# Updated Multilevel Final

Katy Koenig

4/8/2020

## Summary

Below, I compare four regression models (pooled, unpooled and two varying-intercept multilevel models) to understand the relationship between use of a public resource, access and safety. I first provide the large marjority of the code. I then provide corresponding analysis of the output of the code. This analysis demonstrates the utility of multilevel models to examine data that may be problematic using classical regression (either pooled or unpooled) as we are now able to gain information through partially pooling information across our time and spatial groups.

## Code

```
# Read in data
lib_visits <- read.csv("lib_crime_station.csv")
vis_yr <- lib_visits %>% group_by(year) %>% summarise(sum_visits = sum(YTD),
    sum_crime = sum(crime_count))

# Make visualizations for summary
vis1 <- gather(vis_yr, measure, value, -year) %>% ggplot(., aes(year,
    value)) + geom_col(aes(fill = measure)) + expand_limits(y = 0) +
    theme_minimal() + facet_wrap(~measure, scales = "free") +
    ggtitle("Crime Rates Stagnant While Library Visits Fall")

vis2 <- ggplot(lib_visits, aes(YTD)) + geom_density() + theme_minimal() +
    ggtitle("KDE of Library Visits") + xlab("Number of Library Visits") +
    ylab("Count")

vis3_title <- paste0("More Library Visits in Lower Crime Areas: ",
    "Harold Washington Library Excluded")
wo_hw <- subset(lib_visits, YTD < 1e+06)
vis3 <- ggplot(wo_hw, aes(crime_count, YTD)) + geom_point(aes(color = zone,
    size = num_stations, alpha = 0.6)) + facet_wrap(~year) +
    theme_minimal() + labs(y = "library visits") + ggtitle(vis3_title)

# Helper Function
stan_summary <- function(fit) {
    as.data.frame(fit) %>% gather("parameter") %>% group_by(parameter) %>%
        summarize(median = median(value), MAD_SD = mad(value))
}
```

```r
# Transform library vists & crime count to log values to
# ensure that outcomes make sense i.e. avoid negative library
# visits and crime counts
lib_visits$log_visits <- log(ifelse(lib_visits$YTD == 0, 0.1,
    lib_visits$YTD))
lib_visits$log_crime <- log(ifelse(lib_visits$crime_count ==
    0, 0.1, lib_visits$crime_count))

# Pooled Model
pooled <- stan_glm(log_visits ~ log_crime + num_stations, data = lib_visits)
samples_pooled <- as.data.frame(pooled)

# Unpooled Model
unpooled <- stan_glm(log_visits ~ num_stations + as.factor(period),
    data = lib_visits)
samples_unpooled <- as.data.frame(unpooled)

# Varying-intercept model with time period as group
multi_one <- stan_glmer(log_visits ~ (1 | period) + log_crime +
    num_stations, data = lib_visits)

# let's add another level to the varying intercept model
# above using community area
multi_two <- stan_glmer(log_visits ~ (1 | community_area) + (1 |
    period) + log_crime + num_stations, data = lib_visits)

# Function below gets the pooling factor for multilevel
# models
lambda <- function(fit) {
    eta <- as.data.frame(fit) %>% dplyr::select(starts_with("b[(Intercept)"))
    numerator <- var(colMeans(eta))
    denominator <- mean(apply(eta, 1, var))
    1 - numerator/denominator
}

# elpd loo
loo_pooled <- loo(pooled, k_threshold = 0.7)
loo_unpooled <- loo(unpooled, k_threshold = 0.7)
loo_multi_one <- loo(multi_one, k_threshold = 0.7)
loo_multi_two <- loo(multi_two, k_threshold = 0.7)

# Setting up lists for printing below
multi_models <- list("multilevel_one", "multilevel_two")
model_name <- list("pooled", "unpooled", "multilevel_one", "multilevel_two",
    "multilevel_two")
```

