Lecture 4.2 (A)
Crossed Random Effects Models

&

Lecture 4.2 (B)
AIC and Model Selection/Building

Today's Goals

Part A:

Talk about crossed-effect models, where you have a random effect for each of two different groups (e.g., school and neighborhood, if you have kids from different neighborhoods going to different schools)

Part B:

Talk about using AIC/BIC to select from a set of competing models.

Part A
Non-nested
(cross classified) models

Reading: Raudenbush & Bryk Chapter 12

Cross Classified Models

Cross-classified models arise when we have multiple nesting structures, e.g., when each observation can be classified into multiple different level-2 units (where there's no natural hierarchy of those higher order units).

E.g.:

We have a random effect for year

and

We have a random effect for firm (company)

Other cross-classified examples

1. Students are nested within both schools and neighborhoods; students from different neighborhoods attend the same school and students from the same school come from different neighborhoods
2. Students are nested within multiple different classes; if we follow students through elementary school, each student will move through a different sequence of classes, so class membership might be thought of as a cross-classificatory system
3. Responses to individual items on a test can be classified according to the student who is giving the response and the item which is being responded to; that's what we'll look at in this example

Stocks and Bonds

Grunfeld data (RH&S, pg 434)

fn: firm identifier

firmname: name of firm

yr: year

I: annual gross investment (our outcome)

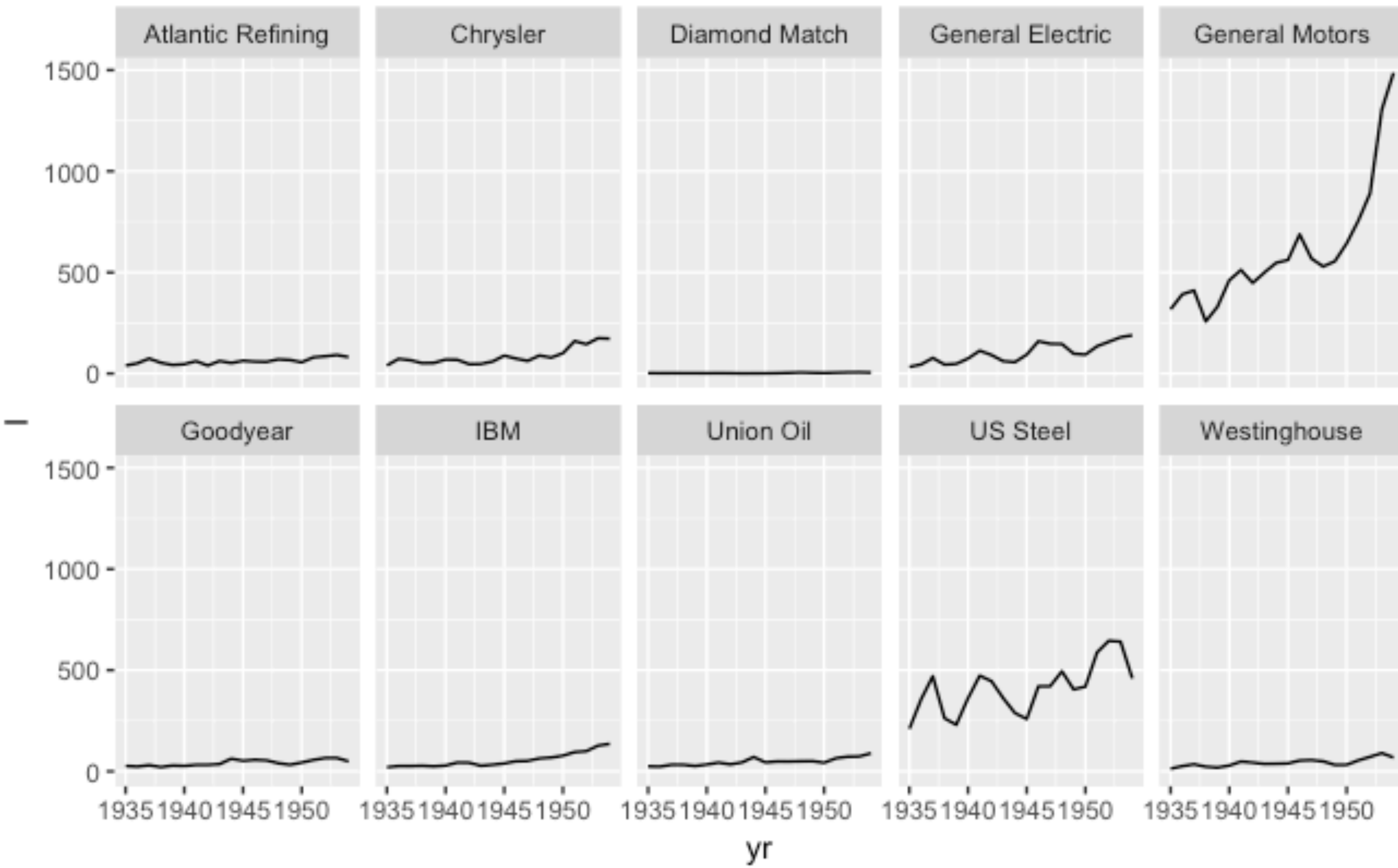
F: market value of firm at beginning of year (which can be thought of as firm's *expected value* at this time)

