Statistics for Linguistics

Adam J.R. Tallman

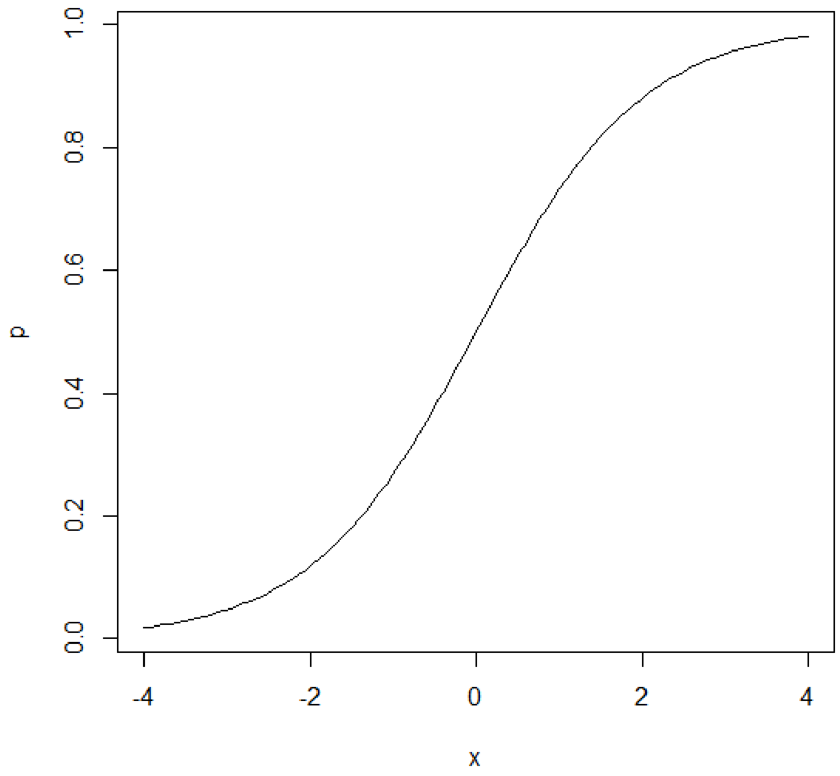
2021-06-08

From last class

- Confounds
  - The fork, the pipe, the collider
- Generalized linear models

Logistic regression

- Logistic regression
- Hierarchical models



Logistic regression

Logistic regression

- Logistic regression is a method for modelling binary data.
- The basic ideas can be extended to non-binary data as long as they are organized into levels.
- It is typically used when the dependent variable is binary and there is an interest in knowing how a change in *x* effects the probability that something is *y*.

Logistic regression (typical uses)

- Psycholinguistic experiments where subjects have to give yes or no answers.

- Various uses in natural language processing

- Predict the risk of developing a specific disease.

- Predict probability that someone will vote for a particular political party

Logistic regression

- A logistic regression or logit model can be represented with the following equation.

$$logit(y) = b_0 + b_1 x_1 + b_2 x_2 ...$$

$$logit(p) = log \frac{p}{1-p}$$

$$Prob\{y = 1 | x\} = \frac{1}{1 + exp(-x\beta)}$$

The logistic function

- But let’s back up a bit to see why this makes sense and what this means…

Bernouilli distribution

- Bernouilli distribution is a discrete distribution with two possible outcomes.
- We want to account for the distribution of *y*.
- You can model it with rbinom()

$$P(y) = \begin{cases} 1-p & \text{for } y=0 \\ p & \text{for } y=1 \end{cases}$$

```
data <- rbinom(20, 1, 0.5)
data
```

$$P(y) = p^y(1-p)^{1-y}$$

Logistic function

- The important point about log odds ratios as that they take any numbers and transform them to a number from 0 to 1 along a sigmoid shape.

- Why is this good?

- Because we are interested in modelling a a binary outcome, 0 or 1, and we want to have a function that translates the effect of predictors into that scale for y.

Logistic function

- That's basically what the following will do.
- Any value of *x* will be transformed into a number that varies from 0 to 1 with ceiling effects.

