Software Engineering for Probabilistic Programming

using CmdstanPy and plotnine

Mitzi Morris
Stan Development Team
Columbia University, New York NY

August 23, 2022

Software Engineering

Aim: develop, deploy, and maintain quality software.

Quality

- usability
- stability
- maintainability
- efficiency and scalability

Best practices

- design
- document
- test

Probabilistic Programming

What is a probabilistic program?

- the program specifies a ***parametric model*** of the ***data generating process***.
- the inference engine uses the data to infer (i.e. learn) the parameters; it ***solves the inverse problem***.

Applications *already in widespread use* try to predict

- student ability from test results,
- player / team abilities from pairwise matchups,
- drug efficacy from clinical trial data,
- population disease prevalence from tests of individuals,
- election outcomes from voter surveys and census data

Aim: develop and maintain quality probabilistic programs.

Quality

- usability, stability, maintainability, efficiency and scalability

Best practices

- design, document, test

What tools do we have?

- *Simulation*
- ***Data Visualization***

Working Example: Radon Levels in the Home

Radon is a radioactive gas known to cause lung cancer which comes from the decay of uranium-containing minerals in the ground.

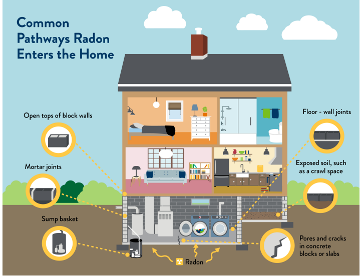


image source: Minnesota Department of Health

The data and models are taken from chapter 12 of the book *Data Analysis Using Regression and Multilevel/Hierarchical Models* by Andrew Gelman and Jennifer Hill, Cambridge Press, 2006.

Know Your Data!

Preliminary data analysis is part of the design process

Estimates and predictions are the result of conditioning the model **on the data**

- What do we want to estimate?
- What data is available?
- What are its size, shape, and tendencies?

EPA Radon Data for Minnesota

Home radon measurements

	floor	county	log_radon	county_id
1	0	AITKIN	0.788457	1
5	0	ANOKA	0.916291	2
58	1	BECKER	1.504077	3

County level measurements

	county	log_uranium	county_id	homes
0	AITKIN	-0.689048	1	4
1	ANOKA	-0.847313	2	52
2	BECKER	-0.113459	3	3

Outcome and predictor variables are on the log scale, per Gelman and Hill

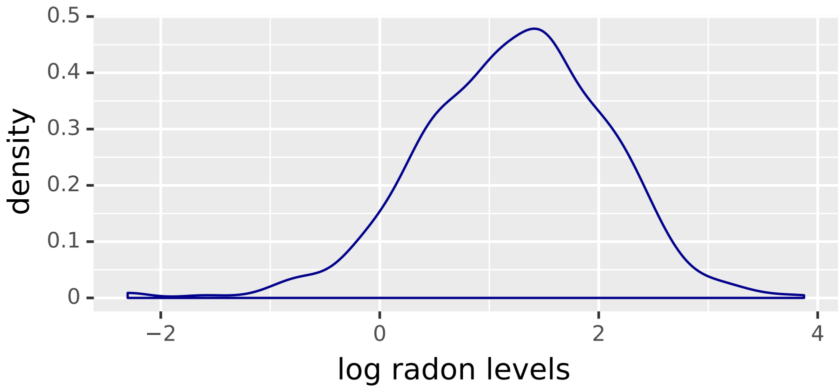
Preliminary Data Analysis


```
amount of data for Minnesota
 919 homes
 85 counties

summary statistics for log_radon
mean      1.224623
std       0.853327
min      -2.302585
25%       0.641854
50%       1.280934
75%       1.791759
max       3.875359
Name: log_radon, dtype: float64
```

