

STATS 211: Statistical Methods II

Lecture 7: Multilevel models

Babak Shahbaba

UCI, Winter 2012

Clustered data

- In linear regression models, we assumed the observations were independent without any underlying structure connecting the subjects.
- This is not a realistic assumption in many data analysis problems.
- For example, consider the “radon” example discussed in Gelman and Hill (2007).
- Radon is a carcinogen responsible for several thousand lung cancer deaths per year in the US.
- The concentration of radon varies in US homes. EPA has started a project to collect radon measurements in a random sample of more than 80,000 houses.

Clustered data

- In general, data are clustered (i.e., dependent) when they have hierarchical structure, or observations are collected on related subjects, or there are multiple observations (possibly over time) on the same subject.
- Clustered data could come from group randomized trials where group of individuals, as opposed to individual themselves, are assigned to different treatments.
- We might also have observational clustered data, where data are collected from related individuals such as families, geographical regions, or in general multiple sources.
- Data are also clustered when we obtain multiple observations over time from the same individuals. In this case, we refer to the data as longitudinal data. We will discuss the analysis of such data later.

Clustered data

- when dealing with clustered data, we expect measurements within a cluster to be more similar compared to measurements in different clusters.
- That is, the structure of the data introduces correlation among measurements within a cluster.
- This violates the independence assumption used in linear regression models.

Clustered data

- Consider the radon example again. We could assume these observations are independent. However, a more realistic assumption is that measurements within a county tend to be more similar compared to measurements from different counties.
- That is, we expect the measurements to form clusters within each county, and the overall radon level varies from one county to another.
- In the above example, the data have a hierarchical structure: houses within counties.
- In general, we could have more than two levels: we could divide counties among states, and we could have multiple observations (i.e., repeated measurements) from each house.
- Clustered data are not always nested (i.e., hierarchical). For example, we can obtain radon measurements over several years. In this case, “years” and “counties” are not nested.

To pool or not to pool!

- To model such data, we could ignore the underlying structure and *pool* all counties together and fit a single model to the whole data. We refer to this approach as *complete pooling*.
- This method, of course, ignores variation between counties.
- ALternatively, we could fit a separate model to data from each county. We refer to this approach as “no pooling”.
- This is an inefficient approach that fits many models (~ 3000 in the radon example), where the estimated parameters in most of them are based on a small sample size, which results in large standard errors.
- A reasonable compromise between complete pooling and no pooling is provided by *multilevel models* that perform *partial pooling*.
- These methods take the underlying structure of the data into account, and they tend to provide more reasonable estimates.

Motivation

- Multilevel models allow us to investigate how some effects vary by group; this is analogous to including interaction terms (here, between covariates and the group indicator) in classical models.
- These models provide a reasonable compromise between complete pooling, which is very restricted, and no pooling, which is overly flexible.
- Multilevel models tend to provide better prediction by capturing the overall trend while taking the group effect into account.
- Using multilevel models, we can model higher level data (e.g., radon level in counties) using their own specific covariates. For the radon problem, we can use, for example, a measurement of soil uranium, which is available at the county level only.

Two-stage models

- Consider the radon data. We would like to estimate the effect of the `floor` variable, which indicates the floor of measurement (basement is coded as 0, first floor as 1, ...).
- To fit a multilevel model, we can use a two-step model for the two levels of the data.
- We use the following model for observations within each county:

$$y_{ij} = \alpha_i + \beta x_{ij} + \epsilon_{ij}$$

where y_{ij} is the measurement for the j^{th} house in the i^{th} county, and x_{ij} is the first-floor indicator.

- Note that in this model, we are assuming that the overall radon level varies from one county to another. However, the floor effect remains the same for all counties. This is a random-intercept model.

Two-stage models

- Next, we model the variation of intercepts among counties as follows:

$$\alpha_i = \mu_\alpha + \eta_i$$

- The model at the higher level can itself depend on some covariates. For example, our model for random intercepts can depend on z_i , county level measurements of soil uranium:

$$\alpha_i = \mu_\alpha + \gamma z_i + \eta_i$$

Mixed effect models

- Alternatively, we can combine the above two model and write the result as a mixed-effect model

$$\begin{aligned}y_{ij}|\alpha_i &\sim N(\alpha_i + \beta x_{ij}, \sigma_y^2) \\ \alpha_i &\sim N(\mu_\alpha, \sigma_\alpha^2)\end{aligned}$$

- If our model for random intercepts depend on a higher level covariate, z_i , we have

$$\alpha_i \sim N(\mu_\alpha + \gamma z_i, \sigma_\alpha^2)$$

- The assumed distribution for α_i shrinks the random intercepts towards their overall mean, μ_α .