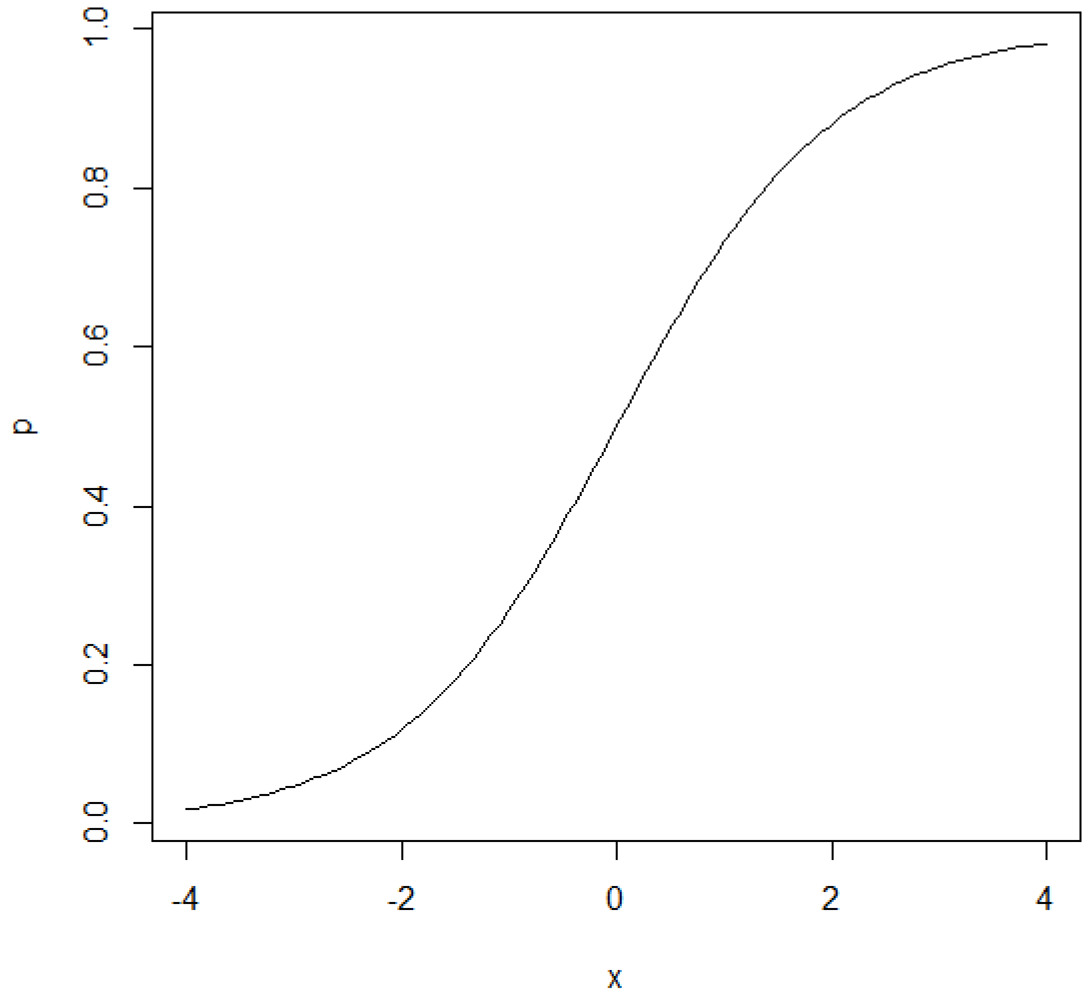
$$Prob\{y = 1|x\} = \frac{1}{1 + exp(-x\beta)}$$

Logistic regression

- Y Is bounded to 0 or 1

- The relationship between x and y has a ceiling effect (like logarithms)

- Let’s run a simulation model to get the feel for it.



Two causative constructions in Dutch

Doen relates to direct causation

Laten relates to indirect causation

(1)	*Hij*	***deed***	*me*	*denken*		*aan*	*mijn*	*vader*
	He	**did**	me	think		at	my	father
	‘He reminded me of my father.’							
(2)	*Ik*	***liet***	*hem*	*mijn*	*huis*	*schilderen*		
	I	**let**	him	my	house	Paint		
	‘I had him paint my house.’							

Interpreting logistic regression cofficients

- It is hard to interpret logistic regression coefficients because the relationship is non-linear.

- The intercept is interpreted assuming 0 for other predictors
  - But sometimes 0 is not interesting
  - Alternatively we can interpret the intercept at the center point
- Rather than consider a discrete change in *x* we can compute the derivative of the logistic curve at the central value (where the relationship is steepest.
  - You get this by dividng the coefficient by 4.

