Likelihood and measurement

Modelling approaches for real data

Dirk Bester

2023-10-02

Introduction

Measurement

"Most neglected subject in all of statistics."

Andrew Gelman

Introduction

Measurement

"Most neglected subject in all of statistics."
Andrew Gelman

Simply, *how* you observe influences *what* you observe.

Introduction

Measurement

"Most neglected subject in all of statistics."

Andrew Gelman

Simply, *how* you observe influences *what* you observe.
Less simply, the measurement process is as important as the data-generating process. People often just ignoring it.

Introduction

Measurement

"Most neglected subject in all of statistics."
Andrew Gelman

Simply, *how* you observe influences *what* you observe.
Less simply, the measurement process is as important as the data-generating process. People often just ignoring it. Well known example: people like to round numbers

Introduction

Measurement

"Most neglected subject in all of statistics."
Andrew Gelman

Simply, *how* you observe influences *what* you observe.
Less simply, the measurement process is as important as the data-generating process. People often just ignoring it. Well known example: people like to round numbers
This influences how the likelihood is constructed.

Problem statement

We want to sell Cyber security insurance.

Problem statement

We want to sell Cyber security insurance.

Data is scarce. Companies aren't willing to share data with each other.

Problem statement

We want to sell Cyber security insurance.
Data is scarce. Companies aren't willing to share data with each other.
One consulting firm collected information from a set of their clients and published it in aggregated form. (See next slide for details).

Problem statement

We want to sell Cyber security insurance.
Data is scarce. Companies aren't willing to share data with each other.
One consulting firm collected information from a set of their clients and published it in aggregated form. (See next slide for details).

Companies were happy for their data to be shared this way, because how could it possibly be useful to their competitors...

Problem statement

I have this data for 2013, 2014, 2015:

Business.Sector	Claims	Min	Median	Mean	Max	year
Education	8	2560	132650	204858	680000	2013
Entertainment	2	1125000	5812500	5812500	10500000	2013
Financial Services	8	20100	166000	1060138	4750000	2013
Healthcare	29	5390	254000	1612343	20000000	2013
...				...		
...				...		
Professional Services	10	6704	29217	329845	2989966	2015
Restaurant	5	4000	16212	75744	250000	2015
Retail	12	91359	455488	1795266	8916432	2015
Technology	11	0	90000	206532	641635	2015
Other	15	708	61339	713133	6700142	2015

... and I want to fit this model:

$$\begin{aligned} y &\sim GLM(g^{-1}(X\beta)) \\ &\sim Dist(\text{mean} = g^{-1}(X\beta), \text{variance}) \end{aligned}$$

Problem statement

I have this data for 2013, 2014, 2015:

Business.Sector	Claims	Min	Median	Mean	Max	year
Education	8	2560	132650	204858	680000	2013
Entertainment	2	1125000	5812500	5812500	10500000	2013
Financial Services	8	20100	166000	1060138	4750000	2013
Healthcare	29	5390	254000	1612343	20000000	2013
...				...		
...				...		
Professional Services	10	6704	29217	329845	2989966	2015
Restaurant	5	4000	16212	75744	250000	2015
Retail	12	91359	455488	1795266	8916432	2015
Technology	11	0	90000	206532	641635	2015
Other	15	708	61339	713133	6700142	2015

... and I want to fit this model:

$$y \sim GLM(g^{-1}(X\beta))$$

$$y_i \sim LogNormal(\text{mean} = \exp\left(\alpha + \beta_{year_i} + \beta_{sector_i}\right), \text{variance})$$

$$y_i \sim Pareto(\text{mean} = \exp\left(\alpha + \beta_{year_i} + \beta_{sector_i}\right))$$

Industry problem: Jargon

y	$\boldsymbol{x} = \{x_1, x_2 \ldots x_k\}$
Variate	Covariate
Dependent variable	Independent variable
Outcome	Predictor
Explained	Explanatory
Regressand	Regressor
Response variable	Controlled variable
Endogenous	Exogenous
Experimental variable	Manipulated variable
	Exposure variable
	Risk factor
Studied variable	Measured variable
Target	Feature
Output	Input

Industry problem: Jargon

y	$\boldsymbol{x} = \{x_1, x_2 \dots x_k\}$
Variate	Covariate
Dependent variable	Independent variable
Outcome	Predictor
Explained	Explanatory
Regressand	Regressor
Response variable	Controlled variable
Endogenous	Exogenous
Experimental variable	Manipulated variable
	Exposure variable
	Risk factor
Studied variable	Measured variable
Target	Feature
Output	Input

Formal definition

A function that includes the data and the parameters that is proportional to the probability of observing the data.

