Mixed Effects Models

NRES 710

Last compiled: 2024-11-12

Any questions from last class?

## Review

Last class I introduced the idea of random-effects and mixed-effects models. Today I am going to go into more detail about the mixed-effects models.

## Note on random-effects models

You may have noticed that last class I didn’t spend too much time talking about the ‘random-effects model’ – the model where we just have one beta (intercept), with two error terms, and no X-variables. The reason why I didn’t talk further about this is because we don’t use this simple model too often in my field of wildlife ecology. We are often use random-effects to control for variation, then move to add fixed effects to test hypotheses using mixed-effects models.

However, in other fields, random-effects models are used more commonly. One example is genetics. For example, let’s say you have a **Y – quantitative trait**. But we don’t have any X-variables. Instead, we have multiple genes:

* **Y – quantitative trait**
* **Gene 1**
* **Gene 2**
* **Gene 3**
* …
* **Gene 20**

Each gene has many different alleles. Which of these genes is driving the response in the the quantitative trait?

This can be analyzed with a random-effects model.

We can look at each error term and see which genes are contributing the most to the response variable.

So basically, random-effects models are used when you are trying to partition the variance. The is variation in Y, and you want to partition it among different causal factors – in this example, genes. If you find that, say, 95% of variation is driven by Gene 1, then you could put those alleles into your model as fixed effect X-variables.

So this is one example of how I’ve seen this model applied.

\*\*Steinhaltz et al. 2018 \*Electronic Journal of Statistics\*\*

## Mixed-effects Models

We have been working with an example:

**Y – Biomass**  
 **– Treatment/control**

*Brian draw 1, 2, 3, … , 15 plots on the board with* ***Treatment*** *and* ***Control****.*

Each block got both treatments; this is called a **Randomized Block Design** and it is a very powerful design. We could analyze this with a simple t-test, such as:

But this would not account for variation within blocks, which could cause swamping, and cause us to lose a ton of power! The best model would be to include a random field effect:

This is also called a ‘paired t-test’!

In the particular dataset we are worked with last class, the true values for the above equation are:

* = 25
* = 10
* = 5
* = 1

One thing I want you to remember is that when you are working with mixed-effects models, you don’t necessarily have to include a random-effect in your model. But if you do not, you lose power: more swamping, larger confidence intervals, larger p-values, less power! All consequences of not using the random-effect to account for this important source of variation.

Another thing I want you to understand is that: in order to include a random effect, each one of those rivers, fields, mountains, forests, etc. **has to have been measured more than once**. For example, let’s say I have 10 fields:

*Brian draw 5 fields on the board as an example, then randomly assign them as T or C but without within-field replication*

I can’t include field as a random effect, because I have only measured each field once. There is no way to understand variation due to field, because I only have one sample per field, cannot measure an average for each field and deviation above and below that average for each field, and thus cannot measure variability among samples in each field with only one sample.

The easiest way to think about this is: you always have to have a column for your random effect in your data. In that column, each groups (e.g., field) has to have more than one row. For example, let’s look at our treatment field data from last class:

# Read in the treatment field data from last class  
datum <- read.csv("lecture\_15\_dataset2.csv")  
  
# Examine the data  
head(datum, 10)

If we only had one row for field, we would not be able to include it as a random effect!

What if we had unequal samples? Generally, not a problem. If we had 10 samples from field 1, 2 samples from field 2, 3 samples from field 3… that would be okay. This is called an **unbalanced design**. You can even have some fields that were only sampled once. You can get away with that, because R uses information from sites with multiple measurements to estimate the standard deviation due to field. However, there are limits. If you have 1 field measured 10 times and the rest only measured once, R will try to fit the model for you because it does what you ask – but the inferences might not be trustworthy.

### Analysis in R

Let’s go back to the analysis we were working on last class and pick up where we left off.

# Fit a t-test  
results <- lm(Biomass ~ Treatment, data = datum)  
summary(results)

##   
## Call:  
## lm(formula = Biomass ~ Treatment, data = datum)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.649 -2.169 -0.817 0.886 8.663   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 25.82 1.01 25.55 < 2e-16 \*\*\*  
## Treatment 9.93 1.43 6.94 1.5e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.91 on 28 degrees of freedom  
## Multiple R-squared: 0.633, Adjusted R-squared: 0.62   
## F-statistic: 48.2 on 1 and 28 DF, p-value: 1.5e-07

Lots of residual error here, all of our field error is being chalked up to be standard noise in our data. This is significant, so we are still learning something – but we can do better.

