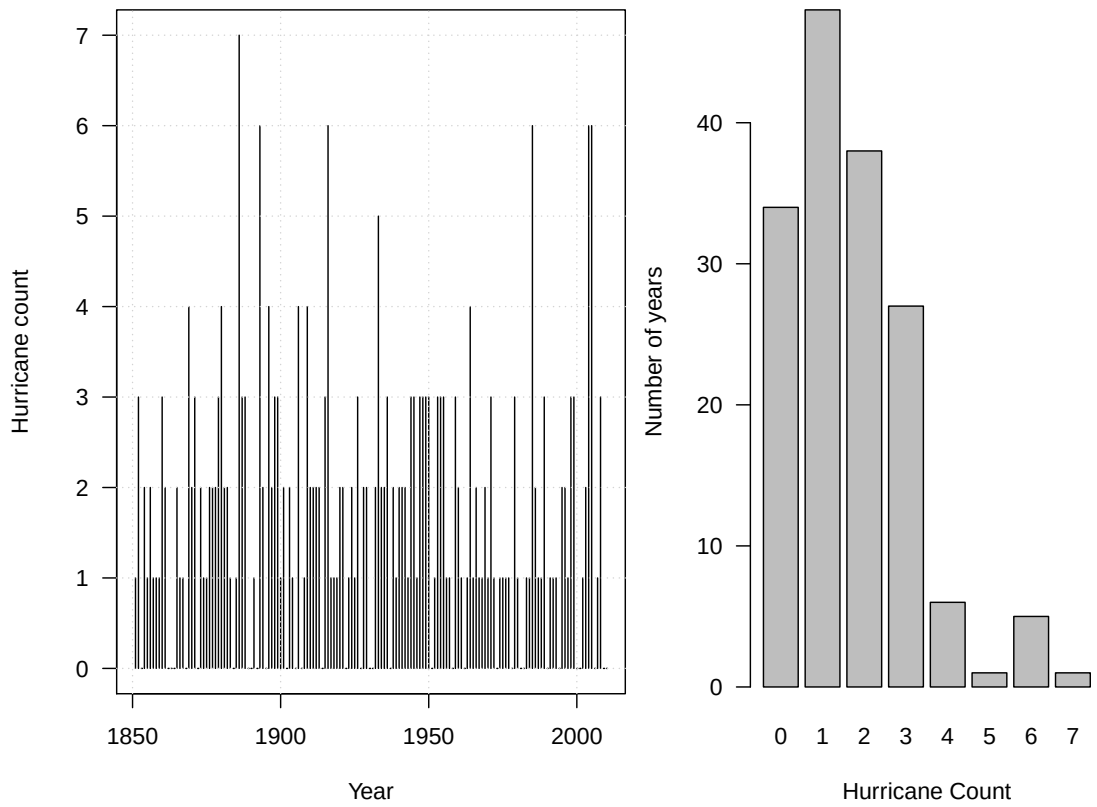
US Landfalling Hurricanes



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

We now model annual hurricane counts using environmental predictors

Number of US Landfalling Hurricanes Per Hurricane Season



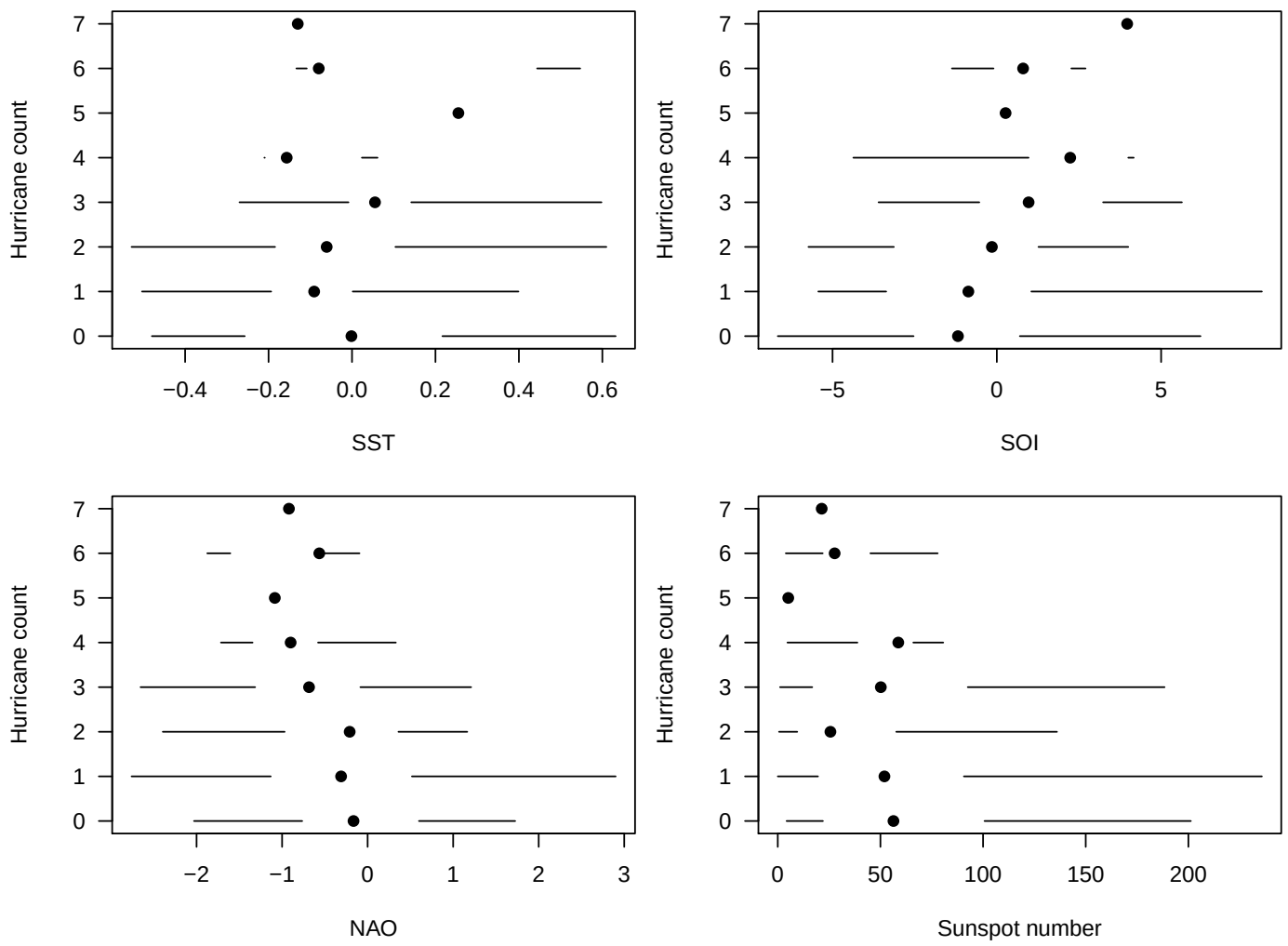
Counts vary substantially across years

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



Relationships are weak and noisy, motivating a regression model rather than visual inspection alone

Poisson Regression

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

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

- Model the logarithm of the mean response as a linear combination of the predictors
- Parameter estimation is carry out using the maximum likelihood method
- Interpretation of $\beta's$: 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 models log-mean counts linearly while guaranteeing positive predicted counts

US Hurricane Count: Poisson Regression Fit

Poisson Regression Model:

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

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

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

Issue with Linear Regression Fit

Linear Regression Model:

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

