On the frequency of America in America…
Now with greater America!

Adam B Kashlak
Mathematical and Statistical Sciences
University of Alberta

14 November 2017

Early 2016

Data mining political speeches

Huffman Encoding

A method for data compression

- The 26 English letters can be uniquely represented with 5 bits.
- Some characters are much more common than others.
- Use fewer bits to represent common characters.
- Use more bits to represent rare characters.
- Can be extended to common substrings.
- Used as part of the ZIP algorithm.

Huffman Encoding

A method for data compression

- The 26 English letters can be uniquely represented with 5 bits.
- Some characters are much more common than others.
- Use fewer bits to represent common characters.
- Use more bits to represent rare characters.
- Can be extended to common substrings.
- Used as part of the ZIP algorithm.

Most common 7 character string

- From 1790-1980, “of the ”, 94.8%

Huffman Encoding

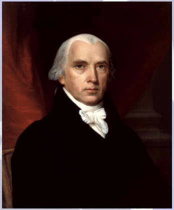
A method for data compression

- The 26 English letters can be uniquely represented with 5 bits.
- Some characters are much more common than others.
- Use fewer bits to represent common characters.
- Use more bits to represent rare characters.
- Can be extended to common substrings.
- Used as part of the ZIP algorithm.

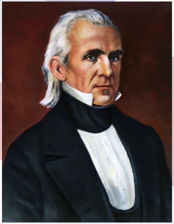
Most common 7 character string

- From 1790-1980, “of the ”, 94.8%
- From 1981-2016, “America”, 94.4%

The usage of America



> *The American people will face it with the undaunted spirit which in their revolutionary struggle defeated his [King George III's] unrighteous projects.*
>
> *James Madison, 1814*



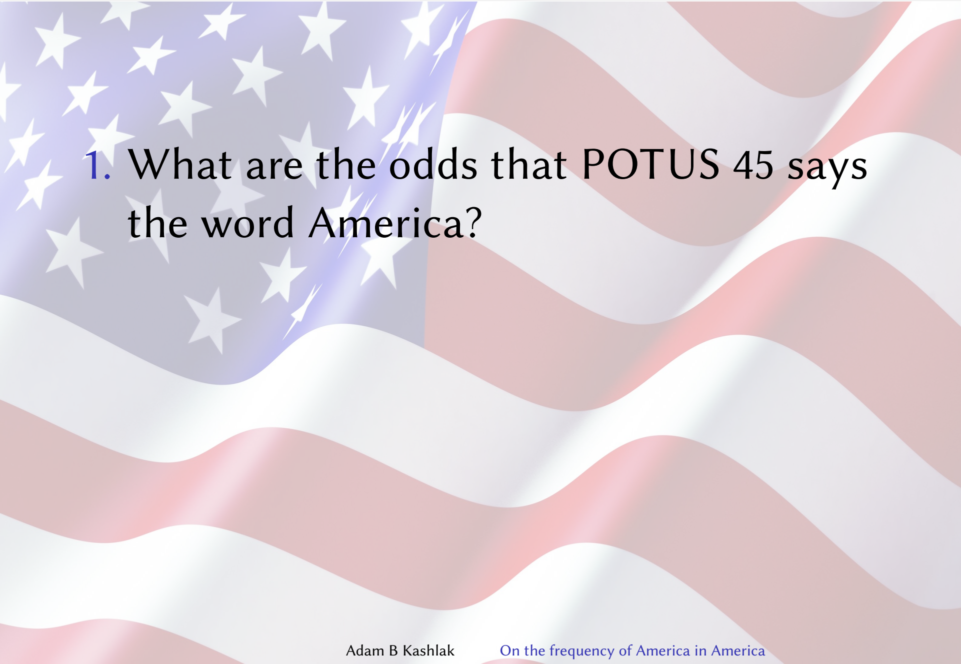
> *The American principle of self-government was sufficient to defeat the purposes of British and French interference.*
>
> *James K. Polk, 1845*