Density Plots

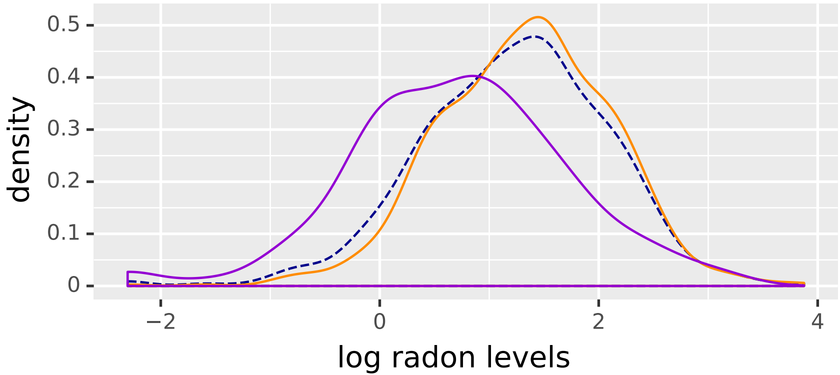
Use plotnine.geom_density to show a smooth density line




```
<ggplot: (8767010100209)>
```

Density Plots

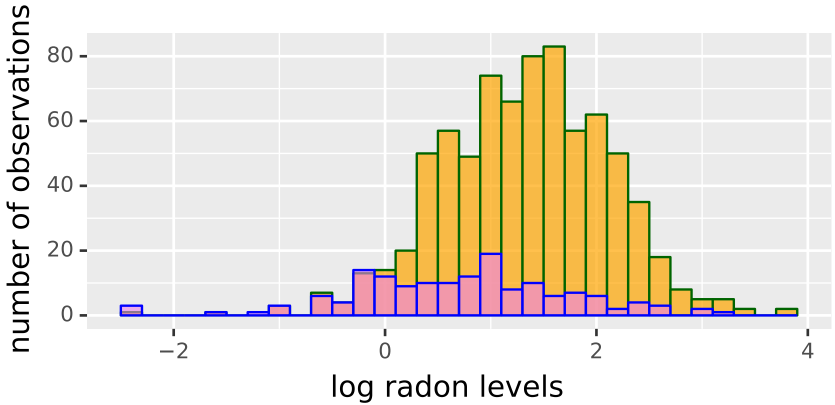
Use multiple plotnine.geom_density layers: basement: orange, ground floor: violet



<ggplot: (8767077850260)>

Histogram Plots

Use plotnine.geom_histogram to show counts by floor: basement: orange, ground floor: violet



Measurements by Floor

The raw counts histogram reveals that most of the observations in the survey were taken on the basement level. Let's compute the exact percentages.

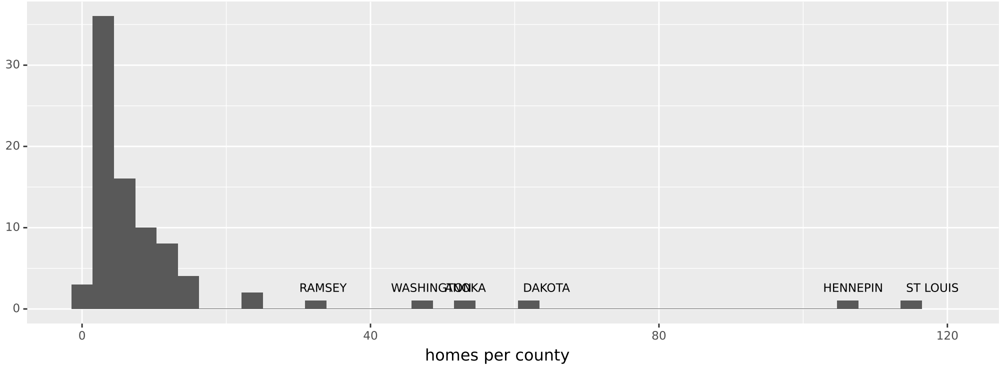
```
floor 0: 83%
floor 1: 17%
```

How many counties have data from both floors?

```
Number of counties: 85
Counties with measurements from floor 0: 85
Counties with measurements from floor 1: 60
```

Homes per County

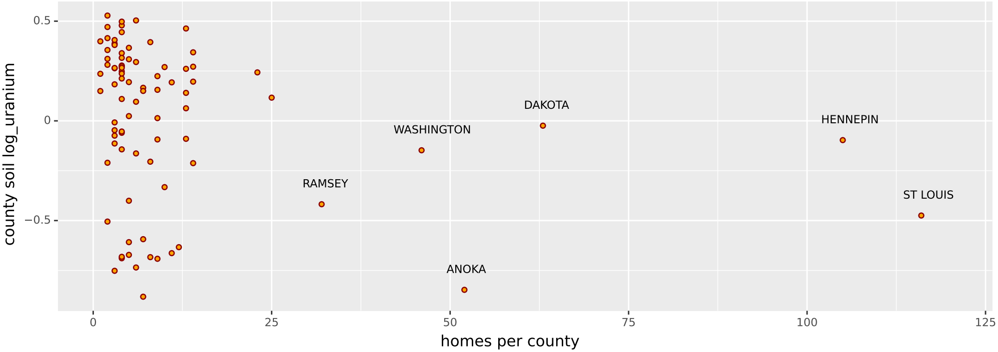
Histogram of homes (observations) per county - a few metropolitan areas have high counts, most counties have fewer than 10 measurements.




```
<ggplot: (8767062080570)>
```