C: real value of stock

Question: How does investment depend on expected profit and capital stock?

Ok, so this is not the most education-related example. But you need money to fund ed, right?

The Data


Atlantic Refining
Chrysler
Diamond Match
General Electric
General Motors
Goodyear
IBM
Union Oil
US Steel
Westinghouse
1500
1000
500
0
1935
1940
1945
1950
yr
I

Modeling Stocks and Bonds

Consider this model for firm *f* in year *y*

These are fixed effects

Three components to our error

$$I_{fy} = \beta_0 + \beta_1 F_{fy} + \beta_2 C_{fy} + \alpha_f + \gamma_y + \epsilon_{fy}$$

$$\epsilon_{fy} \sim N(0, \sigma^2)$$

$$\alpha_f \sim N(0, \sigma_f^2)$$

$$\gamma_y \sim N(0, \sigma_y^2)$$

We now have TWO sets of random effects,
one for the "firm effect"
one for the "year effect"

Q: If we fit it, how many parameters will be estimated?

Thinking about the model

$$I_{fy} = \beta_0 + \beta_1 F_{fy} + \beta_2 C_{fy} + \alpha_f + \gamma_y + \epsilon_{fy}$$

$$\epsilon_{fy} \sim N(0, \sigma^2)$$

$$\alpha_f \sim N(0, \sigma_f^2)$$

$$\gamma_y \sim N(0, \sigma_y^2)$$

Our model says:

Some years have generally more investment than others.

Some firms have generally more investment than others.

Our residuals are idiosyncratic investments above or below expected for a specific firm in a specific year

β_0 The average investment across firms and years

β_1 How much the market value at the beginning of year predicts investment.

β_2 How much the real value of the firm predicts investment.

Aside: Alternate, clean, indexing

$$I_i = \beta_0 + \alpha_{f[i]} + \gamma_{y[i]} + \beta_1 F_i + \beta_2 C_i + \epsilon_i$$

$$\alpha_f \sim \mathcal{N}(0, \sigma_f^2)$$

$$\gamma_y \sim \mathcal{N}(0, \sigma_y^2)$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma_I^2)$$

Note the [] indexing is back. This is a nice way to capture cross-classified models since we don't have hierarchy.

The y[i] says "what year is observation i in?"

This is the "Gelman & Hill" approach

Fitting the model

```
M1 = lmer( I ~ F + C + (1|yr) + (1|firmname),
                      data=grun )

Warning messages:
1: Some predictor variables are on very different scales:
consider rescaling
2: Some predictor variables are on very different scales:
consider rescaling

(Intercept) -58.84    29.51
F             0.11     0.01
C             0.31     0.02

Error terms:
 Groups   Name        Std.Dev.
 yr       (Intercept)  5.40
 firmname (Intercept) 86.07
 Residual             52.47
---
number of obs: 200, groups: yr, 20; firmname, 10
AIC = 2207.8, DIC = 2185.2
deviance = 2190.5
```

Uh-oh.
What do we do about this?

