37457

Advanced Bayesian Methods

Overview of Grouped Data Analysis

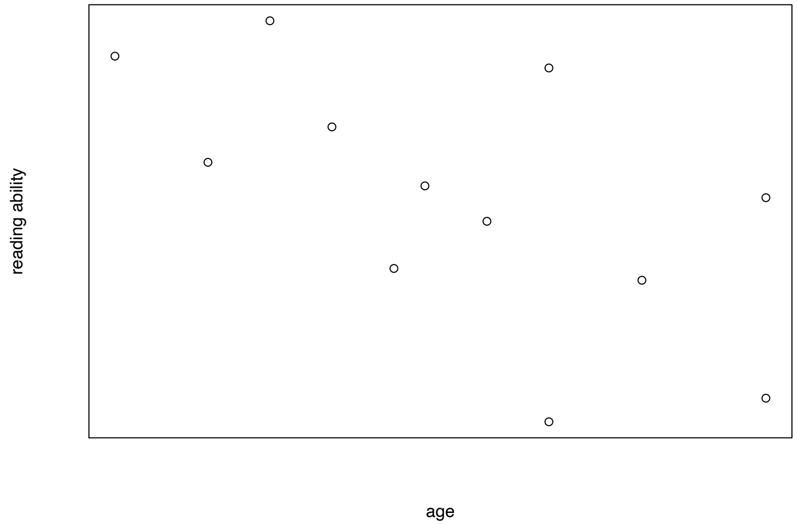
Types of Grouped Data (Common Application Areas)

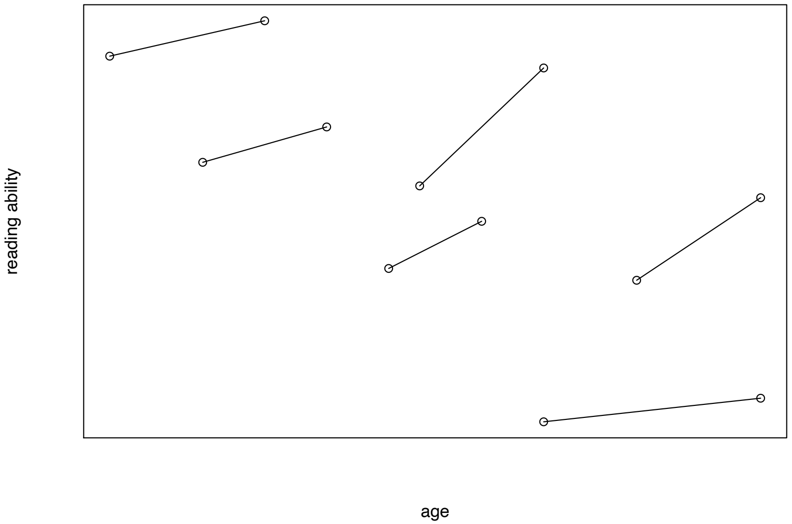
- Longitudinal data (medicine; public health).
- Multilevel data (education; sociology).
- Panel data (economics).
- Small area data (survey sampling).
- Item response data (psychology).