County Level Soil Uranium

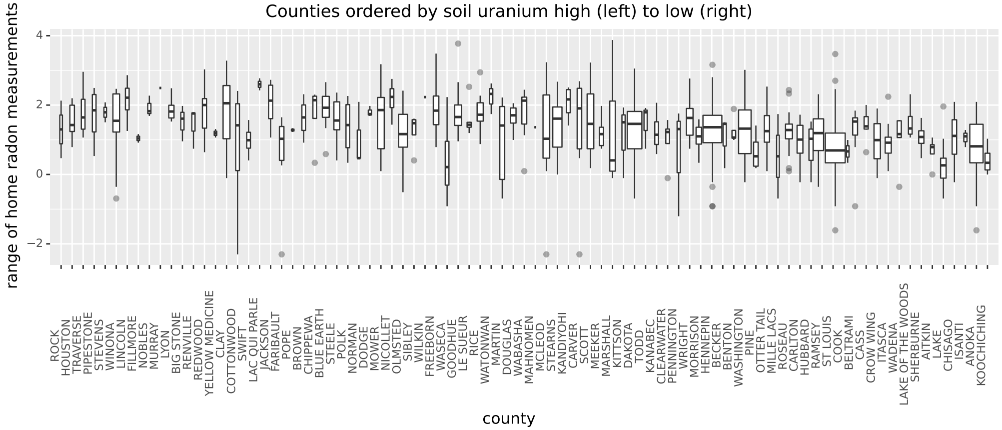
Plot soil log uranium on the y-axis.



<ggplot: (8767045613460)>

Boxplot: Visual Summaries by County

plotnine.geom_boxplot shows central intervals and outliers of home radon measurements.



Preliminary Data Analysis Findings

- Within each county, the range of radon measurements is very wide.
- 83% of the data are measurements taken on the basement level.
- 70% of the counties (60 out of 85) have observations from both floors 0 and 1, the remaining 30% only have observations from floor 0 (basement).
- For most counties, there are fewer than 10 observations; 8 counties in metropolitan areas account for over half of the observations.
- Soil uranium levels vary, very little data from counties with high levels.

Model Building and Model Testing

Systematically test the predictions made by the model.

- Prior Predictive Checking - what does your model predict before it sees any data?
- Posterior Predictive Checking - what does your model predict once it has seen the data?

To see how this works, we use the multi-level model developed in Chapter 12, model 12.2.

Statistical Notation (quick review)

- y - data
- θ - parameters
- $\mathrm{p}(y, \theta)$ - **joint probability distribution** of the data and parameters
- $\mathrm{p}(\theta)$ - **prior probability distribution** - the probability of the parameters before any data are observed
- $\mathrm{p}(\theta \mid y)$ - **posterior probability distribution** - the probability of the parameters conditional on the data
- $\mathrm{p}(y \mid \theta)$ - *probability of the data given the parameters*
  - if y is fixed, this is the **likelihood function**
  - if θ is fixed, this is the **sampling distribution**

Bayesian Estimation (quick review)

Bayesian estimation relates the **conditional probability** of the parameters (θ) given the data (y), i.e., $\mathrm{p}(\theta \mid y)$, to the **joint probability** of parameters and data, $\mathrm{p}(\theta, y)$.

$$\mathrm{p}(\theta \mid \mathrm{y}) = \frac{\mathrm{p}(\mathrm{y}, \theta)}{\mathrm{p}(\mathrm{y})} \qquad \text{[def of conditional probability]}$$

$$= \frac{\mathrm{p}(\mathrm{y} \mid \theta)\,\mathrm{p}(\theta)}{\mathrm{p}(\mathrm{y})} \qquad \text{[rewrite joint probability as conditional]}$$

p(y) doesn't depend on θ - it's a constant for fixed y - we can drop it

$$\mathrm{p}(\theta \mid \mathrm{y}) \propto \mathrm{p}(\mathrm{y} \mid \theta)\,\mathrm{p}(\theta) \qquad \text{[unnormalized posterior density]}$$