Multilevel models

Multilevel regression (varying intercept)

- Multilevel regression is an extension of regression modelling to cases where the data are grouped.

- When we have been talking about statistical models we have modelled the errors, the predictors and the outcomes as *random variables*.

- They are sampled from some probability distribution.

Multilevel regression and random variables

- Here's a normal regression model

$$y = \alpha + \beta x + \epsilon$$

- We can put in *i* to reflect the fact that our model makes predictions about specific data points: for a given *x* you get a specific *y* with a specific error *e*. Everything we can put with an *i* subscript is a random variable

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

Multilevel regression and random variables

- What do we mean by random variable?

- Some distribution with a mean and standard deviation.

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

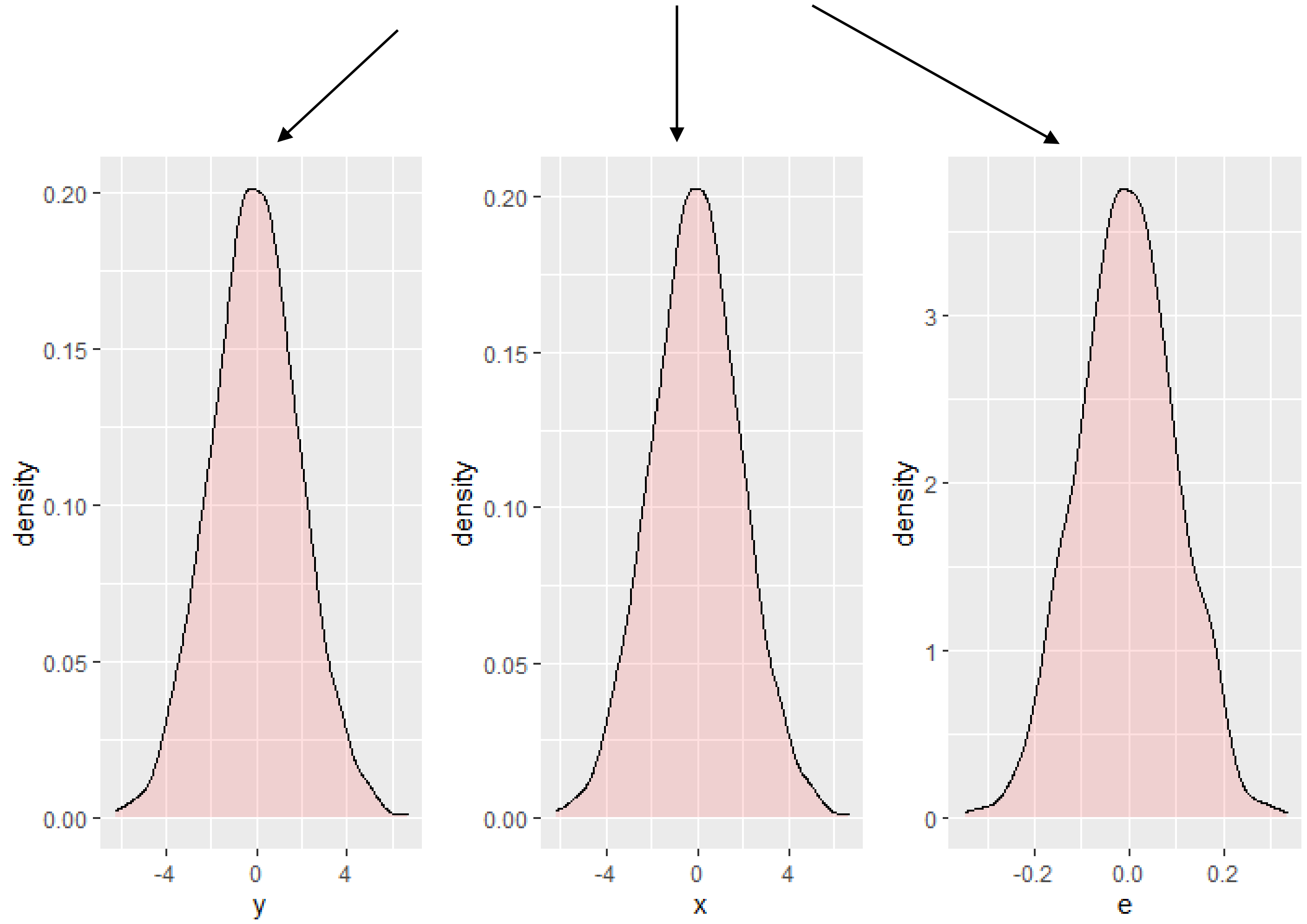
Multilevel regression and random variables

- What do we mean by random variable?

