Non-Parametric Bootstrapping

Statistical Modeling: Deterministic Components

Statistics stands out from other quantitative fields primarily because of the incorporation of probabilistic functions

All statistical modeling incorporates some form of deterministic component:

$$y_i = \beta_0 + \beta_1 x_i$$

Statistical Modeling: Probabilistic Components

And a probabilistic component:

$$... + \epsilon_i$$
$$\epsilon_i \sim [\cdot]$$

Putting these in the context of the full context of the simple linear model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

In Statistics, we accept that uncertainty exists in real and complex systems.

Given that, we model this uncertainty to generate a greater understanding of these real, complex systems.

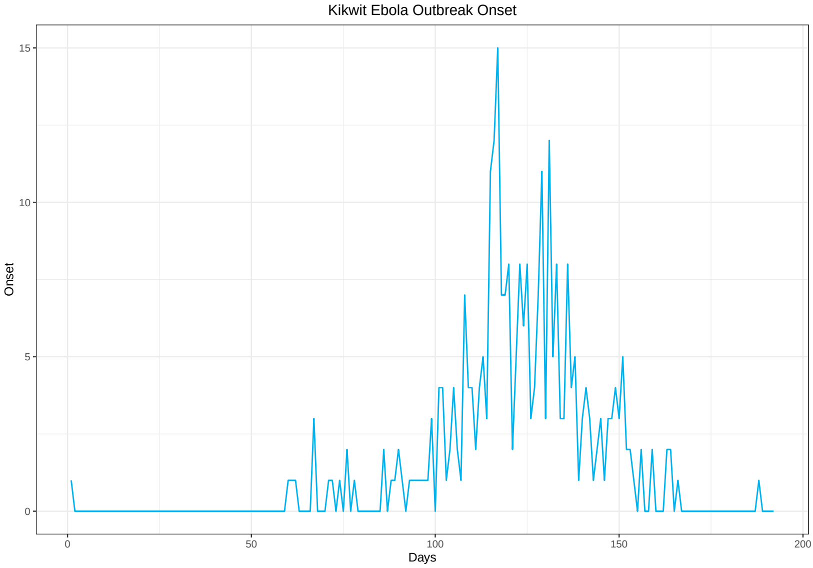
Motivating Example: Zaire Ebolavirus

In 1995, the city of Kikwit in the Democratic Republic of the Congo (formerly Zaire) experienced a devastating outbreak of Ebola virus, resulting in the death of ≈ 236 individuals.

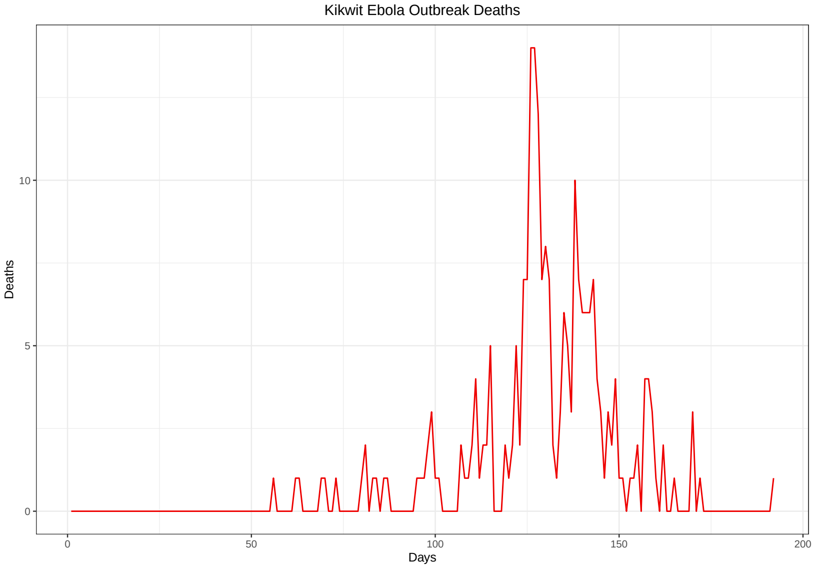
Ebola virus is a fast moving disease; highly pathogenic, contagious, and lethal.

Despite the tragedy that occurred during this outbreak, we've been able to learn a staggering amount about how to improve public health management in disaster scenarios.