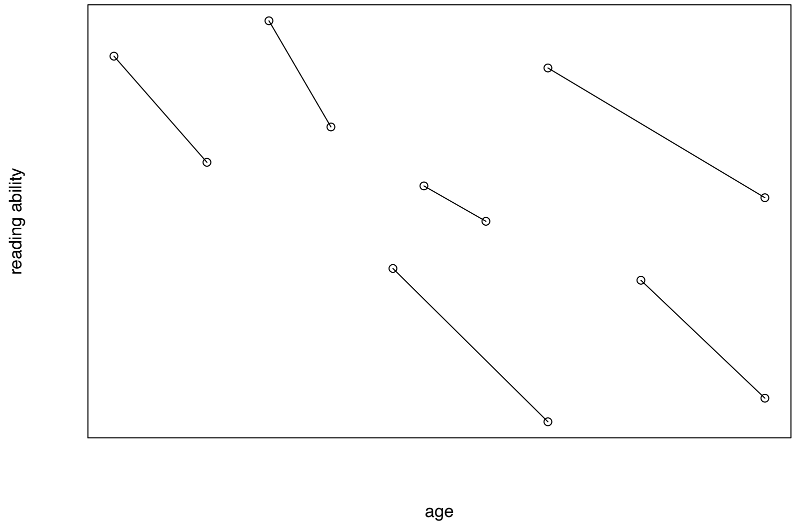
Longitudinal Studies

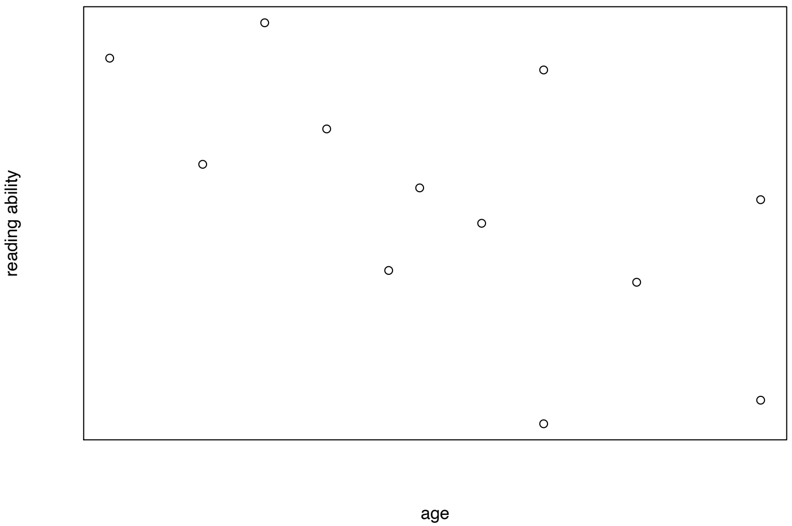
The defining characteristic of longitudinal studies is that subjects are measured repeatedly over time.

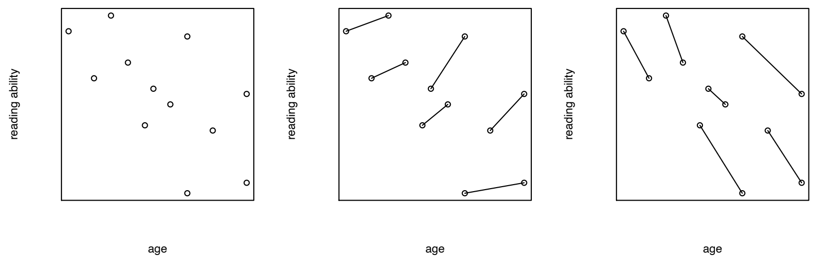
This is in contrast to cross-sectional studies where a single outcome is measured for each individual.