Let’s fit the LME model.

# Load the package  
library(nlme)  
  
# Fit the mixed-effects model  
resultsMix <- lme(Biomass ~ Treatment, data = datum, random = ~1|Field)  
summary(resultsMix)

## Linear mixed-effects model fit by REML  
## Data: datum   
##   
## Random effects:  
## Formula: ~1 | Field  
## (Intercept) Residual  
## StdDev: 3.813 0.888  
##   
## Fixed effects: Biomass ~ Treatment   
## Correlation:   
## (Intr)  
## Treatment -0.16   
##   
## Standardized Within-Group Residuals:  
## Min Q1 Med Q3 Max   
## -1.61959 -0.52825 0.00999 0.42882 1.68140   
##   
## Number of Observations: 30  
## Number of Groups: 15

Have to use ‘lme()’, and have to add in the random effect: tilde 1 blocked by field.

Looking at the results, our beta did not change! Adding an error term into the model should not effect our betas: it’s only influencing the variation in Y, not the average Y. But it does decrease the standard error around that effect.

Note: the mixed-effect model does not test whether the random field effect was significant. We can do that using a ‘partial likelihood ratio test’ (i.e., an f-drop test). This is no longer called ‘f-drop test’ because we are not using F-statistics anymore, we are using maximum likelihood. Different names – but still doing the same thing in R.

It’s now particularly important that the more complicated model goes first, otherwise you will get an error. What does it say?

# Partial likelihood ratio test  
anova(resultsMix, results)

**Q:** Is the random-effect significant?

**Yes** – the more complicated model is a significant improvement to the fit of the data. Before it gave us an F-statistic; now it gives us a ‘L.Ratio’ (likelihood ratio). This variable that we have added is a significant improvement in fit; this is a test of the significance of that random effect.

Let’s examine our mixed-effects model again:

# Re-examine the mixed-effects model  
summary(resultsMix)

## Linear mixed-effects model fit by REML  
## Data: datum   
##   
## Random effects:  
## Formula: ~1 | Field  
## (Intercept) Residual  
## StdDev: 3.813 0.888  
##   
## Fixed effects: Biomass ~ Treatment   
## Correlation:   
## (Intr)  
## Treatment -0.16   
##   
## Standardized Within-Group Residuals:  
## Min Q1 Med Q3 Max   
## -1.61959 -0.52825 0.00999 0.42882 1.68140   
##   
## Number of Observations: 30  
## Number of Groups: 15

**Q:** What if we wanted to know if the **treatment variable**, as a whole, was significant? Do we have to do an F-drop test here?

**No!** It tells us right here in the results (P=0). In the past, we have used F-drop tests to look at the significance of variables as a ‘whole’ when they were categorical variables with more 2 groups. For example, if we had three treatment groups – control, treatment 1, treatment 2 – just looking at this output would **not** necessarily tell us if there is a significant difference because it doesn’t show us differences for all pairwise comparisons. So, in general, if you want to know if a multi-category variable is significant, you need to do an F-drop test.

But we are going to do one here now anyway because I want to show you something about F-drop tests when we use a maximum-likelihood model.

Let’s compare: the mixed-effects model, **resultsMix**, to a reduced model, **resultsRed**:

(Basically, a random-effects model)

# Reduced model (basically, a random-effects model)  
resultsRed <- lme(Biomass ~ 1, data = datum, random = ~1|Field)  
summary(resultsRed)

## Linear mixed-effects model fit by REML  
## Data: datum   
##   
## Random effects:  
## Formula: ~1 | Field  
## (Intercept) Residual  
## StdDev: 0.0003457 6.347  
##   
## Fixed effects: Biomass ~ 1   
##   
## Standardized Within-Group Residuals:  
## Min Q1 Med Q3 Max   
## -1.7107 -0.8822 0.2077 0.5843 2.1470   
##   
## Number of Observations: 30  
## Number of Groups: 15

# Now let's test whether the complex model is significantly better than the reduced model  
anova(resultsMix, resultsRed)

## Warning in anova.lme(resultsMix, resultsRed): fitted objects with different fixed  
## effects. REML comparisons are not meaningful.