Rescaling the variables

```
> grun = mutate( grun, F.stand = scale( F ),
                       C.stand = scale( C ) )

> M1.v2 = lmer( I ~ F.stand + C.stand +
                  (1|yr) + (1|firmname), data=grun )
> display( M1.v2 )
            coef.est coef.se
(Intercept) 145.96   27.50
F.stand     144.68   13.94
C.stand      93.53    5.25

Error terms:
 Groups   Name        Std.Dev.
 yr       (Intercept)  5.40
 firmname (Intercept) 86.07
 Residual             52.47
---
number of obs: 200, groups: yr, 20; firmname, 10
AIC = 2182, DIC = 2211
deviance = 2190.5
```

Getting the random effects

```
> ranefs = ranef( M1.v2 )
> names( ranefs )
[1] "yr"        "firmname"

> ranefs$yr
     (Intercept)
1935  3.29100755
1936  1.77877330
1937  0.04989363
1938 -0.01463782
1939 -2.66755903


> ranefs$firmname
                    (Intercept)
Atlantic Refining   -55.019629
Chrysler             30.434183
Diamond Match        51.316634
General Electric   -173.622137
General Motors      -11.362805
```

Calculating the ICCs

```
> M1.v2 = lmer( I ~ F.stand + C.stand +
                   (1|yr) + (1|firmname), data=grun )
> display( M1.v2 )
            coef.est coef.se
(Intercept) 145.96    27.50
F.stand     144.68    13.94
C.stand      93.53     5.25

Error terms:
 Groups   Name        Std.Dev.
 yr       (Intercept)  5.40
 firmname (Intercept) 86.07
 Residual             52.47
---
number of obs: 200, groups: yr, 20; firmname, 10
AIC = 2182, DIC = 2211
deviance = 2190.5
```

Calculate year ICC to 2 decimal places

Part B
AIC and model search

Further reading on Model Building: R&B Chapter 9, pgs 252-276

First the reading covers the assumptions behind the models.

Then the reading discuses model building as a staged process:

- ★ Building the level 1 model
- ★ Building the level 2 model
- ★ Etc.

Much advice on thinking this through, and the possible consequences of misspecification.

Options for Model Selection

Likelihood Ratio Tests

- ★ Compare a simpler model to a more complex model
- ★ Requires models to be *nested*

Model inspection/evaluation

- ★ Check *model fit* and reject models that do not capture core structure of data.
- ★ To evaluate, compare results to a more unrestricted model: if results are very different, one should worry

AIC/BIC/etc

- ★ Compare a family of models to each other.
- ★ Evaluates **trade-off** of model fit (evaluated by likelihood) and complexity (evaluated by number of parameters)
- ★ Can compare non-nested models.

Complexity vs. Model Fit

As you add more parameters to your model:

- ★ The model is **more flexible,** even if the parameters are meaningless/not truly related to the data
- ★ The ability to find a good fit is **improved**, meaning
  - your **likelihood will get better**
  - your residuals will decrease
  - your R^2 will increase
- ★ But the model is **more complicated**

The Tradeoff

For each element of complexity, is the model fit better than the “cost” of that additional complexity?

Akaike Information Criterion (AIC)

"a measure of the relative quality of statistical models" - Wikepedia!

$$AIC = -2\ln(L) + 2k$$

Log Likelihood

parameters (a penalty)

AIC is a tradeoff:

- ★ More complicated models fit data better -- so the log likelihood is lower.
- ★ The "penalty" of $2k$ offsets this expected gain in fit due to a more flexible model.