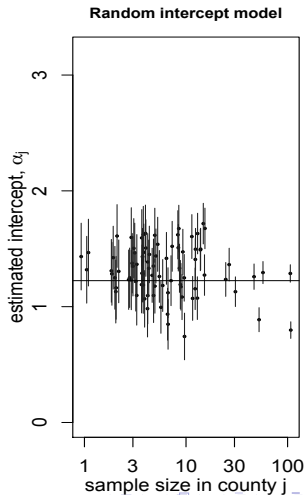
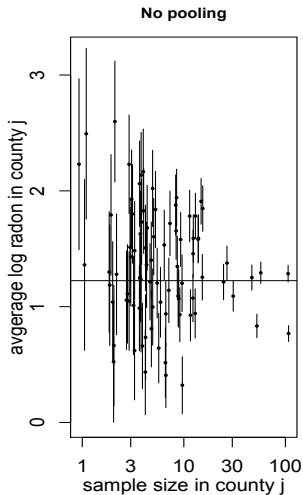
Mixed effect models

- It is now easy to see how multilevel models can be considered as a compromise between complete pooling and no pooling.
- As $\sigma_\alpha^2 \rightarrow 0$, the estimates for α_i shrink all the way to μ_α . Therefore, in limit, as $\sigma_\alpha^2 \rightarrow 0$, the multilevel model becomes equivalent to the model with complete pooling.
- On the other hand, as $\sigma_\alpha^2 \rightarrow \infty$, the multilevel model converges to the model with no pooling.

Radon example

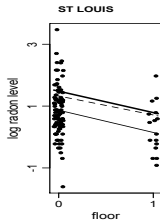
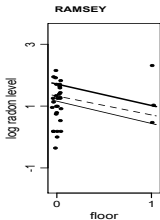
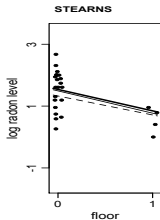
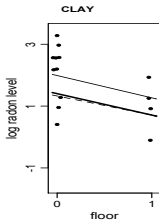
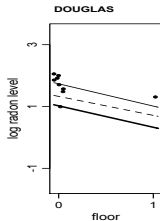
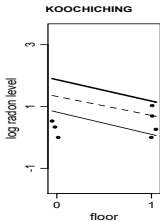
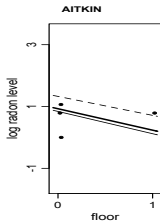
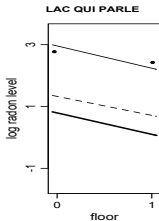
- We now discuss the application of mixed effects models for analyzing the radon data.
- We focus on the counties in Minnesota.
- First, we fit a model without any predictor. Next, we include the variable “floor” as a predictor in our model.
- To fit mixed effects models, we use the function `lmer` from the `lme4` package.
- Alternatively, we can use the `lme` function from the `nlme` package.

Random intercept and no predictor for the counties in Minnesota



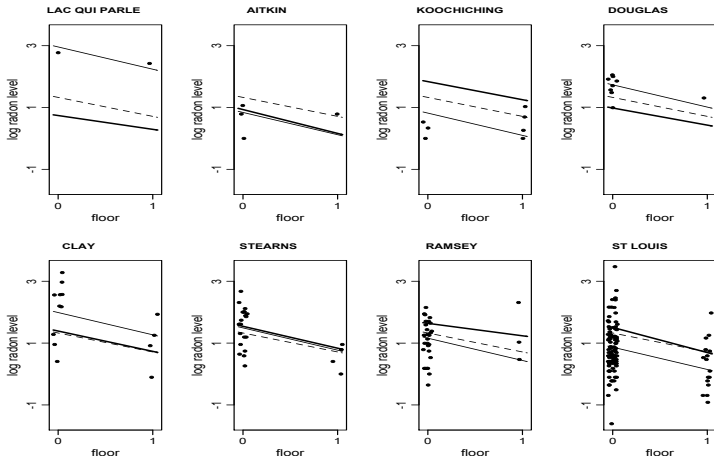
Random intercept and fixed coefficient for “floor”

dashed line: complete pooling; solid thin line: no pooling; solid thick line: partial pooling



Random intercept and random coefficient for “floor”

dashed line: complete pooling; solid thin line: no pooling; solid thick line: partial pooling



A general form

- Consider the following mixed effects model

$$y = x\beta + z\gamma + \epsilon$$

where β and γ are vectors of fixed and random effects respectively.