The usage of America

> “The program [The New Deal] itself comes from the American people.”
>
> Franklin D. Roosevelt, 1934

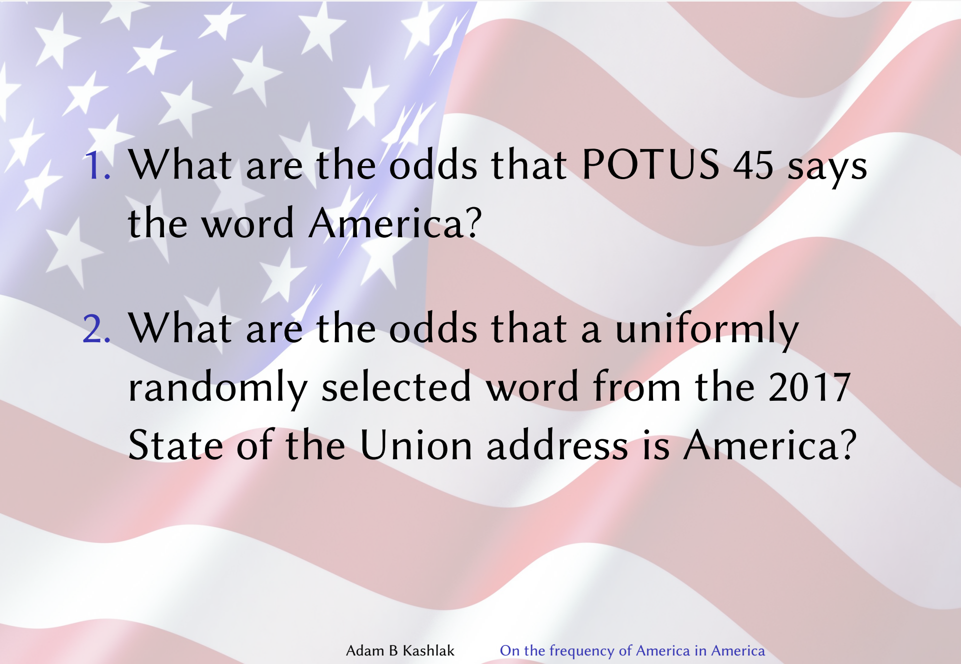
> “American freedom is threatened so long as the world Communist conspiracy exists.”
>
> Dwight D. Eisenhower, 1954

Driving Questions



1. What are the odds that POTUS 45 says the word America?

Driving Questions



1. What are the odds that POTUS 45 says the word America?
2. What are the odds that a uniformly randomly selected word from the 2017 State of the Union address is America?

The State of the Union

> *"The President shall from time to time give to the Congress Information of the State of the Union, and recommend to their Consideration such measures as he shall judge necessary and expedient."*
>
> *Article II, Section 3, US Constitution*

The State of the Union: History

Historical Highlights

- Began with Washington (1790)
- Written reports from Jefferson (1801) until Wilson (1913)
- First radio broadcast, Coolidge (1923)
- First TV broadcast, Truman (1947)
- First evening broadcast, Johnson (1965)
- First Internet webcast, Bush (2002)

The State of the Union: Data

Corpus of speeches and writings

- 227 years of data
- 42 of 44 presidents
- 1,752,383 words
- 127 of 227 are written reports

Most common 7 character string

- From 1790-1980, “of the ”, 94.8%
- From 1981-2016, “America”, 94.4%

Peters, G., and Woolley, J. T. The American Presidency Project. http://www.presidency.ucsb.edu/sou.php, 2016. [Online; accessed 13-March-2016].

Logistic Regression: Modelling Frequencies

Consider a linear model with binary response:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$$

with $y_i \sim \text{Bernoulli}\,(\pi_i)$. This does not follow the usual linear regression assumptions.

Logistic Regression: Modelling Frequencies

Consider a linear model with binary response:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$$

with $y_i \sim \text{Bernoulli}(\pi_i)$. This does not follow the usual linear regression assumptions.