This allows us to compare models with different numbers of parameters

AIC is easy to get for a model

```
> M2 = lmer( mathach ~ ses + meanses + (1|id), data=dat
)
> display( M2 )
lmer(formula = mathach ~ ses + meanses + (1 | id), data
= dat)
            coef.est coef.se
(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
AIC = 46578.6, DIC = 46559
deviance = 46563.8
```

AIC is a measure of model fit balanced against model complexity

AIC’s implementation & promises

The game:

- ★ Take a collection of possible models.
- ★ Calculate AIC for all the models.
- ★ Pick the model with the **lowest** (best) AIC

Notes:

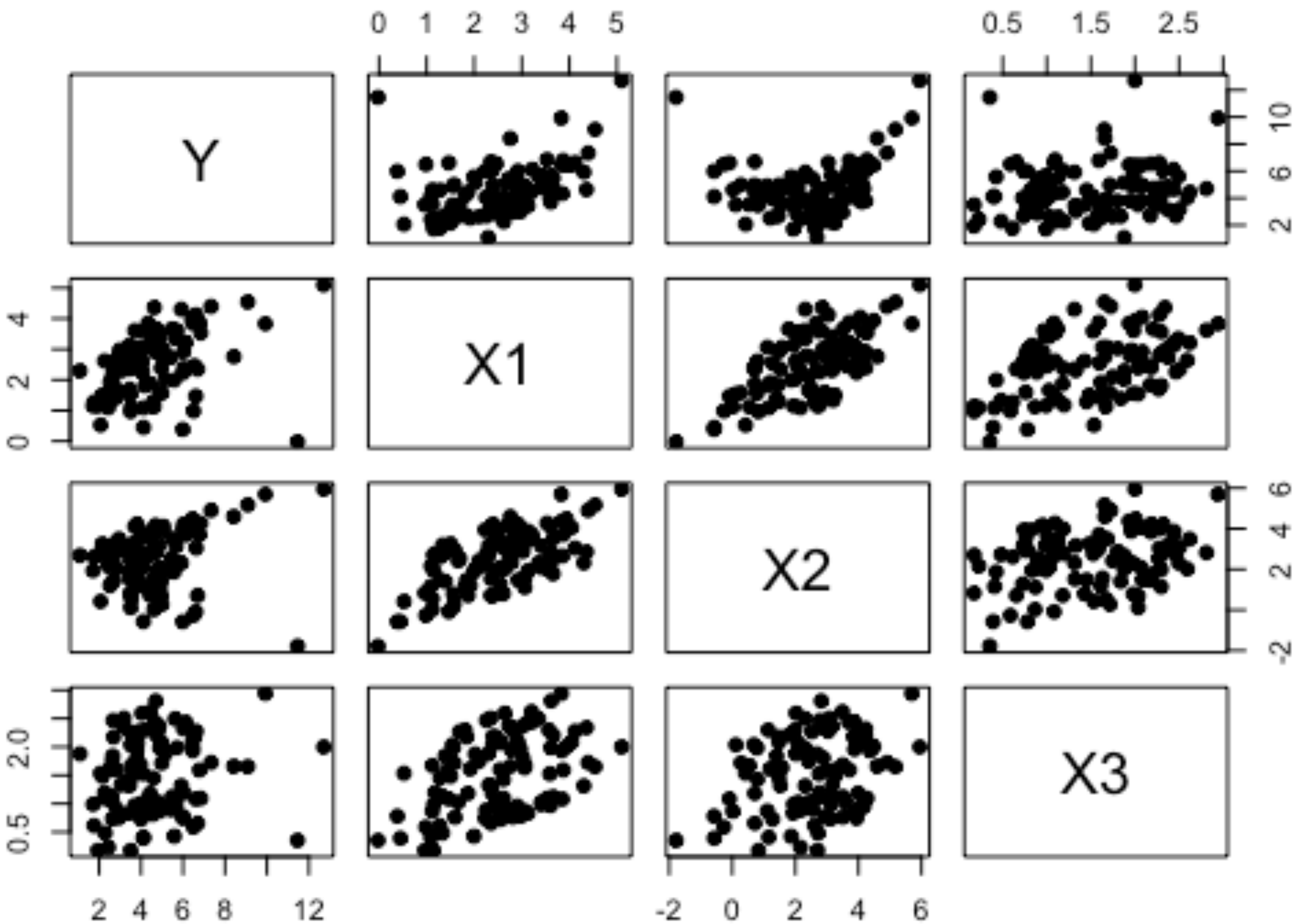
- ★ Asymptotically, AIC will give you the “right” model, if there is a true “right” model.
- ★ “Mallows's Cp” is AIC in the OLS case.

