ECON5120 Bayesian Data Analysis

Week 4: Normal Linear Regression and Hierarchical Models

Santiago Montoya-Blandón

University of Glasgow
Adam Smith Business School

February 4, 2026

Outline

Review: Normal Linear Regression

- Goal: Uncover relationship between outcome y and data features $x_1, \ldots, x_p$
- Example: relationship of wage (y) with education (x_1) and age (x_2)?
- How to quantify this relationship statistically? *Conditional expectation* of y on $x = (x_1, \ldots, x_p)$
- Main assumption in linear regression: conditional expectation is linear

$$\mathbb{E}[y \mid x_1, \ldots, x_p] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = x^\top \beta$$

- Thanks to linearity, β_j represents marginal effect of x_j:

$$\frac{\partial \mathbb{E}[y \mid x_1, \ldots, x_p]}{\partial x_j} = \beta_j$$

- What does β_0 represent?
- How to learn about the parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$?

Review: Normal Linear Regression

- Sample n individuals at random from population
- Access to random sample on variables of interest: $\{y_i, x_{1,i}, \ldots, x_{p,i}\}_{i=1}^n$
- Our model is not exact and errors exist: $u_i := y_i - \mathbb{E}[y_i \mid x_{1,i}, \ldots, x_{p,i}]$
- Replacing the linear conditional mean form, we can write for $i \in \{1, \ldots, n\}$

$$y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i} + u_i = x_i^\top \beta + u_i$$

- u are all factors aside from x that influence y
- Main assumption in *normal* linear regression:

$$u_i \sim \mathcal{N}(0, \sigma^2) \text{ independently across } i$$

- Equivalent to probabilistic model directly on y:

$$y_i \mid x_{1,i}, \ldots, x_{p,i} \sim \mathcal{N}(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i}, \sigma^2) \text{ independently across } i$$

Frequentist Estimation of Normal Linear Regression

- Using the probabilistic model on y, we have

$$p(y_1, \ldots, y_n \mid x_1, \ldots, x_n; \beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2 \right\}$$

- (Log) Likelihood function given data $y = (y_1, \ldots, y_n)$ and $X = (x_1, \ldots, x_n)$

$$\begin{aligned} \ell(\beta, \sigma^2) &= \log p(y_1, \ldots, y_n \mid x_1, \ldots, x_n; \beta, \sigma^2) \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} (y - X\beta)^\top (y - X\beta) \end{aligned}$$

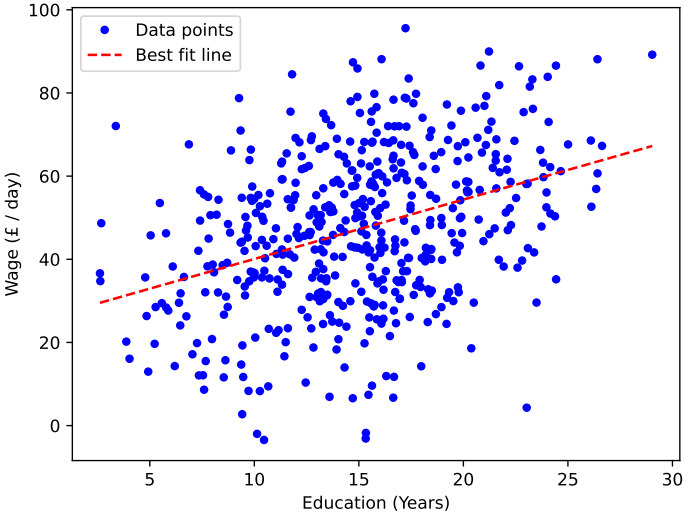
- Frequentist solution is the MLE: $(\widehat{\beta}_{\text{MLE}}, \widehat{\sigma}^2_{\text{MLE}}) = \arg\max \ell(\beta, \sigma^2)$

$$\begin{aligned} \widehat{\beta}_{\text{MLE}} &= (X^\top X)^{-1} X^\top y = \widehat{\beta}_{\text{OLS}} \\ \widehat{\sigma}^2_{\text{MLE}} &= \frac{1}{n} (y - X^\top \widehat{\beta}_{\text{MLE}})^\top (y - X^\top \widehat{\beta}_{\text{MLE}}) \end{aligned}$$

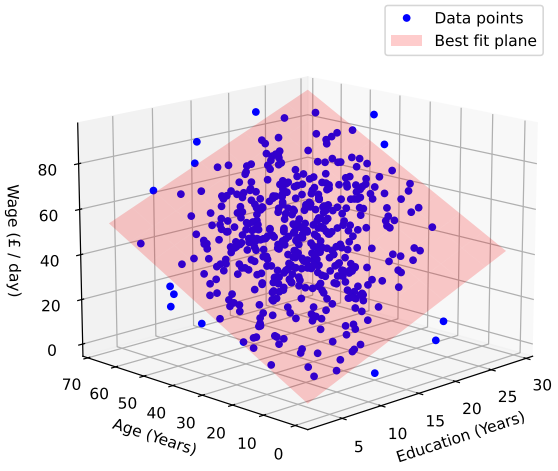
Frequentist Estimation of Normal Linear Regression