$$\mathcal{L}(\theta; x) \propto P_\theta(X = x)$$

Formal definition

A function that includes the data and the parameters that is proportional to the probability of observing the data.

$$\begin{aligned}\mathcal{L}(\theta; x) &\propto P_\theta(X = x)\\ &= P(X = x \mid \theta).\end{aligned}$$

Formal definition

A function that includes the data and the parameters that is proportional to the probability of observing the data.

$$
\begin{aligned}
\mathcal{L}(\theta; x) &\propto P_\theta(X = x) && \text{(Frequentist)} \\
&= P(X = x \mid \theta). && \text{(Bayesian)}
\end{aligned}
$$

Formal definition

Discrete

$$\begin{aligned}\mathcal{L}(\beta; x) &\propto P_\beta(X = x)\\ &= f_\beta(x)\end{aligned}$$

Formal definition

Discrete

$$\begin{aligned} \mathcal{L}(\beta; x) &\propto P_\beta(X = x) \\ &= f_\beta(x) \end{aligned}$$

Continuous

$$\begin{aligned} \mathcal{L}(\beta; x) &\propto P_\beta(X = x) \\ &= P_\beta(x < X \leq x + dx) \\ &= f_\beta(x)dx \\ &\propto f_\beta(x) \end{aligned}$$

Formal definition

Discrete

$$\begin{aligned}\mathcal{L}(\beta; x) &\propto P_\beta(X = x)\\ &= f_\beta(x)\end{aligned}$$

Continuous

$$\begin{aligned}\mathcal{L}(\beta; x) &\propto P_\beta(X = x)\\ &= P_\beta(x < X \leq x + dx)\\ &= f_\beta(x)dx\\ &\propto f_\beta(x)\end{aligned}$$

n i.i.d. observations

$$\mathcal{L}(\beta; x) = \prod_i^n f_\beta(x_i)$$

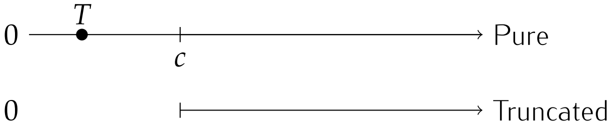
Measurement

Truncation

Under certain conditions, you don't see anything.

Example: You only see events if they occur after $c = 30$ days.
Likelihood contribution for event at time t.

Pure	Truncation present
$f(t)$	$\dfrac{f(t)}{1 - F(30)}$



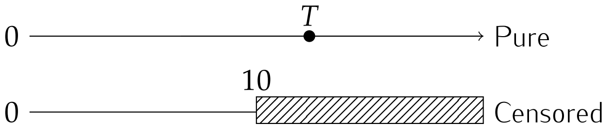
Measurement

Censoring

Under certain conditions, you only get limited information.

Example: You only have data for a 10 year window. If you don't see an event in this period, you know it occurs at $T > 10$.
Likelihood contribution for no event in your period of 10 years (the real event happened, say, at time 12.)

Pure	Censoring present
$f(12)$	$P(T > 10) = S(10)$



Our example

$$\begin{aligned} \text{stddev} &= cv \times \text{mean} \\ \sigma^2 &= \log(1 + (\text{stddev}^2)/(\text{mean}^2)) \\ \mu &= \log(\text{mean}) - \frac{1}{2}\sigma^2. \end{aligned}$$

Hence the following is equivalent:

$$y \sim \text{lognormal2}(\text{mean}_i, cv)$$

$$y \sim \text{lognormal}\Big(\mu = \log(\text{mean}) - \frac{1}{2}\log(1 + ((cv \times \text{mean})^2)/(\text{mean}^2))^2,$$
$$\sigma^2 = \log(1 + (cv \times \text{mean}^2)/(\text{mean}^2))\Big).$$

That is, we have reparametrised the lognormal distribution so that we can think in terms of its mean and coefficient of variation, instead of the unintuitive μ and σ.

If we have a random variable X_i, independently and identically distribution, with a cumulative distribution function (CDF) $F(x)$ and a density $f(x)$, we can derive the distribution of $Y = \max\left(\{X_1, X_2, X_3, \dots, X_n\}\right)$,

$$\begin{aligned}
F_Y(x) &= P\left(\max\left(\{X_1, X_2, X_3, \dots, X_n\}\right) < x\right) \\
&= P\left(\text{all } X_i \text{ less than } x\right) \\
&= P\left(X_1 < x, X_3 < x, X_3 < x, \dots, X_n < x\right) \\
&= P(X_1 < x)P(X_2 < x)P(X_3 < x)\dots P(X_n < x) \\
&= \prod_{i=1}^{n} P(X_i < x) \\
&= \prod_{i=1}^{n} F_X(x) \\
&= \left(F_X(x)\right)^n
\end{aligned}$$

and thus

$$\begin{aligned}
f_Y(x) &= \frac{d}{dx}\left(F_X(x)\right)^n \\
&= n\left(F_X(x)\right)^{n-1} f_X(x).
\end{aligned}$$