Hence, consider the logistic response function:

$$\mathrm{E}(y|x) = \frac{e^{\langle x,\beta\rangle}}{1+e^{\langle x,\beta\rangle}} \quad \text{or} \quad \log\left(\frac{\mathrm{E}(y|x)}{1-\mathrm{E}(y|x)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

Log odds ratio: $\log\left(\frac{\mathrm{E}(y|x)}{1-\mathrm{E}(y|x)}\right)$

Logistic Regression: Modelling Frequencies

Also used for modelling data with $y_i \sim \mathrm{Binomial}\left(N, \pi_i\right)$ with

$$\log\left(\frac{\mathrm{E}\left(y|x\right)}{N - \mathrm{E}\left(y|x\right)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

Here, we model the log odds in terms of the frequencies y_i/N.

Poisson Regression: Modelling Raw Counts

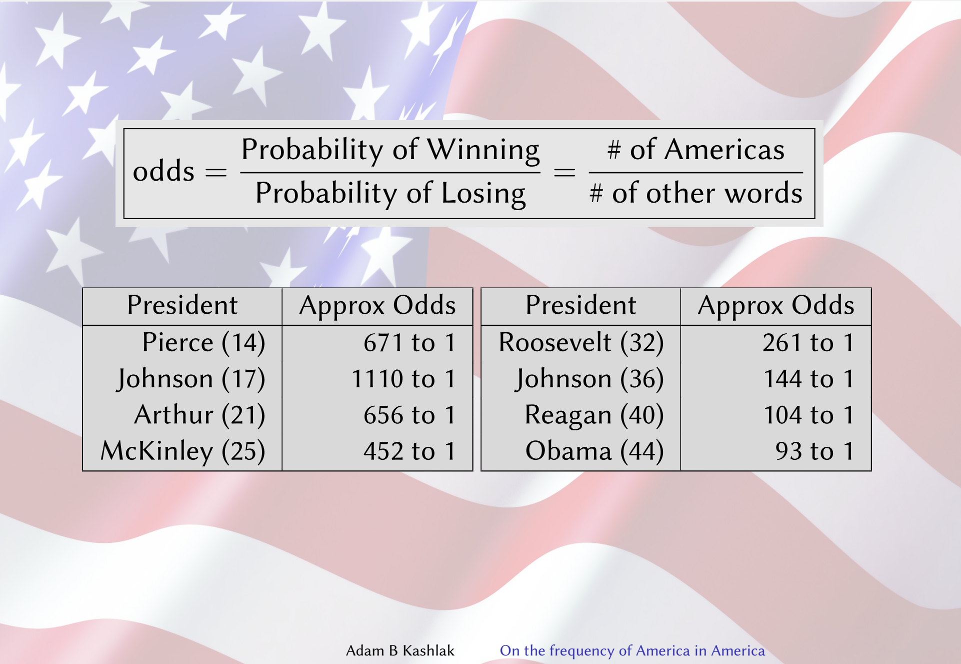
Alternatively, we can consider $y_i \sim \text{Poisson}\left(\lambda_i\right)$ where

$$f(y) = \frac{\lambda_i^{y_i} \mathrm{e}^{-\lambda_i}}{y_i!}, \quad y \in \mathbb{Z}^+.$$

The Poisson regression with a log link is

$$\log \mathrm{E}\left(y|x\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

What are the odds?



$$\boxed{\text{odds} = \frac{\text{Probability of Winning}}{\text{Probability of Losing}} = \frac{\text{\# of Americas}}{\text{\# of other words}}}$$

President	Approx Odds
Pierce (14)	671 to 1
Johnson (17)	1110 to 1
Arthur (21)	656 to 1
McKinley (25)	452 to 1