- Some distribution with a mean and standard deviation.

- But the *coefficients* alpha and beta are fixed.

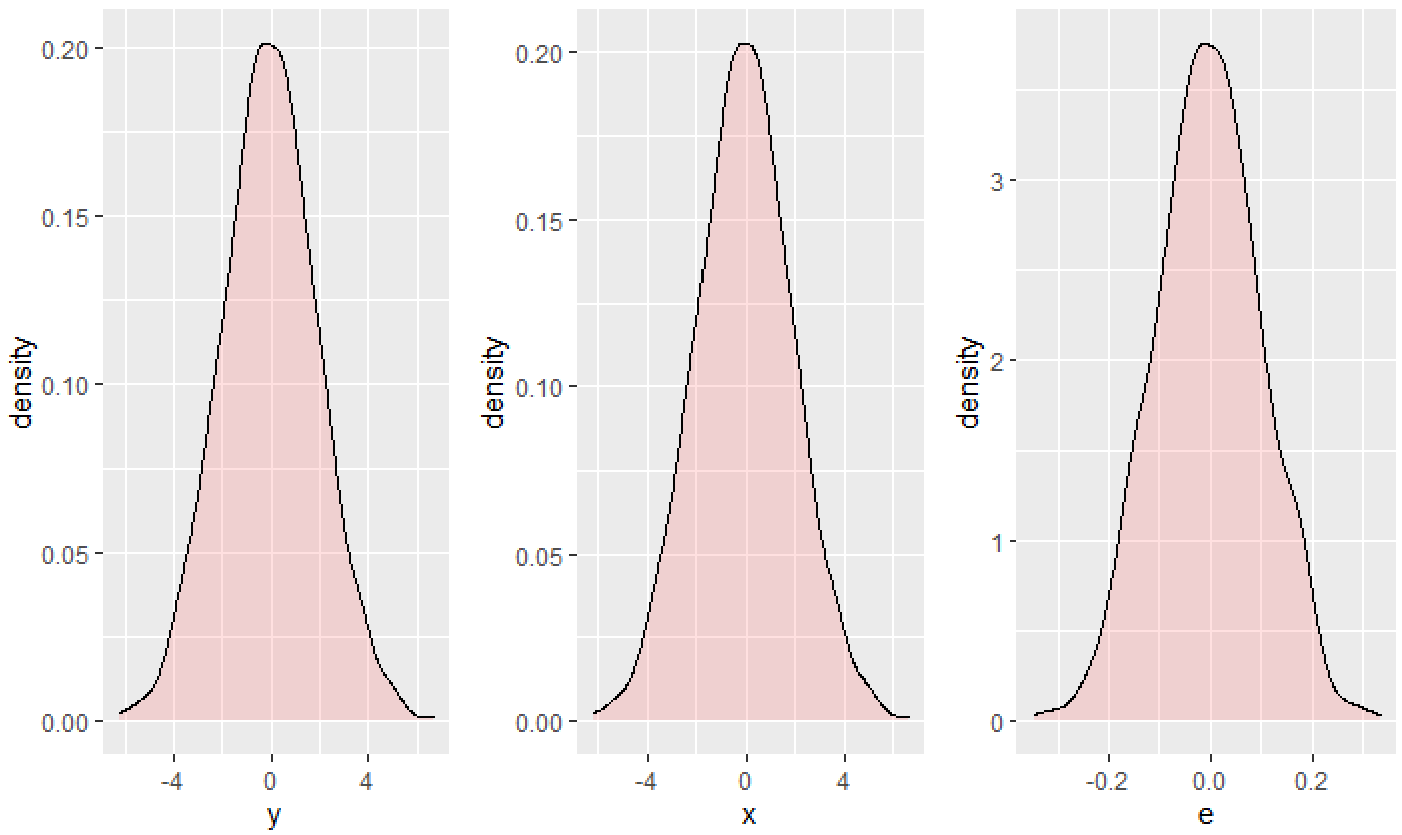
$$y_i = \alpha + \beta x_i + \epsilon_i$$



Multilevel regression and random variables

$$y_i = \boxed{\alpha} + \boxed{\beta} x_i + \epsilon_i$$

- What do we mean by random variable?
- Some distribution with a mean and standard deviation.
- But the *coefficients* alpha and beta are **fixed numbers**



Multilevel regression and random variables

- But let’s say you are running the model repeatedly over different groups within a population. Do you expect the coefficients to be the same?

