S-043/Stat-151
Analysis for Clustered and Longitudinal Data
(Multilevel & Longitudinal Models)

Unit 2, Lecture 5: Intuition for Maximum Likelihood Estimation plus Standard Errors and Confidence Intervals

Instructor: Prof. Luke W. Miratrix
lmiratrix@g.harvard.edu
Larsen 603

PARTICIPATE

Ask questions
Answer questions
Take risks

More generally: Participation grade -- self assessment at end of the term, ideally with something like evidence.

Quiet questions are not keeping our TFs busy

Lecture Goals

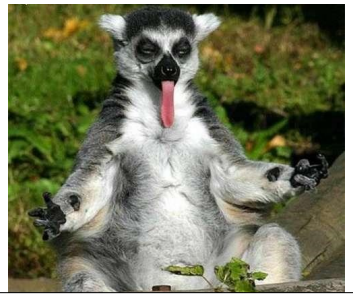
Part I:
Learn about Within vs Between Centering

Part II:
Introduce the concept of maximum likelihood estimation

Part III:
Give a taxonomy of things we might estimate with a MLM.

Discuss how to generate confidence intervals for some of these things.

Within vs Between Variation (which leads to more centering)



Dataset: Birthweight and smoking

Multiple births from same mothers

Outcome: birthweight

Primary covariate: smoking during pregnancy

Other covariates: education, married, age, history of pre-natal care, etc.

See RH&S Chapter 3

Remarks:

Some mothers always (or never) smoke, while others change their behavior between pregnancies

Our Core Issue

We want to regress weight onto smoking, controlling for other effects (heading towards a causal argument)

But it is “unrealistic to assume that birthweights of children born to the same mother are uncorrelated given the observed covariates.”

What do we do?

(RH&S 3.3)

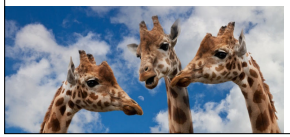
Potential comparisons

Between Mother:

Take two mothers with the same covariates, but one who smokes and one who does not. What is the expected difference in birthweight of their babies?

Within Mother:

Take a mother. What is the expected difference in birthweight between her babies from when she was smoking vs. from when she was not smoking?



Turn and talk to your neighbors: Come up with a concrete reason/ illustration/thought experiment to show why each of these could give biased (confounded) estimates.

Estimating between (only)

Collapse data to the group level

- ★For each mother, calculate average birthweight of her babies, average smoking status, average everything else
- ★Regress average birthweight onto these averages

Estimating within (only)

Take out between variation and then regress

- ★Add a fixed-effect (completely unpooled) intercept for each mother.
- ★Then all between mother variation is taken up by these unrestricted fixed effects
- ★Any other variation explained by covariates must be within-mother variation
- ★Question: What about the mothers with no variation in smoking?

Alternative: ***recentering***

- ★Subtract mother-level means from outcomes and all covariates (i.e., subtract the between-mother model from the overall model). See RH&S pg 145

Estimating Both with MLM!

We allow for different within and between effects. So, for birth *i* of mother *j:*

$$y_{ij} = \beta_1 + \beta^W(s_{ij} - \bar{s}_{\cdot j}) + \beta^B \bar{s}_{\cdot j} + \xi_j + \epsilon_{ij}$$

with s_{ij} smoking status, and $\bar{s}_{\cdot j}$ average smoking status for mother *j*.

Remarks:

- ★Our recentered variable is *not correlated* with our random effect by construction.
- ★We may want to recenter all our level-1 covariates so none of them are correlated, so as to improve validity of inference for our target β^W.