- Maximum likelihood estimator (MLE) = Ordinary Least Squares (OLS)!
- Intuitively: best fitting line that minimizes sum of squared residuals
- To predict at new value $\widetilde{x}$: use $\widetilde{y} = \widetilde{x}^\top \widehat{\beta}_{\text{MLE}}$
- Estimator of the variance is biased: denominator is n instead of $n - p - 1$
- Estimated coefficients are interpreted as effects on y from unit change in x
- All standard results of OLS apply (see BDA Ch. 14):
  1. $\widehat{\beta}_{\text{MLE}}$ is random and normally distributed with $\text{Var}(\widehat{\beta}_{\text{MLE}} \mid X) = \sigma^2 (X^\top X)^{-1}$
  2. Estimation only requires no multicollinearity in X (linearly independent columns)
  3. Regressors $x_1, \ldots, x_p$ can be continuous, binary, categorical, etc.
  4. Regressors can include non-linear transformations of the others

Example: Linear Regression in 2D



Example: Linear Regression in 3D



Bayesian Linear Regression

- Frequentist assumption: there are true but unknown values (β^*, σ^{2*}) for (β, σ^2) that govern the *population* relationship between y and x
- Observed data $\{y_i, x_i\}_{i=1}^n$ though as a sample of the population relationship
- While true values are fixed, $(\widehat{\beta}_{\text{MLE}}, \widehat{\sigma}^2_{\text{MLE}})$ are random
- Given normality of errors, we have exactly: $\widehat{\beta}_{\text{MLE}} \mid X \sim \mathcal{N}(\beta^*, \sigma^{2*}(X^\top X)^{-1})$

Bayesian Linear Regression

- Frequentist assumption: there are true but unknown values (β^*, σ^{2*}) for (β, σ^2) that govern the *population* relationship between y and x
- Observed data $\{y_i, x_i\}_{i=1}^n$ though as a sample of the population relationship
- While true values are fixed, $(\widehat{\beta}_{\text{MLE}}, \widehat{\sigma}^2_{\text{MLE}})$ are random
- Given normality of errors, we have exactly: $\widehat{\beta}_{\text{MLE}} \mid X \sim \mathcal{N}(\beta^*, \sigma^{2*}(X^\top X)^{-1})$

- Bayesian assumption: as (β, σ^2) are unknown, we will assign them a probability distribution
- Observed data are fixed and help us to learn about (β, σ^2)
- Posterior of (β, σ^2) likely related to distribution of $(\widehat{\beta}_{\text{MLE}}, \widehat{\sigma}^2_{\text{MLE}})$?
- Importantly, interpretation of parameters remains the same

Likelihood

- Our linear model remains the same:

$$y_i \mid x_i; \beta, \sigma^2 \sim \mathcal{N}(x_i^\top \beta, \sigma^2) \text{ independently across } i$$

- Likelihood is therefore still the same:

$$p(y \mid X; \beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)^\top (y - X\beta)\right\}$$

- Write in terms of MLE estimates (see next slide):

$$p(y \mid X; \beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(\beta - \widehat{\beta}_{\text{MLE}})^\top X^\top X(\beta - \widehat{\beta}_{\text{MLE}})\right\}$$
$$\times \exp\left\{-\frac{n \cdot \widehat{\sigma}^2_{\text{MLE}}}{2\sigma^2}\right\}$$

Likelihood Details

- Note $\widehat{\beta}_{\text{MLE}} = (X^\top X)^{-1} X^\top y \implies (X^\top X)\widehat{\beta}_{\text{MLE}} = X^\top y$
- Residuals are orthogonal to regressors:

$$X^\top (y - X\widehat{\beta}_{\text{MLE}}) = X^\top y - X^\top X \widehat{\beta}_{\text{MLE}} = X^\top y - X^\top y = 0$$

- Finally, use this to write:

$$\begin{aligned}
(y - X\beta)^\top (y - X\beta) &= [y - X(\beta - \widehat{\beta}_{\text{MLE}} + \widehat{\beta}_{\text{MLE}})]^\top [y - X(\beta - \widehat{\beta}_{\text{MLE}} + \widehat{\beta}_{\text{MLE}})] \\
&= [(y - X\widehat{\beta}_{\text{MLE}}) - X(\beta - \widehat{\beta}_{\text{MLE}})]^\top [(y - X\widehat{\beta}_{\text{MLE}}) - X(\beta - \widehat{\beta}_{\text{MLE}})] \\
&= (y - X\widehat{\beta}_{\text{MLE}})^\top (y - X\widehat{\beta}_{\text{MLE}}) - 2(\beta - \widehat{\beta}_{\text{MLE}})^\top X^\top (y - X\widehat{\beta}_{\text{MLE}}) \\
&\quad + (\beta - \widehat{\beta}_{\text{MLE}})^\top X^\top X (\beta - \widehat{\beta}_{\text{MLE}}) \\
&= (\beta - \widehat{\beta}_{\text{MLE}})^\top X^\top X (\beta - \widehat{\beta}_{\text{MLE}}) + n \cdot \widehat{\sigma}^2_{\text{MLE}}
\end{aligned}$$

Non-informative Prior

- Having defined the likelihood, we need a joint prior over (β, σ^2)
- Non-informative prior is uniform over real line for $(\beta, \log \sigma^2)$
- This translates to the following joint prior: $p(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$