Multilevel regression and random variables

- You can build a model where the intercept is also a random variable reflecting the variation between groups, where *j* is the group.

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

- Or a model where the slope is also a random variable.

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

- Or a model where both the intercept and the slope are random variables.

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

Multilevel regression and random variables

- In such cases you have a *random variable* for the effects of your model.

- That's why multilevel models are sometimes called 'random effects models'

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

Multilevel regression and random variables

- In such cases you have a *random variable* for the effects of your model.

- That's why multilevel models are sometimes called 'random effects models'

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

$$\begin{cases} y_i = \alpha + \beta x_i + \epsilon_i \\ y_i = \alpha + \beta x_i + \epsilon_i \\ y_i = \alpha + \beta x_i + \epsilon_i \\ y_i = \alpha + \beta x_i + \epsilon_i \\ \ldots \end{cases}$$

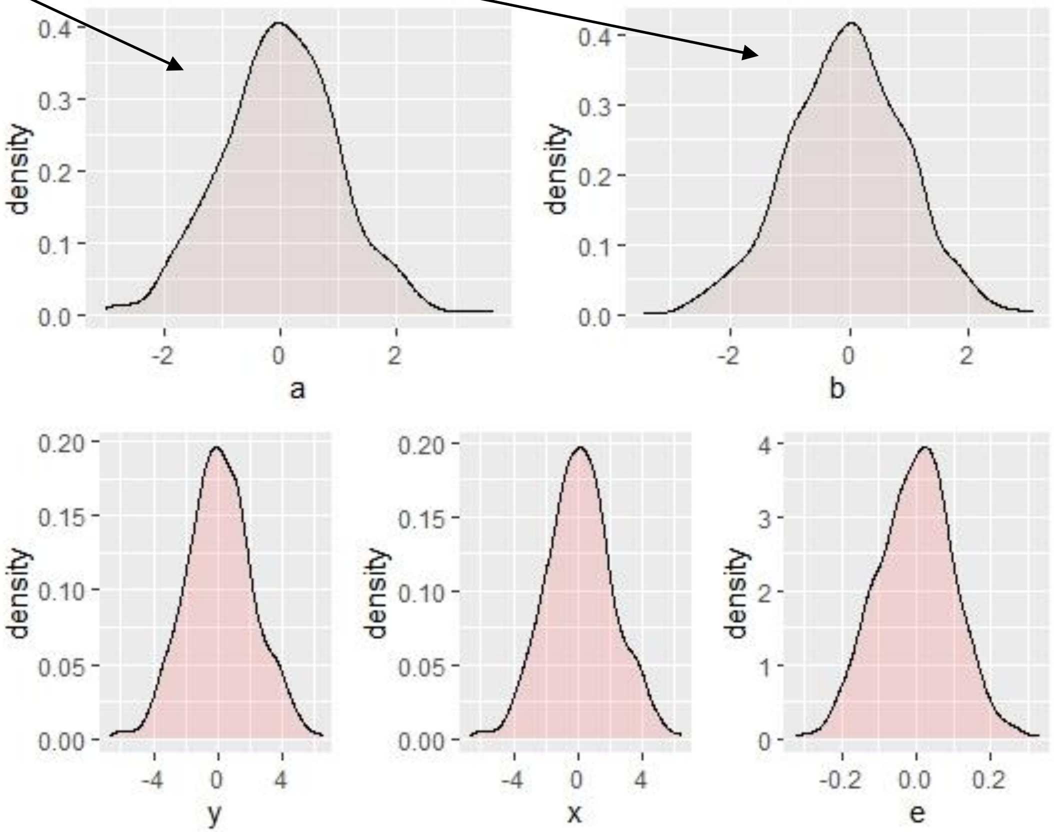
Multilevel regression and random variables

- In such cases you have a *random variable* for the effects of your model.