RH&S pg145
"Table 3.2"

Rand Int. (w/ means)

	Random effects Rand Int. (Simple) $\hat{\beta}^{ML}$		Between effects OLS (aggregated) $\hat{\beta}^{B}$		Within effects FE for mothers $\hat{\beta}^{W}$		Random effects +clust. mean $\hat{\beta}^{ML}$	
	Est	(SE)	Est	(SE)	Est	(SE)	Est	(SE)
Fixed part								
β_1 [_cons]	3,117	(41)	3,241	(46)	2,768	(86)	3,238	(46)
β_2 [smoke]	−218	(18)	−286	(23)	−105	(29)	−105	(29)
β_3 [male]	121	(10)	105	(19)	126	(11)	126	(11)
β_4 [mage]	8	(1)	4	(2)	23	(3)	23	(3)
β_5 [hsgrad]	57	(25)	59	(26)			56	(25)
β_6 [somecoll]	81	(27)	85	(28)			83	(28)
β_7 [collgrad]	91	(28)	100	(29)			98	(29)
β_8 [married]	50	(26)	42	(26)			42	(26)
β_9 [black]	−211	(28)	−218	(29)			−219	(28)
β_{15} [m_smok]							−183	(37)
β_{16} [m_male]							20	(32)
β_{21} [m_pretri3]							96	(117)
Random part								
$\sqrt{\psi}$	339				440		338	
$\sqrt{\theta}$	371				369		369	

Some output truncated for space

These are not parameter estimates, but sd of raw residuals and Yes

Aside: Violations of level 2 endogeneity of level 1 covariates

Recall the assumption: $Cov(X_{qij}, u_{q'j}) = 0$

Level 1 covariates Level 2 random effects

Fixed Effect Regression:

- ★All level 2 covariates (unobserved and observed) are implicitly included.
- ★Everything is within-cluster variation

Group mean centering & inclusion of cluster mean covariates

- ★Deviations from cluster means are automatically uncorrelated with cluster means and random offsets
- ★So this makes our core assumption true by construction!

Back to our friend, HS&B

Math achievement and SES, again

First, our model:

mathach ~ ses + meanses + (1|id)

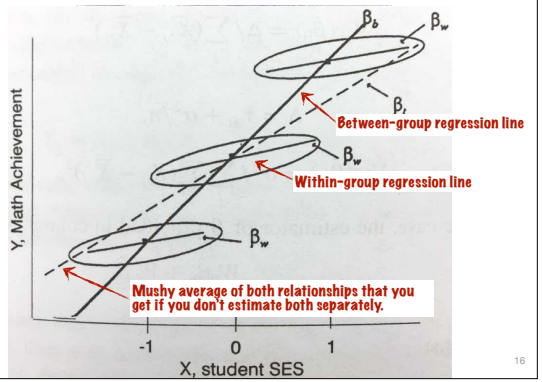
ses is an *individual-level* covariate

meanses is a *school-level* covariate

Questions:

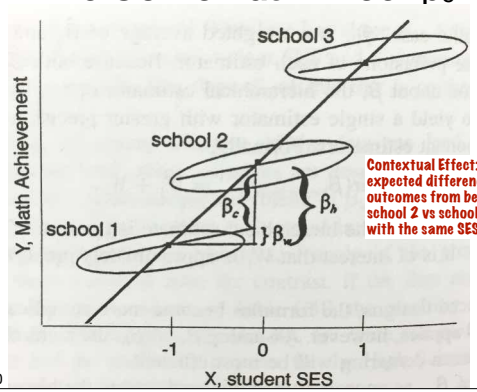
- ★What model are we estimating?
- ★What coefficients will we get?

Within vs Between Relationships



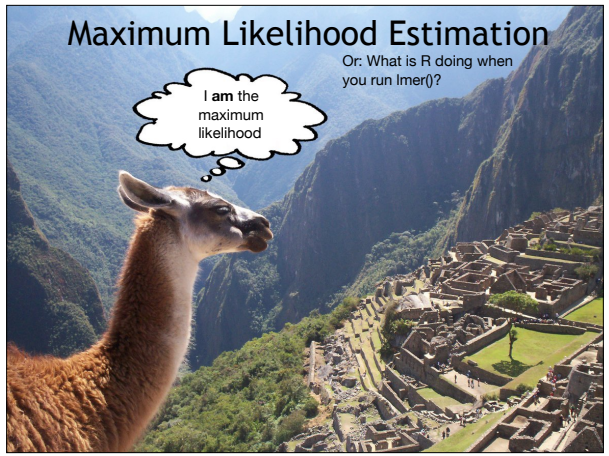
R&B pg 137

The **Contextual Effect** β_c



R&B pg 140

Maximum Likelihood Estimation



What does "fitting a model" mean?