Cautions:

- ★ The AIC values of the candidate models must all be computed with the same data set.
- ★ You cannot compare transformed models with transformed outcomes to non-transformed, without some extra work.

Toy example: Three Variables

Consider three variables that are all potentially related to Y (and each other)



If you plot a data frame you get all pairs of variables.

plot(df)

Some possible linear models

call	r.squared	adj.r.squared	df
Y ~ X1	0.18	0.17	2
Y ~ X2	0.05	0.04	2
Y ~ X1 + X2	0.18	0.17	3
Y ~ X1 + X2 + I(X2^2)	0.75	0.74	4
Y ~ X1 + I(X1^2) + X2 + I(X2^2)	0.75	0.74	5
Y ~ X2 + I(X1^2)	0.30	0.29	3
Y ~ X1 + I(X2^2)	0.27	0.25	3
Y ~ X1 + X2 + X3	0.19	0.16	4

call	logLik	AIC	BIC
Y ~ X1	-199.0050	404.0101	411.8256
Y ~ X2	-206.3778	418.7556	426.5711
Y ~ X1 + X2	-198.5588	405.1176	415.5383
Y ~ X1 + X2 + I(X2^2)	-139.5070	289.0140	302.0398
Y ~ X1 + I(X1^2) + X2 + I(X2^2)	-139.4894	290.9788	306.6098
Y ~ X2 + I(X1^2)	-190.8022	389.6044	400.0251
Y ~ X1 + I(X2^2)	-193.1599	394.3197	404.7404
Y ~ X1 + X2 + X3	-198.4825	406.9649	419.9908

ANOVA/lr testing is nonsensical

```
> anova( M1, M2, M4, M8 )

Model 1: Y ~ X1
Model 2: Y ~ X2 + X3
Model 3: Y ~ X1 + X2 + I(X2^2)
Model 4: Y ~ X1 + X2 + X3
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     98 313.37
2     97 359.95  1   -46.577
3     96  95.34  1   264.612 266.45 < 2.2e-16 ***
4     96 310.11  0  -214.776
```

This is all invalid since the models are not all nested!!!

Other versions of this game

$$AIC = -2\log(p(y|\hat{\theta})) + 2k$$

$$BIC = -2\log(p(y|\hat{\theta})) + k\log(n)$$

$$DIC = -2\log(p(y|\hat{\theta})) + k_D, \text{ where}$$

$$k_D = E_{\theta|y}[-2\log(p(y|))] + 2\log(p(y|\hat{\theta}))$$

For observed data *y*, parameter θ, MLE θ-hat, number of model parameters *k*

For DIC, one can think of the effective number of parameters, k_D, as the posterior mean deviance minus the deviance measured at the posterior mean of the parameters.

Sample size for BIC with MLMs?

In MLMs, it is not clear what sample size n should be used to calculate BIC. (The number of units at the lowest level? The number of units at the highest level? Some weighted average?)

R and SPSS use the level 1 sample size to calculate BIC.

SAS PROC MIXED uses the number of independent sampling units as n .

MLMs & AIC/BIC/DIC

In MLMs, the number of independent parameters included in the model is difficult to determine, which makes the use of AIC and BIC problematic.

The Deviance Information Criterion (developed by David Spiegelhalter et al. in 2002) was meant to address this, but in 2012, Spiegelhalter et al. published another article admitting the many flaws of DIC and hinted that DIC maybe isn't such a great model selection criterion to use after all.

So what do we do?

AIC will get the gross ordering of good vs. bad models right.

Specifically testing whether components of a model are needed with likelihood ratio tests is good.

Looking at plots (e.g., do your growth curves follow the data) is also very important.

Recap

Recap

http://cs179.org/lec42

Part A:

- ★ We can have random effects for different things.
- ★ We just add them all to our model, and it works!

Part B:

- ★ We can score model fit using various measures.
- ★ We can then take the best model given the score!
- ★ Scores can differ, which gives us some ambiguities…