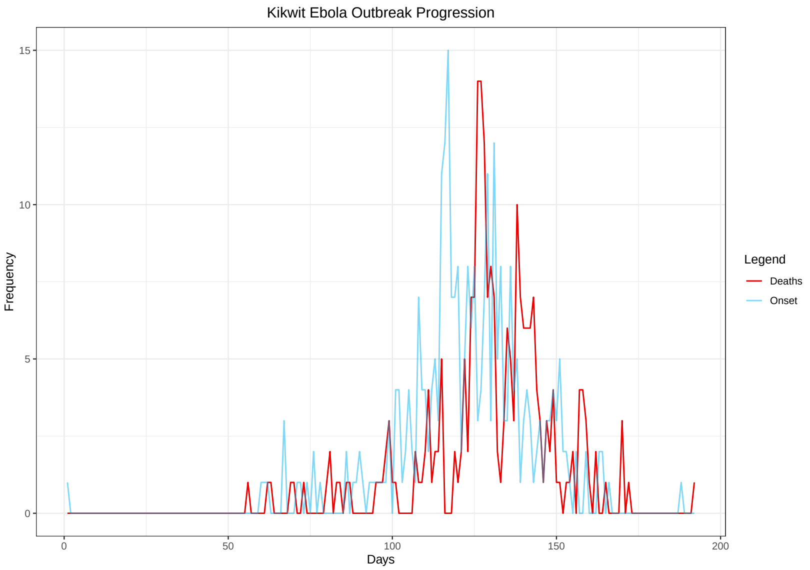
Time series of outbreak: Disease onset



Time series of outbreak: Deaths



Time series of outbreak: Full Progression



Model Proposals: Simple Linear Model

The primary predictors for all disease spread are space and time.

Since this is isolated to one location, time is our only predictor of interest.

$$y_i = \beta_0 + \beta_1 t_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

Does this seem like an appropriate model?

Model Proposals: Generalized Linear Model

Generalized Linear Model Framework:

$$y \sim [y|\mu, \psi]$$
$$g(\mu) = X\beta$$

Since positive cases are a discrete count of occurrences, it's justifiable that the Poisson distribution holds:

$$[y|\mu, \psi] = \mathsf{Pois}(\mu)$$

Model Proposals: Final Model

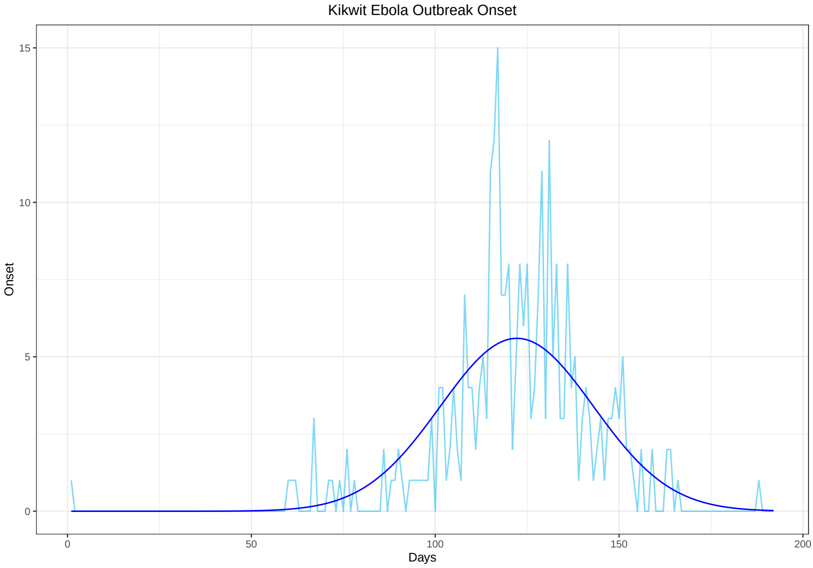
Resolving on a distributional assumption of Poisson for my generalized linear model, the final model is:

$$y_i \sim \mathsf{Pois}(\lambda_i)$$

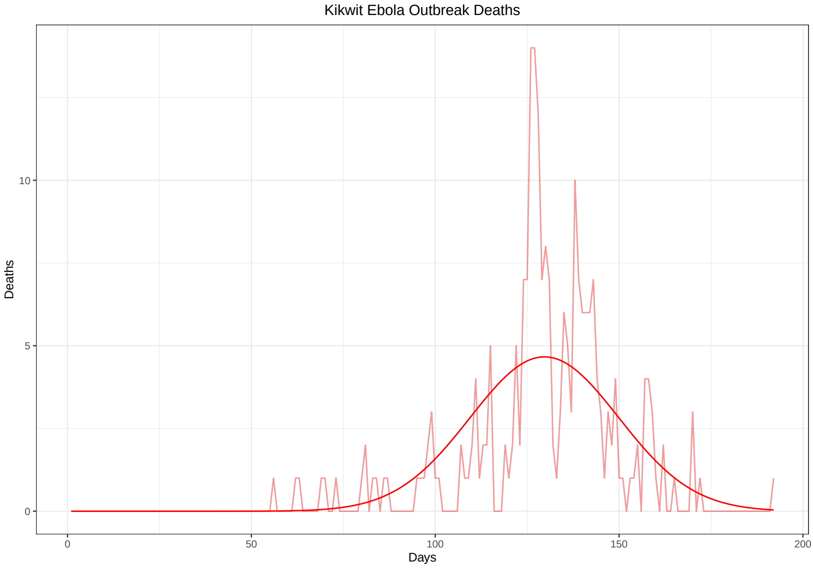
$$\lambda_i = \gamma_0 + \gamma_1 t_i + \gamma_2 t_i^2$$

Why the polynomial?

Model Fitting: Onset



Model Fitting: Deaths



Confidence Intervals

We've previously learned that confidence intervals are a method for calculating uncertainty in our parameters:

$$P(L(x) \leq x \leq U(x)) = p$$

Where:

$$L(x) \equiv \text{Lower interval of } x$$

$$U(x) \equiv \text{Upper interval of } x$$

And p is loosely, and *arbitrarily* defined as $= 0.95$

The Delta Method

1. Obtain the estimate for your variable

   i. Least-squares regression

2. Obtain the standard error for your variable

   i. Calculate the jacobian matrix of the inverse link function of $X\beta, J$
   ii. Get the variance-covariance matrix, V
   iii. Sandwich multiply the matrices: $J^T V J$

3. Add/Subtract the standard error multiplied by your interval value from the variable estimate

The Delta Method

$$\text{Var}[P(X\beta)] = \left(\frac{d(P(X_1\beta))}{d(X\beta)}\right)^T V \left(\frac{d(P(X_1\beta))}{d(X\beta)}\right)$$

$$U(\theta)_i = \theta_i + 1.96 * SE_\theta$$

$$L(\theta)_i = \theta_i - 1.96 * SE_\theta$$

Pros:

- Consistent process, works well when it works
- Hypothetically computable by hand

Cons:

- Becomes less reliable as distributions change
- Falls apart when the model become non-linear

Non-parametric Bootstrapping

Bootstrapping is a computational algorithm for obtaining confidence intervals for a wider range of models than the Delta method.

1. For a data set of n size, take a sample of size n **with replacement**
2. Estimate the parameters for a statistical model using the sampled data from step 1.
3. Save the estimates of interest.
4. Repeat steps (1-3) m times.

- App Example: https://rmsholl.shinyapps.io/bootstrap_showcase/

Bootstrapping Algorithm Syntax in R

```r
# set seed for reproducibility
set.seed(1)

# repeat m times
m_boot <- 1000

# initialize matrix for saving results
save_matrix <- matrix(,m_boot,1)

# for loop iterated by m from 1 to m_boot value
for(m in 1:m_boot) {

  # sample size of n with replacement
  samples <- sample(1:nrow(data),replace=TRUE)

  # temporarily save the samples from the data
  boot_data <- data[samples,]

  # run the model with this sampled data
  model_boot <- lm(y ~ x, data=boot_data)

  # save the outputs
  save_matrix[m,] <- coef(model_boot)[1]
}
```