The **model**: A description or recipe of how the data might have come to be (a data-generating process). The model tells us what predictions are possible.

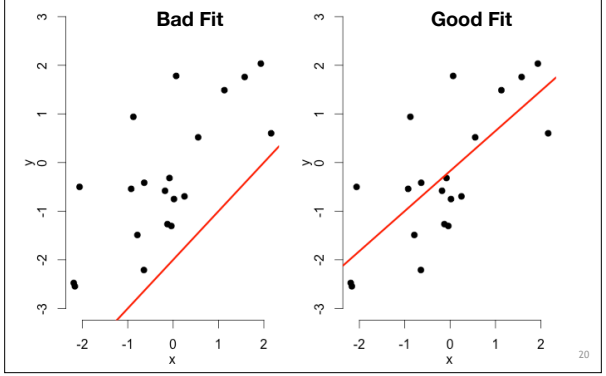
The **parameters**: Numbers that make this general recipe more specific.

The **data:** What you actually see.

Given **data** and a **model**, we "fit a model" by finding the parameters that **best fit** the data given the assumption of the model.

When doing (OLS) linear regression, we identify the **best fit** as the set of parameters which minimize the variance of the residual

One Classic Picture of "Fitting": OLS minimizes the variance of the residuals



Another is "Likelihood of Data"

Say you have the following data:

$$Y = [1,\ 2,\ 1,\ -3,\ -2,\ 1]$$

What parameters for your model make this data MOST LIKELY?

Your model is

$$Y_i \sim N(\mu, \sigma^2)$$

with all of the Y_i independent of each other

Q1: What are the parameters of our model?

Q2: How can you get reasonable parameter estimates?

The MLE approach

Given any set of parameters, we can calculate the probability/likelihood of observing the particular sample that we happened to take (this is not exactly right, but is pretty close).

Define "best fit" as the set of parameters which make the observed data as likely as possible (i.e., "maximum likelihood").

Gaussian Densities (Likelihoods)

The Density of a Gaussian:

$$p(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$$

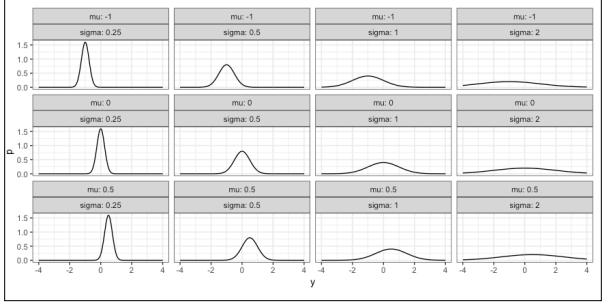
The log of the density:

$$\log p(y|\mu,\sigma^2) = -\frac{1}{2\sigma^2}(y-\mu)^2 - \frac{1}{2}\log(2\pi\sigma^2)$$

The exponential went away!

Some densities given our parameters

$$p(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$$



Our overall Log Likelihood

$$\ell(\mu, \sigma^2; Y) = \sum_{i=1}^{n} \log p(Y_i|\mu, \sigma^2)$$

$$= \sum_{i=1}^{n} -\frac{1}{2\sigma^2}(Y_i - \mu)^2 - \frac{1}{2}\log(2\pi\sigma^2)$$

$$= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mu)^2 + \frac{n}{2}\log(2\pi\sigma^2)$$

This is your classic sum of square error from "least squares" view of OLS

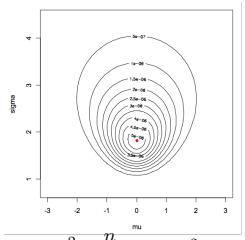
We can calculate our likelihood for any values of our parameters (and the data)

Finding the maximum of our likelihood function (given our data)

MLE is easy with models like ours. We can easily calculate the likelihood for each μ and σ^2.