# Notice at the bottom! This is a warning we should not ignore.

I ran this to show you all this warning message. ‘fitted objects with different fixed effects’. We have different x-variables in our two models. ‘REML comparisons are not meaningful’. We should not ignore this! This models are fit with **restricted maximum likelihood (REML)**. It is a hierarchical (ordered) model-fitting approach:

1. It first fits the fixed effects and calculates
2. Then it fits the random effects.
3. REML produces the best estimates of

These are calculated separately: it can’t partition out the variance until after it has fit the fixed effects. If we fit two models with REML that have different effects, our variances will be different. We don’t get a fair comparison of effects.

Instead, we want to use standard **maximum likelihood**:

1. First fits the random effects
2. Then, it fits the fixed effect

By using this approach, we guarantee the random effects are the same, and then when we compare the results between two models with different fixed effects we can make inferences about which one is better. So we want to use regular-old ‘maximum likelihood’ when we use F-drop tests for mixed-effects models.

### ‘Partial likelihood ratio test’ for models with mixed-effects models with different fixed effects

To do a **‘partial likelihood ratio test’** with two models with different fixed effects, we have to re-run the models with a different specification in R:

# Mixed model  
resultsMix <- lme(Biomass ~ Treatment, data = datum, random = ~1|Field, method = "ML")  
summary(resultsMix)

## Linear mixed-effects model fit by maximum likelihood  
## Data: datum   
##   
## Random effects:  
## Formula: ~1 | Field  
## (Intercept) Residual  
## StdDev: 3.683 0.8579  
##   
## Fixed effects: Biomass ~ Treatment   
## Correlation:   
## (Intr)  
## Treatment -0.16   
##   
## Standardized Within-Group Residuals:  
## Min Q1 Med Q3 Max   
## -1.67644 -0.54679 0.01034 0.44387 1.74041   
##   
## Number of Observations: 30  
## Number of Groups: 15

# Didn't really change much. Effects still the same. 'Linear mixed-effects model fit by maximum likelihood'.  
  
# Reduced model  
resultsRed <- lme(Biomass ~ 1, data = datum, random = ~1|Field, method = "ML")  
summary(resultsRed)

## Linear mixed-effects model fit by maximum likelihood  
## Data: datum   
##   
## Random effects:  
## Formula: ~1 | Field  
## (Intercept) Residual  
## StdDev: 0.0003037 6.24  
##   
## Fixed effects: Biomass ~ 1   
##   
## Standardized Within-Group Residuals:  
## Min Q1 Med Q3 Max   
## -1.7400 -0.8972 0.2113 0.5942 2.1837   
##   
## Number of Observations: 30  
## Number of Groups: 15

# One change here is that the standard deviation due to Field reduces to ~zero. This happens because Treatment explains a lot of variation in Biomass (P = 0), but we haven't accounted for that effect. This variance gets absorbed into the Residual Error, and then the random Field intercept variance becomes almost zero because the model fails to capture the between-field variation effectively.  
  
# Now let's test whether the complex model is significantly better than the reduced model  
anova(resultsMix, resultsRed)  
# No longer a WARNING at the bottom.

With this result in hand, we can now say: we want to use the more complex model (fixed + random effects), because there is a significant improvement in fit by adding the fixed effect; there is a significant effect of treatment.

After that, we would want to go back and re-run our best model using REML. To get the more accurate results for betas. Only use maximum likelihood for F-drop comparisons. Otherwise use REML for everything else. REML is better at estimation, particularly when sample sizes are small.

### Variation estimates for ‘random effects’

The other thing I want to talk about is: the variation estimate for the effect of field.

# Mixed model  
resultsMix <- lme(Biomass ~ Treatment, data = datum, random = ~1|Field)  
summary(resultsMix)

## Linear mixed-effects model fit by REML  
## Data: datum   
##   
## Random effects:  
## Formula: ~1 | Field  
## (Intercept) Residual  
## StdDev: 3.813 0.888  
##   
## Fixed effects: Biomass ~ Treatment   
## Correlation:   
## (Intr)  
## Treatment -0.16   
##   
## Standardized Within-Group Residuals:  
## Min Q1 Med Q3 Max   
## -1.61959 -0.52825 0.00999 0.42882 1.68140   
##   
## Number of Observations: 30  
## Number of Groups: 15

# Error for field: 3.81

One thing that’s interesting to me is that people never report the results from their random effects. They include the random effect, but they don’t report the result. But this is a real number, and it has a real meaning, and I want you to understand what that meaning is.