reading ability
age


reading ability
age


reading ability
age


reading ability
age
reading ability
age
reading ability
age

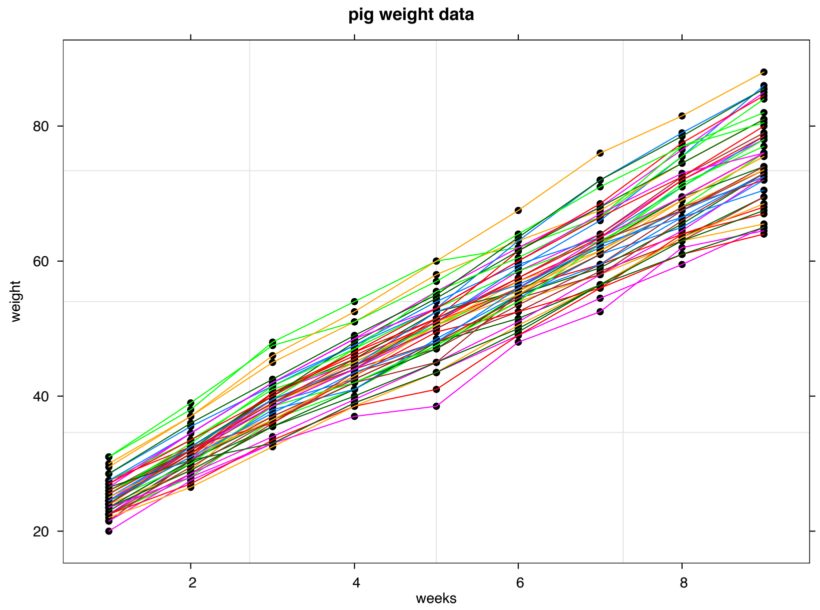
Cross-sectional versus Longitudinal

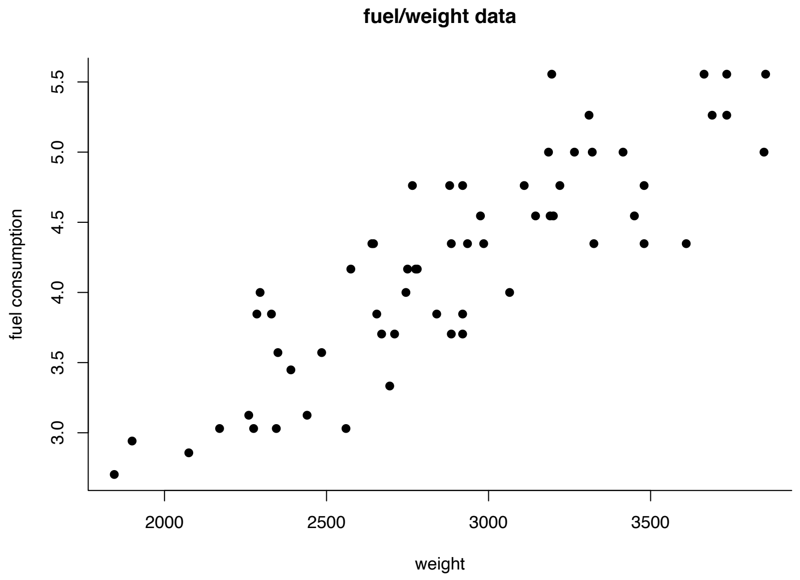
The next three slides show real data examples of

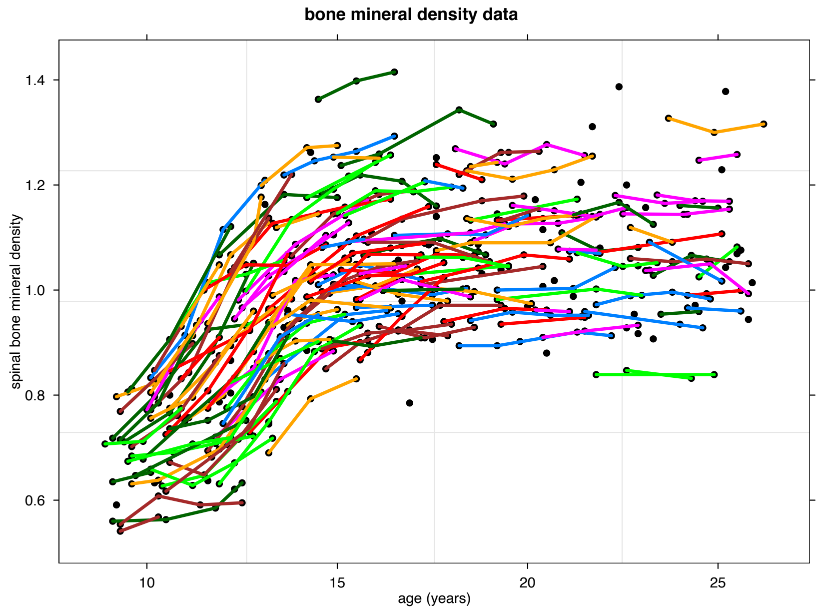
Cross-sectional data (1st slide)

and

Longitudinal data (2nd & 3rd slides)