The posterior is **proportional** to the **prior** times the **likelihood**

Linear Regression (quick review)

Linear regression finds the linear function - a non-vertical straight line, plane, hyperplane - which best relates a scalar outcome (the dependent variable "y") to one or more predictors (the independent variable "x").

$$y_i = \alpha + \beta x_i + \epsilon_i$$

- α is the *intercept*, the offset from zero on the x-axis
- β is the *slope*, the multiplier applied to x.
- ϵ_i is the error term, assuming independent errors drawn from a normal distribution with mean 0, standard deviation σ.

Simple Linear Regression in Stan


```
data {
  int<lower=0> N;
  vector[N] x;
  vector[N] y;
}
parameters {
  real alpha;
  real beta;
  real<lower=0> sigma;
}
model {
  y ~ normal(alpha + beta * x, sigma);
}
```

This model needs priors on the regression parameters! See the Stan prior choice wiki for recommendations.

Fitting the radon data

$$\text{log\_radon}_i = \alpha + \beta\,\text{floor}_i + \epsilon_i$$

- outcome "y" is the log radon level
- the predictor "x" is the floor on which the measurement was taken.
- the basement level is coded as $\text{floor} = 0$, thus for basement measurements

$$\text{log\_radon}_i = \alpha + \epsilon_i$$

Multilevel Regression

Multilevel regression models the *dependency structures in the data* in addition to the relation between outcome and predictors, allowing for ***partial pooling*** of information. In this example

- Houses are located within counties
- Counties are drawn from a common distribution

Multi-level model for the intercept term α, for I homes across J counties:

$$\mathrm{log\_radon}_i \sim \mathrm{N}(\alpha_{j[i]} + \beta\,\mathrm{floor}_i,\, \sigma^2) \quad \text{for} \;\; i = 1, \ldots, \mathrm{I}$$
$$\alpha_j \sim \mathrm{N}(\mu_\alpha,\, {\sigma_\alpha}^2), \quad \text{for} \;\; j = 1, \ldots, \mathrm{J}$$

County-level and home-level distributions are estimated jointly.

Multilevel Radon Model

Data:

- `N`: number of houses
- `J`: number of counties
- `y`: log_radon measurements (size N)
- `x`: floor (size N)
- `county`: county index (size N, values 1:J)

Parameters:

- `alpha`: a vector of per-county intercept terms (size J)
- `mu_alpha`, `mu_sigma`: the mean and variance for the distribution over `alpha`
- `beta`: the coefficient for the floor level
- `sigma`: common variance

Multilevel Radon Model

Model:

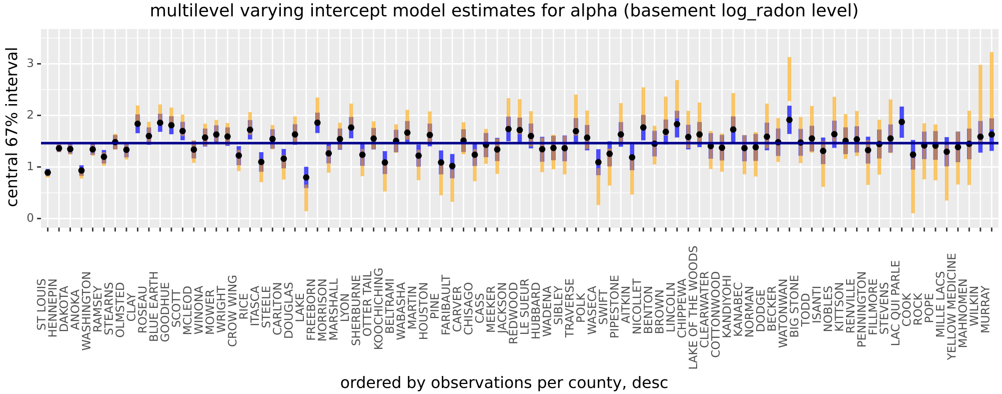
```
y ~ normal(alpha[county] + beta * x, sigma);
alpha ~ normal(mu_alpha, sigma_alpha); // partial-pooling
beta ~ normal(0, 10);  // weakly informative priors
sigma ~ normal(0, 10);
mu_alpha ~ normal(0, 10);
sigma_alpha ~ normal(0, 10);
```