This number is a standard deviation in y-variation caused by field. For example:

Graphically, we have: *brian draw this on the board*

* Y – biomass
* X – treatment: control vs. treatment (beta0 = 25, beta1 = 10)

Within these differences, we have field effects: different averages for different fields. The variation due to field has a standard deviation of **3.81**. Given the rule of thumb that 95% of data lie within 2 standard deviations, we would expect most data to be within the 95% confidence interval of **+/-7.62**. This means that we observed **a 7.62 kg of biomass variation due to field** on either side of the average. That’s what that number means: a standard deviation due to variation of field; multiple that by two and we get some sense of ***how much variation your random effect is contributing to the response***.

### Example application

This could be really useful for **drugs**! What if you gave your control drug (placebo) to individuals, then two weeks later you gave them the drug itself (treatment). But everyone has to get both. Then you could include the random effect of individual, and then you could learn how much the drug *varies from individual to individual*. You could talk about the variation – but not many researchers do this. (Although Jeff Falke does.)

## Random intercept

The next thing I want to talk about is the ~1|Field element. I mentioned previously in the last class is that by including the random effect we are basically creating a random intercept. We are creating variation in the Y due to ‘field’. Using the experimental design on fields again:

*Brian draw five plots, each with Treatment and Control blocked within each plot*

Right now we are assuming the treatment effect is the same across all fields, and that fields contribute some variation due to Y. But what if there is variation in the effect of the treatment due to the field…? This would be an interaction between field and treatment.

To do this, we need more replicates within each field. In order to do this interaction, we need to have at least 2 treatment x field samples; i.e., each field has two treatment and two control samples.

**An interaction between a fixed and a random effect is a random effect.**

This error term is not a variation in Y, but rather a that influences the effect of

### Analysis in R

We can briefly look at ‘Truth’ to see how the data were analyzed.

***Examine Truth using the R-code for the lecture***

These data are mostly the same as the data from last class, as we have 15 fields with an experimental treatment and control in each field, a y-intercept of 25, a treatment effect of 10, error due to field of 5, and residual error of 1. But now (1) we have 4 samples per field (2 control, 2 treatment), and (2) there is an interaction effect between treatment and field, which has a standard deviation of 3. This means that we have variation in the field effect with a standard deviation of 3. So if the average treatment effect is 10, the treatment effect might be as low as 4 or as high as 16 in any given field!!

Let’s analyze this in R.

Read in the data

# Read in the data  
datum <- read.csv("lecture\_16\_dataset1.csv")  
  
# Examine it  
head(datum, 10)  
  
# Fit a simple linear model  
results <- lm(Biomass ~ Treatment, data = datum)  
summary(results)

##   
## Call:  
## lm(formula = Biomass ~ Treatment, data = datum)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.085 -2.618 0.347 2.386 8.021   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 25.808 0.731 35.30 <2e-16 \*\*\*  
## Treatment 9.371 1.034 9.06 1e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4 on 58 degrees of freedom  
## Multiple R-squared: 0.586, Adjusted R-squared: 0.579   
## F-statistic: 82.1 on 1 and 58 DF, p-value: 1.05e-12

This is an acceptable result. It gives us an accurate estimate of the effect on average: 9.3 plus or minus 2.06. Truth is within the confidence intervals. But, it doesn’t give us any information about variation in the treatment.

Let’s run the LME().

# Fit a linear mixed-effects model with a 'block' of field (field as random effect)  
results <- lme(Biomass ~ Treatment, data = datum, random=~1|Field)  
summary(results)

## Linear mixed-effects model fit by REML  
## Data: datum   
##   
## Random effects:  
## Formula: ~1 | Field  
## (Intercept) Residual  
## StdDev: 3.582 1.911  
##   
## Fixed effects: Biomass ~ Treatment   
## Correlation:   
## (Intr)  
## Treatment -0.25   
##   
## Standardized Within-Group Residuals:  
## Min Q1 Med Q3 Max   
## -1.975557 -0.659613 0.003324 0.451057 1.781538   
##   
## Number of Observations: 60  
## Number of Groups: 15

This is the same mixed-effects model that we ran earlier in class. The error around the effect is halved! Now the effect is 9 plus/minus 1. Much more precise estimate. Truth is still contained. And now we have a measure of variation due to field: 3.58.