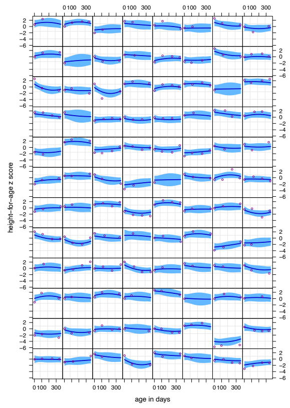






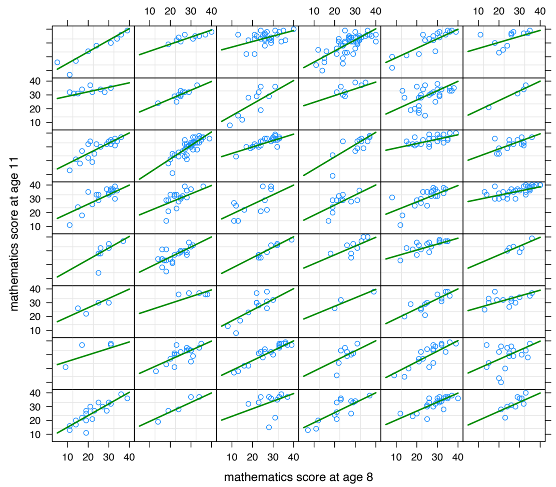
SOME OTHER GROUPED DATA EXAMPLES

"BIG DATA": 45 Thousand Infants

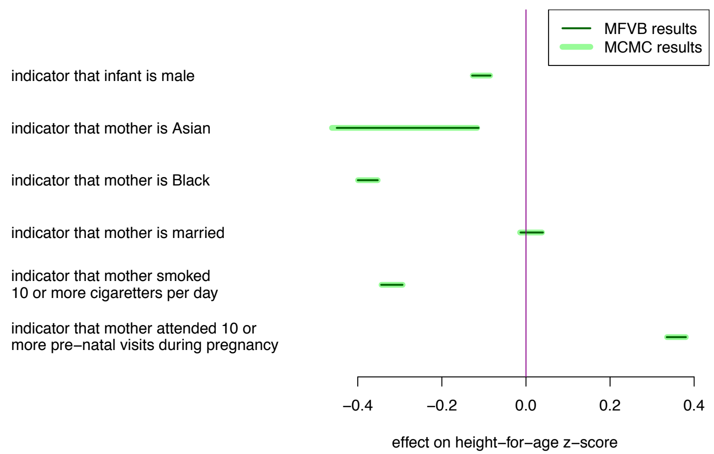


The above plot is only ONE FIVE HUNDREDTH of the full data.

Grouping By Schools



But We Still Want Regression-Type Results for the Other Predictors



Double Subscript Notation

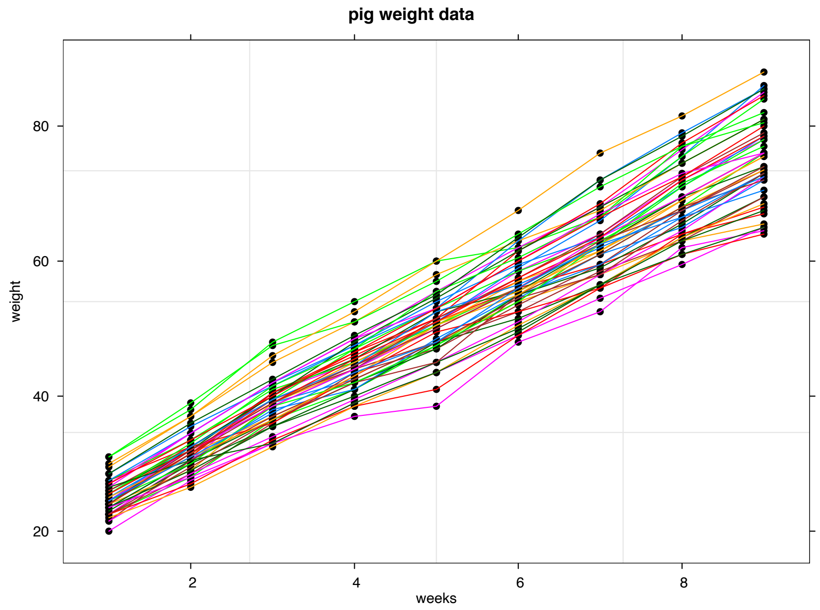
The letter y is usually used for a generic response or outcome variable.

Let

$$\begin{aligned} m &= \text{number of groups} \\ n_i &= \text{number of measurements for group } i\ (1 \le i \le m) \end{aligned}$$

Then

$$y_{ij} = \text{response for } j\text{th measurement on group } i.$$



Whiteboard Interlude

This is to illustrate the ideas of

double subscripting.