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 negative 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
```

AIC suggests removing `SST` improves model parsimony

Generalized Linear Models (GLMs)

Model	Response	Distribution	Link Function
Linear	Continuous	$Y \sim N(\mu, \sigma^2)$	$\mu = \boldsymbol{X}^T\boldsymbol{\beta}$
Logistic	Binary	$Y \sim \text{Bernoulli}(\pi)$	$\log\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{X}^T\boldsymbol{\beta}$
Poisson	Count	$Y \sim \text{Poisson}(\lambda)$	$\log(\lambda) = \boldsymbol{X}^T\boldsymbol{\beta}$

These models fall into the family of generalized linear models (GLMs) [Nelder and Wedderburn (1972); McCullagh and Nelder (1989)]

Key takeaway: GLMs unify regression models through:

- a probability distribution for the response
- a linear predictor $\boldsymbol{X}^T\beta$
- a link function connecting them

Summary

These slides cover:

- Logistic Regression
- Poisson Regression

Both of these, along with linear regression, can be unified under the framework of generalized linear models (GLMs)

`R` functions to know:

- **Logistic and Poisson Regressions**: `glm` with `family` being `"binomial"` and `"poisson"`, respectively
- Many `lm` utility functions can still be used; for example, `predict` can still be used for prediction, and `step` can still be used for model selection