The Effectiveness of Preventative Measure on the Spread of Malaria in sub-Saharan Africa

Julie Matthias, Ksenia Lepikhina

May 3, 2020

1 Introduction

Malaria is caused by *Plasmodium* parasites which are transmitted to people through the bites of infected female *Anopheles* mosquitoes. People who contract malaria can face severe illness including death if they are not treated quickly. Though the mosquito borne infectious disease is primarily prevalent in Africa, cases of the disease have been found on each continent. The disease is preventable and curable, however, in 2018, "there were an estimated 228 million cases of malaria worldwide" [1]. Even more alarming, there were approximately 405,000 deaths in 2018 [1]. In order to prevent the spread of malaria, vector control interventions have been implemented across Africa. There are two forms of vector control — long-lasting insecticidal nets (LLINs) and indoor residual spraying (IRS). These vector controls are primarily focused on protecting people when they are indoors or in bed [2]. IRS is done by spraying the inside of housing structures with an insecticide while LLINs are used to protect people in their sleep by providing a physical insecticidal barrier. While these methods seem proven, it is being discovered that mosquitoes are developing resistance to these insecticides which highlights the need to study and improve the tools to combat malaria [1]. This need for continued research is a main reason why we chose this paper topic.

Malaria is frequently researched across multiple disciplines ranging from biomedical and epidemic research to network models [3] to research on deforestation [4]. The research conducted in this paper is inspired by Sherrard-Smith et al.'s paper "Mosquito feeding behavior and how it influences residual malaria transmission across Africa" (July 2019). Their research focused largely on estimating the number of outdoor mosquito bites across sub-Saharan Africa, exploring temporal trends of the epidemiological spread and estimating the significance to the public and finally estimating the "residual transmission" across Africa to make a claim about the relationship of outdoor biting and the spread of the disease. This is an important relationship to explore because even with LLINs and IRS populations are still seeing malaria spread. This means that vector controls are not enough to prevent this disease and other ways must be explored [2].

The main focus of this paper was to recreate the temporal analysis that was done in the paper by Sherrard-Smith et al. In particular, we explored the temporal aspects of mosquito bites indoors and in bed (within insecticide-treated bed nets) and aimed to verify if the paper took into account that time is a frequent confounder that may cause spurious relationships. As a starting point, we investigated if the paper's results violated any regression assumptions. In this investigation, we found some potential violations and conducted our own analysis to try to account for these.

In addition to exploring Sherrard-Smith's research [2], we looked into other malaria studies and factors that could cause outdoor biting to increase. Some of these involved looking into a specific country's population, GDP, average humidity, etc. to see if these factors were also contributing to the proportion of bites measured. These new data points lead us to discover a new model that we believe may tell the story of the data more accurately. This data was collected from many different sources accourding to Sheppard-Smith et al. [2] and was purely observational which prohibited us from deriving a causal effect. However, we were able to identify a relationship between the proportion of bites and some attributes of a country.

2 Data

For this project, we worked with a couple of different data sources. Our primary data source was the dataset provided in the appendix of the paper: "Mosquito feeding behavior and how it influences residual malaria transmission across Africa" [2]. The data was pulled by downloading the necessary Excel spreadsheet files from the appendix of the paper at this source: `https://www.pnas.org/content/116/30/15086/tab-figures-data`.

The data, as downloaded from the appendix, came formatted as an Excel spreadsheet with multiple sheets. In order to be able to analyze the data comfortably, we manually split the spreadsheets into separate files and converted them to CSVs. The primary CSVs we looked at were the sheets containing information about the proportion of humans inside at each hour of the day[1], the proportion of humans in bed at each hour of the day[2], the proportion of mosquito bites indoors (ϕ_I), the proportion of mosquito bites in bed (ϕ_B), and country level estimates of ϕ_I and ϕ_B over time (1989-2017) by mosquito species.

The columns of each proportion dataset contained data from various papers (over 20 datasets) [2]. Each row indicated the hour of day and the data itself consisted of proportions with ranges from 0 to 1.

The country data consisted of data from 12 African countries, the specific sites that were used to find estimates for ϕ_I and ϕ_B, and which study the data came from. For our purposes, we ignored the specific sites and which study the data came from and instead grouped the data by country and year and averaged the ϕ values. The data also included whether IRS or LLINs were used and what type along with which species that was studied.

In addition to the data provided in the paper, we pulled data on the population (in millions of people) of each country and year that correspond to the data provided by Sherrard-Smith, et al., the GDP (in billions of USD) for each country and year pair, the average yearly precipitation (in millimeters) in each country, the average yearly humidity (out of 100 percent) in each country, and the average yearly temperature (in Fahrenheit) for each country. The first two (population and GDP) are specific to a country and year and the last few (precipitation, humidity and temperature) are only specific to a given country.

The dataset mentioned in the previous paragraph was manually created. To create it, for each country, we determined the years that it had data for and manually created a CSV with columns: country, year, population, GDP, humidity, temperature, and precipitation. The new CSV ended up having 60 rows. For analysis, this data was joined to the Sherrard-Smith, et al. data.

All of the data from Sherrard-Smith, et al. are pre-processed meaning that we did not have access to the raw data. The data that was used from the paper is all observational. From that dataset, we were given the normalized proportion of people's indoor and in bed hours, the normalized proportion of indoor mosquito biting and in bed mosquito biting, the country name, the year and mosquito species. Each of the predictors are observational.