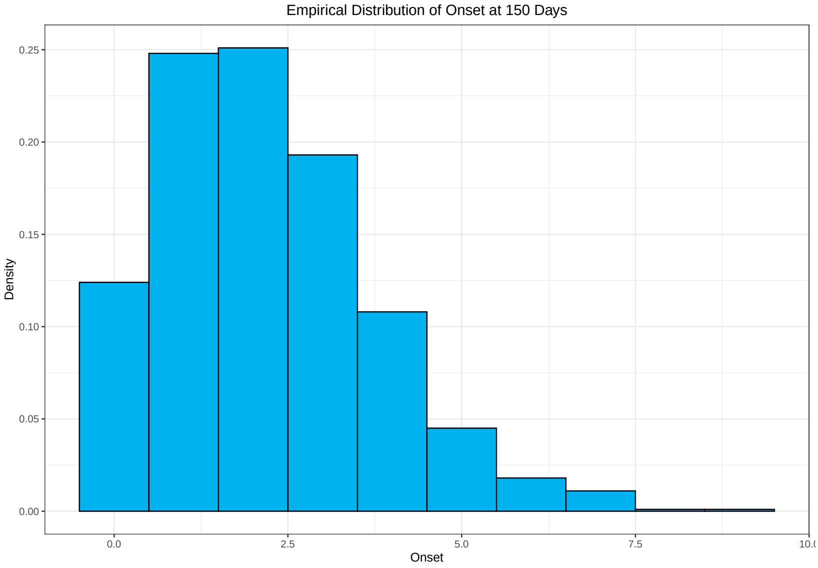
Bootstrapping Algorithm Psuedocode

$n \leftarrow$ data
$M \leftarrow$ statistical model
$m \leftarrow x$ where $x \geq 500$
$s \leftarrow$ empty list $\triangleright$ List for saved samples of the data
$Q \leftarrow$ empty queue $\triangleright$ Queue for saved samples
$S_\theta \leftarrow \vec{0}, m, \vec{1}$ $\triangleright$ θ is some statistic of interest
for each m **do**
    $n(s) \leftarrow$ resample n where replacement $=$ TRUE
    enqueue s into Q
    fit M to Q
    $M(s) \leftarrow M(Q)$
    append $\theta(M(s))$ to S_θ at row m
end for

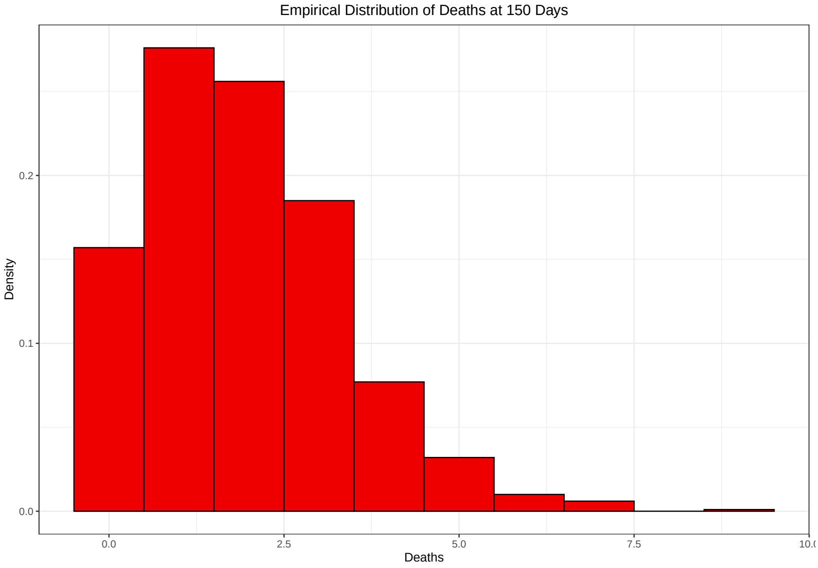
Bootstrapping with Mosaic

```
library(mosaic)

predict_ps_o <- function(){
  m1 <- glm(onset ~ days, family = poisson,data=resample(ebola))
  p <- predict(m1,newdata=data.frame(days=150),type="response")
  y <- rpois(1,p)
  y
}

bootstrap_onset <- do(1000)*predict_ps_o()

predict_ps_d <- function(){
  m1 <- glm(death ~ days, family = poisson,data=resample(ebola))
  p <- predict(m1,newdata=data.frame(days=150),type="response")
  y <- rpois(1,p)
  y
}

bootstrap_deaths <- do(1000)*predict_ps_d()
```