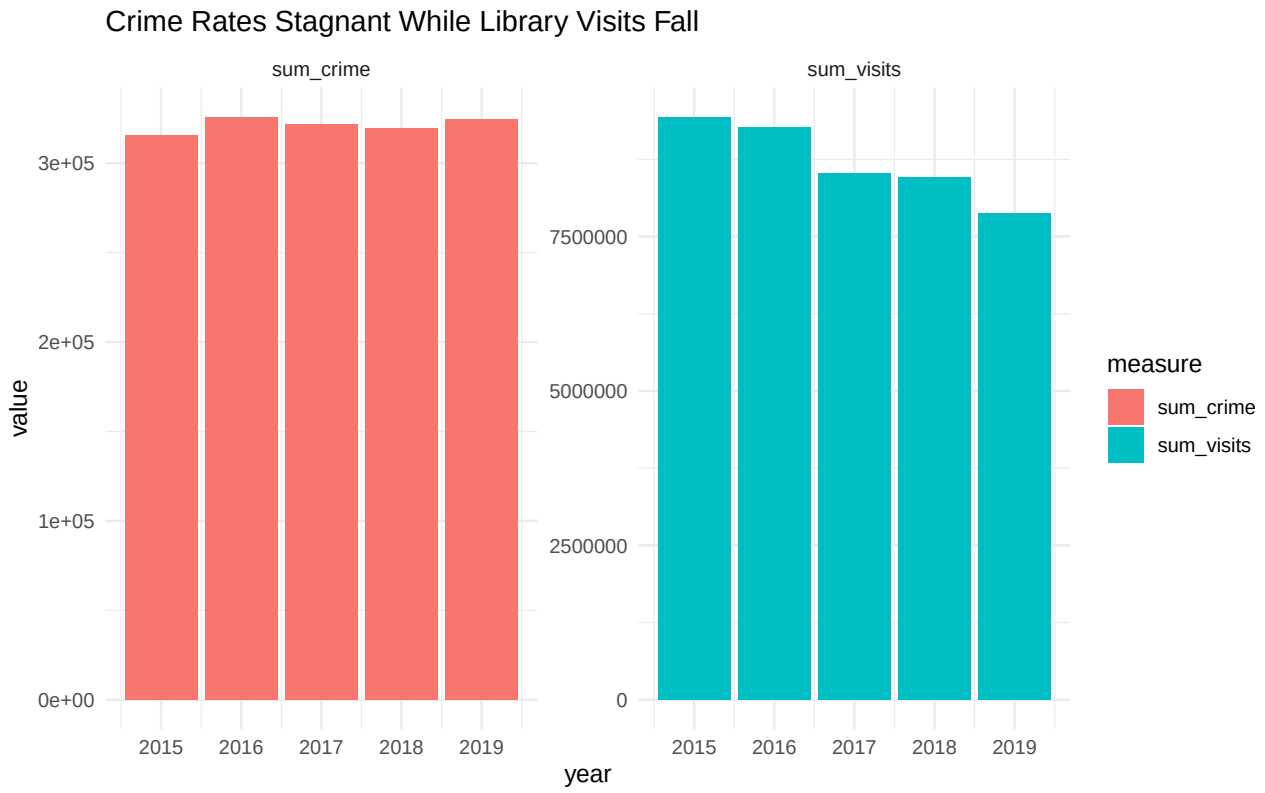
# Analysis

## Data Description & Summary

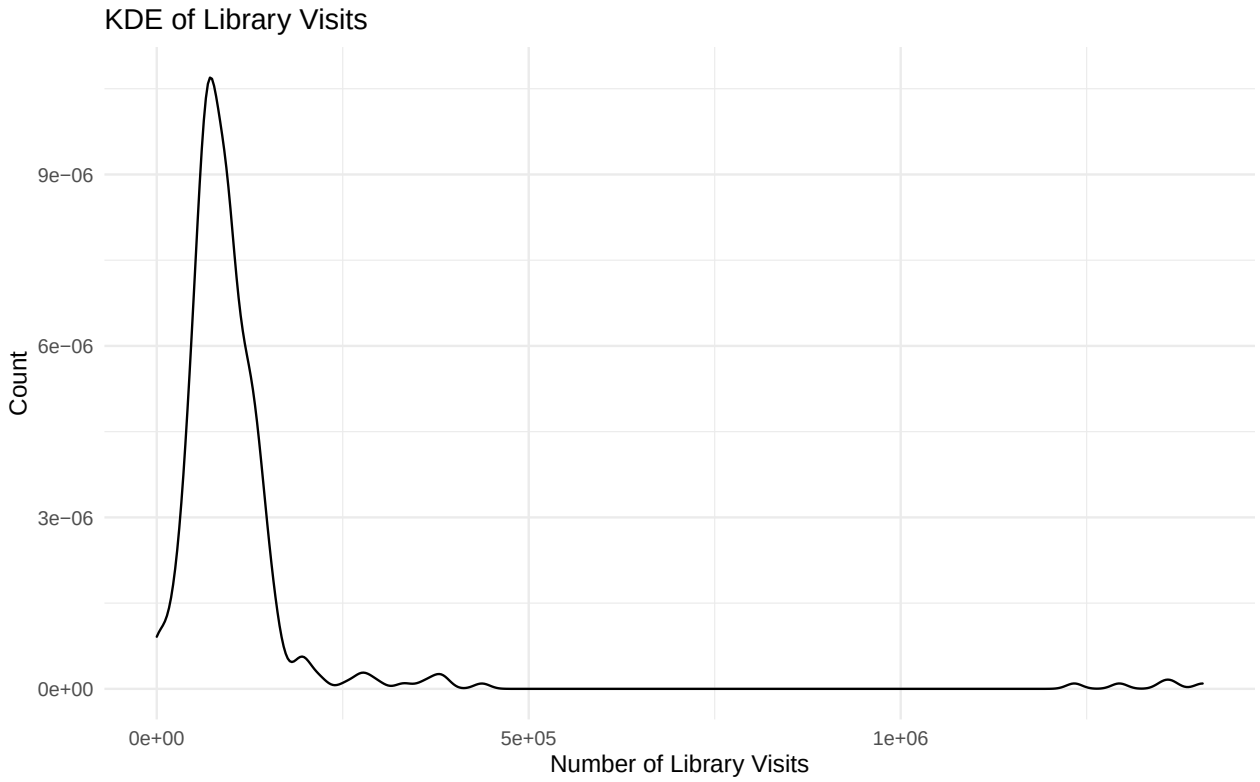
In my analysis below, I analyze the impact of access and safety on the use of public utility. To do so, I collected data from the City of Chicago data portal regarding library visits (a public utility), reported

crime (a proxy for safety) and L stop locations (a proxy for access) for 2015-2019. I aggregated the data by year and community area and merged this information with library visit information, so that each row consists of a library branch, year (group: period), number of library visits that year (outcome variable: YTD), number of crimes in the community area in which the library is located for a given period (group-level predictor: crime_count), and number of L stations in the community area in which the library is located (observation-level predictor: num_stations). For more information regarding the data cleaning process, see this notebook. As I used data from 2015 to 2019, there are five groups in this dataset, with each group having 79-80 observations. (Note: I used reported crime data. Any crimes that occur which are not reported are therefore missing from this dataset. Conversely, if someone falsely reports a crime, it would be included in this dataset.)

