Introduction to Bayesian Inference and Hierarchical Models

Sergio E. Betancourt
(sergio.betancourt@mail.utoronto.ca)

University of Toronto

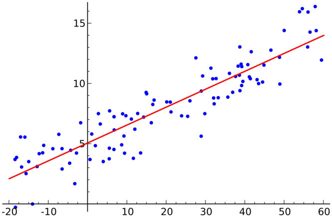
August 15, 2019

Agenda

1. Linear Models and Inference
2. Case Study: Road Safety in the City of Toronto
3. Conclusion

Linear Models and Inference
Linear Regression

- Starting point for most modeling problems
- Two flavors: ML (loss-based) and classical (model/likelihood-based)
- The latter offers solid inference framework based on probability and statistical theory
- Former is algorithmic (more flexible in certain cases) and perhaps more popular in industry



Linear Models and Inference
Maximum Likelihood Estimation

$$
\begin{aligned}
y &= X'\beta + \epsilon \\
\epsilon &\sim N(0, \Sigma)
\end{aligned}
\tag{1}
$$

- We wish to solve for the set of parameters that maximize the likelihood of the data
  - i.e., the parameters that best explain the underlying phenomenom

$$
\begin{aligned}
l(D;\theta) &= \log\left(\prod_{i=1}^{n} P(y^i|x^i,\theta)\right) \\
\hat{\theta} &= \text{argmax}_{\theta\in\mathcal{R}^p} l(D;\theta)
\end{aligned}
\tag{2}
$$

- Then we build CIs for our parameters, our predictions, etc.

Linear Models and Inference

Bayesian Estimation

- MLE above gives us distributions and CIs for the parameters βs
- Why not imbue them with a prior probability distribution right out of the bat? (Wakefield, 2013)
  - $\beta \sim P$?
- This is done for regularization, numerical stability, more rigurous inference, better science, expensive data acquisition

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{\int P(D|\theta)P(\theta)d\theta} \propto P(D|\theta)P(\theta) \tag{3}$$

- MCMC! (Rosenthal, 2009)

Linear Models and Inference

5 Mins on Bayesian Inference

- Imagine you encounter an arbitrary coin on the ground and you wonder whether it is fair or not.
- The outcome (Y) is binary and we can encode it as following:

$$Y = \begin{cases} 1 & \text{if "Heads"} \\ 0 & \text{if "Tails"} \end{cases}$$

- In fact, we can assert $Y \sim \text{bernoulli}(\theta)$ (our likelihood) and we indicate "fairness" as $\theta = 50\%$.
- We have a model of reality! Now we proceed to experiment and observe a sample of Y's.

Linear Models and Inference

5+ Mins on Bayesian Inference

- **Behold!** We throw the coin K=4 times and we get {0,0,0,0} (all "Tails").
- As frequentists, we can easily devise the maximum likelihood estimator for θ as

$$\hat{\theta}_{ML} = \sum_{i}^{K} \frac{y_i}{K}$$

  (the sample proportion!)
- Then $\hat{\theta}_{ML} = 0$ in our scenario.
  - 0 probability of attaining "Heads"?
  - This is too extreme! Both sides of the coin should have similar surface area...

Linear Models and Inference

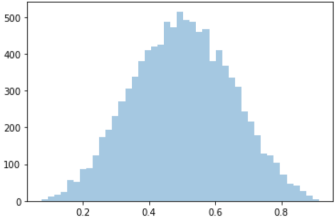
Ok, 5++ Mins on Bayesian Inference

- Let's adopt an alternative approach by expressing our **prior** belief about this coin.
- I believe the coin is most likely to be fair (from experience), but I believe there to be a smaller chance that it is not.
- Our "fairness" parameter θ must lie in $[0, 1]$
  - Use a beta prior! $\theta \sim \text{beta}(a, b)$
  - This prior is appropriate because it is defined on $[0, 1]$

Linear Models and Inference

End of Coin Toss

- Note how for $a = b \in \mathrm{R}^+$, $P(\theta)$ is symmetric at 0.5 (Fairness)
- I set $a = b = 5$ and sample from my posterior dist $P(\theta|D)$ (3)



- **Stan** gives me a posterior mean of $0.36 > 0$
  - However, 95% credibility interval $(0.14, 0.61) \ni 0.5$
  - We can tighten the CI with more observations and "smarter" priors

Linear Models and Inference

Priors!