Naïve Model for Pig Weights

The slopes look about the same.

But each pig seems to have his/her own intercept

$$\Longrightarrow \quad y_{ij} = \beta_{0i} + \beta_1 x_{ij} + \varepsilon_{ij}$$

for $1 \le i \le 48$, with $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$.

But this model has 50 parameters!:

$$\beta_{01}, \beta_{02}, \ldots, \beta_{0,48}, \beta_1 \text{ and } \sigma_\varepsilon^2.$$

And only the last 2 are interpretable.

Random Intercept Model

An better model is:

$$y_{ij} = U_i + \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}$$

where

$$U_i \overset{\text{ind.}}{\sim} N(0, \sigma_U^2)$$

are independent of the

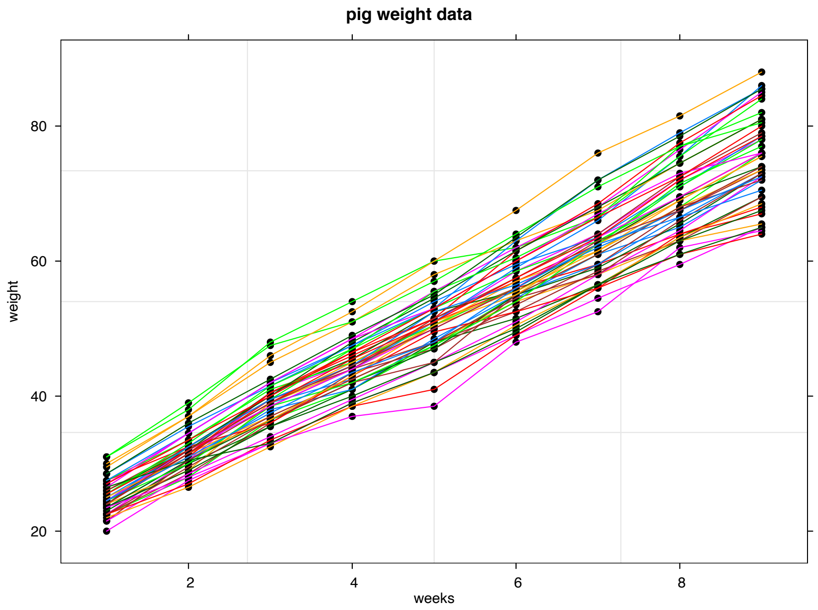
$$\varepsilon_{ij} \overset{\text{ind.}}{\sim} N(0, \sigma_\varepsilon^2).$$

Random Effects

The U_i are called

random effects

By design they are centred around zero and correspond to the ith pig's devation from the 'average' intercept β_0.



Mixed Model Terminology

$$y_{ij} = \underbrace{U_i}_{\text{random effect}} + \underbrace{\beta_0 + \beta_1 x_{ij}}_{\text{fixed effects}} + \varepsilon_{ij}$$

The right-hand side has a mixture of random effects and fixed effects and so is called a

(LINEAR) MIXED MODEL

TECHNICAL ASIDE

HOW MIXED MODELS INDUCE

WITHIN GROUP

CORRELATION STRUCTURE

We now describe the essential properties of the *random intercept model*

$$y_{ij} = U_i + \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}$$

$$U_i \overset{\text{ind.}}{\sim} N(0, \sigma_U^2), \qquad \varepsilon_{ij} \overset{\text{ind.}}{\sim} N(0, \sigma_\varepsilon^2)$$

and the U_i and ε_{ij} are independent of each other.

For $j \neq j'$, $\mathrm{Cov}(y_{ij}, y_{ij'})$ is the covariance between different measurements on the same group:

$$\begin{aligned}
\mathrm{Cov}(y_{ij}, y_{ij'}) &= \mathrm{Cov}(U_i + \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}, \\
&\qquad U_i + \beta_0 + \beta_1 x_{ij'} + \varepsilon_{ij'}) \\
&= \mathrm{Cov}(U_i + \varepsilon_{ij}, U_i + \varepsilon_{ij'}) \\
&= \mathrm{Cov}(U_i, U_i) + \mathrm{Cov}(U_i, \varepsilon_{ij'}) + \mathrm{Cov}(U_i, \varepsilon_{ij}) \\
&\quad + \mathrm{Cov}(\varepsilon_{ij}, \varepsilon_{ij'}) \\
&= \sigma_U^2 + 0 + 0 + 0 \\
&= \sigma_U^2
\end{aligned}$$