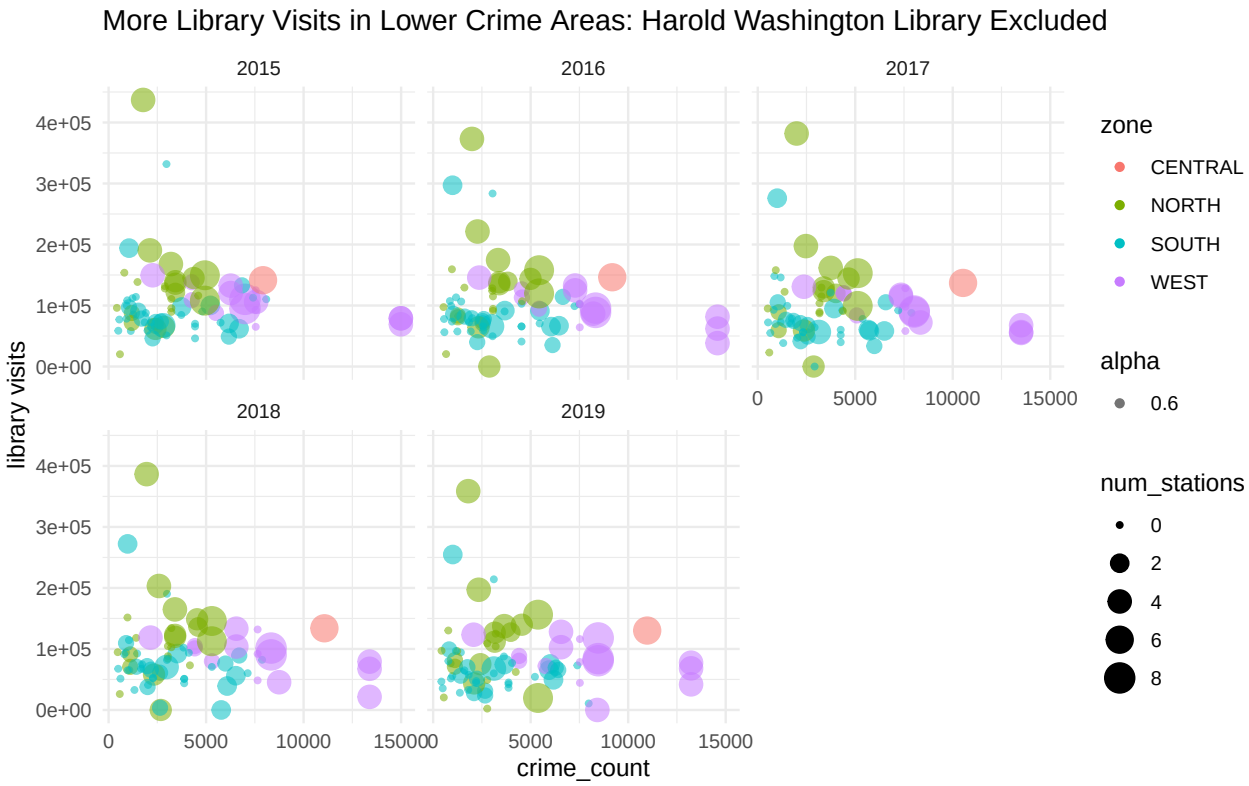
```
vis1
```




```
vis2
```




```
vis3
```




As we can see in our first plot above, library visits overall decrease slightly in our five year period whereas crime rates seem to hold pretty steady. In the second plot, we see that our data is not exactly normal: most libraries have a similiar number of visits, we have a few extreme outliers.

Note: in the third data visualization, Harold Washington Library was not included as it drown out almost all other data points. We also include a 'zone' variable corresponding to the "sides" of Chicago (north/south/west/central), e.g. Hyde Park would be in the south zone whereas the Loop comprised almost all of the central zone. This is only for illustrative purposes and is not used in the analysis above. Additionally, in this plot we see that zones are somewhat related to number of L stations: the south zone has considerably smaller dots on average, meaning there are fewer L stops in that community area. We also see here that crime is somewhat related to library visits: areas with a high amount of library visits also have lower than average crime rates.

For my analysis, I compare varying regression models including a pooled model, an unpooled model, and two varying-intercept multilevel models. Each model is described in greater detail below.

## Models

### Description & Output of Models

### Model 1: Pooled Model

I first run the following standard OLS model:

$y \sim N(\alpha + \beta_1 x_i + \beta_2 u_i, \sigma_y^2)$

using our variables, we have:

$log\_visits \sim N(\alpha + \beta_1 num\_stops_i + \beta_2 log\_crime_i, \sigma_y^2)$

```
stan_summary(pooled)
```

```
## # A tibble: 4 x 3
##   parameter    median MAD_SD
##   <chr>         <dbl>  <dbl>
## 1 (Intercept)  12.3   0.976
## 2 log_crime    -0.161 0.125
## 3 num_stations  0.112 0.0402
## 4 sigma         1.78  0.0626
```

From our summary above, we see that the number of L stations has a positive effect on library visits: with each additional L stations in the community area in which the library is located, we see approximattely 11 percent more library visits. Conversely, crime has a negative effect on library visits: for each one percent increase in crime rates, we see a decrease in library visits of approximately 16 percent.