But now if we want to include that treatment by field interaction, here is how we do that:

# Fit a linear mixed-effects model with a treatment x field interaction  
results <- lme(Biomass ~ Treatment, data = datum, random=~Treatment|Field)  
summary(results)

## Linear mixed-effects model fit by REML  
## Data: datum   
##   
## Random effects:  
## Formula: ~Treatment | Field  
## Structure: General positive-definite, Log-Cholesky parametrization  
## StdDev Corr   
## (Intercept) 4.0192 (Intr)  
## Treatment 2.9972 -0.403  
## Residual 0.8909   
##   
## Fixed effects: Biomass ~ Treatment   
## Correlation:   
## (Intr)  
## Treatment -0.413  
##   
## Standardized Within-Group Residuals:  
## Min Q1 Med Q3 Max   
## -2.01457 -0.47589 0.04183 0.37200 1.84020   
##   
## Number of Observations: 60  
## Number of Groups: 15

Before we had ~1|Field – just a random effect of field. Every field would have a random intercept.

But now we are saying that we want a **random intercept due to field** and a **random treatment effect due to field**. In other words, the treatment effect can vary randomly within the random effect of field. The treatment effect can be larger or smaller depending on the field.

What’s different about our results? The estimate is the same, but the standard error went up!! We lost a lot of power.

But now we can say that field accounts for a standard deviation of 4.01 in the response, the treatment effect can vary with a standard deviation of 2.99, and there is residual, within-field error with standard deviation of 0.89.

**This last bit is the key difference**. The most simple model said that this within-field error was 1.91 – but here is now 0.89 and much closer to truth (1). **We have decreased this residual and instead partitioned that error correctly to being due to treatment.**

But we actually lost power. **This is a standard consequence of including random effects in our model.** Random effects make the model more complicated, which causes you to lose power.

## Reasons to include a random interaction

There are two reasons to include random interactions:

1. You are specifically **interested in the variation of the treatment effect**. If you want to know how the treatment effect varies among the random-effect of plot, you will want to include this. Make sure you have enough samples within each plot/treatment so you can test for that random interaction.
2. The second reason relates to **unbalanced designs**. For our example, certain fields are measured A LOT and those certain fields have strong effects. If you don’t include a random interaction, that will bias the results so that your treatment estimate will look more similar to the random field\*treatment effect. (We could demonstrate this with data simulation, but I didn’t have time to put this together.)

Let’s examine an example with an unbalanced design.

### Truth

***Examine Truth using the R-code for the lecture***

### Analysis

# Read in the data  
datum <- read.csv("lecture\_16\_dataset2.csv")  
head(datum, 20)  
tail(datum, 20)  
  
# Fit a linear model  
results <- lm(Biomass ~ Sunlight, data = datum)  
summary(results)

##   
## Call:  
## lm(formula = Biomass ~ Sunlight, data = datum)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.88 -4.49 1.55 4.74 10.31   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 25.45 1.12 22.72 < 2e-16 \*\*\*  
## Sunlight 14.12 1.92 7.37 2.5e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.98 on 118 degrees of freedom  
## Multiple R-squared: 0.315, Adjusted R-squared: 0.309   
## F-statistic: 54.3 on 1 and 118 DF, p-value: 2.53e-11

# Fit a linear mixed-effects model with a random-effect of field (no interaction)  
results <- lme(Biomass ~ Sunlight, data = datum, random=~1|Field)  
summary(results)

## Linear mixed-effects model fit by REML  
## Data: datum   
##   
## Random effects:  
## Formula: ~1 | Field  
## (Intercept) Residual  
## StdDev: 5.805 1.908  
##   
## Fixed effects: Biomass ~ Sunlight   
## Correlation:   
## (Intr)  
## Sunlight -0.207  
##   
## Standardized Within-Group Residuals:  
## Min Q1 Med Q3 Max   
## -2.14470 -0.60559 -0.03805 0.69076 2.22458   
##   
## Number of Observations: 120  
## Number of Groups: 15

# Fit a linear mixed-effects model with a random field-x-sunlight effect  
results <- lme(Biomass ~ Sunlight, data = datum, random=~Sunlight|Field)  
summary(results)

