Lesson 16

Clark

Awhile ago I collected some data on burglaries in the southside of Chicago. The data are available here:

```
library(tidyverse)

chi_data <- read.csv("https://raw.githubusercontent.com/nick3703/Chicago-Data/master/crime.csv")
```

If we explore this a bit, we see that the data are presented as a 552×73 matrix where each row corresponds to a different Census Block Group and each column corresponds to a month of the year. What I want to do, just to kick us off, is to pair up, go to the boards and come up with a statistical model for this data, you may assume that we also have data on the percent of the population that is unemployed and our main statistical question is determining the relationship between unemployment and burglaries.

If we think about data, such as spatio-temporal data, there are mulitple ways for us to consider the data. We could consider our observations to be vectors in time, vectors in space, an entire spatio-temporal matrix, or to be univariate. However, we certainly would think that there are unique aspects of each spatial location here that should be addressed somehow.

Perhaps, just to start, we will think of each observational unit as a vector of length 73, that is, each observational unit is a unique Census block group. Again, our covariates that we want to consider are the percent unemployed.

One way for us to model this is through a marginal model. Here, we will let y_{ij} denote the number of burglaries in month j for location i. Our model is:

Here we see that each time period has a separate relationship with number of burglaries. Our x_i terms don't change though. This looks like a typical GLM, however we are left with trying to figure out what the joint distribution for y_i is.

If, however, we modify our model and shrug our shoulders and allow our data to come from a linear regression model, we have:

Though in this case that may not be appropriate.

An alternative approach is to capture the unique aspects of each spatial location through a random effect. That is we write:

This is an example of a *mixed effects* model. The random effects, u_i are specific to each location and typically are assumed to be distributed as:

If, we allow $z_{ij} = 1$ we are left with what is sometimes called a *random intercept* model. That is, each location has it's own intercept.

Let's assume, for a minute, that we can model our data with a linear model. Further we will assume that $\Sigma_u = \sigma^2 I$. Here we can rewrite everything as:

This implies that the correlation between observations within the same census block group are:

Let's consider perhaps a simpler example

```
library(faraway)
data(pulp)
```

Here we can write out our model two different ways:

```
library(lme4)
model <- lmer(bright ~ 1+(1|operator),pulp,REML=FALSE)
summary(model)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: bright ~ 1 + (1 | operator)
##    Data: pulp
##
##      AIC      BIC   logLik deviance df.resid
##     22.5     25.5     -8.3     16.5       17
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.50554 -0.78116 -0.06353  0.65850  1.56232
##
```

```
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  operator (Intercept) 0.04575  0.2139
##  Residual             0.10625  0.3260
## Number of obs: 20, groups:  operator, 4
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  60.4000     0.1294   466.7
```

Let's write out the fitted model:

Let's talk, just a bit, about this notion of `REML=FALSE`. Recall (from MA476) when we estimate σ^2 we typically use:

Why do we have $n-1$ in our denominator?

If we were to use MLEs to estimate σ^2 we would instead use n, which means that the MLEs are biased. While this mostly doesn't matter, for mixed effects models we may have lots of variance components that we are estimating and n may not be very bit. To get around this, typically REML gets used, or *restricted maximum likelihood estimators*, or sometimes people call it residual maximum likelihood estimates. The general idea is, the reason our MLEs are biased is due to the fixed effect parameter being used in the estimate of σ^2. That is, in our estimate of σ^2 we have an $\bar{x}$ term. To get around this, REML finds independent linear combinations of

the response such that $k^t X = 0$. For example we could have:

We then use this data to estimate our variance via:

Now, our mean is zero so we don't have to estimate a $\bar{X}$ to estimate σ^2

```
library(lme4)
model2 <- lmer(bright ~ 1+(1|operator),pulp,REML=TRUE) #Default is TRUE
summary(model2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: bright ~ 1 + (1 | operator)
##    Data: pulp
##
## REML criterion at convergence: 18.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.4666 -0.7595 -0.1244  0.6281  1.6012
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  operator (Intercept) 0.06808  0.2609
##  Residual             0.10625  0.3260
## Number of obs: 20, groups:  operator, 4
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  60.4000     0.1494   404.2
```

Here our fitted model is:

We are going to go much more in depth into mixed effects models and I may deviate from the syllabus to talk more about this and talk less about machine learning algorithms that you likely already saw in MA477. The reason these become difficult is, unlike for a GLM, we will no longer have the iterative reweighted least squares algorithm to maximize our likelihood. Nor can we, in general, coerce our models to become something like a negative binomial or ZIP model. For example, for our Chicago data, perhaps we want to build a model like: