# ECON5529
# Bayesian Theory

Lecture 7
Ellis Scharfenaker

## Hierarchical Modeling

- Hierarchical data structures are frequently encountered in economics since measurement often takes place at different levels of aggregation.
- For example, firm level data gathered by the SEC can be considered as nested within various levels of industry data, county level data is nested in state level data which is nested in country level data, etc.
- Hierarchical data is simply “multilevel” data. In econometrics the most frequently encountered hierarchical data is “panel” or “longitudinal” data that contains information at both the spacial, or cross-sectional level, as well temporal level. We will see that models for panel data are a special subset of hierarchical models.
- The structure of hierarchical models is that observed data depend on parameters *which are themselves* described in terms of probability distributions called *hyperpriors.*
- The question is, “how do we treat different levels of variables in the same statistical model?”
- If we ignore the aggregate or “group level” information we may exclude potentially important effects of the whole on the individual parts. In this case too many “independent” parameters tend to overfit the data.
- Similarly, if we treat the aggregate as an individual component (pooled data) we will miss out on important between group variation. That is, if the data are hierarchical, then estimates using only a few parameters don’t capture the heterogeneity and hence fit poorly.
- The solution is to use hierarchical modeling that recognizes the different groupings as well as their relationship to the aggregate, or ***population distribution***, i.e. the dependency between the parameters, and to model all parameters in way consistent with Bayesian methods.

## Hierarchical Bayes

- We have the following data on dropouts in various public schools in a particular district:

| school | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| No. of Students (n) | 47 | 148 | 119 | 810 | 211 | 196 | 148 | 215 | 207 | 97 | 256 | 360 |
| No. of Dropouts (k) | 0 | 18 | 8 | 46 | 8 | 13 | 9 | 31 | 14 | 8 | 29 | 24 |

- One approach to this problem would be to model the dropout rate in each school separately. Notice that this data is distributed as a Binomial distribution.

$$x_i \sim \text{Binomial}[\theta_i] \propto \theta_i^{k_i}(1-\theta_i)^{n_i-k_i} \quad \text{for } i = A, B, \ldots, L$$

- Using an ignorance Beta[1, 1] prior, our posterior for the unknown dropout rate $\theta_i$ for school $i$ is Beta[$k_i + 1$, $n_i - k_i + 1$].

In R

## Hierarchical Bayes

- Alternatively, we may ignore the nested structure of the data and estimate the underlying dropout probability of the entire district. In this case we ignore the individual school index.

$$x \sim \text{Binomial}[\theta] \propto \theta^{\sum_i k_i}(1-\theta)^{\sum_i n_i - \sum_i k_i} \quad \text{for } i = A, B, \ldots, L$$

In R

- Ideally we would like to find a balance between two and use information from the district level distribution in our estimates of the school level distributions.

## Exchangeability

- Our posterior distribution is $p[\theta_i \mid x]$ for $i = A, \ldots, L$ where each $\theta_i$ is a parameter for an individual school. If we have no information to distinguish the $\theta_i$'s, one must assume symmetry (exchangeability) *a priori.*
- We need to create a joint probability model for all the parameters $\theta_i$ and the notion of exchageability is important here.
- The simplest form of an exchangeable distribution says the $\theta_i$'s are i.i.d. from some distribution with parameters $\phi$:

$$p[\theta \mid \phi] = \prod_{i=1}^{N} p[\theta_i \mid \phi]$$

- However, $\phi$ is generally unknown and thus the key hierarchical part of the model is that $\phi$ has its own prior distribution $p[\phi]$. From the laws of probability the posterior distribution now includes an extra term:

$$\begin{aligned} p[\theta, \phi \mid x] &\propto p[x \mid \theta, \phi]\, p[\theta, \phi] \\ &= p[x \mid \theta]\, p[\theta \mid \phi]\, p[\phi] \end{aligned}$$

- Note the likelihood simplification $p[x \mid \theta, \phi] = p[x \mid \theta]$ says the data only depends on $\theta$ and not the distribution of $\theta$ defined by $\phi$. This is because the hyperparameters $\phi$ affect $x$ only through $\theta$.
- In order to derive the posterior, we need a form for the joint distribution $p[\theta, \phi]$ which means we need to specify a prior to $p[\phi]$.
- In this example we are interested in inference about $\theta_i$, the dropout rate of school $i$, which we may take as a measure of performance for school $i$.