### Model 2: Unpooled Model

We then run the following unpooled intercept model:

$y \sim N(\alpha_{j[i]} + \beta_1 x_i, \sigma_y^2)$

using our variables, we have:

$log\_vists \sim N(\alpha_{j[i]} + \beta_1 num\_stops_i, \sigma_y^2)$ where J is our time periods and $j \in J$.

```
stan_summary(unpooled)
```

```
## # A tibble: 7 x 3
##   parameter           median MAD_SD
##   <chr>                <dbl>  <dbl>
## 1 (Intercept)        11.3    0.210
## 2 as.factor(period)2 -0.191  0.278
## 3 as.factor(period)3 -0.427  0.280
```

```
## 4 as.factor(period)4 -0.489  0.289
## 5 as.factor(period)5 -0.493  0.289
## 6 num_stations        0.0894 0.0362
## 7 sigma               1.78   0.0599
```

Again, we see that number of L stops has a positive effect on number of library visits. Specifically, when there is one additional L stop in the library's area, we see an increase in library visits of approximately nine percent.

We also see above that time period has a negative effect on library visits, implying that over time, we see a decrease in overall library visits, although this may not be true year to year. For example, in period 4 (or 2018), we see a coefficient of roughly -0.5, meaning there was roughly 0.5 percent fewer visits in 2018 than our first time period, 2015), ceteris paribus. For period 5, 2019, we see a coefficient also of roughly -0.5, also meaning that there was roughly 0.5 percent fewer visits in 2019 than our first time period, 2015), ceteris paribus. Therefore, we expect a very similar number of visit in 2018 and 2019 (not a decrease in visits), all else being equal.

### Model 3: Multilevel Model, Varying Intercept by Time Period

Below we present a summary for the following model: $y_i \sim N(\alpha_{j[i]} + \beta x_i, \sigma_y^2)$

Second level regression: $\alpha_j \sim N(\gamma_0 + \gamma_1 u_1, \sigma_\alpha^2)$

Using our variables our model becomes:

$log\_visits_i \sim N(\alpha_{j[i]} + \beta num\_stop_i, \sigma_y^2)$

Second level regression: $\alpha_j \sim N(\gamma_0 + \gamma_1 log\_crime, \sigma_\alpha^2)$

```
stan_summary(multi_one)
```

```
## # A tibble: 10 x 3
##    parameter                                median MAD_SD
##    <chr>                                     <dbl>  <dbl>
##  1 (Intercept)                              12.3    0.983
##  2 b[(Intercept) period:1]                   0.0854 0.137
##  3 b[(Intercept) period:2]                   0.0264 0.108
##  4 b[(Intercept) period:3]                  -0.0207 0.105
##  5 b[(Intercept) period:4]                  -0.0361 0.112
##  6 b[(Intercept) period:5]                  -0.0373 0.114
##  7 log_crime                                -0.160  0.125
##  8 num_stations                              0.113  0.0418
##  9 sigma                                     1.77   0.0641
## 10 Sigma[period:(Intercept),(Intercept)]     0.0248 0.0341
```

In this model, we again see positive effects of number of L stations on library visits. Namely, that an increase in one L station results in an increase in library visits of approximately 11.5 percent. We also see a negative affect of crime rates: an increase in crime of one percent results in a decrease in library visits of approximately 16 percent. We now also see varying effects of time on library visits. Initially, the number of library visits increase as time goes by, then library visits decrease as time goes by.

Above, we also see that there is much more variation within our time period groups ($\sigma_y$) than between our time periods ($\sigma_{alpha}$):

$\sigma_\alpha \approx \sqrt{0.0275} \approx 0.1658$

$\sigma_y \approx 1.77$

### Model 3: Multi-level Model, Varying Intercept by Community Area & Time Period

Expanding on the multilevel model above: $y_i \sim N(\alpha_{j[i]}^{period} + \alpha_{k[i]}^{community\_area} + \beta * num\_stop_i, \sigma_y^2)$

Second level regression lines: $\alpha_j^{period} \sim N(\gamma_0 + \gamma_1 log\_crime, \sigma_{period}^2)$