The bands on this contour plot show us curves of equal likelihood.

It's easy to find the coordinates of the top of this likelihood surface (the red dot)!



Our Data: $Y = [1,\ 2,\ 1,\ -3,\ -2,\ 1]$

Our Model: $Y_i \sim N(\mu, \sigma^2)$

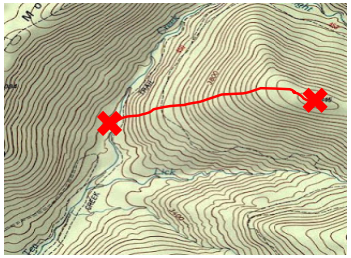
Our Likelihood: $\ell(\mu, \sigma^2; Y) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mu)^2 - \frac{n}{2}\log(2\pi\sigma^2)$

An analogy

You need to get to the top of a hill, but it's misty out.

You

1. Feel around to see which way is up.
2. If there is no up, you are at the top. Stop.
3. If there is, walk in that direction a few paces.
4. Repeat until you reach the top.



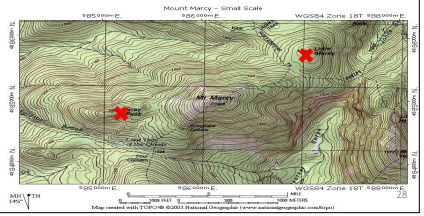
What are some pitfalls?

You may get stuck in a local maximum

It can be a long, long trek up the hill

In a few cases, there is no maximum likelihood

Sometimes we need more complex approaches



The take-away intuition

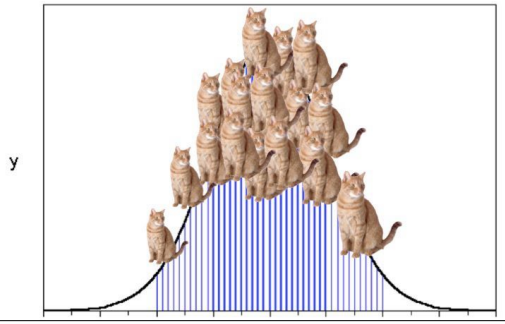
lm() or lmer() are "fitting a model"

They take your data, and a statement of a model, and determine estimates for your parameters that give the ***highest probability of generating your data***

This model, which you store in a variable, can then be examined in various ways (e.g., via fixef(), ranef(), etc.)

29

Standard errors and normal assumption-based confidence intervals



30

Three Things We Estimate

Our MLM give us three types of things:

- ★Estimated fixed effects
- ★Estimates of the variances and covariances of the hyperparameters

With Empirical Bayes *as a second step* we can obtain

- ★Estimates of the random effects for each group.

We have seen how to get **point estimates.**

We also need **uncertainty** for those estimates.

The Normal Approximation

Maximum Likelihood Estimation provides some nice things:

- ★**Consistency**
  (as the sample grows, uncertainty shrinks towards 0)
- ★**Asymptotically Normally** distributed point estimates
  (in big samples, point estimates follow a normal distribution across repeated random samples)
- ★Approximate (asymptotically correct) **standard errors**
  (as the sample grows, our estimates of the standard errors will approach the correct ones)

I.e., roughly speaking, for any estimand β we have:

$$\sqrt{n}\left(\hat{\beta}-\beta\right) \rightarrow N(0,\tau^2)$$

Or "as the sample grows, the estimated parameters will be normally distributed about the true parameters, and the variance will shrink with the sample size towards an estimable value."

