Gelman – Chapter 1 – Why?

Data Analysis Using Regression and Multilevel/Hierarchical Models

Andrew Gelman
Jennifer Hill

June 13, 2012

Gelman – Chapter 1 – Why?

Structure

- What is multilevel regression modelling?
- Some examples from our own research
- Motivations for multilevel modelling
- Computing

Gelman – Chapter 1 – Why?

What is multilevel regression modelling?

- Example – an educational study predicting in each school the students' grades y on a standardised test given their scores on a pre-test x and other information

Gelman – Chapter 1 – Why?

What is multilevel regression modelling?

- Example – an educational study predicting in each school the students' grades y on a standardised test given their scores on a pre-test x and other information
- A multilevel model is a regression (a linear or generalised linear model) in which the parameters – the regression coefficients – are given a probability model

Gelman – Chapter 1 – Why?

What is multilevel regression modelling?

- Example – an educational study predicting in each school the students' grades y on a standardised test given their scores on a pre-test x and other information
- A multilevel model is a regression (a linear or generalised linear model) in which the parameters – the regression coefficients – are given a probability model
- In our student example the second-level model – the school model – has parameters of its own – the hyperparameters of the model – which are also estimated from data

Gelman – Chapter 1 – Why?

Models for regression coefficients

- Keep our example simple – one student-level predictor x – pre-test score and one school-level predictor u – average parents' income

Gelman – Chapter 1 – Why?

Models for regression coefficients

- Keep our example simple – one student-level predictor x – pre-test score and one school-level predictor u – average parents' income
- **Varying-intercept model** – i for individual student and $j[i]$ for the school j containing student i

$$y_i = \alpha_{j[i]} + \beta x_i + \epsilon_i, \text{ for students } i = 1, \ldots, n$$
$$\alpha_j = a + bu_j + \eta_j, \text{ for schools } j = 1, \ldots J$$

Gelman – Chapter 1 – Why?

Models for regression coefficients

- Keep our example simple – one student-level predictor x – pre-test score and one school-level predictor u – average parents' income
- **Varying-intercept model** – i for individual student and $j[i]$ for the school j containing student i

$$y_i = \alpha_{j[i]} + \beta x_i + \epsilon_i, \text{ for students } i = 1, \ldots, n$$
$$\alpha_j = a + bu_j + \eta_j, \text{ for schools } j = 1, \ldots J$$

- **Varying-intercept, varying slope model**

$$y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i, \text{ for students } i = 1, \ldots, n$$
$$\alpha_j = a_0 + b_0 u_j + \eta_{j1}, \text{ for schools } j = 1, \ldots J$$
$$\beta_j = a_1 + b_1 u_j + \eta_{j2}, \text{ for schools } j = 1, \ldots J$$

Gelman – Chapter 1 – Why?

Home radon measurement and remediation

- Random sample of more than 80,000 homes throughout the U.S.

Gelman – Chapter 1 – Why?

Home radon measurement and remediation

- Random sample of more than 80,000 homes throughout the U.S.
- Goal – estimate the distribution of radon levels in each of the approximately 3,000 counties in the U.S.

Gelman – Chapter 1 – Why?

Home radon measurement and remediation

- Random sample of more than 80,000 homes throughout the U.S.
- Goal – estimate the distribution of radon levels in each of the approximately 3,000 counties in the U.S.
- Hierarchical structure – houses within counties

Gelman – Chapter 1 – Why?

Home radon measurement and remediation

- Random sample of more than 80,000 homes throughout the U.S.
- Goal – estimate the distribution of radon levels in each of the approximately 3,000 counties in the U.S.
- Hierarchical structure – houses within counties
- House-level predictor – floor on which measurement was taken (0 for basement, 1 for first floor)

Gelman – Chapter 1 – Why?

Home radon measurement and remediation

- Random sample of more than 80,000 homes throughout the U.S.
- Goal – estimate the distribution of radon levels in each of the approximately 3,000 counties in the U.S.
- Hierarchical structure – houses within counties
- House-level predictor – floor on which measurement was taken (0 for basement, 1 for first floor)
- County-level predictor – measurement of the soil uranium

Gelman – Chapter 1 – Why?

Home radon measurement and remediation

- Random sample of more than 80,000 homes throughout the U.S.
- Goal – estimate the distribution of radon levels in each of the approximately 3,000 counties in the U.S.
- Hierarchical structure – houses within counties
- House-level predictor – floor on which measurement was taken (0 for basement, 1 for first floor)
- County-level predictor – measurement of the soil uranium
- Our varying-intercept model

  y_i is the log of the radon measurement in house i

Gelman – Chapter 1 – Why?

Home radon measurement and remediation

- Random sample of more than 80,000 homes throughout the U.S.
- Goal – estimate the distribution of radon levels in each of the approximately 3,000 counties in the U.S.
- Hierarchical structure – houses within counties
- House-level predictor – floor on which measurement was taken (0 for basement, 1 for first floor)
- County-level predictor – measurement of the soil uranium
- Our varying-intercept model

  y_i is the log of the radon measurement in house i

  x is the floor of the measurement

Gelman – Chapter 1 – Why?

Home radon measurement and remediation

- Random sample of more than 80,000 homes throughout the U.S.
- Goal – estimate the distribution of radon levels in each of the approximately 3,000 counties in the U.S.
- Hierarchical structure – houses within counties
- House-level predictor – floor on which measurement was taken (0 for basement, 1 for first floor)
- County-level predictor – measurement of the soil uranium
- Our varying-intercept model

  y_i is the log of the radon measurement in house i
  x is the floor of the measurement
  u is the uranium measurement at the county level

Gelman – Chapter 1 – Why?

Home radon measurement and remediation

- Random sample of more than 80,000 homes throughout the U.S.
- Goal – estimate the distribution of radon levels in each of the approximately 3,000 counties in the U.S.
- Hierarchical structure – houses within counties
- House-level predictor – floor on which measurement was taken (0 for basement, 1 for first floor)
- County-level predictor – measurement of the soil uranium
- Our varying-intercept model

  y_i is the log of the radon measurement in house i

  x is the floor of the measurement

  u is the uranium measurement at the county level

  ϵ_i within-county variation: measurement error, natural variation in radon levels within a house over time, variation between houses

Gelman – Chapter 1 – Why?

Home radon measurement and remediation

- Random sample of more than 80,000 homes throughout the U.S.
- Goal – estimate the distribution of radon levels in each of the approximately 3,000 counties in the U.S.
- Hierarchical structure – houses within counties
- House-level predictor – floor on which measurement was taken (0 for basement, 1 for first floor)
- County-level predictor – measurement of the soil uranium
- Our varying-intercept model

  y_i is the log of the radon measurement in house i

  x is the floor of the measurement

  u is the uranium measurement at the county level

  ϵ_i within-county variation: measurement error, natural variation in radon levels within a house over time, variation between houses

  η_j between-county variation: beyond what is explained by the county-level uranium predictor

Gelman – Chapter 1 – Why?

Forecasting presidential elections

- Predicted outcomes – the Democratic Party's share of the two-party vote in each of the 11 elections (1948–1988), yielding 511 points

Gelman – Chapter 1 – Why?

Forecasting presidential elections

- Predicted outcomes – the Democratic Party's share of the two-party vote in each of the 11 elections (1948–1988), yielding 511 points
- Various predictors $(1 \cdots k)$including:

  The performance of the Democrats in the previous election

Gelman – Chapter 1 – Why?

Forecasting presidential elections

- Predicted outcomes – the Democratic Party's share of the two-party vote in each of the 11 elections (1948–1988), yielding 511 points
- Various predictors $(1 \cdots k)$including:

  The performance of the Democrats in the previous election
  Measures of state-level and national level economic trends

Gelman – Chapter 1 – Why?

Forecasting presidential elections

- Predicted outcomes – the Democratic Party's share of the two-party vote in each of the 11 elections (1948–1988), yielding 511 points
- Various predictors $(1 \cdots k)$including:

  The performance of the Democrats in the previous election
  Measures of state-level and national level economic trends
  National opinion polls up to two months before the election

Gelman – Chapter 1 – Why?

Forecasting presidential elections

- Predicted outcomes – the Democratic Party's share of the two-party vote in each of the 11 elections (1948–1988), yielding 511 points
- Various predictors $(1 \cdots k)$including:

  The performance of the Democrats in the previous election
  Measures of state-level and national level economic trends
  National opinion polls up to two months before the election
- Goal – make prediction for 1992 election for the 50 states

Gelman – Chapter 1 – Why?

Forecasting presidential elections – Model

- $y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots + X_{ik}\beta_k + \eta_{t[i]} + \delta_{r[i],t[i]} + \epsilon_i$, for $i = 1, ..., n$, where

Gelman – Chapter 1 – Why?

Forecasting presidential elections – Model

- $y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots + X_{ik}\beta_k + \eta_{t[i]} + \delta_{r[i],t[i]} + \epsilon_i$, for $i = 1, ..., n$, where

  $t[i]$ is an time indicator (election year)
  $r[i]$ is an region indicator (Northeast, Midwest, South or West)
  $n = 511$ is the number of state-years used to fit the model
  For each election year, η_t is a nationwide error and the $\delta_{r,t}$'s are four independent regional errors

Gelman – Chapter 1 – Why?

Forecasting presidential elections – Model

- $y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots + X_{ik}\beta_k + \eta_{t[i]} + \delta_{r[i],t[i]} + \epsilon_i$, for $i = 1, ..., n$, where

  $t[i]$ is an time indicator (election year)
  $r[i]$ is an region indicator (Northeast, Midwest, South or West)
  $n = 511$ is the number of state-years used to fit the model
  For each election year, η_t is a nationwide error and the $\delta_{r,t}$'s are four independent regional errors

- Normal distributions are used for error terms:

  $\eta_t \sim N(0, \sigma_\eta^2)$, for $t = 1, \cdots, 11$
  $\delta_{r,t} \sim N(0, \sigma_\delta^2)$, for $r = 1, \cdots, 4, t = 1, \cdots, 11$
  $\epsilon_i \sim N(0, \sigma_\epsilon^2)$, for $i = 1, \cdots, 511$

Gelman – Chapter 1 – Why?

Forecasting presidential elections – Model

- $y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots + X_{ik}\beta_k + \eta_{t[i]} + \delta_{r[i],t[i]} + \epsilon_i$, for $i = 1, ..., n$, where

  $t[i]$ is an time indicator (election year)
  $r[i]$ is an region indicator (Northeast, Midwest, South or West)
  $n = 511$ is the number of state-years used to fit the model
  For each election year, η_t is a nationwide error and the $\delta_{r,t}$'s are four independent regional errors

- Normal distributions are used for error terms:

  $\eta_t \sim N(0, \sigma_\eta^2)$, for $t = 1, \cdots, 11$
  $\delta_{r,t} \sim N(0, \sigma_\delta^2)$, for $r = 1, \cdots, 4, t = 1, \cdots, 11$
  $\epsilon_i \sim N(0, \sigma_\epsilon^2)$, for $i = 1, \cdots, 511$

- All the parameters $\beta, \sigma_\eta, \sigma_\delta, \sigma_\epsilon$ are estimated from the data.

Gelman – Chapter 1 – Why?

The Central Limit Theorem

- The Central Limit Theorem of probability states that the sum of many small independent random variables will be a random variable with an approximate normal distribution. (Page 14)

Gelman – Chapter 1 – Why?

The Central Limit Theorem

- The Central Limit Theorem of probability states that the sum of many small independent random variables will be a random variable with an approximate normal distribution. (Page 14)
- We write the summation of independent components as

$$z = \sum_{i=1}^{n} z_i$$

Gelman – Chapter 1 – Why?

The Central Limit Theorem

- The Central Limit Theorem of probability states that the sum of many small independent random variables will be a random variable with an approximate normal distribution. (Page 14)
- We write the summation of independent components as
  $z = \sum_{i=1}^{n} z_i$
- Then the mean and variance of z are the sums of the means and variances of the z_i's:
  $\mu_z = \sum_{i=1}^{n} \mu_{z_i}$ and $\sigma_z = \sqrt{\sum_{i=1}^{n} \sigma_{z_i}^2}$.

Gelman – Chapter 1 – Why?

The Central Limit Theorem

- The Central Limit Theorem of probability states that the sum of many small independent random variables will be a random variable with an approximate normal distribution. (Page 14)
- We write the summation of independent components as

  $$z = \sum_{i=1}^{n} z_i$$

- Then the mean and variance of z are the sums of the means and variances of the z_i's:

  $$\mu_z = \sum_{i=1}^{n} \mu_{z_i} \text{ and } \sigma_z = \sqrt{\sum_{i=1}^{n} \sigma_{z_i}^2}.$$

- We write this as:

  $$z \sim N(\mu_z, \sigma_z^2).$$

Gelman – Chapter 1 – Why?

Forecasting presidential elections – 1992 Prediction

- We can then make a prediction by simulating the election outcome in the 50 states in the next election year, $t = 12$:

$$y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots + X_{ik}\beta_k + \eta_{12} + \delta_{r[i],12} + \epsilon_i, \text{ for } i = n+1, ..., n+50$$

Gelman – Chapter 1 – Why?

Forecasting presidential elections – 1992 Prediction

- We can then make a prediction by simulating the election outcome in the 50 states in the next election year, $t = 12$:

  $y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots + X_{ik}\beta_k + \eta_{12} + \delta_{r[i],12} + \epsilon_i$, for $i = n+1, ..., n+50$

- For this we need as before:

  $y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots + X_{ik}\beta_k$

  State-level errors ϵ

Gelman – Chapter 1 – Why?

Forecasting presidential elections – 1992 Prediction

- We can then make a prediction by simulating the election outcome in the 50 states in the next election year, $t = 12$:

$$y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots + X_{ik}\beta_k + \eta_{12} + \delta_{r[i],12} + \epsilon_i, \text{ for } i = n+1, \ldots, n+50$$

- For this we need as before:

$$y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots + X_{ik}\beta_k$$

  State-level errors ϵ

- and new from our distributions:

  A new national error η_{12}

  4 new regional errors $\delta_{r,12}$

Gelman – Chapter 1 – Why?

Multilevel Models v Classical Regression

- **Learning about treatment effects that vary.** How does y change when some x is varied, with all other inputs held constant? E.g., a particular educational innovation may be more effective for girls than for boys.

Gelman – Chapter 1 – Why?

Multilevel Models v Classical Regression

- **Learning about treatment effects that vary.** How does y change when some x is varied, with all other inputs held constant? E.g., a particular educational innovation may be more effective for girls than for boys.

  Multilevel models: Allow us to study effects that vary by group, e.g., an intervention that is more effective in some schools than others (because of some unmeasured school-level factor)

Gelman – Chapter 1 – Why?

Multilevel Models v Classical Regression

- **Learning about treatment effects that vary.** How does y change when some x is varied, with all other inputs held constant? E.g., a particular educational innovation may be more effective for girls than for boys.

  Multilevel models: Allow us to study effects that vary by group, e.g., an intervention that is more effective in some schools than others (because of some unmeasured school-level factor)

  Classical regression: Estimates of varying effects can be noisy, especially when there are few observations per group

Gelman – Chapter 1 – Why?

Multilevel Models v Classical Regression (cont.)

- **Analysis of structured data.** Some datasets are collected with an inherent multilevel structure, e.g., students within schools, patients within hospitals.

Gelman – Chapter 1 – Why?

Multilevel Models v Classical Regression (cont.)

- **Analysis of structured data.** Some datasets are collected with an inherent multilevel structure, e.g., students within schools, patients within hospitals.

  Multilevel modelling: A direct way to include indicators for clusters at all levels of a design.

Gelman – Chapter 1 – Why?

Multilevel Models v Classical Regression (cont.)

- **Analysis of structured data.** Some datasets are collected with an inherent multilevel structure, e.g., students within schools, patients within hospitals.

  Multilevel modelling: A direct way to include indicators for clusters at all levels of a design.

  Classical regression: Problems of overfitting with large numbers of parameters.

Gelman – Chapter 1 – Why?

R & Bugs

- R – a free open-source version of S

Gelman – Chapter 1 – Why?

R & Bugs

- **R – a free open-source version of S**

  A program developed in the '70s and '80s at Bell Laboratories

Gelman – Chapter 1 – Why?

R & Bugs

- **R – a free open-source version of S**

  A program developed in the '70s and '80s at Bell Laboratories
  It is excellent for graphics and classical statistical modelling – most relevant here are **lm()** and **glm()**

Gelman – Chapter 1 – Why?

R & Bugs

- **R – a free open-source version of S**

  A program developed in the '70s and '80s at Bell Laboratories
  It is excellent for graphics and classical statistical modelling – most relevant here are **lm()** and **glm()**
  Bugs can be called from R.

Gelman – Chapter 1 – Why?

R & Bugs

- **R – a free open-source version of S**

  A program developed in the '70s and '80s at Bell Laboratories
  It is excellent for graphics and classical statistical modelling – most relevant here are **lm()** and **glm()**
  Bugs can be called from R.

- **Bugs – Bayesian Inference using Gibbs Sampling**

Gelman – Chapter 1 – Why?

R & Bugs

- **R – a free open-source version of S**

  A program developed in the '70s and '80s at Bell Laboratories
  It is excellent for graphics and classical statistical modelling – most relevant here are **lm()** and **glm()**
  Bugs can be called from R.

- **Bugs – Bayesian Inference using Gibbs Sampling**

  Program developed by statisticians at the Medical Research Council in Cambridge

Gelman – Chapter 1 – Why?

R & Bugs

- **R – a free open-source version of S**

  A program developed in the '70s and '80s at Bell Laboratories
  It is excellent for graphics and classical statistical modelling – most relevant here are **lm()** and **glm()**
  Bugs can be called from R.

- **Bugs – Bayesian Inference using Gibbs Sampling**

  Program developed by statisticians at the Medical Research Council in Cambridge
  Most powerful versions available – WinBugs 1.4 & OpenBugs

Gelman – Chapter 1 – Why?

R & Bugs

- **R – a free open-source version of S**

  A program developed in the '70s and '80s at Bell Laboratories

  It is excellent for graphics and classical statistical modelling – most relevant here are **lm()** and **glm()**

  Bugs can be called from R.

- **Bugs – Bayesian Inference using Gibbs Sampling**

  Program developed by statisticians at the Medical Research Council in Cambridge

  Most powerful versions available – WinBugs 1.4 & OpenBugs

  The Bugs modelling language has a modular form that allows the user to put together all sorts of Bayesian models, including multilevel models

Gelman – Chapter 1 – Why?

R & Bugs

- **R – a free open-source version of S**

  A program developed in the '70s and '80s at Bell Laboratories
  It is excellent for graphics and classical statistical modelling – most relevant here are **lm()** and **glm()**
  Bugs can be called from R.

- **Bugs – Bayesian Inference using Gibbs Sampling**

  Program developed by statisticians at the Medical Research Council in Cambridge
  Most powerful versions available – WinBugs 1.4 & OpenBugs
  The Bugs modelling language has a modular form that allows the user to put together all sorts of Bayesian models, including multilevel models
  Methodology – set up data in R, fit models in Bugs, then go back to R for further statistical analysis using the fitted models.

Gelman – Chapter 1 – Why?

Where next?

Gelman – Chapter 1 – Why?

Where next?

- The deep end!

Gelman – Chapter 1 – Why?

Where next?

- The deep end!
- Part 2B, Chapter 16, Page 345 – Multilevel modelling in Bugs and R: the basics