$\alpha_k^{community\_area} \sim N(\mu_\alpha^{community\_area}, \sigma_{community\_area}^2)$

```
head(stan_summary(multi_two))
```

```
## # A tibble: 6 x 3
##   parameter                           median MAD_SD
##   <chr>                                <dbl>  <dbl>
## 1 (Intercept)                         12.6      1.69
## 2 b[(Intercept) community_area:1]      0.413    0.591
## 3 b[(Intercept) community_area:10]     0.00719  0.480
## 4 b[(Intercept) community_area:12]     0.128    0.643
## 5 b[(Intercept) community_area:13]     0.544    0.602
## 6 b[(Intercept) community_area:14]    -0.311    0.578
```

```
tail(stan_summary(multi_two))
```

```
## # A tibble: 6 x 3
##   parameter                                          median MAD_SD
##   <chr>                                               <dbl>  <dbl>
## 1 b[(Intercept) period:5]                           -0.0675 0.133
## 2 log_crime                                         -0.205  0.220
## 3 num_stations                                       0.129  0.0703
## 4 sigma                                              1.48   0.0595
## 5 Sigma[community_area:(Intercept),(Intercept)]      1.20   0.312
## 6 Sigma[period:(Intercept),(Intercept)]              0.0366 0.0453
```

While it would be difficult to efficiently explain the intercept coefficients for each of our 77 community areas, we can again say that the number of L stations has a positive effect on the number of library visits and that crime rates have a negative effect on library visits. It is also important to note that we see a significant decrease in the standard errors for these two coefficients in our second multilevel model than our first multilevel model.

Again, we see more variation within groups ($\sigma_y$) than between either of our two groups ($\sigma_\alpha^{period}$, $\sigma_\alpha^{community\_area}$), but we do see significantly more variation between our community area groups than our time period groups:

$\sigma_y \approx 1.5$

$\sigma_\alpha^{period} \approx \sqrt{0.04} \approx 0.2$

$\sigma_\alpha^{community\_area} \approx \sqrt{1.2} \approx 1.1$

### Evaluation of Models & Conclusion

```
# get pooling factor multilevel models
i <- 1
for (model in list(multi_one, multi_two)) {
    print(multi_models[i])
    print(lambda(model))
    i = i + 1
}
```

```
## [[1]]
## [1] "multilevel_one"
##
## [1] 0.7167722
## [[1]]
## [1] "multilevel_two"
##
## [1] 0.270755
```

As we can see above, there is significantly more pooling between groups in our first multilevel regression model, meaning that this model looks a lot more similar to our classical, pooled model than our second multilevel model.

```
# Bayesian R Squared for all models
i <- 1
for (model in list(pooled, unpooled, multi_one, multi_two)) {
    print(model_name[i])
    print(quantile(bayes_R2(model), c(0.25, 0.5, 0.75)))
    i = i + 1
}
```

```
## [[1]]
## [1] "pooled"
##
##        25%        50%        75%
## 0.01357218 0.02193134 0.03270436
## [[1]]
## [1] "unpooled"
##
##        25%        50%        75%
## 0.02615403 0.03676126 0.04928816
## [[1]]
## [1] "multilevel_one"
##
##        25%        50%        75%
## 0.01882458 0.02884891 0.04103296
## [[1]]
## [1] "multilevel_two"
##
##       25%       50%       75%
## 0.2954630 0.3242047 0.3520595
```

```
# elpd loo
loo_compare(loo_pooled, loo_unpooled, loo_multi_one, loo_multi_two)
```

```
##           elpd_diff se_diff
## multi_two   0.0       0.0
## pooled    -34.3      36.8
## multi_one -34.6      36.8
## unpooled  -36.3      36.3
```

In the Bayesian $R^2$ values printed above, we seee that our second multilevel model (which includes varying intercept values for each time period and each community area) performs significantly better than our unpooled, pooled and first multilevel model (with a varying intercept for time period only) in that our second model explains much more of the variance in our data than our other models, representing the importance of community area on library visits.

To conduct cross validation, we also print the leave-one-out log scores of each of our models above. Again, we see that our second model again outperforms our other three models significantly.