Standard Errors and MLE

We're not going to show this in detail, but we can get standard errors based on the curvature of the likelihood function evaluated at the MLE estimate

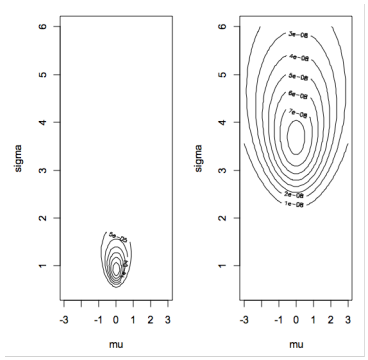
If there's **lots of curvature** (the steepness of the hill is changing quickly near the top), **our estimates are more precise**.

- ★Only a small set of parameter values have good values, so we can easily locate what parameters are best.

If there's **not much curvature** (the hill is really flat near the top), our estimates are **less precise**.

- ★Lots of parameter values are all nearly as good as the top and we can't distinguish between then.

A steep likelihood and a flat likelihood



The 95% Confidence Interval

If the sampling distribution is relatively symmetric and bell-shaped, a 95% confidence interval can be estimated using

$$statistic \pm 2 \times \widehat{SE}$$

Sample Language:

"We are 95% confident that the true proportion of all Americans that considered the economy a 'top priority' in January 2012 is between 0.84 and 0.88"

Why I'm ok with these approximate confidence intervals

1) We are making a lot of modeling assumptions

2) At the end of the day, our estimators usually follow a rough normal distribution

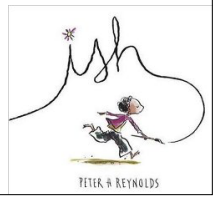
$$\hat{\theta} \sim N(\theta, SE^2) \textit{ ish}$$

3) We don't even have this. We instead have

$$\hat{\theta} \overset{?}{\sim} N(\theta, \widehat{SE}^2)$$

4) So just use for a rough measure:

$$\theta \in \hat{\theta} \pm 2\widehat{SE}$$



Standard Errors and Confidence Intervals for the Fixed Effects

"Fixed Effects": se.fixef()

From the "arm package

```
> M1 = lmer( mathach ~ ses + (ses|id), data=dat )

> fixef( M1 )
(Intercept)         ses
      12.67        2.39

> se.fixef( M1 )
(Intercept)         ses
      0.190       0.118

> # lower and upper bounds of a 95% confidence interval
(Normal Approximation)
> fixef( M1 ) - 2 * se.fixef( M1 )
(Intercept)         ses
      12.29        2.16
> fixef( M1 ) + 2 * se.fixef( M1 )
(Intercept)         ses
      13.04        2.63
```

Or use critical values from a t-distribution with df calculated by the lmerTest package; may be slightly better, especially if there are few level-2 units.

Confidence Intervals for variance and covariance estimates

An unfortunate hiccup: Standard Errors for Hyperparameters

We can get our estimates:

```
> VarCorr( M1 )
 Groups   Name        Std.Dev. Corr
 id       (Intercept) 2.197
          ses         0.643    -0.11
 Residual             6.069
```

But getting uncertainty on these estimates is **hard**, and what you get is **unreliable.**

Generally avoid doing it.

(We will test against them being 0 later on.)

Why are there no standard errors for the variance parameters?

From the creator of the lme4 package:
a) "because they are awkward to calculate"
and
b) "because I don't think they make sense"

He then goes on to say:
"Quoting an estimate and a standard error of the estimate for a parameter is useful if you think that the distribution of the estimator is more-or-less symmetric."

Variance parameters tend to have skewed distributions, with lots of samples having estimates which are slightly below the true value, and a handful having estimates which are far above.

Related issue: bad boundaries

For the variance parameters you will often see statements such as

"parameter estimates are often on the boundary of the parameter space."

What does this mean?

Why does this matter for our variance estimation issues?