Non-informative Prior

- Having defined the likelihood, we need a joint prior over (β, σ^2)
- Non-informative prior is uniform over real line for $(\beta, \log \sigma^2)$
- This translates to the following joint prior: $p(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$
- Posterior under non-informative prior:

$$p(\beta, \sigma^2 \mid y, X) \propto \frac{1}{\sigma^2} \times (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(\beta - \widehat{\beta}_{\text{MLE}})^\top X^\top X(\beta - \widehat{\beta}_{\text{MLE}})\right\} \times \exp\left\{-\frac{n \cdot \widehat{\sigma}^2_{\text{MLE}}}{2\sigma^2}\right\}$$

- Note that we condition on X as done in frequentist analysis
- We do not need to model the full p-dimensional joint distribution of X

Posterior with Non-Informative Prior

- Re-arrange previous expression to

$$p(\beta, \sigma^2 \mid y, X) \propto \exp\left\{-\frac{1}{2}(\beta - \widehat{\beta}_{\text{MLE}})^\top \left[\sigma^2 (X^\top X)^{-1}\right]^{-1} (\beta - \widehat{\beta}_{\text{MLE}})\right\}$$
$$\times (\sigma^2)^{-n/2-1} \exp\left\{-\frac{n \cdot \widehat{\sigma}^2_{\text{MLE}}}{2\sigma^2}\right\}$$

- We can decompose $p(\beta, \sigma^2 \mid y, X) = p(\beta \mid \sigma^2; y, X) p(\sigma^2 \mid y, X)$ with

$$p(\beta \mid \sigma^2; y, X) = \mathcal{N}\left(\beta \mid \widehat{\beta}_{\text{MLE}}, \sigma^2 (X^\top X)^{-1}\right)$$
$$p(\sigma^2 | y) = \text{Inv-Gamma}\left(\sigma^2 \,\middle|\, \frac{n}{2}, \frac{n \cdot \widehat{\sigma}^2_{\text{MLE}}}{2}\right)$$

Posterior with Non-Informative Prior

$$p(\beta, \sigma^2 \mid y, X) = \mathcal{N}\left(\beta \mid \widehat{\beta}_{\text{MLE}}, \sigma^2 (X^\top X)^{-1}\right) \times \text{Inv-Gamma}\left(\sigma^2 \,\middle|\, \frac{n}{2}, \frac{n \cdot \widehat{\sigma}^2_{\text{MLE}}}{2}\right)$$

- Posterior for β with non-informative prior is centered at the MLE estimate
- Posterior variance for β equals the sampling variance of the MLE estimator
- Posterior for σ also dependent on the MLE value
- We can now summarize β and σ^2 using the posterior

Posterior with Non-Informative Prior

$$p(\beta, \sigma^2 \mid y, X) = \mathcal{N}\left(\beta \mid \widehat{\beta}_{\text{MLE}}, \sigma^2 (X^\top X)^{-1}\right) \times \text{Inv-Gamma}\left(\sigma^2 \,\middle|\, \frac{n}{2}, \frac{n \cdot \widehat{\sigma}^2_{\text{MLE}}}{2}\right)$$

- Posterior for β with non-informative prior is centered at the MLE estimate
- Posterior variance for β equals the sampling variance of the MLE estimator
- Posterior for σ also dependent on the MLE value
- We can now summarize β and σ^2 using the posterior
- As in the simple normal model, we have the marginal posteriors:

$$\beta \mid y, X \sim t_{n-1}\left(\widehat{\beta}_{\text{MLE}}, \widehat{\sigma}^2_{\text{MLE}} (X^\top X)^{-1}\right)$$

$$\sigma^2 \mid y, X \sim \text{Inv-Gamma}\left(\frac{n-1}{2}, \frac{(n-1) \cdot \widehat{\sigma}^2_{\text{MLE}}}{2}\right)$$

Conjugate Prior

- Shape of conjugate prior has the same structure as in simple normal model:

$$p(\beta, \sigma^2) = p(\beta \mid \sigma^2) p(\sigma^2)$$

- We can parametrize this prior as

$$p(\beta, \sigma^2) = \mathcal{N}\left(\beta \mid \underline{\beta}, \sigma^2 \underline{B}\right) \times \text{Inv-Gamma}\left(\sigma^2 \mid \underline{\alpha}/2, \underline{\delta}/2\right)$$

- $\underline{\beta}$ controls the center of the prior
- $\underline{B}$ controls the prior covariance
- $\underline{\alpha}$ are the prior degrees of freedom
- $\underline{\delta}$ controls the scale of the variance

Posterior with Conjugate Prior

- Using the conjugate prior, the shape remains the same

$$p(\beta, \sigma^2 \mid y, X) = \mathcal{N}\left(\beta \mid \bar{\beta}, \sigma^2 \bar{B}\right) \times \text{Inv-Gamma}\left(\sigma^2 \middle| \frac{\bar{\alpha}}{2}, \frac{\bar{\delta}}{2}\right)$$