## The Model

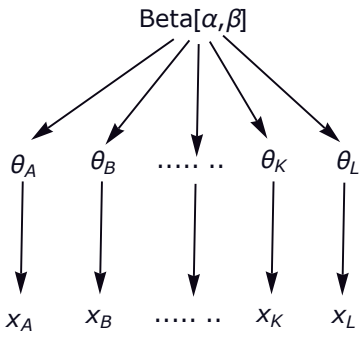
- Let's start with a simple noninformative prior and seek to add more information to it as we progress. Let our data model continue to be:

$$x_i \sim \text{Binomial}[\theta_i] \propto (\theta_i)^{k_i}\,(1-\theta_i)^{n_i - k_i}$$

- With the number of dropouts in school $i$, $k_i$, known.
- Further, let us assume the $\theta_i$ parameters are independent samples from a Beta distribution:

$$\theta_i \sim \text{Beta}[\alpha, \beta]$$

- So we can see that all $\theta$'s are nested into a *parent,* or *population* distribution from which they are independently drawn.




- Notice this also says that the value of $\theta$ depends on the value of $\alpha$ and $\beta$ which represent the aggregate structure of the parameters, in this case the overarching school district.
- Some important properties of the Beta distribution that we can use are:

$$\text{Mean}: \mu = \frac{\alpha}{\alpha+\beta}, \rightarrow \alpha = \mu\,\kappa$$

$$\text{Variance}: \sigma = \frac{\alpha\beta}{(\alpha+\beta+1)}(\alpha+\beta)^2$$

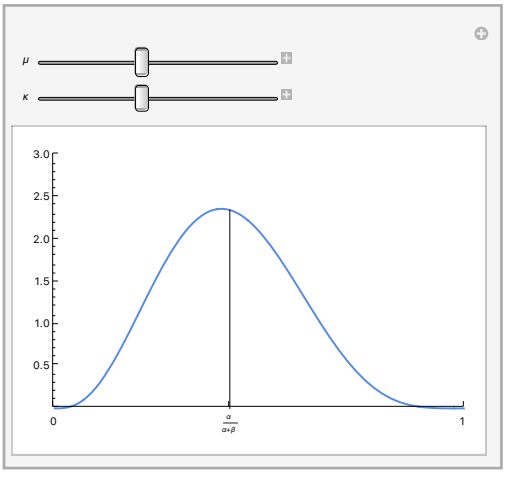
$$\text{Concentration}: \kappa = \alpha+\beta = \mu(1-\mu)/(\sigma-1)$$

$$\text{Mode}: \omega = \frac{\alpha-1}{\alpha+\beta-2}$$

- This implies we can write the Beta distribution in a number of ways. In terms of the mean and concentration:

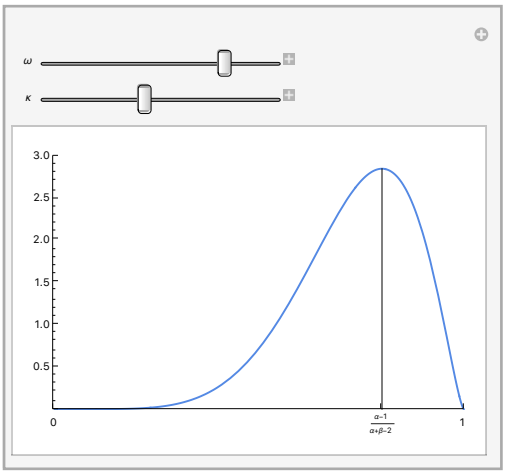
$$\text{Mean}: \mu = \frac{\alpha}{\alpha+\beta}, \rightarrow \alpha = \mu\,\kappa,\ \beta = (1-\mu)\,\kappa$$

$$\theta_i \sim \text{Beta}[\alpha = \mu\,\kappa,\ \beta = (1-\mu)\,\kappa]$$




- Or in terms of the mode and concentration:

$$\theta_i \sim \text{Beta}[\alpha = \omega(\kappa-2)+1,\ \beta = (1-\omega)(\kappa-2)+1]$$




- If this is say, $\omega = 0.25$, then all $\theta_i$'s will be near, i.e. shrunk towards, 0.25 and the spread or concentration of the beta distribution, $\kappa$, governs how near $\theta_i$'s are to the mode. Thus, the magnitude of $\kappa$ is an expression of our prior certainty regarding the dependence of $\theta_i$ on $\alpha$ and $\beta$.
- The idea is that now instead of thinking of $\alpha$, $\beta$, or $\omega$ and $\kappa$, as fixed by prior knowledge, we think of it as another parameter to be estimated which is at the population level.
- Because school districts are have common socio-economic characteristics we would think that schools in a particular district have a bias near $\omega$.
- The smaller $\kappa$ is, the more variability there is between schools. The larger $\kappa$ is the more consistently school dropout rates are near $\omega$.
- When the data from different schools show very similar proportions of dropouts, we have evidence that $\kappa$ is high. When the data from different schools show diverse proportions of dropouts, then we have evidence that $\kappa$ is small.
- Our prior distribution $p[\omega, \kappa]$ expresses what we believe about the school district.
- The key idea is that hierarchical models produce ***shrinkage*** to the population mean, because the individual estimates are informed both by the data from the individual schools and by the data from all other schools via the overarching district estimate.