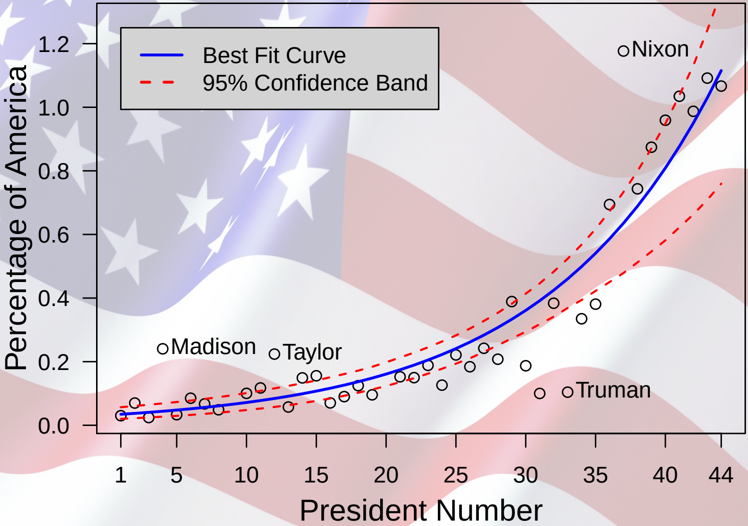
President	Approx Odds
Roosevelt (32)	261 to 1
Johnson (36)	144 to 1
Reagan (40)	104 to 1
Obama (44)	93 to 1

Logistic Regression with Binomial Link

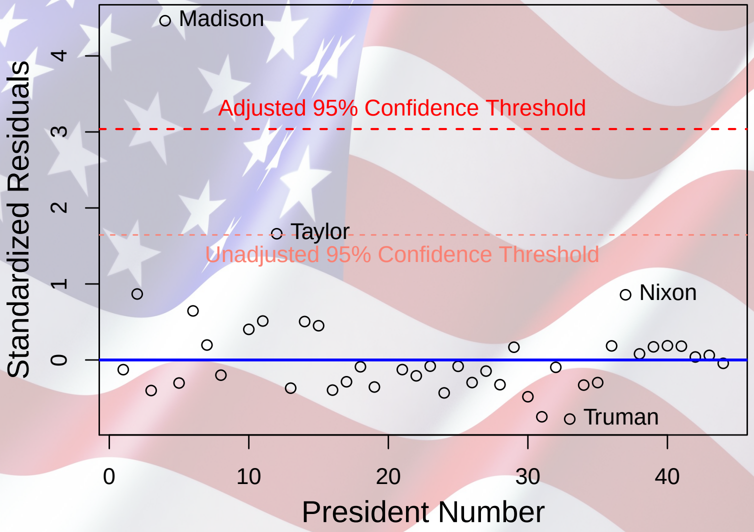
$$\boxed{\log\left(\text{expected odds}\right) = a + b \times (\text{President's Number})}$$

- Estimated $\hat{b} = 0.081$
- 95% confidence interval $[0.067, 0.096]$
- Roughly a 7% to 10% increase in the odds with each president
- Odds cut in half every 8-10 presidents
- Predicted odds for POTUS 45 between 102 and 67 to 1
- Predicted frequency between 1.0% and 1.5%

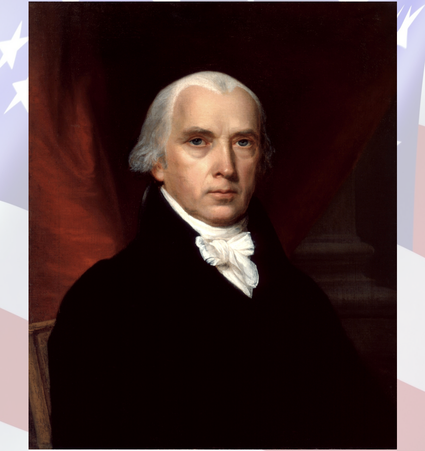
Logistic Regression: 95% confidence bands



Logistic Regression: Standardized Residuals



★ ★ ★ Winner: Most American President! ★ ★ ★



> *"And in the instance in which skill and bravery were more particularly tried with those of the enemy, the American flag had an auspicious triumph."*
>
> *James Madison, 1812*