- Posterior hyperparameters:

$$\begin{aligned}
\bar{B} &= \left(\underline{B}^{-1} + X^\top X\right)^{-1} \\
\bar{\beta} &= \bar{B}\left(\underline{B}^{-1}\underline{\beta} + X^\top X \widehat{\beta}_{\text{MLE}}\right) \\
\bar{\alpha} &= \underline{\alpha} + n \\
\bar{\delta} &= \underline{\delta} + n \cdot \widehat{\sigma}^2_{\text{MLE}} + (\underline{\beta} - \widehat{\beta}_{\text{MLE}})^\top [\underline{B} + (X^\top X)^{-1}]^{-1} (\underline{\beta} - \widehat{\beta}_{\text{MLE}})
\end{aligned}$$

- Posterior quantities are linear combinations of prior and data

Marginal Posteriors with Conjugate Prior

- As before, we can obtain the marginal posteriors
- For β, integrate out σ^2:

$$\beta \mid y, X \sim t_{\bar{\alpha}}\left(\bar{\beta}, \frac{\bar{\delta}}{\bar{\alpha}}\bar{B}\right)$$

- For σ^2, integreate out β:

$$\sigma^2 \mid y, X \sim \text{Inv-Gamma}\left(\frac{\bar{\alpha}-1}{2}, \frac{\bar{\delta}-1}{2}\right)$$

- Can provide summary and inference based on these distributions
- Non-informative model recovered by setting $\underline{\beta} = 0$, $\underline{B}^{-1} = 0$, $\underline{\alpha} = 0$, $\underline{\delta} = 0$

Bayesian Linear Regression Example

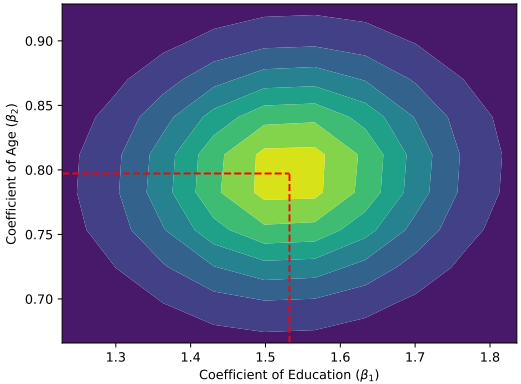


Figure: Exact marginal posterior of β

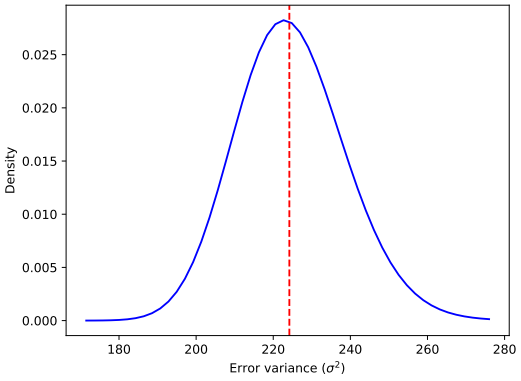


Figure: Exact marginal posterior of σ^2

Posterior Predictive Distribution

- Suppose we have m new observations $\widetilde{X}$ that we want to use to predict $\widetilde{y}$
- Example: we want to predict the wage of someone given their education level and work experience
- From our model:

$$p(\widetilde{y} \mid \widetilde{X}; \beta, \sigma^2) = \mathcal{N}(\widetilde{y} \mid \widetilde{X}\beta, \sigma^2 I_m)$$

Posterior Predictive Distribution

- Suppose we have m new observations $\widetilde{X}$ that we want to use to predict $\widetilde{y}$
- Example: we want to predict the wage of someone given their education level and work experience
- From our model:

$$p(\widetilde{y} \mid \widetilde{X}; \beta, \sigma^2) = \mathcal{N}(\widetilde{y} \mid \widetilde{X}\beta, \sigma^2 I_m)$$

- Integrate out the uncertainty in β and σ^2 according to their posterior
- This leads to the *posterior predictive distribution*

$$p(\widetilde{y} \mid \widetilde{X}, y, X) = \int \int p(\widetilde{y} \mid \widetilde{X}; \beta, \sigma^2) p(\beta, \sigma^2 \mid y, X) d\beta d\sigma^2$$

Posterior Predictive Distribution

- First integrate out β:

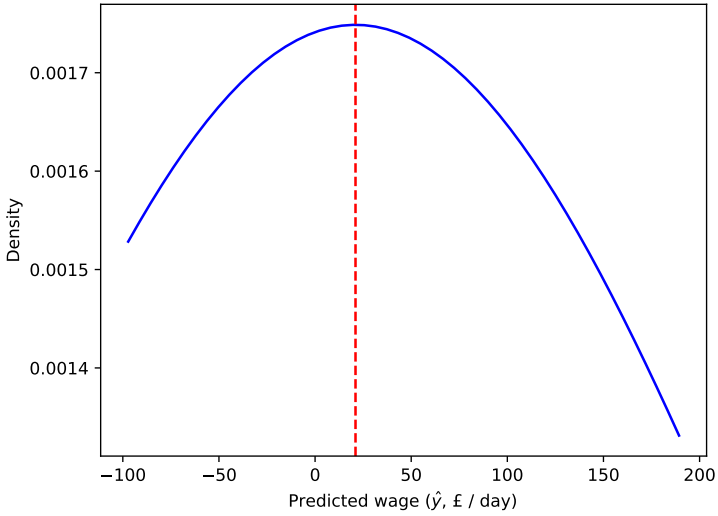
$$\widetilde{y} \mid \sigma^2; \widetilde{X}, y, X \sim \mathcal{N}(\widetilde{X}\bar{\beta}, \sigma^2[I_m + \widetilde{X}\bar{B}\widetilde{X}^\top])$$