- What are good priors?
  - Tractability? Conjugates!
  - Scientific relevance
  - Regularization
- Ideally, priors should come without ever looking at data
  - Uninformative (flat) priors
  - Weakly informative priors
  - Informative priors (Usually based on expert opinion or solid scientific knowledge)
- **PC Framework!** (Simpson et al, 2017)
  - KL discrepancy to measure the increased complexity introduced by $\psi > 0$
  - $P(\mu > u) = \alpha$ for base $N(0, 1)$ and upgrade $N(\mu, 1)$, $\mu > 0$

Linear Models and Inference

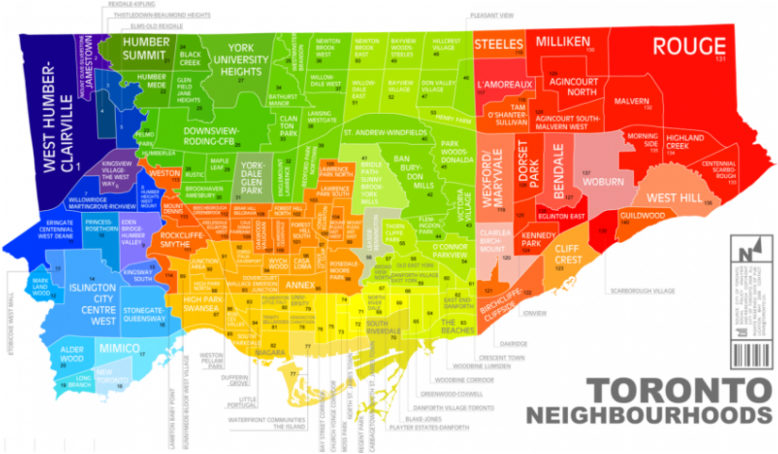
Hierarchical Models

- Data is not always homogenenous, nor every class of interest is represented in a balanced way
  - Individuals have different biologies
  - Schools have different demographies, funding, quality of teaching, etc.
  - Neighborhoods have different densities, infrastructure, etc.
  - Firms have different idiosyncrasies
- In real life independence assumption usually goes out the window
- We can allow for and make use of these differences in hierarchies/clusters/groupings
  - **Latent variables!**

Introduction

- Road safety is close to all of us
  - At any point in a given day we commute as bikers, pedestrians, drivers, etc.
- In the last five years, in the City of Toronto, 190 pedestrians and 16 cyclists were killed in collisions with vehicles
- We examined road safety in the City of Toronto from 2007 to 2017, exploring the areas with highest risk of a traffic incident, controlling for different fixed factors, neighborhoods, and time

Introduction



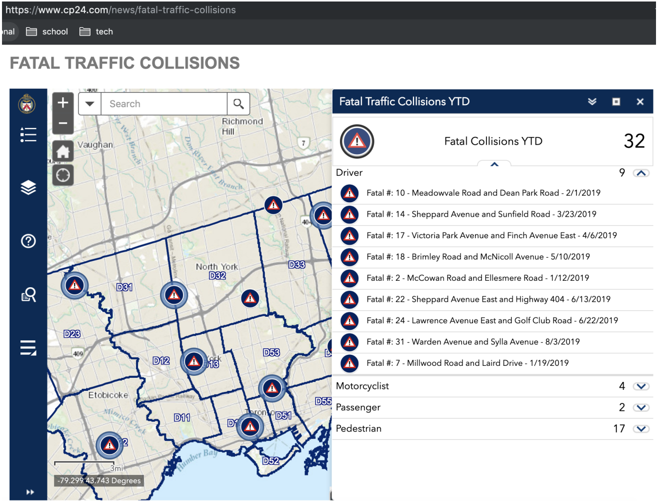
Toronto Safety

Dataset

Toronto Police Service PUBLIC SAFETY DATA PORTAL	Automobile accident-level data each row representing a person involved from 2007 - 2017
Statistics Canada / Statistique Canada	2016 and 2011 Census population by neighborhood
Government of Canada / Gouvernement du Canada	Daily weather measurements for Toronto taken in University of Toronto
GEOTAB // DATA	Rolling annual average of harsh braking incidents and number of accidents

