Intro to hierarchical models

Why multilevel regression modeling?

Consider a housing dataset that contains information about sales of 2000 houses across 100 different zipcodes.

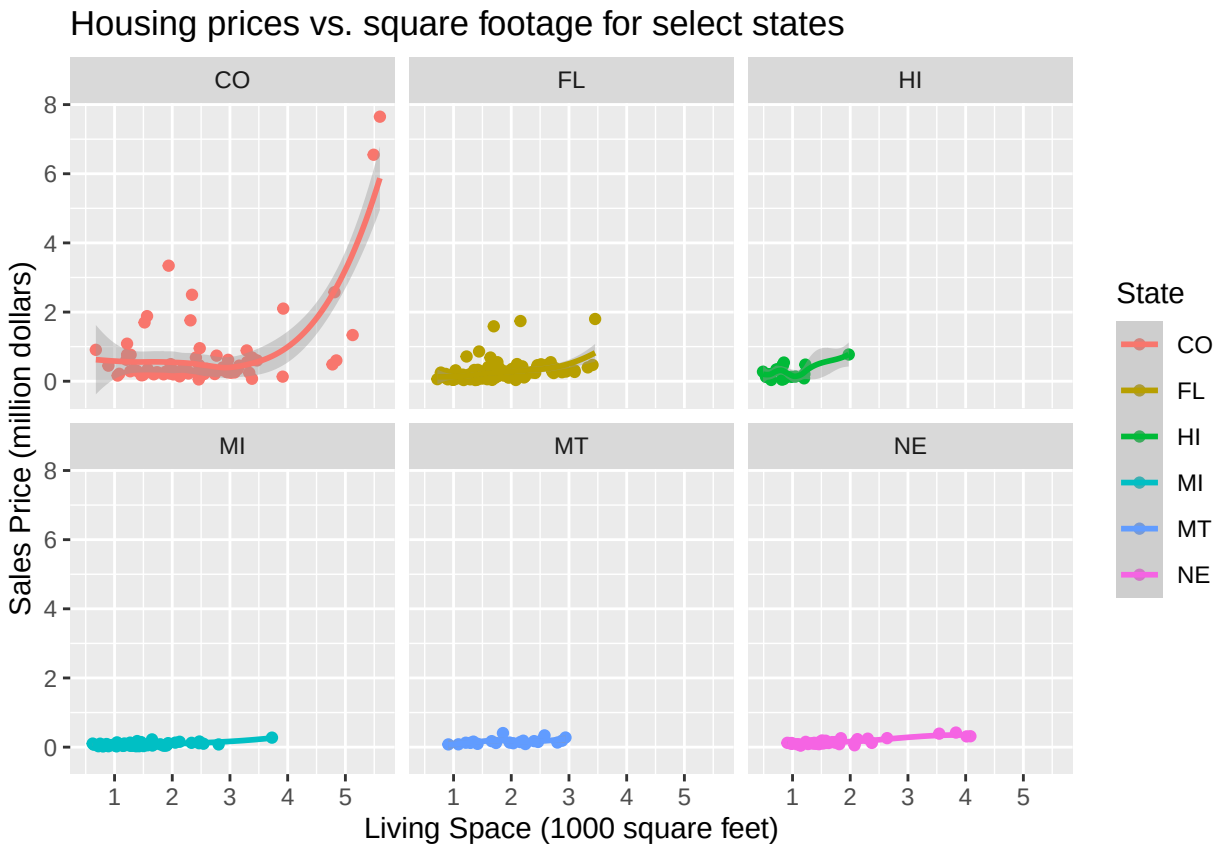
```
housing_sales <- read_csv('http://math.montana.edu/ahoegh/teaching/stat408/datasets/HousingSales.csv')
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   City = col_character(),
##   State = col_character(),
##   Zip_Code = col_double(),
##   Living_Sq_Ft = col_double(),
##   Closing_Price = col_double()
## )
```

Q: Do you expect to see the same relationship between the size of the home `Living_Sq_Ft` and the sales price for all cities? Note a few cities in this dataset include Lincoln, NE; Lewistown, MT; Miami, FL; Honolulu, HI; Snowmass, CO; and Flint, MI.

```
housing_sales %>% filter(State %in% c("NE", "MT", "CO", 'HI','FL','MI')) %>%
  mutate(sales_price = Closing_Price / 1000000, thousand_sq_ft = Living_Sq_Ft / 1000) %>%
  ggplot(aes(y = sales_price, x = thousand_sq_ft, color = State)) +
  geom_point() + geom_smooth(method = 'loess') +
  xlab('Living Space (1000 square feet)') +
  ylab('Sales Price (million dollars)') + facet_wrap(.~State) +
  ggtitle('Housing prices vs. square footage for select states')
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Another option would be to fit separate models for each zipcode. What are some of the implications for this type of model?

A multilevel, or hierarchical model, contains another level that models the covariates from each individual level model.

Thus rather than

$$Y_i = \alpha + \beta X_i + \epsilon_i,$$

where $i = 1, ..., n$ corresponds to the n houses, the model can be written as

Terminology

These multilevel or hierarchical models carry this designation for two reasons:

1. There are multiple levels in the data structure. In this case, consider houses nested in zipcodes.

2. The model also has multiple levels.

Multilevel models could also be applied for several layers...

About Mixed Models / Random Effects The authors intentionally avoid the term "random effects" and hence, mixed models. More on this later...

GH include several interesting applications of hierarchical models from their own research. Read through these in Chapter 1.2.

Motivations for using hierarchical models **Learn about treatment effects that may vary:**

Use all of the data to perform inferences for groups with small sample size:

Prediction:

Analysis of Structured Data:

More efficient inference for regression parameters:

Including predictors at multiple levels

Getting the right standard error accurately accounting for uncertainty in prediction and estimation:

Multilevel Models

For multilevel models, observations fall into groups and coefficients can vary by the group.

Assume there are J groups and $j[i]$ denotes that observation i falls into group j

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

$$\begin{aligned} y_{[1]i} &= \alpha_1 + \beta_1 x_{[1]i} + \epsilon_i \\ y_{[2]i} &= \alpha_2 + \beta_2 x_{[2]i} + \epsilon_i \\ &\vdots \\ y_{[J]i} &= \alpha_J + \beta_J x_{[J]i} + \epsilon_i \end{aligned}$$

$$y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i$$

Shrinkage Equation Assume that the multilevel model only includes group averages, then the partial pooling estimate of the mean (or intercept) is

Thus the estimate value for a group is a weighted average from the data in that group and the overall data.

The weights are:

Correlation structure A common assumption in regression models is that the observations are independent. There are a few common data types that violate this assumption and can be addressed with hierarchical models.

Repeated Measurements:

Cross Sectional Data (Longitudinal/Time series):

"Fixed vs. Random" These type of models are commonly referred to as "mixed models" that include "fixed" and "random" effects.

Random Effects:

Fixed Effects:

Some general advice about when to use fixed/random effects focuses on the research goal; however, GH suggest *always* using multilevel models.

Furthermore, given the inconsistencies in the meaning of fixed/random, GH (and I) prefer using multilevel or hierarchical models.

Multilevel Modeling Pros and Cons

Classical Regression Overview

Multilevel Modeling