- Then integrate out σ^2 for full result:

$$\widetilde{y} \mid \widetilde{X}, y, X \sim t_{\bar{\alpha}}\left(\widetilde{X}\bar{\beta}, \frac{\bar{\delta}}{\bar{\alpha}}[I_m + \widetilde{X}\bar{B}\widetilde{X}^\top]\right)$$

- Remember: we have a full distribution over our prediction!
- Accounting for uncertainty in β and σ results in heavy-tailed distribution
- Center is similar to MLE, but *contains additional variation*
- Usual result: Bayesian accounts for additional prediction uncertainty compared to standard frequentist predictions

Posterior Predictive Distribution



Independent Priors

- What if we do not want to assume any prior dependence between β and σ^2?
- Posterior under following prior: $p(\beta, \sigma^2) = p(\beta)p(\sigma^2)$

$$p(\beta, \sigma^2 \mid y, X) \propto (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(\beta - \widehat{\beta}_{\text{MLE}})^\top X^\top X(\beta - \widehat{\beta}_{\text{MLE}})\right\}$$
$$\times \exp\left\{-\frac{n \cdot \widehat{\sigma}^2_{\text{MLE}}}{2\sigma^2}\right\} \times p(\beta)p(\sigma^2)$$

- Unless $p(\beta)$ is uniform and improper, joint posterior is not of known form
- Standard choice uses similar priors as before

$$p(\beta) = \mathcal{N}\left(\beta \mid \underline{\beta}, \underline{B}\right)$$
$$p(\sigma^2) = \text{Inv-Gamma}\left(\sigma^2 \mid \underline{\alpha}/2, \underline{\delta}/2\right)$$

Conditional Posteriors under Independent Priors

- What to do when the posterior is not analytically available?
- We use simulation (e.g., MCMC) methods as in previous classes
- *Conditional posteriors* are analytically available $\Longrightarrow$ Easy Gibbs sampler!

$$\beta \mid \sigma^2; y, X \sim \mathcal{N}\left(\widetilde{\beta}, \widetilde{B}\right) \qquad \text{with hyperparameters}$$

$$\widetilde{B} = \left(\underline{B}^{-1} + \frac{1}{\sigma^2} X^\top X\right)^{-1}$$

$$\widetilde{\beta} = \widetilde{B}\left(\underline{B}^{-1}\underline{\beta} + \frac{1}{\sigma^2} X^\top y\right)$$

$$\sigma^2 \mid \beta; y, X \sim \text{Inv-Gamma}\left(\widetilde{\alpha}/2, \widetilde{\delta}/2\right) \qquad \text{with hyperparameters}$$

$$\widetilde{\alpha} = \underline{\alpha} + n$$

$$\widetilde{\delta} = \underline{\delta} + (y - X\beta)^\top (y - X\beta)$$

Bayesian Linear Regression Gibbs Sampler

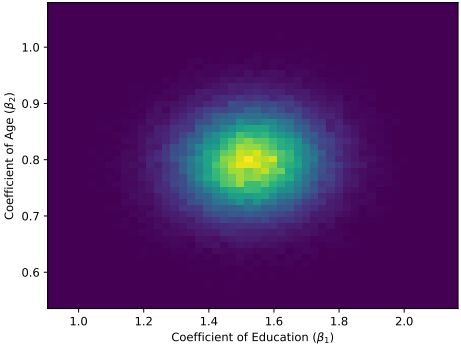


Figure: Empirical marginal posterior of β

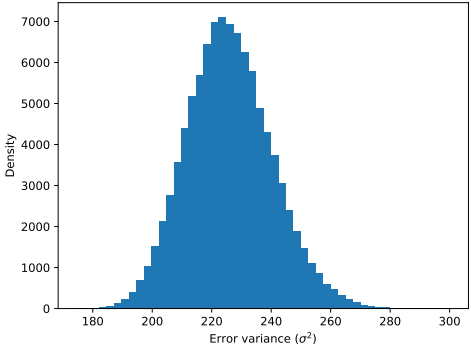


Figure: Empirical marginal posterior of σ^2

Hierarchical models

- As the name suggests, a model is hierarchical when it can be conceptualized in terms of levels or stages

Hierarchical models

- As the name suggests, a model is hierarchical when it can be conceptualized in terms of levels or stages
- Simple example:
  1. Observations given parameters: $y|\theta$
  2. Parameters given hyperparameters: $\theta|\tau$
  3. Hyperparameters given more hyperparameters: $\tau|\alpha$
- Posterior distribution:

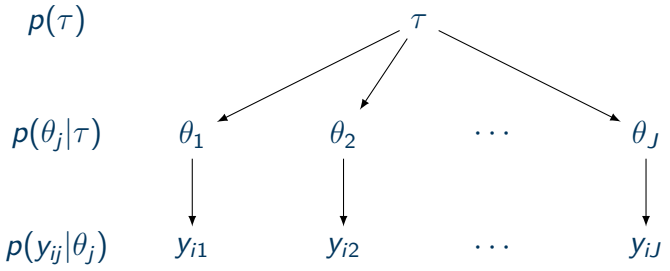
$$\begin{aligned} p(\theta, \tau, \alpha|y) &\propto p(y|\theta, \tau, \alpha)p(\theta, \tau, \alpha) \\ &\propto p(y|\theta)p(\theta|\tau)p(\tau|\alpha)p(\alpha) \end{aligned}$$

Hierarchical models

- Data on $i = 1, \ldots, n$ units across $j = 1, \ldots, J$ groups

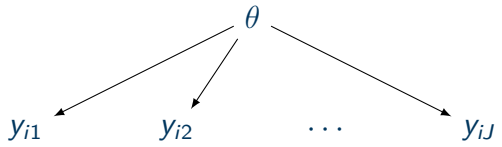
Level 1: Observations given group-specific parameters: $p(y_{ij}|\theta_j)$

Level 2: Parameters from a common distribution parameterized by τ: $p(\theta_j|\tau)$

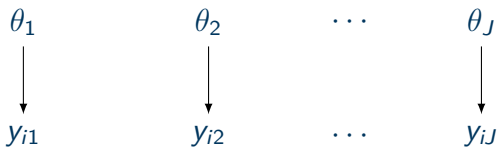


Hierarchical models: Comparison to alternatives

- Joint model: model with a common effect / pooled model



- Separate model: model with separate / independent effects



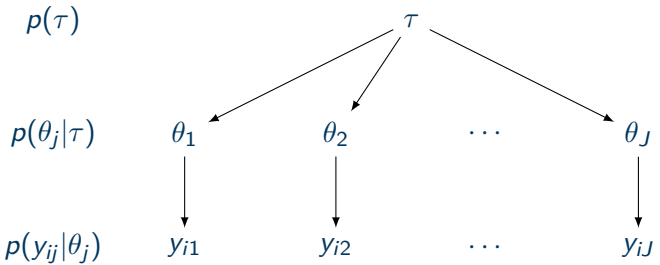
- Can only estimate separate models with many observations per group
- Can still estimate hierarchical model without grouped observations

Hierarchical model: Example

- Example: CVD treatment effectiveness
  - In hospital j the survival probability is θ_j
  - Observations y_{ij} tell whether patient i survived in hospital j

Hierarchical model: Example

- Example: CVD treatment effectiveness
  - In hospital j the survival probability is θ_j
  - Observations y_{ij} tell whether patient i survived in hospital j
  - Sensible to assume that θ_j are similar



  - Natural to think that θ_j have common population distribution
  - θ_j is not directly observed and the population distribution is unknown

Hierarchical model: Example

- Heterogeneous treatment effects in cross-sectional data

$$outcome_i = \alpha + \beta_i treatment_i + \delta' controls_i + u_i$$

- Treatment is binary with potential outcomes $outcome_i^{(0)}$ and $outcome_i^{(1)}$
- β_i has a causal interpretation as a treatment effect assuming unconfoundedness conditional on the controls, i.e.,

$$treatment_i \perp\!\!\!\perp outcome_i^{(0)}, outcome_i^{(1)} \mid controls_i$$

- Hierarchical model: $\beta_i \sim \mathcal{N}(\beta_0, \tau^2) \implies \beta_i = \beta_0 + \tau \cdot v_i$, with $v_i \sim \mathcal{N}(0,1)$
- Intuition: treatment effects are different across units, but arise as deviations from a common (average) treatment effect β_0

Hierarchical models: Binomial example with rats data

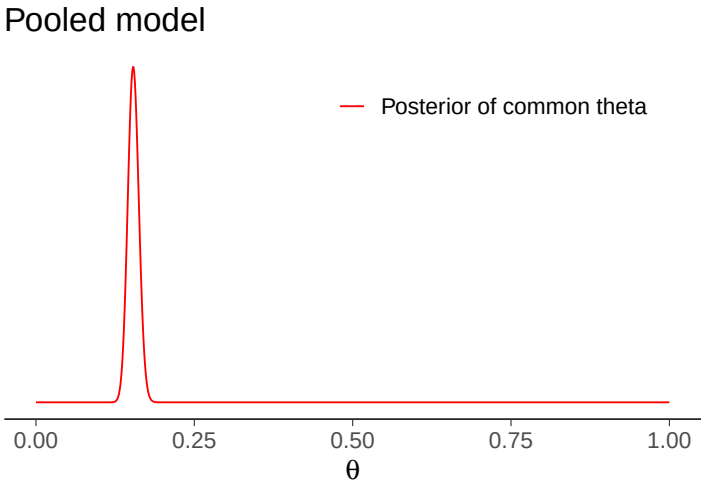
- Medicine testing example on rats across many labs
- In every lab, some rats can either carry or not a particular disease
- Outcome y_j measures how many rats out of n_j got the disease, where $j = 1, \ldots, J$ indexes the lab
- Binomial model: $p(y_j|\theta_j, n_j) = \text{Bin}(\theta_j, n_j)$