Summer 2016

The National Conventions

Democratic National Convention

> *“America’s destiny is ours to choose. So let’s be stronger together, my fellow Americans. . . . And when we do, America will be greater than ever.”*
>
> *Hillary Clinton, 2016*

Republican National Convention

> *We will make America strong again.*
> *We will make America proud again.*
> *We will make America safe again.*
> *And we will make America great again!*
>
> *Donald J. Trump, 2016*

Frequency Comparison

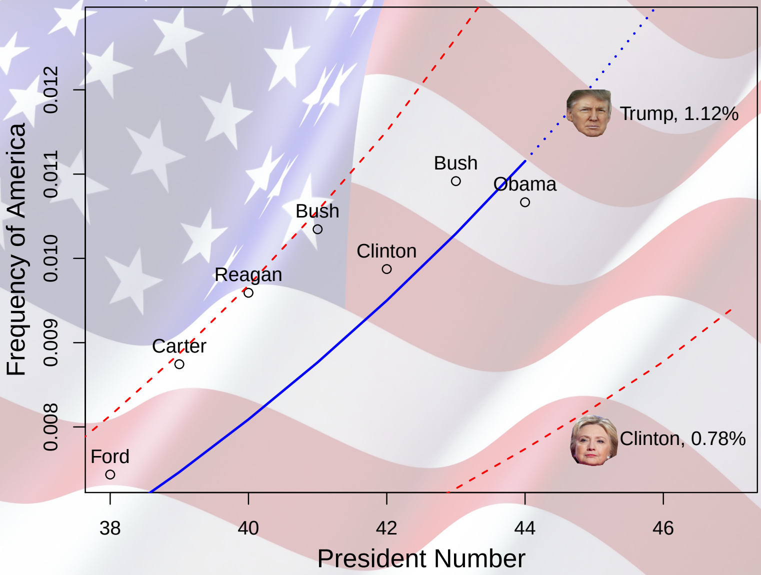
Letter Z	0.074%
Letter Q	0.095%
Letter X	0.150%
Letter J	0.153%
Madison	0.241%
Letter K	0.772%
Clinton	0.784%
Letter V	0.978%
Obama	1.067%

Trump	1.117%
Letter B	1.492%
Letter P	1.929%
Letter Y	1.974%
Letter G	2.015%
Letter F	2.228%
Hughes(*)	2.355%
Letter W	2.361%
Letter M	2.406%

(*) The frequency of America in the Langston Hughes poem “Let America be America again.”

Lewand, Robert. Cryptological mathematics. MAA, 2000.