Compare to unmodeled per-county intercepts

```
y ~ normal(alpha[county] + beta * x, sigma);
beta ~ normal(0, 10);  // weakly informative priors
sigma ~ normal(0, 10);
```

Model Comparison

- orange - estimate from regular regression model; no-pooling
- blue - estimate from multilevel model; partial-pooling




```
<ggplot: (8767062009242)>
```

Posterior Predictive Check

"Posterior predictive checks are the unit tests of probabilistic programming." – Ben Goodrich

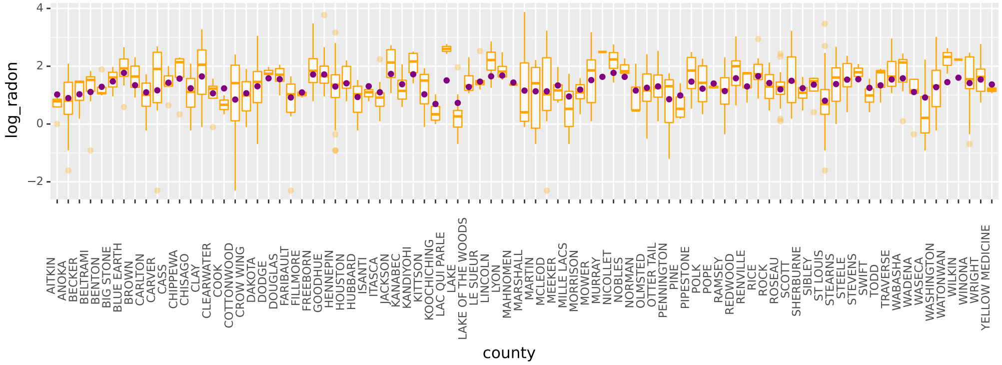
Simulate a new data set using the fitted model parameters:

```
generated quantities {
  array[N] real y_rep =
      normal_rng(alpha[county] + beta * x, sigma);
}
```

- Each iteration of the sampler generates a new dataset `y_rep` (replicate)
- Compute statistics on dataset `y_rep`; compare with those on `y`.

Posterior Predictive Test

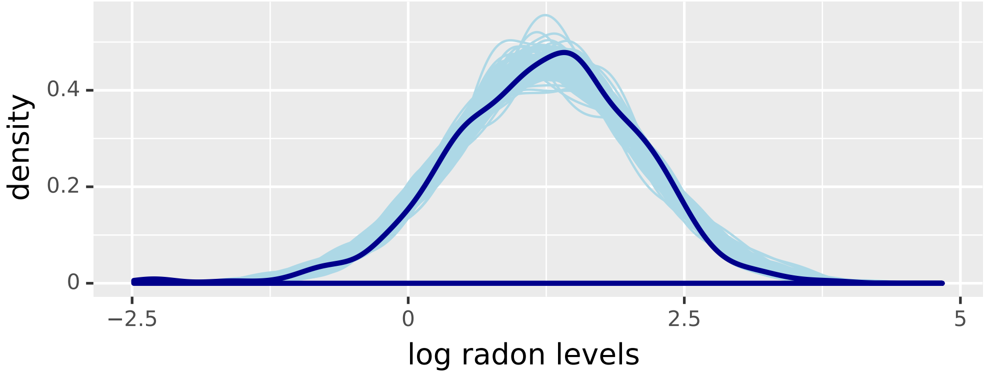
Estimated mean home radon level per county overlaid on boxplot of y




```
<ggplot: (8767078380535)>
```

Posterior Predictive Density Plot

Overlay the density plots of a random sample of `y_rep` with density plot of `y`




```
<ggplot: (8767035328600)>
```

Concluding remarks

Data visualization drives design, testing, and documentation.

Prior and posterior predictive checks validate the model specification.

Plotnine provides a rich set of visualizations to demonstrate and communicate data, inferences, and predictions.

We welcome feedback and ideas for ways to turn these plots into easy-to-use tools.

Many Thanks!

Thanks! to members of the Stan Team

Thanks! to Reshama and Data Umbrella

Thanks! to everyone for watching

Any Questions?