## Inference for the hierarchical model

- The inferential strategy for hierarchical models is similar to multiparameter models, but have the added complexity of substantially more parameters.
- The parameters at different levels in a hierarchical model are all merely parameters that ***coexist*** in a ***joint parameter space***.
- Since we typically cannot plot the joint posterior distribution $p[\theta, \phi]$ we need to follow a simulation-based approach. This will follow from the analytic derivation of the conditional and marginal distributions which we get by:
  - 1) Writing the joint posterior density as the product of the hyperprior distribution ($p[\phi]$), the overarching population distribution ($p[\theta \mid \phi]$), and the likelihood ($p[x \mid \theta]$):

    $p[\theta, \phi \mid x] \propto p[x \mid \theta]\, p[\theta \mid \phi]\, p[\phi]$
  - 2) Determining analytically the conditional posterior density of $\theta$ given the hyperparameters $\phi$ for fixed observed $x$, which will be a function of $\phi$:

    $p[\theta \mid \phi, x]$

- 3) Estimating $\phi$ by obtaining its marginal posterior distribution:

$$p[\phi \mid x] = \frac{p[\theta, \phi \mid x]}{p[\theta \mid \phi, x]}$$

- Once we have $p[\phi \mid x]$ and $p[\theta \mid \phi, x]$ we can simulate the posterior by drawing a sample vector of hyperparameters from $p[\phi \mid x]$, then use that sample to draw from $p[\theta \mid \phi, x]$.

## School Example

- 1) The joint posterior distribution for all parameters:

$$p[\theta, \alpha, \beta \mid x] \propto p[x \mid \theta]\, p[\theta \mid \alpha, \beta]\, p[\alpha, \beta]$$

$$\propto \left( \prod_{i=1}^{N} \theta_i^{x_i} (1-\theta_i)^{n_i - x_i} \right) \left( \prod_{i=1}^{N} \frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]\,\Gamma[\beta]}\, \theta_i^{\alpha-1} (1-\theta_i)^{\beta-1} \right) p[\alpha, \beta]$$

$$= \left( \prod_{i=1}^{N} \frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]\,\Gamma[\beta]}\, \theta_i^{\alpha+x_i-1} (1-\theta_i)^{\beta+n_i-x_i-1} \right) p[\alpha, \beta]$$

- 2) Given $\alpha$ and $\beta$ the individual components of $\theta$ have independent posterior distributions of the Beta form. Since the prior for each $\theta_i$ is conjugate, the full posterior conditional density is also Beta distributed:

$$p[\theta \mid \alpha, \beta, x] = \prod_{i=1}^{N} \left( \Gamma[\alpha+\beta+n_i] \,/\, (\Gamma[\alpha+x_i]\,\Gamma[\beta+n_i-x_i]) \right) \theta_i^{\alpha+x_i-1} (1-\theta_i)^{\beta+n_i-x_i-1}$$

- 3) Determine the marginal posterior for $(\alpha, \beta)$ from the identity $p[\alpha, \beta \mid x] = \frac{p[\theta, \alpha, \beta \mid x]}{p[\theta \mid \alpha, \beta, x]}$:

$$p[\alpha, \beta \mid x] = \left( \left( \prod_{i=1}^{N} \frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]\,\Gamma[\beta]}\, \theta_i^{\alpha+x_i-1} (1-\theta_i)^{\beta+n_i-x_i-1} \right) p[\alpha, \beta] \right) \Big/ \left( \prod_{i=1}^{N} \left( \Gamma[\alpha+\beta+n_i] \,/\, (\Gamma[\alpha+x_i]\,\Gamma[\beta+n_i-x_i]) \right) \theta_i^{\alpha+x_i-1} (1-\theta_i)^{\beta+n_i-x_i-1} \right)$$

$$= p[\alpha, \beta] \prod_{i=1}^{N} \frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]\,\Gamma[\beta]} \left( (\Gamma[\alpha+x_i]\,\Gamma[\beta+n_i-x_i]) \,/\, \Gamma[\alpha+\beta+n_i] \right)$$