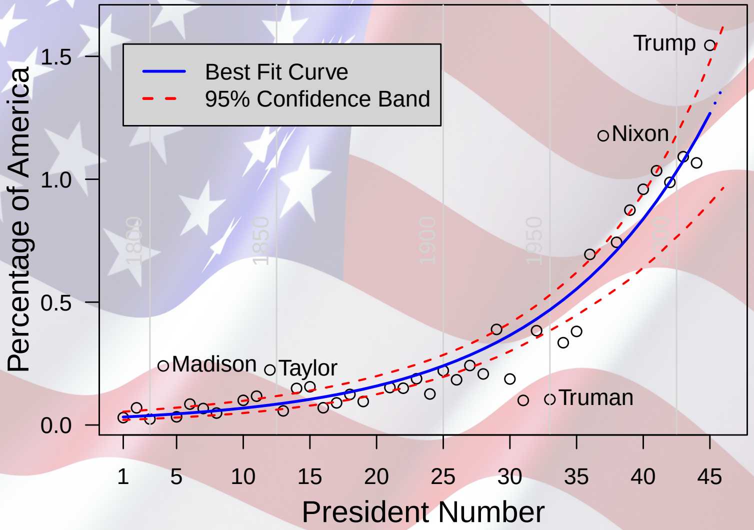
Logistic Regression: Prediction 2017



February 2017

Trump’s first SOTU

Logistic Regression: Observation 2017

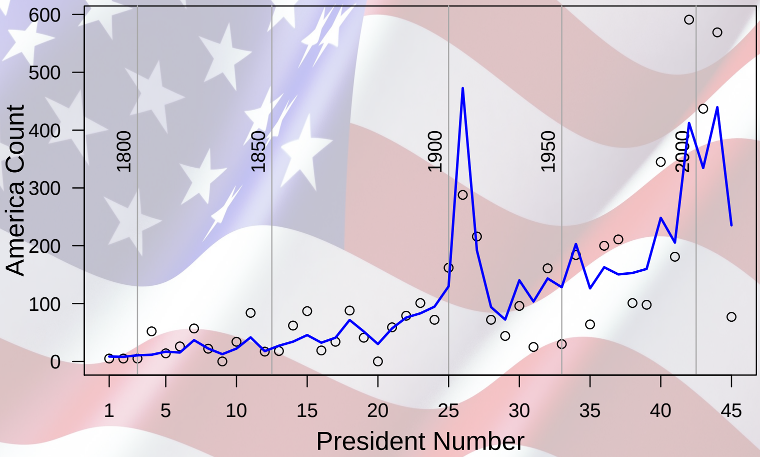


Take Two: Poisson Regression

$$\boxed{\log\left(\mathrm{E}\left(\text{count}\right)\right) = a + b \times (\text{President's \#}) + c \times (\text{Word Count})}$$

- Estimated $\hat{b} = 0.083$
- 95% confidence interval $[0.080, 0.086]$
- Roughly an 8% to 9% increase in the number of Americas with each president
- Estimated $\hat{c} = 1.37 \times 10^{-5}$
- $\mathrm{E}\left(\text{count}\right) \propto \exp\left((\text{Word Count})/73000\right)$
- Poor prediction for POTUS 45 due to low word count.

Poisson Regression: Observation 2017



Inauguration Addresses

Another dataset to consider

- 40 of the 45 presidents gave at least one inaugural address.
- 135,124 words in total
- Fit to the first 44 presidents, the logistic regression…
  - Rate parameter $\hat{b} = 0.076$
  - Approximately 5.6% to 10% increase in the odds per president.
  - Predicts frequency of 1.11% for Trump.

Inauguration Addresses

Another dataset to consider

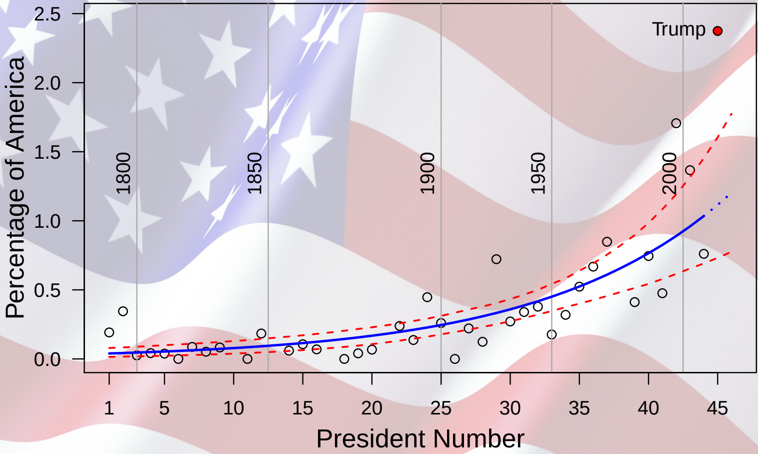
- 40 of the 45 presidents gave at least one inaugural address.
- 135,124 words in total
- Fit to the first 44 presidents, the logistic regression…
  - Rate parameter $\hat{b} = 0.076$
  - Approximately 5.6% to 10% increase in the odds per president.
  - Predicts frequency of 1.11% for Trump.
- Fit to the first 44 presidents, the Poisson regression…
  - Rate parameter $\hat{b} = 0.09$
  - Approximately 8.4% to 11% increase in the America count.
  - Predicts 20.3 America's for Trump (1.42%).