Boundaries distort normality

If a variance *parameter* is equal (or close) to 0, the distribution of the variance *estimate* can't be normal; a normal distribution is symmetric and an estimated variance parameter can't be negative.

This might not seem like a big deal, but it makes standard errors much less meaningful.

When this happens

$$\hat{\sigma}^2 \pm 2\widehat{SE}(\hat{\sigma}^2)$$

← Bad when we are close to zero

is not a valid approach here; we'll get meaningless confidence intervals, and they won't have good coverage rates.

But we can get confidence intervals

```
> display( M2 )
lmer(formula = mathach ~ 1 + ses + sector + (1 | id),
data = dat)
            coef.est coef.se
(Intercept) 11.72     0.23
ses          2.37     0.11
sector       2.10     0.34
```

```
> confint( M2 )

               2.5 % 97.5 %
.sig01         1.653  2.194
.sigma         5.986  6.188
(Intercept) 11.272 12.165
ses          2.167  2.588
sector       1.433  2.770
```

These are *profile likelihood based confidence intervals.* We will see where these come from later on.

Recap

Recap

Check-In
http://cs179.org/lec25

Part I:
Within vs. between is a way of identifying whether a covariate is associated with outcome due to grouping of individuals by that covariate, or a more direct link

This is a slippery concept, and it takes careful thinking to get it right. Examples are helpful.

Part II:
Models are estimated with Maximum Likelihood, which asks "what parameter values make my data most likely?"

Part III:
Confidence intervals can be generated by doubling the standard error (or doing fancy stuff—the profile confidence interval—that is a black box).

Appendix
Fitting within vs. between for HS&B Example

Fitting our model with between and within effects

```
> dat = dat %>% group_by( id ) %>%
      mutate( meanses = mean( ses ) )

> M2 = lmer( mathach ~ ses + meanses + (1|id), data=dat )
> display( M2 )

(Intercept) 12.66     0.15
ses          2.19     0.11
meanses      3.68     0.38

Error terms:
 Groups   Name        Std.Dev.
 id       (Intercept) 1.64
 Residual             6.08
---
number of obs: 7185, groups: id, 160
```

What would you predict average math achievement to be for a school with an average SES 0.5 above average?
What about for a student with 0 SES in such a school?

Alternative fitting (note group recentering)

```
> dat = dat %>% group_by( id ) %>%
        mutate( ses.cent = ses - mean( ses ) )
> M3 = lmer( mathach ~ ses.cent + meanses + (1|id),
                       data=dat )

> display( M3 )
            coef.est coef.se
(Intercept) 12.68     0.15
ses.cent     2.19     0.11
meanses      5.87     0.36

Error terms:
 Groups   Name        Std.Dev.
 id       (Intercept) 1.64
 Residual             6.08
---
number of obs: 7185, groups: id, 160
AIC = 46578.6, DIC = 46559
deviance = 46563.8
```

Consider our prior questions from the earlier model. Easier to answer?

Contextual vs. Overall Effects

```
> stargazer( M2, M3, type = "text" )

================================================
                         Dependent variable:
                     ---------------------------
                               mathach
                          (1)            (2)
------------------------------------------------
ses                     2.191***
                        (0.109)

ses.cent                                2.191***
                                        (0.109)

meanses                 3.675***       5.866***
                        (0.378)        (0.362)

Constant               12.680***      12.680***
                        (0.149)        (0.149)

------------------------------------------------
Observations             7,185          7,185
Log Likelihood       -23,284.000    -23,284.000
Akaike Inf. Crit.     46,579.000     46,579.000
```

grand mean cent **Group centered**

Slope is same for both models

Note: 5.866 - 2.191 = 3.675

Aside: You can't include Level-2 variables in fixed effect models

Consider doing this the "fixed effects" way:

```
a = lm( mathach ~ ses + meanses + id,
                                  data=dat )
```

This will **fail** due to **collinearity.**

When covariates are collinear, we cannot get unique estimates for their coefficients in the linear model.

G&H Pg 269