For $j = j'$, $\mathrm{Cov}(y_{ij}, y_{ij'}) = \mathrm{Var}(y_{ij})$ is the variance of the jth measurement on group i, and has expression:

$$\begin{aligned}
\mathrm{Cov}(y_{ij}, y_{ij}) &= \mathrm{Var}(U_i + \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}) \\
&= \mathrm{Var}(U_i + \varepsilon_{ij}) \\
&= \mathrm{Var}(U_i) + \mathrm{Var}(\varepsilon_{ij}) \\
&= \sigma_U^2 + \sigma_\varepsilon^2
\end{aligned}$$

For $i \neq i'$ we get

$$\mathrm{Cov}(y_{ij}, y_{i'j'}) = 0.$$

This says that observations of different individuals are uncorrelated (e.g. Anne's blood pressure is not correlated with Bill's blood pressure).

Consider the sample sizes corresponding to the example:

$$m = 3, \qquad n_1 = 2, \qquad n_2 = 3, \qquad n_3 = 2.$$

The covariance matrix of

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \end{bmatrix}$$

is

$$\left[\begin{array}{cc|ccc|cc}
\sigma_U^2+\sigma_\varepsilon^2 & \sigma_U^2 & 0 & 0 & 0 & 0 & 0 \\
\sigma_U^2 & \sigma_U^2+\sigma_\varepsilon^2 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & \sigma_U^2+\sigma_\varepsilon^2 & \sigma_U^2 & \sigma_U^2 & 0 & 0 \\
0 & 0 & \sigma_U^2 & \sigma_U^2+\sigma_\varepsilon^2 & \sigma_U^2 & 0 & 0 \\
0 & 0 & \sigma_U^2 & \sigma_U^2 & \sigma_U^2+\sigma_\varepsilon^2 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & \sigma_U^2+\sigma_\varepsilon^2 & \sigma_U^2 \\
0 & 0 & 0 & 0 & 0 & \sigma_U^2 & \sigma_U^2+\sigma_\varepsilon^2
\end{array}\right]$$

The correlation matrix is then

$$\left[\begin{array}{cc|ccc|cc}
1 & \rho & 0 & 0 & 0 & 0 & 0 \\
\rho & 1 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 1 & \rho & \rho & 0 & 0 \\
0 & 0 & \rho & 1 & \rho & 0 & 0 \\
0 & 0 & \rho & \rho & 1 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 1 & \rho \\
0 & 0 & 0 & 0 & 0 & \rho & 1
\end{array}\right] \quad \text{where} \quad \rho = \frac{\sigma_U^2}{\sigma_U^2+\sigma_\varepsilon^2}.$$

Note the blocking structure corresponding to within-group correlation. The first block is for group 1, the second is for group 2 and the third is for group 3.

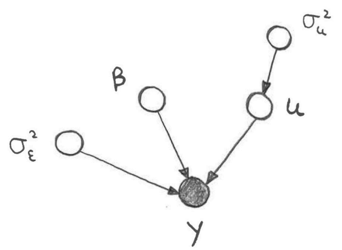
Remarks

1. The random intercept U_i invokes correlation between measurements on same group.
2. A shortcoming of the random intercept model is that the correlation is the same for each group; e.g. Anne's ρ is the same as Bill's ρ.
3. Another shortcoming is that the within-group correlation is constant over time; e.g. the correlation between Anne's blood pressure measurements 2 days apart is the same as those taken 10 days apart.

BAYESIAN VERSION OF LINEAR MIXED MODELS

$$y_{ij}|\beta_0, \beta_1, u_i, \sigma_\varepsilon^2 \overset{\text{ind.}}{\sim} N(\beta_0 + \beta_1\, x_{ij} + u_i, \sigma_\varepsilon^2)$$

$$u_i|\sigma_u^2 \overset{\text{ind.}}{\sim} N(0, \sigma_u^2)$$



♦ ♦ ♦