Inauguration Addresses

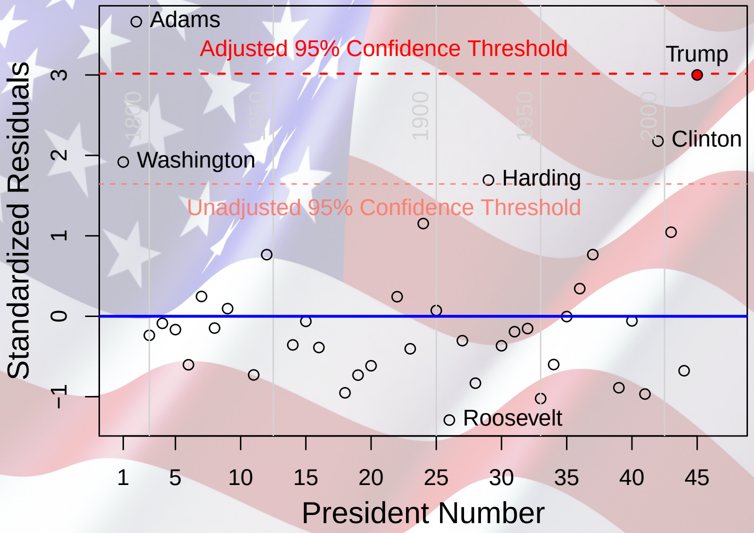
Another dataset to consider

- 40 of the 45 presidents gave at least one inaugural address.
- 135,124 words in total
- Fit to the first 44 presidents, the logistic regression…
  - Rate parameter $\hat{b} = 0.076$
  - Approximately 5.6% to 10% increase in the odds per president.
  - Predicts frequency of 1.11% for Trump.
- Fit to the first 44 presidents, the Poisson regression…
  - Rate parameter $\hat{b} = 0.09$
  - Approximately 8.4% to 11% increase in the America count.
  - Predicts 20.3 America's for Trump (1.42%).
- Actual frequency: 2.37%
- Actual Count: 34
- Or as Bush allegedly said, it "was some weird s***".

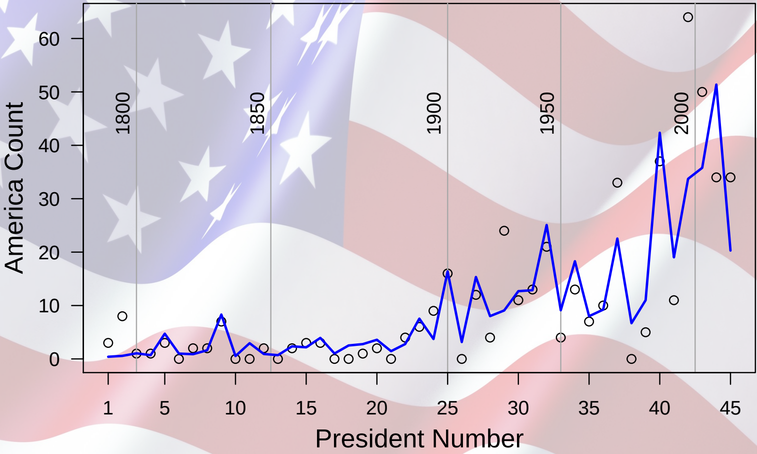
Logistic Regression: Inauguration Data



Logistic Regression: Inauguration Data



Poisson Regression: Inauguration Data



The Future of America

Powerful Rhetoric, or

Silly Cliché?

Mme./Mr. Speaker, the President of the United States

at the twilight's last gleaming
bright stars through the perilous
watch'd were so gallantly
US
UNITED
AMERICA
INDIVISIBLE
SINCE 1776
America
E PLURIBUS UNUM
THE REDWOOD FOREST TO THE GULF
THIS LAND WAS MADE FOR YOU
ERTY & JUSTICE FOR A