- In this course, we assume normal distributions for γ and ϵ , and write the above model as follows:

$$\begin{aligned} y|\gamma &\sim N(x\beta + z\gamma, \Sigma) \\ \gamma &\sim N(0, \Lambda) \end{aligned}$$

A general form

- The mean and covariance matrix of y are

$$\begin{aligned} E(y) &= E(E(y|\gamma)) \\ &= E(x\beta + z\gamma) \\ &= x\beta \end{aligned}$$

$$\begin{aligned} \text{Cov}(y) &= E(\text{Cov}(y|\gamma)) + \text{Cov}(E(y|\gamma)) \\ &= \Sigma + \text{Cov}(x\beta + z\gamma) \\ &= \Sigma + \text{Cov}(z\gamma) \\ &= \Sigma + z\Lambda z^T \end{aligned}$$

- Therefore,

$$y \sim N(x\beta, V) \text{ with } V = \Sigma + z\Lambda z^T$$

which is known as the *marginal model*.

Statistical inference for LME models

- Using the above marginal model, we can write down the likelihood function and find the maximum likelihood estimates for model parameters.

$$f(y) = \frac{1}{(2\pi)^{n/2} |V|^{1/2}} \exp \left[-\frac{1}{2} (y - x\beta)^T V^{-1} (y - x\beta) \right]$$

- The log-likelihood (up to a constant) can be written as

$$L(\beta, \theta) = -\frac{1}{2} \log(|V|) - \frac{1}{2} (y - x\beta)^T V^{-1} (y - x\beta)$$

where θ represents the set of parameters for the covariance part of the model.

- Setting the derivatives to zero, we obtain the following results (for more details, see Searle et al., 1992)

$$\hat{\beta} = (x^T \hat{V}^{-1} x)^{-1} x^T \hat{V}^{-1} y$$

where $\hat{V} = V(\hat{\theta})$.

Statistical inference for LME models

- \hat{V} itself is obtained from

$$\text{tr}\left(V^{-1} \frac{\partial V}{\partial \theta_r}\right) = y^T P \frac{\partial V}{\partial \theta_r} P y$$

for $r = 1, \dots, q$, and

$$P = V^{-1} - V^{-1} X (X^T V^{-1} X)^{-1} X^T V^{-1}$$

Statistical inference for LME models

- As usual, under some regularity conditions, the MLE has an asymptotic normal distribution with covariance matrix equal to the inverse of the Fisher information.
- Let $\phi = (\beta, \theta)$. The element in row i and column j of the Fisher information is obtained as follows:

$$-E\left(\frac{\partial^2 L}{\partial \phi_i \partial \phi_j}\right)$$

- For the above mixed effects model, we have

$$E\left(\frac{\partial^2 L}{\partial \beta \partial \beta^T}\right) = -x^T V^{-1} x$$

$$E\left(\frac{\partial^2 L}{\partial \beta \partial \theta_r}\right) = 0$$

$$E\left(\frac{\partial^2 L}{\partial \theta_r \partial \theta_s}\right) = -\frac{1}{2} \text{tr}\left(V^{-1} \frac{\partial V}{\partial \theta_r} V^{-1} \frac{\partial V}{\partial \theta_s}\right)$$

Statistical inference for LME models

- Based on the joint distribution of y and γ , we can find γ as follows:

$$\hat{\gamma} = \Lambda z^T V^{-1}(y - x\hat{\beta})$$

- This is known as the “best linear unbiased predictor” (BLUP).
- Henderson (1953) proposed the following compact equations for estimating β and γ

$$\begin{bmatrix} x^T \Sigma^{-1} x & x^T \Sigma^{-1} z \\ z^T \Sigma^{-1} x & z^T \Sigma^{-1} z + \Lambda^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} x^T \Sigma^{-1} y \\ z^T \Sigma^{-1} y \end{bmatrix}$$

- See Searle et. al. (1992) for more details.

Statistical inference for LME models

- For hypothesis testing regarding the fixed effects, we can use the likelihood-based tests as before.
- For random effects, however, Stram and Lee (1994) showed that this test is conservative when we evaluate the null hypothesis regarding the random effects.
- In this case, under the null hypothesis, we need to set the corresponding variance of a random effect to zero in order to exclude it from the full model. However, zero is on the boundary of the parameter region.
- As the result, the asymptotic distribution of the likelihood ratio test under H_0 is not χ_1^2 anymore. In fact, it is a 50:50 mixture of χ_1^2 and point mass at 0.
- Therefore, the asymptotic distribution of the likelihood ratio test should be adjusted. In practice, we obtain the p -value form χ_1^2 and divide it by 2.