Toronto Safety

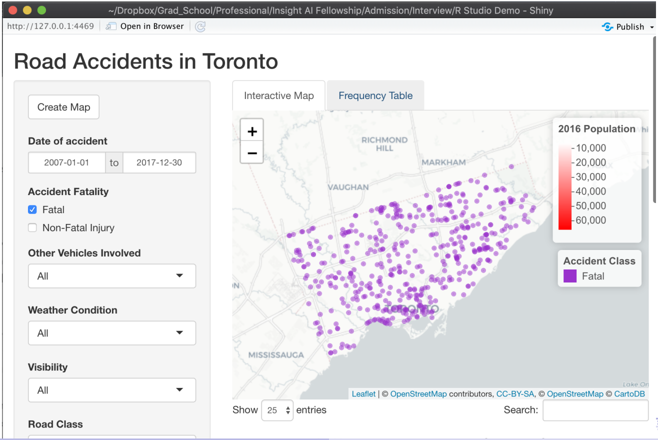
IDE: Existing Visualization



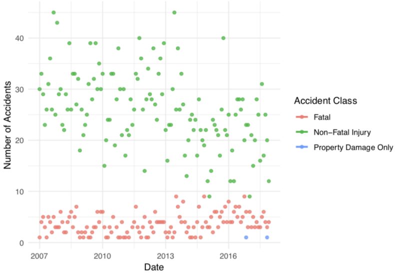
Toronto Safety

IDE: Our App

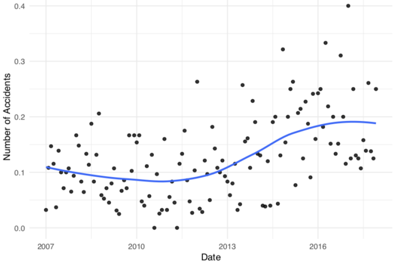
- Created a webapp with Shiny for visualizing accidents by different filters (time, parties involved, weather, etc)



Toronto Safety
IDE



(a) Accidents by Class



(b) Total Number of Accidents

Modeling

Bayesian Hierarchical GAM

$$
\begin{gathered}
Y_{ijt} \sim \text{bernoulli}(\pi_{ijt}) \\
\text{logit}(\pi_{ijt}) = X_{ijt}\beta + U_j + V_t + f(W_t) \\
U_j \sim N(0, \sigma_U^2) \quad \text{(Residual Neighborhood Component)} \\
V_t \sim N(0, \sigma_V^2) \quad \text{(Residual Time Component)} \\
W_{t+1} - W_t \sim N(0, \sigma_W^2) \quad \text{(RW1 - Time Trend Component)}
\end{gathered}
$$

Results

Fixed Effects

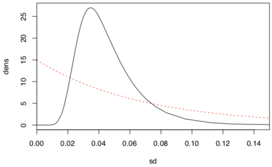
Objective: Estimate the odds of fatality, subject to being in a vehicular accident, accounting for differences across neighborhoods and time

Results:

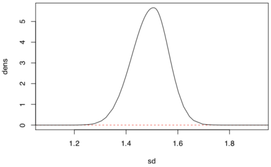
- Odds of fatality increasing till mid-2016, then slowly falling (Vision Zero?)
- Expressway: +73% odds of fatality
- Traffic Sign (stop, pedestrian crossing): -13% odds of fatality
- Traffic Light: -46% odds of fatality
- Pedestrian not involved: -38% odds of fatality
- +1 mm of precipitation: -2% odds of fatality

Results

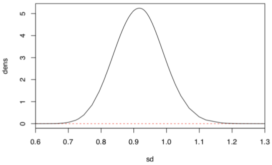
Priors-Posteriors on Hierarchical Parameters



