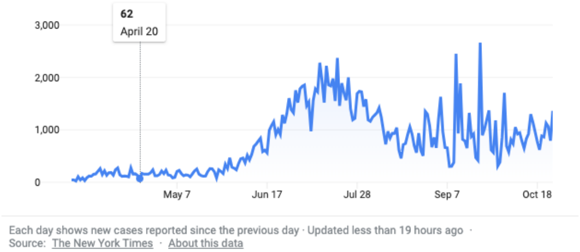
Lecture 19
Poisson Regression

STAT 8020 Statistical Methods II
October 27, 2020

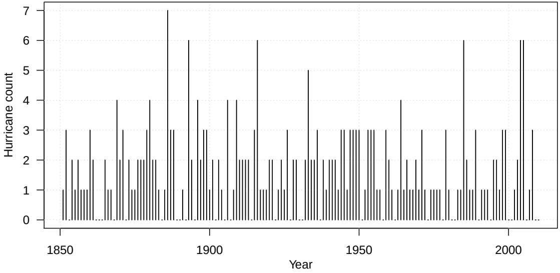
Whitney Huang
Clemson University

Count Data

- Daily COVID-19 Cases in South Carolina



- Number of landfalling hurricanes per hurricane season



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 could use Poisson Regression to model count data

Poisson Distribution

- If Y follow a Poisson distribution, then we have

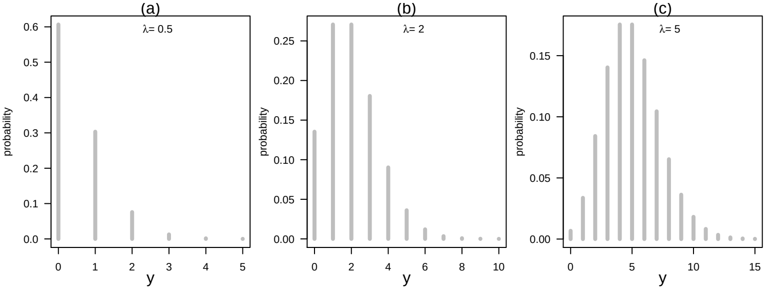
$$\mathrm{P}(Y=y)=\frac{e^{-\lambda}\lambda^{y}}{y!},\quad y=0,1,2,\cdots,$$

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

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

- 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



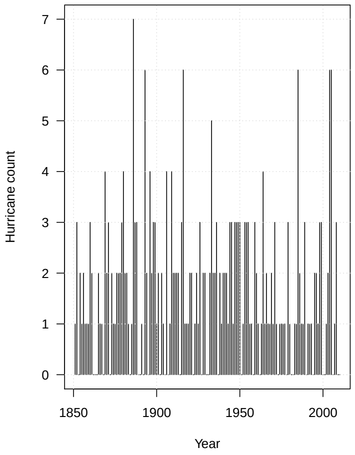
- (a), $\lambda = 0.5$: distribution gives highest probability to $y = 0$ and falls rapidly as y $\uparrow$
- (b), $\lambda = 2$: a skew distribution with longer tail on the right
- (c), $\lambda = 5$: distribution become more normally shaped

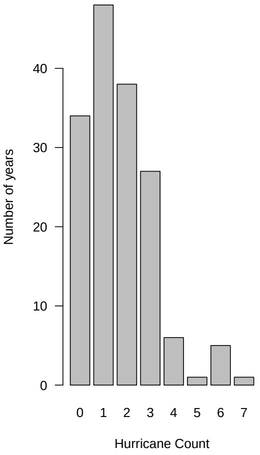
Flying-Bomb Hits on London During World War II [Feller, 1957]

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

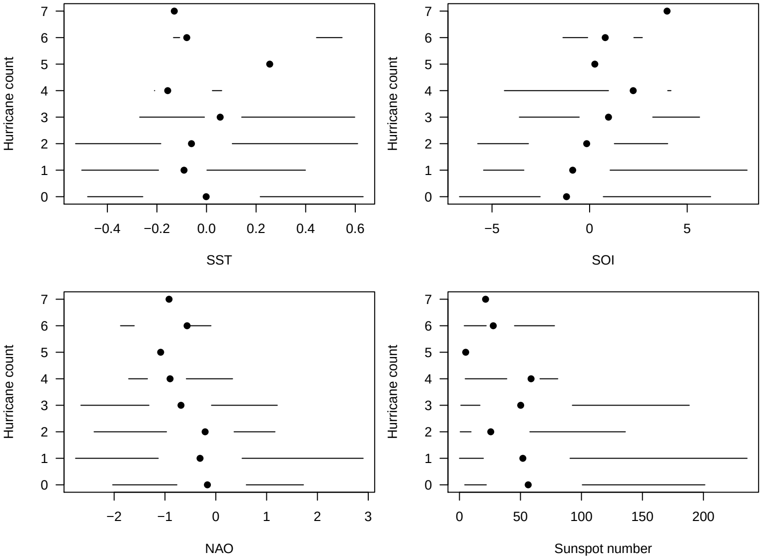
Number of US Landfalling Hurricanes Per Hurricane Season





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

Hurricane Count vs. Environmental Variables



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 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 $\exp(\beta_j)$

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%.

CLEMSON UNIVERSITY

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 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
```