We can do the same for the minimum of a set of observations, or for any order statistic[1] Let $X_{(k)}$ be the kth order statistic, then

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!}[F_X(x)]^{k-1}[1-F_X(x)]^{n-k}f_X(x).$$

We can also write down the joint distribution for pairs of order statistics! For the pair of order statistics[2] $(X_{(j)}, X_{(k)})$ from n observations, the joint distribution is:

$$f_{X_{(j)},X_{(k)}}(x,y) = \frac{n!}{(j-1)!(k-j-1)!(n-k)!}\times$$
$$[F_X(x)]^{j-1}[F_X(y)-F_X(x)]^{k-1-j}[1-F_X(x)]^{n-k}f_X(x)f_X(y)$$

[1] `https://en.wikipedia.org/wiki/Order_statistic`

[2] `https://en.wikipedia.org/wiki/Order_statistic`

Let's carry on. For the three order statistics. Their distribution is

$$
\begin{aligned}
f_{X_{(j)},X_{(k)},X_{(l)}}(x,y,z) = \quad & \text{Big Ugly Equation} \\
& \frac{n!}{(j-1)!(k-j-1)!(l-k-1)!(n-k)!} \\
& \times [F_X(x)]^{j-1}[F_X(y)-F_X(x)]^{k-j-1} \\
& \times [F_X(z)-F_X(y)]^{l-k-1}[1-F_X(z)]^{n-l} \\
& \times f_X(x)f_X(y)f_X(z)
\end{aligned}
$$

Conclusion: Measurement and Likelihood

Each row in our dataset can be thought of as an observation from a joint distribution. For row i we have

$$\boldsymbol{y}_i = \{y_{\text{min}}, y_{\text{med}}, y_{\text{max}}, n\}_i$$

and its likelihood

$$\mathcal{L}(\theta; \boldsymbol{y}_i) = f_{X_{(1)}, X_{(n_i/2)}, X_{(n_i)}}(y_{i\text{min}}, y_{i\text{med}}, y_{i\text{max}})$$

and then we have the likelihood for all the data D in the N rows,

$$\mathcal{L}(\theta; \boldsymbol{D}) = \prod_{i=1}^{N} f_{X_{(1)}, X_{(n_i/2)}, X_{(n_i)}}(y_{i\text{min}}, y_{i\text{med}}, y_{i\text{max}})$$

Recall that we believe our data to be from a process:

$$y_i \sim LogNormal(\text{mean} = \exp\left(\alpha + \beta_{year_i} + \beta_{sector_i}\right), \text{variance})$$

so we will use $F(x)$ and $f(x)$ (the CDF and pdf) of the LogNormal distributoin to write down the Big Ugly Equation.

Estimation

We can write down the likelihood. Now we just have to maximise it! How?

Custom code	?	Hard to debug, error prone
Tensorflow	python	requires “hacks”
Pytorch	python	requires “hacks”
statsmodels	python	GenericLikelihoodModel
PyMC	python	syntax quite bad
Stan	stan	R + library(rstan)
		python + import cmdstanpy

A lot of the above make use of automatic differentiation libraries. There are a lot of these out there as well. Check which one is closest to what you are comfortable with.

Stan

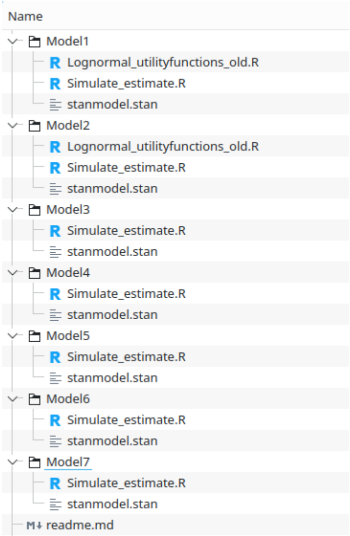
https://mc-stan.org/docs

R Code:

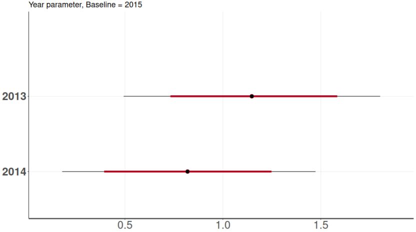
```r
library(rstan)
theta <- 0.1
n <- 1000
y <- rexp(n, theta)
my_dat <- list(N = n, y = y)
fit <- stan(
    file = 'stanmodel.stan',
    data = my_dat,  iter = 10000)
summary(fit)
plot(fit)
```