- That's why multilevel models are sometimes called 'random effects models'

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

$$\left\{\begin{array}{l} y_i = \alpha + \beta x_i + \epsilon_i \\ y_i = \alpha + \beta x_i + \epsilon_i \\ y_i = \alpha + \beta x_i + \epsilon_i \\ y_i = \alpha + \beta x_i + \epsilon_i \\ \ldots \end{array}\right.$$



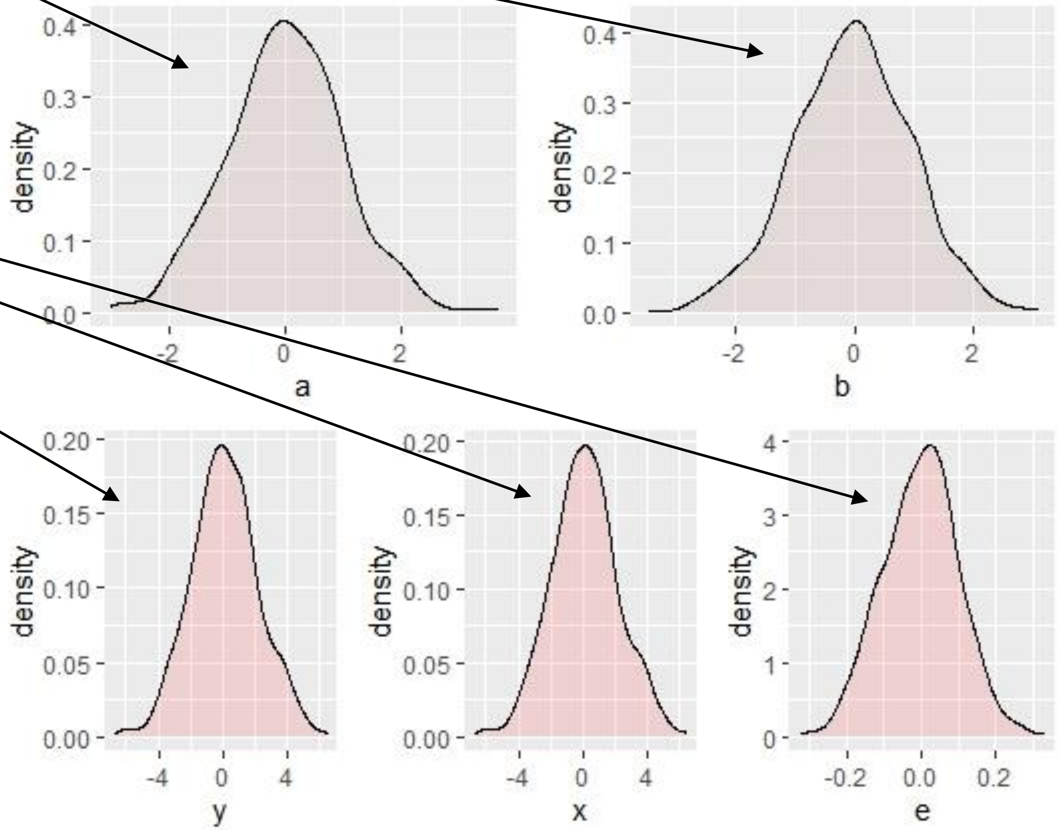
Multilevel regression and random variables

- In such cases you have a *random variable* for the effects of your model.

- That’s why multilevel models are sometimes called ‘random effects models’