Hierarchical models: Binomial example with rats data

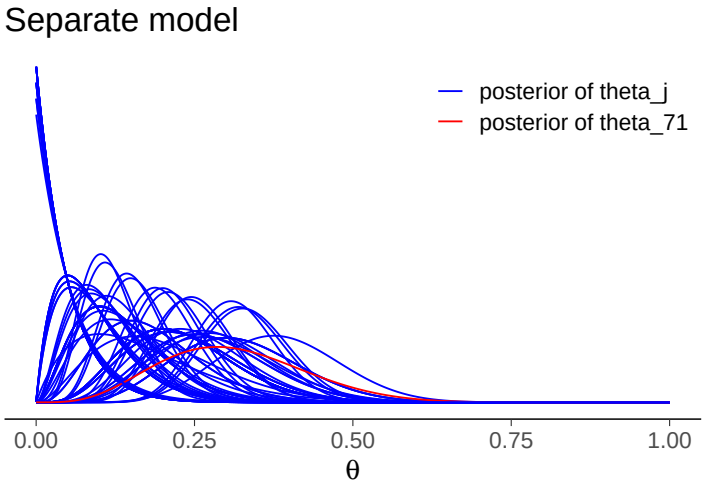
- Medicine testing example on rats across many labs
- In every lab, some rats can either carry or not a particular disease
- Outcome y_j measures how many rats out of n_j got the disease, where $j = 1, \ldots, J$ indexes the lab
- Binomial model: $p(y_j|\theta_j, n_j) = \text{Bin}(\theta_j, n_j)$
- Experiment has been repeated in $J = 71$ labs

0/20	0/20	0/20	0/20	0/20	0/20	0/20	0/19	0/19	0/19
0/19	0/18	0/18	0/17	1/20	1/20	1/20	1/20	1/19	1/19
1/18	1/18	2/25	2/24	2/23	2/20	2/20	2/20	2/20	2/20
2/20	1/10	5/49	2/19	5/46	3/27	2/17	7/49	7/47	3/20
3/20	2/13	9/48	10/50	4/20	4/20	4/20	4/20	4/20	4/20
4/20	10/48	4/19	4/19	4/19	5/22	11/46	12/49	5/20	5/20
6/23	5/19	6/22	6/20	6/20	6/20	16/52	15/46	15/47	9/24
4/14									

Hierarchical models: Binomial example with rats data



Hierarchical models: Binomial example with rats data



Hierarchical exchangeability

How can we justify a hierarchical model in this setting?

Exchangeability

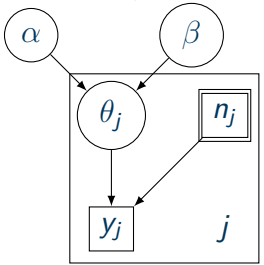
Joint distribution of random variables is independent of their order.

- All rats are not exchangeable
- In a single laboratory rats are exchangeable
- Laboratories are exchangeable (perhaps conditional on other characteristics; i.e., conditional exchangeability)
- Under these conditions, we have a hierarchical model with units given by the rats and groups given by laboratories

Hierarchical models: Binomial example with rats data

Hierarchical binomial model for rats: prior parameters α and β are unknown

- Level 1: $y_j|n_j, \theta_j \sim \text{Bin}(y_j|n_j, \theta_j)$
- Level 2: $\theta_j|\alpha, \beta \sim \text{Beta}(\alpha, \beta)$



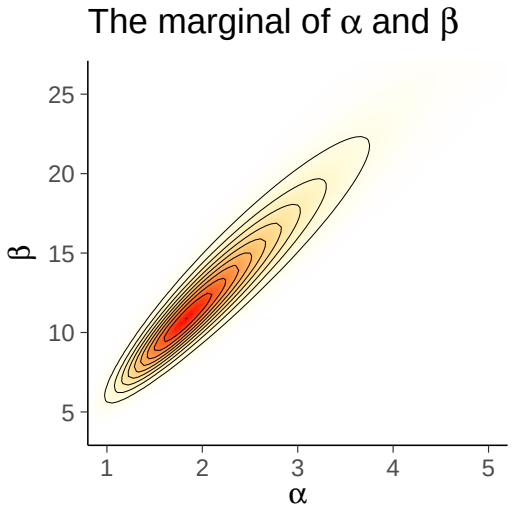
Hierarchical models: Binomial example with rats data

- Joint posterior under exchangeability:

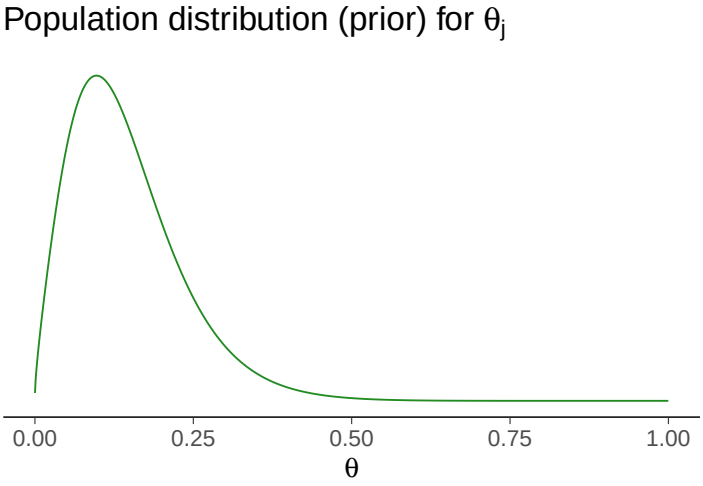
$$\begin{aligned} p(\theta_1, \ldots, \theta_J, \alpha, \beta | y) &\propto p(y | \theta_1, \ldots, \theta_J, \alpha, \beta) p(\theta_1, \ldots, \theta_J, \alpha, \beta) \\ &\propto p(y_1, \ldots, y_J | \theta_1, \ldots, \theta_J) p(\theta_1, \ldots, \theta_J | \alpha, \beta) p(\alpha, \beta) \\ &\propto \prod_{j=1}^{J} p(y_j | \theta_j) p(\theta_j | \alpha, \beta) p(\alpha, \beta) \end{aligned}$$

- Posterior of α and β does not directly depend on data, only on $\theta_1, \ldots, \theta_J$
- The likelihood is given by the first term relating (lab-specific) number of diseased rats y_j to probability of carrying the disease θ_j
- Standard prior for each lab: $p(\theta_j | \alpha, \beta) = \text{Beta}(\theta_j | \alpha, \beta)$
- How to set the hyperprior $p(\alpha, \beta)$? BDA sets $p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$

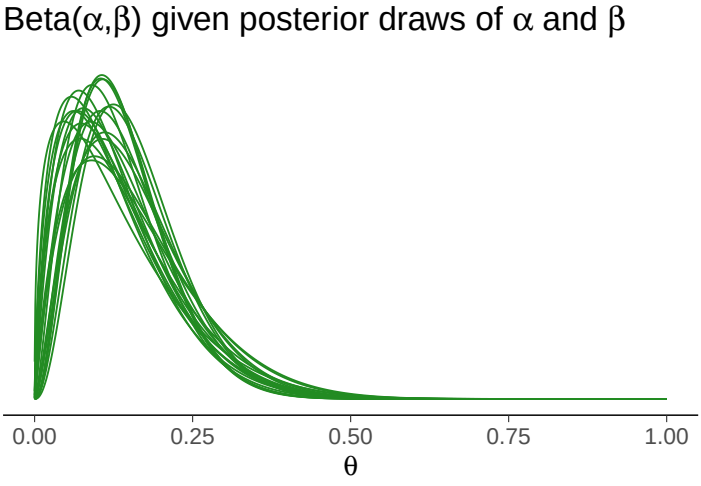
Hierarchical models: Binomial example with rats data



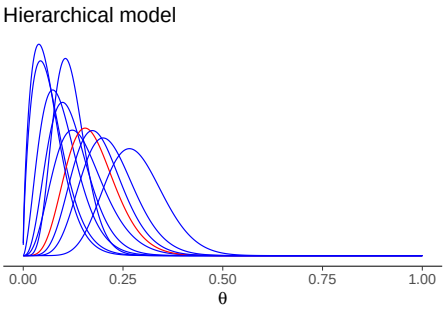
Hierarchical models: Binomial example with rats data

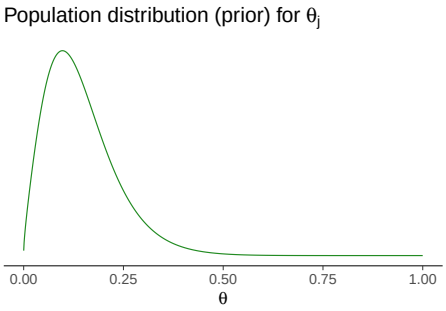


Hierarchical models: Binomial example with rats data

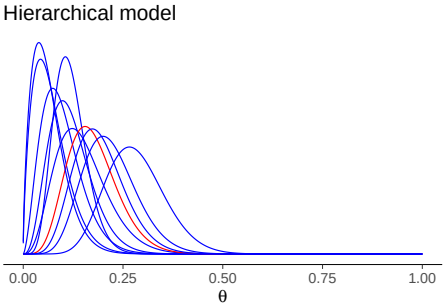


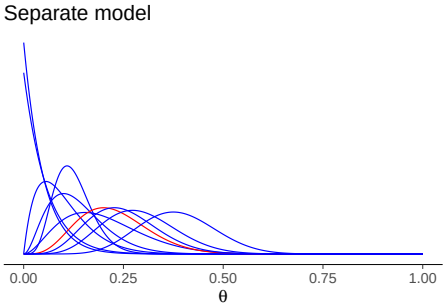
Hierarchical models: Binomial example with rats data





Hierarchical models: Binomial example with rats data





Hierarchical models: Normal example

- Each group has its own mean θ_j and own variance σ_j^2
- As before, we model the parameters as coming from a common distribution

- Level 1: $y_{ij}|\theta_j, \sigma_j^2 \sim \mathcal{N}(\theta_j, \sigma_j^2)$
- Level 2:
  - $\theta_j|\mu_0, \tau_0^2 \sim \mathcal{N}(\mu_0, \tau_0^2)$
  - $\sigma_j^2|\alpha_0, \delta_0 \sim$ Inv-Gamma(α_0, δ_0)