- It turns out that an ignorance prior on $\alpha$ and $\beta$ results in an improper posterior distribution. Hence, Gelman suggests a diffuse hyperprior density $p[\alpha, \beta] \propto (\alpha+\beta)^{-5/2}$, see Gelman pp. 110 for details.
- Posterior mean dropout rate, $\mu = \frac{\alpha}{\alpha+\beta}$, "shrinks" sample proportions toward population mean:
  - Shrinkage tends to be greater for smaller studies
  - Smaller studies "borrow strength" from other studies

In R

## Linear Model: Fixed and Random Effects

- In frequentist econometrics, multilevel models are often used in panel data analysis with a linear model. In these panel models, fixed-effects, random-effects, or mixed-effects models are typically used.
- The regression coefficients that are being modeled are called random effects, because they are considered random outcomes of a process identified with the model that is predicting them.
- On the other hand, fixed effects models usually refer to parameters that do not vary, such as a constant slope, but varying intercept.

- The term fixed effects is used in contrast to random effects, but typically this is done in an inconsistent way. ***The statistical literature is full of confusing and contradictory advice.***
- From the Bayesian perspective we avoid the terms “fixed” and “random” entirely, and focus on the description of the model itself. Since all parameters are modeled, it may be appropriate to think of Bayesian panel analysis as always dealing with “random effects.”

## Hierarchical Regression

- We can easily apply hierarchical modeling to linear regression models. This is the Bayesian equivalent to panel regression.
- Recall that ***cross-section data*** consists of observations on a group of comparable entities, individuals, households, states, or economies, typically collected in one time period. Using the subscript $j$ to distinguish the entities, each data point is of the form $\{y_j, x_j\}$ where $x_j$ may be a whole list of measured qualitative or quantitative characteristics of a household.
- Cross-section data is typically regarded as *exchangeable*, in that the numbering of the entities can be changed arbitrarily without affecting the information in the data.
- ***Time series data*** takes the form of a series of observations on one or more entities, such as a household or economy, over time and may be in the form $\{y_t, y_{t+1}, \ldots\}$, or $\{y_t, x_t\}$. Time series data is not exchangeable unless we "stack" some number of successive observations in overlapping groups.
- ***Panel data*** takes the form of a group of successive observations over time on a group of entities. Each data point thus has both an entity identifier and a time identifier, $\{y_{jt}, x_{jt}\}$.
- Let’s take the example of estimating a neoclassical production function from the Extended Penn World Tables.
- In this case the data is observations of capital per worker ($\frac{K_{jt}}{L_{jt}} = k_{jt}$) and output per worker $\left(\frac{X_{jt}}{L_{jt}} = x_{jt}\right)$ where $j$ indicates the country and $t$ indicates the year. The Cobb-Douglas, constant elasticity model is:

  $$x = f[k] = A\,k^{\alpha}$$

- Taking the log we get the linear relationship:

  $$\text{Log}[x] = \text{Log}[A] + \alpha\,\text{Log}[k]$$

- We have several options for making posterior inferences into the relationship between output per worker and capital per worker.

## No Pooling

- The first is to ignore the group structure and estimate the production function ***independently*** for each country.

  $$\text{Log}[x_{jt}] = \text{Log}[A_{jt}] + \alpha_{jt}\,\text{Log}[k_{jt}] + \epsilon_{ij}$$

In R

## Complete Pooling

- Second, we could ***pool the data***, that is, regard each observation on each country and year as a separate sample from an experiment designed to measure the neoclassical relation between capital per worker and output worker.

- In this case we regard all the observations as ***exchangeable***, and fit some model to them. Pooling implicitly regards differences among countries and over time as part of the "noise" to be filtered out of the data by the model. In this case the model is:

$$\text{Log}[x_{j\,t}] = \text{Log}[A] + \alpha\,\text{Log}[k_{j\,t}] + \epsilon_{i\,j}$$

- Where the coefficients $A$ and $\alpha$ are constrained to be the same for each country and time period. In this case each country follows the same line.

In R

## The Multilevel or Hierarchical Model

- Multilevel regression can be thought of as a method for compromising between the two extremes of excluding a categorical predictor from a model (complete pooling), or estimating separate models within each level of the categorical predictor (no pooling).
- We can think of hierarchical modeling as a regression that includes a categorical input variable representing group membership.
- From this perspective, the group index is a factor with $J$ levels, corresponding to $J$ predictors in the regression model.
- $J - 1$ linear predictors are added to the model and these $J$ coefficients are then themselves given a model.