(a) SD of time trend

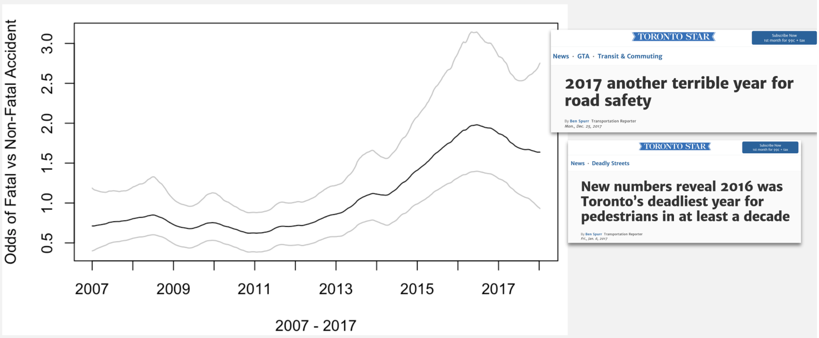


(b) SD of residual time variation



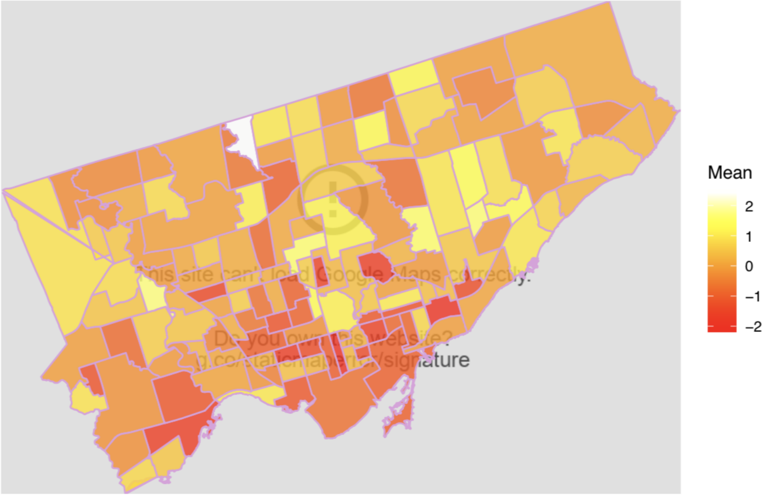
(c) SD of neighborhood random effect

Results
Time Trend Effect



Results

Neighborhood Random Effects



Results
Discussion

- Interesting time modelling!
  - Trend component shows semblance of seasonality, but the imbalance and small amount of data did not allow us to fit this
- Not great in explaining neighborhood variations
  - No inputs to model describing road density, road infrastructure, neighborhood density, traffic intensities through the day
- We can strenghten inference with a spatio-temporal model (HARD!)
  - We tried this as a (Log-Cox Gaussian) point process but current spatial data not available (very expensive)
  - Satellite data?
- Need better data for both

Conclusion

Some thoughts

- Bayesian inference and computation have recently enjoyed a splendid renaissance with better computing HW/SW
  - more involved than reg MLE or many ML implementations
- LOTS of active research into Bayesian DL and RL
  - Causal Inference
  - Bayesian optimization of hyperparams (AlphaGo)
  - Multi-task learning
  - Exploration in RL
  - Efficient, computationally stable MCMC (HMC)
- Most people are exposed to MLE only in their undergraduate studies and in industry...
- For full Bayesian methodology, and greater flexibility, use **stan** (Carpenter et al., 2017)

Conclusion

References

- Wakefield, J. (2013). Bayesian and Frequentist Regression Methods. Springer Series in Statistics. Springer New York. doi: 10.1007/978-1-4419-0925-1.
- Faraway, Julian J, Xiaofeng Wang, and Yu Yue Ryan (2018). Bayesian Regression Modeling with INLA. Chapman and Hall/CRC.
- Rue, Havard, Sara Martino, and Nicolas Chopin (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations". In: Journal of the Royal Statistical Society B 71.2, pp. 319–392. doi: 10.1111/j.1467-9868.2008.00700.x.
- Simpson, Daniel, Haavard Rue, Andrea Riebler, Thiago G. Martins, and Sigrunn H. Sorbye (Feb. 2017). "Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors". In: Statistical Science 32.1, pp. 1–28. doi: 10.1214/16-STS576.
- Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. Stan: A probabilistic programming language. Journal of Statistical Software 76(1). DOI 10.18637/jss.v076.i01.