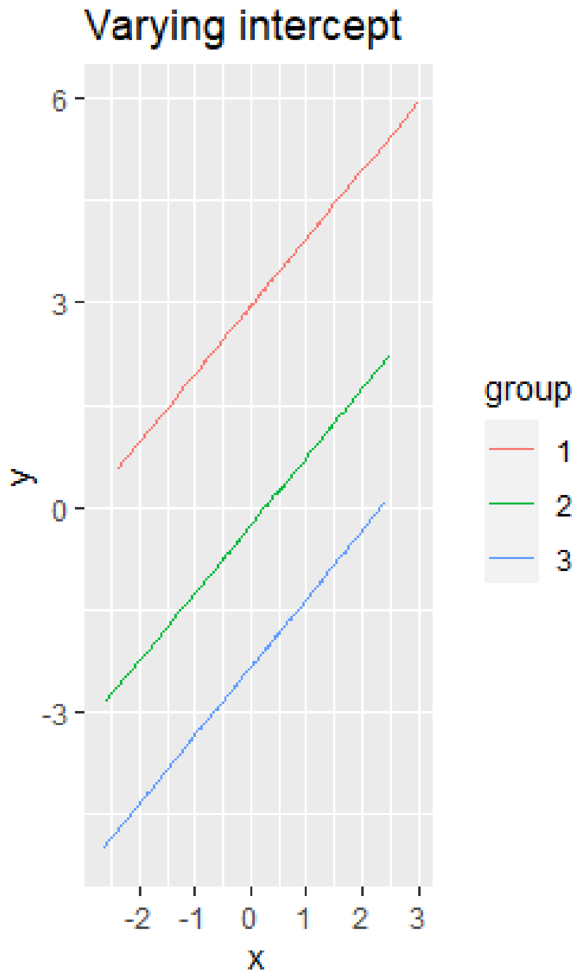
$$y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i$$

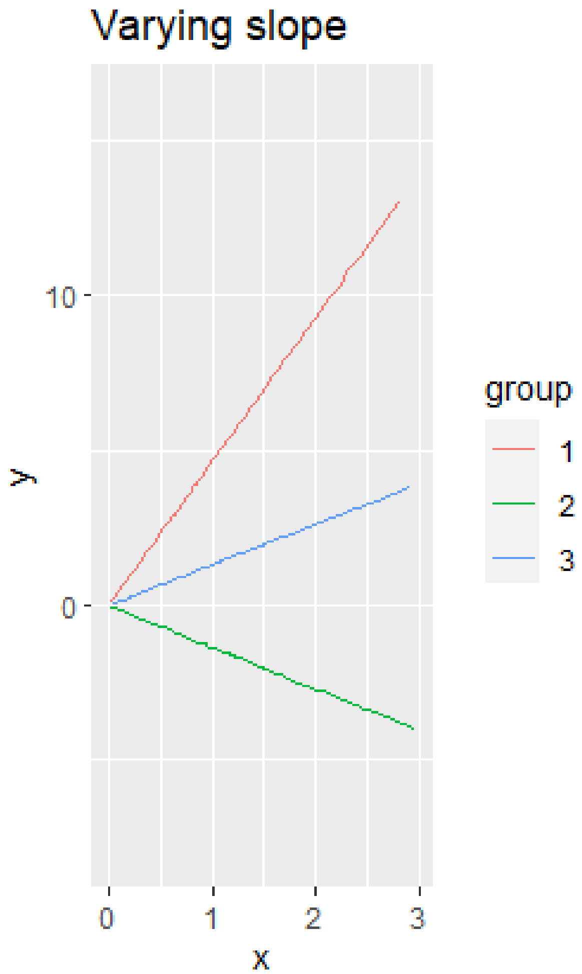
$$\left\{ \begin{array}{l} y_i = \alpha + \beta x_i + \epsilon_i \\ y_i = \alpha + \beta x_i + \epsilon_i \\ y_i = \alpha + \beta x_i + \epsilon_i \\ y_i = \alpha + \beta x_i + \epsilon_i \\ \ldots \end{array} \right.$$