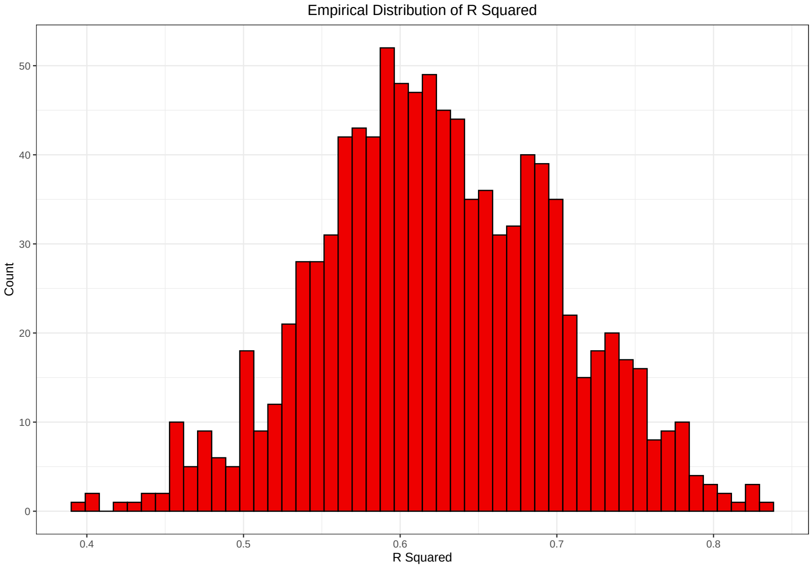
Bootstrap Histograms: Onset



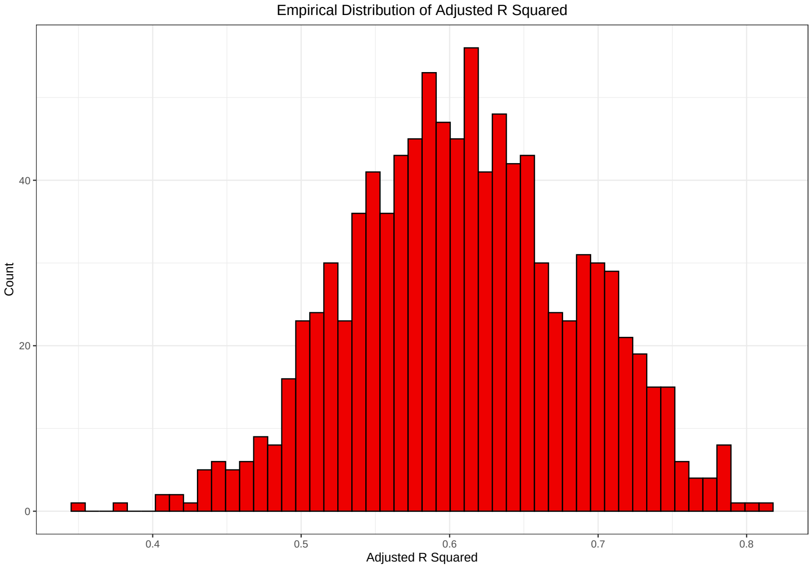
Bootstrap Histograms: Deaths



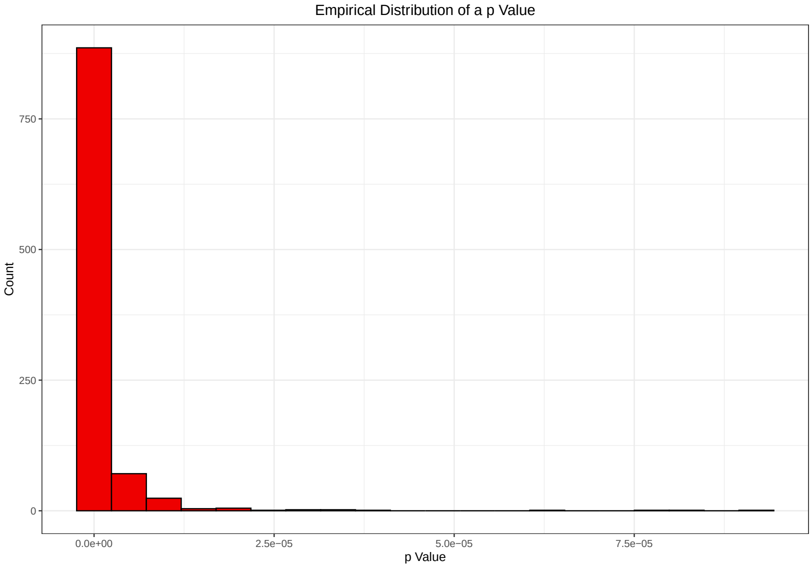
Fun with Bootstrapping



More Fun with Bootstrapping



Even more fun with Bootstrapping



References

1. Hall RC, Hall RC, Chapman MJ. The 1995 Kikwit Ebola outbreak: lessons hospitals and physicians can apply to future viral epidemics. Gen Hosp Psychiatry. 2008 Sep-Oct;30(5):446-52. doi: 10.1016/j.genhosppsych.2008.05.003. Epub 2008 Jul 23. PMID: 18774428; PMCID: PMC7132410.
2. Ver Hoef, Jay M. 2012. "Who Invented the Delta Method?" The American Statistician 66 (2): 124–27.
3. Efron, Bradley, and Robert J Tibshirani. 1994. An Introduction to the Bootstrap. CRC press.