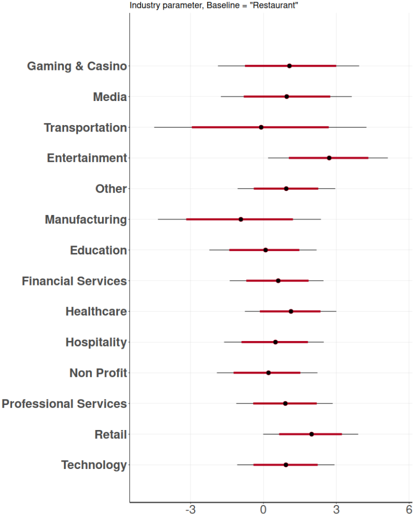
Stan code:

```stan
data {
  int<lower=0> N;
  vector[N] y;
}
  parameters {
  real<lower=0> theta;
}
model {
  y ~ exponential(theta);
  theta ~ lognormal(0, 1000);
}
```



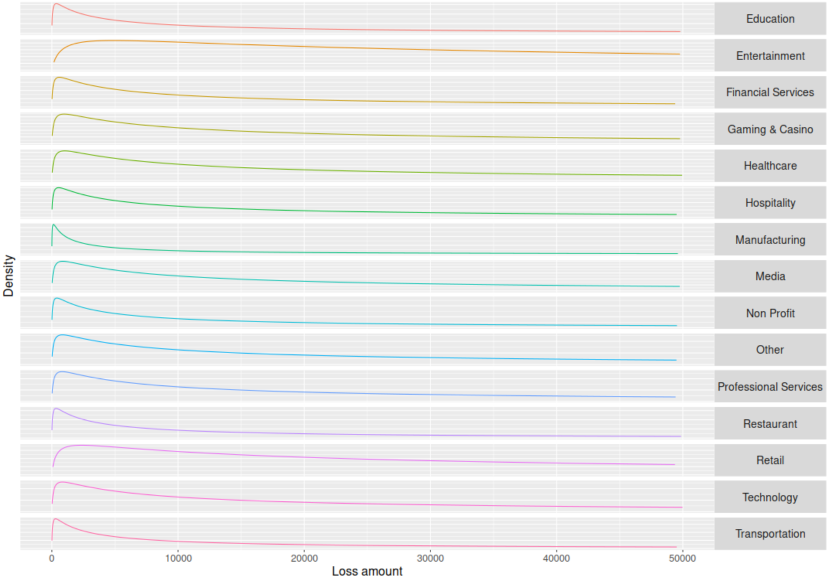
Test the code by simulating real data and checking that we can estimate the true parameters. Start simple add more and more complexity.





Recall that y_i is the claim severity.

$$y_i \sim LogNormal(\text{mean} = \exp\left(\alpha + \beta_{year_i} + \beta_{sector_i}\right), \text{variance})$$



$$y_i \sim LogNormal(\text{mean} = \exp\left(\alpha + \beta_{year_i} + \beta_{sector_i}\right), \text{variance})$$

```r
deductible <- 10e3
limit <- 10e6

payout <- function(x, deductible, limit){
  if (x < deductible)  return(0)
  else if (deductible <= x & x < limit )  return(x - deductible)
  else if (limit <= x )  return(limit - deductible )
}
payout <- Vectorize(payout, "x")

montecarlo_expected_claim <- function(mu, sigma, N=10000, deductible=deductible, limit=limit){
  sim_claim <- rlnorm(N, mu, sigma)
  sim_payouts <- payout(sim_claim, deductible, limit)
  return(mean(sim_payouts))
}
```

sector	mean_payout	claim_frequency	premium
Gaming & Casino	502,125	0.1	50,212.5
Media	435,513	0.1	43,551.3
Transportation	179,929	0.1	17,992.9
Entertainment	1,613,356	0.1	161,335.6
Other	449,043	0.1	44,904.3
Manufacturing	77,390	0.1	7,739.0
Education	195,717	0.1	19,571.7
Financial Services	327,431	0.1	32,743.1
Healthcare	519,231	0.1	51,923.1
Hospitality	291,054	0.1	29,105.4
Non Profit	238,819	0.1	23,881.9
Professional Services	418,962	0.1	41,896.2
Retail	1,032,428	0.1	103,242.8
Technology	432,961	0.1	43,296.1
Restaurant	202,587	0.1	20,258.7