## Varying-Intercept Model

- We could also regard each country as a separate observation on the time series relation of $x$ to $k$, and allow the absolute level of productivity to vary from country to country. The first level of the model is:

$$\text{Log}[x_{j\,t}] = \text{Log}[A_j] + \alpha\,\text{Log}[k_{j\,t}] + \epsilon_{i\,j}$$

- In this model we get a “constrained” measure of the elasticity of real output with respect to real capital stock that is based on cross-country comparisons, but we are flexible to let the level of the production function to vary from country to country.
- The term “fixed effects” is used for the regression coefficients that do not vary by group, in this case $\alpha$. The second level of the model is:

$$\text{Log}[A_j] \sim \mathcal{N}(\mu_A,\, \sigma_A^2)$$

- This model partially pools the group-level parameters $A_j$ toward their mean level, $\mu_A$. There is more pooling when the group-level standard deviation $\sigma_A$ is small, and more smoothing for groups with fewer observations.
- The multilevel-modeling estimate of $A_j$ can be expressed as a weighted average of the no-pooling estimate for it’s group $(\overline{x_j} - \alpha\,\overline{k_j})$ and the mean $\mu_A$, where $x_j = \text{Log}[x_j]$ and $k_j = \text{Log}[k_j]$

$$\hat{A}_j \approx \frac{\frac{n_j}{\sigma_x^2}}{\frac{n_j}{\sigma_x^2} + \frac{1}{\sigma_A^2}} (\overline{x_j} - \alpha\,\overline{k_j}) + \frac{\frac{1}{\sigma_A^2}}{\frac{n_j}{\sigma_x^2} + \frac{1}{\sigma_A^2}} \mu_A$$

```
Limit[ n/σx^2 / (n/σx^2 + 1/σA^2) (x̄j - α k̄j) + 1/σA^2 / (n/σx^2 + 1/σA^2) μA, n → 0]
```

$\mu_A$

$$\text{Limit}\left[\frac{\frac{n}{\sigma_x^2}}{\frac{n}{\sigma_x^2}+\frac{1}{\sigma_A^2}}\left(\overline{x_j}-\alpha\,\overline{k_j}\right)+\frac{\frac{1}{\sigma_A^2}}{\frac{n}{\sigma_x^2}+\frac{1}{\sigma_A^2}}\mu_A,\ n\to\infty\right]$$

$$-\alpha\,\overline{k_j}+\overline{x_j}$$

In R

## Varying-Slope Model

- We can also continue regard each country as a separate observation on the time series relation of $x$ to $k$, and allow the elasticity of real output with respect to real capital stock to vary by country while keeping the absolute level of productivity constant.

  $$\text{Log}\left[x_{j\,t}\right] = \text{Log}[A] + \alpha_j\,\text{Log}\left[k_{j\,t}\right] + \epsilon_{i\,j}$$

- Almost always, when a slope is allowed to vary, it makes sense for the intercept to vary also. It would not make much sense if the productivity levels in all countries were all identical, but they differed in their elasticity of real output with respect to real capital stock.

## Varying Slope Varying Intercept: Mixed Effects Model

- Lastly, we can use a hierarchical regression that allows variation in the intercepts and slopes for each group where the group level parameters do not vary independently, but are constrained by an overarching global distribution.
- That is, each country has its own estimated production function line, but the parameters defining these lines come from a common distribution.
- The varying-intercept, varying-slope model is:

  $$\text{Log}[x_t] \sim \mathcal{N}\left(\text{Log}\left[A_j\right] + \alpha_j\,\text{Log}\left[k_{j\,t}\right],\ \sigma_x^2\right)\ \text{for}\ t = 1,\ \ldots,\ n$$

  $$\begin{pmatrix} A_j \\ \alpha_j \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_A \\ \mu_\alpha \end{pmatrix}, \begin{pmatrix} \sigma_A^2 & \rho\sigma_A\,\sigma_\alpha \\ \rho\sigma_A\,\sigma_\alpha & \sigma_\alpha^2 \end{pmatrix}\right)$$

- With variation in the $A_j$'s and the $\alpha_j$'s and also a between-group correlation parameter $\rho$.
- The hierarchical structure involves drawing the coefficients from a prior that is also estimated with the data. The hierarchically estimated prior determines the amount of pooling.
- If the data in each level are very similar, strong pooling will be reflected in low hierarchical variance. If the data in the levels are dissimilar, weaker pooling will be reflected in higher hierarchical variance.
- In essence, the hierarchical model reduces the estimation variability (***shrinkage to the overarching mean***) and exploits coefficient similarities across countries without imposing the same population structure.

In R