Statistical inference for LME models

- Obviously, the adjustment is more complicated for hypotheses that involve multiple random effects.
- However, as long as we are comparing nested random effects models with r and $r + 1$ random effects (same fixed effects), the distribution of the likelihood ratio test statistic is a 50-50 mixture of two χ^2 distributions with r and $r + 1$ degrees of freedom.
- For simplicity, we can use $\chi^2(r + 1)$ to be conservative.

Restricted maximum likelihood

- The MLE for V is biased the same way that the MLE for σ^2 is biased in linear regression models.
- To improve the estimator, Patterson and Thompson (1971) proposed the method of *restricted maximum likelihood* (REML).
- This is analogous to using $RSS/(n - p)$ instead of RSS/n in linear regression models. In fact $RSS/(n - p)$ is the REML estimator for σ^2 .
- To find REML, we need to transform the observed response variable, $y^* = Ay$ so the distribution of y^* does not depend on β .
- One way to do this, is to use $A = I - x(x^T x)^{-1}x^T$.
- y^* will have a singular multivariate Gaussian distribution with mean zero.
- To obtain a non-singular distribution, we can use only $n - p$ rows (does not matter which rows) of A .

ML vs. REML

- Note that we cannot use the likelihood ratio test for inference regarding β anymore if we use REML estimates since the commonly used distributional assumption for this test does not hold.
- We could however use the Wald test.

Statistical inference for the “radon” data

- Random intercept model without covariates

```
> MO <- lmer(y ~ 1 + (1 | county), REML = FALSE)
> summary(MO)
```

Linear mixed model fit by maximum likelihood

Formula: $y \sim 1 + (1 \mid \text{county})$

	AIC	BIC	logLik	deviance	REMLdev
	2261	2276	-1128	2255	2259

Random effects:

Groups	Name	Variance	Std.Dev.
county	(Intercept)	0.09340	0.30562
	Residual	0.63663	0.79789

Number of obs: 919, groups: county, 85

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	1.31226	0.04856	27.02

```
> fixef(MO)
```

```
(Intercept)
  1.312260
```

```
> se.fixef(MO)
```

```
(Intercept)
  0.04856454
```

Statistical inference for the “radon” data

- Random intercept and a fixed covariate

```
> M1 <- lmer(y ~ x + (1 | county), REML = FALSE)
> summary(M1)
```

Linear mixed model fit by maximum likelihood

Formula: $y \sim x + (1 \mid \text{county})$

AIC BIC logLik deviance REMLdev

2172 2191 -1082 2164 2171

Random effects:

Groups	Name	Variance	Std.Dev.
--------	------	----------	----------

county	(Intercept)	0.10528	0.32446
--------	-------------	---------	---------

Residual		0.57027	0.75516
----------	--	---------	---------

Number of obs: 919, groups: county, 85

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	1.46115	0.05124	28.517
x	-0.69263	0.07036	-9.844

Correlation of Fixed Effects:

(Intr)

x -0.290

```
> fixef(M1)
```

(Intercept)	x
1.4611543	-0.6926333

```
> se.fixef(M1)
```

(Intercept)	x
0.05123725	0.07035995

Statistical inference for the “radon” data

- Random intercept and a random covariate

```
> M2 <- lmer(y ~ x + (x | county), REML = FALSE)
> summary(M2)
```

Linear mixed model fit by maximum likelihood

Formula: $y \sim x + (x \mid \text{county})$

AIC	BIC	logLik	deviance	REMLdev
-----	-----	--------	----------	---------

2173	2202	-1081	2161	2168
------	------	-------	------	------

Random effects:

Groups	Name	Variance	Std.Dev.	Corr
county	(Intercept)	0.11852	0.34427	
	x	0.10776	0.32828	-0.339
Residual		0.55702	0.74634	

Number of obs: 919, groups: county, 85

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	1.46244	0.05346	27.354
x	-0.68246	0.08635	-7.903

Correlation of Fixed Effects:

(Intr)

x -0.382

```
> fixef(M2)
```

(Intercept)	x
1.4624361	-0.6824609

```
> se.fixef(M2)
```

(Intercept)	x
0.05346374	0.08635283