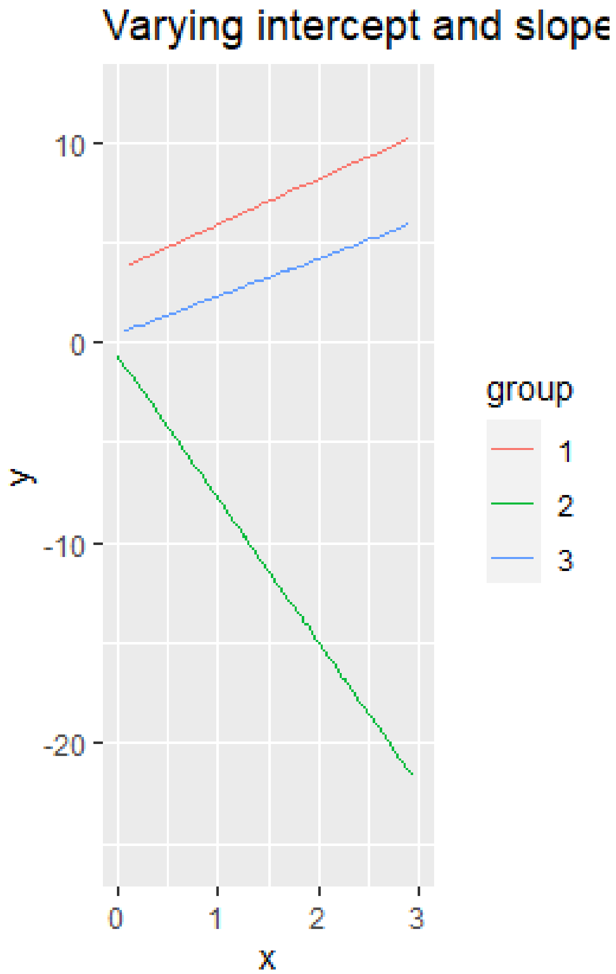


Multilevel regression and random variables

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







Why?

- Why would we ever do this?
- Why no just have lots of separate models for each group?

Schools

- Imagine trying to assess the effectiveness of some new education curriculum of teaching style.
- You have a treatment group and a control group and then you assess the students’ results.

Schools

- But you know that there will be variation between schools.
- Some schools won’t be able to effectively administer the training/treatment because they have less resources.
- Furthermore, you have variation between schools with respect to how many students participated.
- Schools vary in terms of their culture, socioeconomic conditions teachers, size, quality and style of education.
- Yet, the students are all from the same population.

Schools

- You have a measurement for aptitude *y* and you have a treatment variable (trained or not trained) *x*.

- What do you do with the schools variable?

- Complete pool
  - Run a regression ignoring the variation between schools

- No pooling
  - Run a regression for each school
  - Run a regression with school as a factor

Complete pooling

- Complete pooling has the obvious danger of ignoring the variation between schools.
- If one school has more data points it could be an outlier with respect to the effects, but overwhelm the data across cases.
- The results might be biased towards with more data points.

No pooling

- No pooling could tend to exaggerate the variation between schools.

- For schools that do not have very many data points, there is a higher likelihood of variation simply appearing by chance.
  - Think of the law of large numbers

Partial pooling

- Multilevel modelling basically compromises between complete pooling and no pooling.
- ‘Multilevel modeling partially pools the group-level parameters α_j toward their mean μ_α. There is more pooling when the group-level standard deviation σ_α is small, and more smoothing for groups with fewer observations.’
  - Gelman & Hill (2009: 258)

$$\text{estimate of } \alpha_j \approx \frac{\frac{n_j}{\sigma_y^2}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}} (\overline{y} - \beta \overline{x}_j) + \frac{\frac{1}{\sigma_\alpha^2}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}} \mu_\alpha$$

Partial pooling

- Partial pooling results in *shrinkage* of variance in the coefficients of each group towards the overall mean as a function of their in-group sample size (n_j)

$$\text{estimate of } \alpha_j \approx \frac{\boxed{\frac{n_j}{\sigma_y^2}}}{\frac{n_j}{\sigma_y^2}+\frac{1}{\sigma_\alpha^2}}(\bar{y}-\beta\bar{x}_j)+\boxed{\frac{\frac{1}{\sigma_\alpha^2}}{\frac{n_j}{\sigma_y^2}+\frac{1}{\sigma_\alpha^2}}\mu_\alpha}$$

- Shrinkage is less as your in-group sample size is larger.
- Partial pooling towards the mean.

Multilevel models

- Classical regression models can be viewed as a special case of multilevel models

- As $\sigma_\alpha \rightarrow 0$ the model is more like a complete pooling model.

- As $\sigma_\alpha \rightarrow \infty$ the model is more like a no-pooling model.

Multilevel models

- Extension of regression to grouped data
- Varying slope model
- Varying intercept model
- Individual vs. group level models
- Indicator variables
- Fixed or random effects
- Complete pooling vs. no pooling
- Multilevel weighted average
- Classical regressions as a special case

Odds,log odds, odds ratios and log odds ratios

- **Odds**: simple ration of the probability of one event to the probability of another event (frequency of a / frequency b) are the odds of a over b.

- **Log odds**: Logarithmically transformed odds.

- **Odds ratio**: ratio of two odds.

- **Log odds ratio**: Logarithmically transformed log odds.