## Linear mixed-effects model fit by REML  
## Data: datum   
##   
## Random effects:  
## Formula: ~Sunlight | Field  
## Structure: General positive-definite, Log-Cholesky parametrization  
## StdDev Corr   
## (Intercept) 4.917 (Intr)  
## Sunlight 3.886 0.272   
## Residual 1.022   
##   
## Fixed effects: Biomass ~ Sunlight   
## Correlation:   
## (Intr)  
## Sunlight 0.17   
##   
## Standardized Within-Group Residuals:  
## Min Q1 Med Q3 Max   
## -2.0151 -0.6053 -0.1749 0.5413 2.9131   
##   
## Number of Observations: 120  
## Number of Groups: 15

The effect of sunlight was still 10.

The linear model and the mixed-effects model without the proper field-x-sun interaction have biased estimates of ~14.5 +/- 1.2, which do not overlap with truth (10). We are getting an inaccurate estimate because Field 1 has huge **‘leverage’**, which is dragging the estimate up, because Field 1 has a huge effect of treatment and it has a greater number of sample, and thus has too much of an effect on the estimation results.

Adding in the random effect of field (random intercept) still is biased, because it treats every field equally with respect to it’s contribution to the sunlight effect.

Correctly adding in the sunlight x field random interaction effect produces unbiased estimates. R will now calculate (1) the effect of sunlight for every field individually, and then (2) calculates the average based on those individual samples. Now our estimate is accurate; truth is contained within our confidence limits.

Where might this be valuable? Radiotelemetry of animals. We collar animals; some we get 10,000 datapoints, others we get 5. When we **have vastly unbalanced designs** with different sample sizes for individuals, we need to be careful about how individuals with large samples might bias our results.

We should be aware of this.

## Truth

Here is the code to create Truth for the two datasets analyzed here today.

################### 'Truth' ####################   
### Lecture 16: code to simulate data for class  
  
### Dataset 1  
# Set the seed for reproducibility  
set.seed(123)  
  
# Simulate data  
n\_fields <- 15  
Field <- rep(1:n\_fields, each = 4)  
Treatment <- rep(c(0,0,1,1), n\_fields)  
  
# Field error  
FieldError <- rep(rnorm(n\_fields, 0, 5), each = 4)  
  
# FieldXTreatment error  
FieldXTreatError <- rep(rnorm(n\_fields, 0, 3), each = 4)  
  
# Within-field error  
Error <- rnorm(length(Field), 0, 1)  
  
# Response variable  
Biomass <- 25 + 10\*Treatment + Treatment\*FieldXTreatError + FieldError + Error  
  
# Save the data  
datum <- data.frame(Field = Field, Treatment = Treatment, FieldXTreatError = FieldXTreatError, FieldError = FieldError, Error = Error, Biomass = Biomass)  
  
# Save the data  
write.csv(datum, "lecture\_16\_dataset1.csv", row.names = FALSE)  
  
  
### Dataset 2  
# Everything is the same as above in terms of truth, but now the sampling design  
# is unbalanced. Field 1 has a lot of samples and a large effect of treatment.  
# Set the seed for reproducibility  
set.seed(123)  
  
# Simulate data  
n\_fields <- 15  
Field <- sort(c(rep(1, each = 50), rep(2:n\_fields, 5)))  
  
# Continuous X-variable  
Sunlight <- runif(length(Field), 0, 1)  
  
# Field error (matching the number of samples for each field)  
FieldError <- c(rep(rnorm(1, 0, 5), 50), rep(rnorm(n\_fields - 1, 0, 5), each = 5))  
  
# Field X Sunlight interaction error (matching the number of samples for each field)  
FieldXSunError <- c(rep(10, 50), rep(rnorm(n\_fields - 1, 0, 3), each = 5))  
  
# Within-field error  
Error <- rnorm(length(Field), 0, 1)  
  
# Response variable  
Biomass <- 25 + 10\*Sunlight + Sunlight\*FieldXSunError + FieldError + Error  
  
# Save the data  
datum <- data.frame(Field = Field, Sunlight = Sunlight, FieldXSunError = FieldXSunError, FieldError = FieldError, Error = Error, Biomass = Biomass)  
  
# Save the data  
write.csv(datum, "lecture\_16\_dataset2.csv", row.names = FALSE)

[–go to next lecture–](lecture_17.html)