[1] These value were implicitly used in the ϕ values

[2] See the previous note.

The data that was pulled (using a simple google search) population, GDP, etc., was also observational.

The dataset from Sherrard-Smith, et al. consisted of multiple discrete parameters such as country, site, year, spray or no spray, mosquito species, as well as two primary continuous parameters (ϕ_I and ϕ_B, the proportion of indoor biting and in bed biting, respectively). Note, as mentioned previously, we chose to ignore site as that was too granular of a measure. We also chose to ignore spray versus no spray for our analysis and simply focused on the proportion of biting indoors and in bed. Additionally, it's important to note that ϕ_I and ϕ_B are not observed values but are rather derived values:

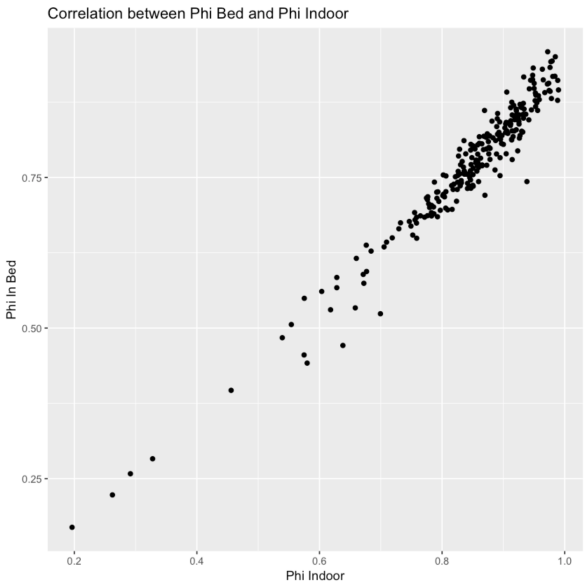
$$\phi_I = \frac{\sum_t p_I(t)\lambda_I(t)}{\sum_t \Big((1-p_I(t))\lambda_O(t) + p_I(t)\lambda_I(t)\Big)}$$

$$\phi_B = \frac{\sum_t p_B(t)\lambda_I(t)}{\sum_t \Big((1-p_I(t))\lambda_O(t) + p_I(t)\lambda_I(t)\Big)}$$

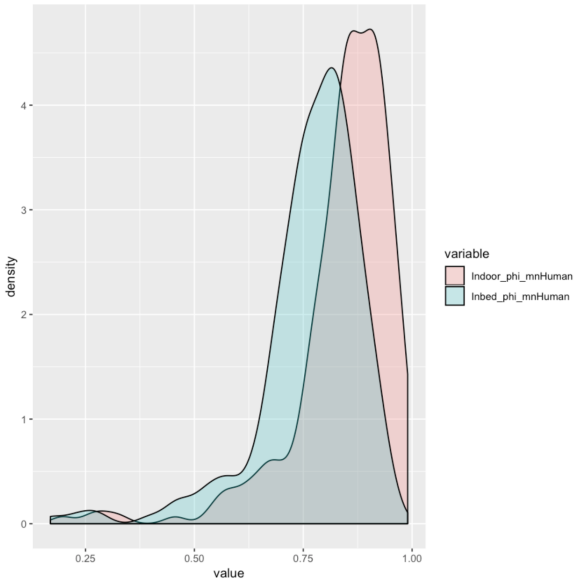
where $p_I(t)$ is the proportion of people inside at hour t, $\lambda_I(t)$ is the sum of bites inside at hour t over the sum of bites for all hours for both indoor and outside (i.e. the biting rate indoors) and similarly for $p_O(t)$, $\lambda_O(t)$, and $\lambda_B(t)$.

3 Methods

Our original goal was to recreate the results of the data. First, we wanted to verify whether or not the proportion of biting indoor and the proportion of biting in bed were highly correlated:



The figure above demonstrates that clearly, there is a strong positive relationship between ϕ_I (proportion of indoor bites) and ϕ_B (proportion of in bed bites). We can see that though they are correlated, they have slightly different means:

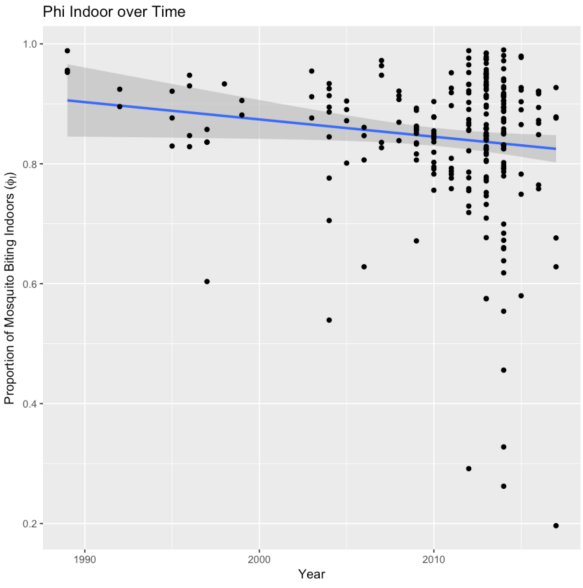


The density figure above shows us that on average we see more biting indoors than in bed. This is reasonable because people are

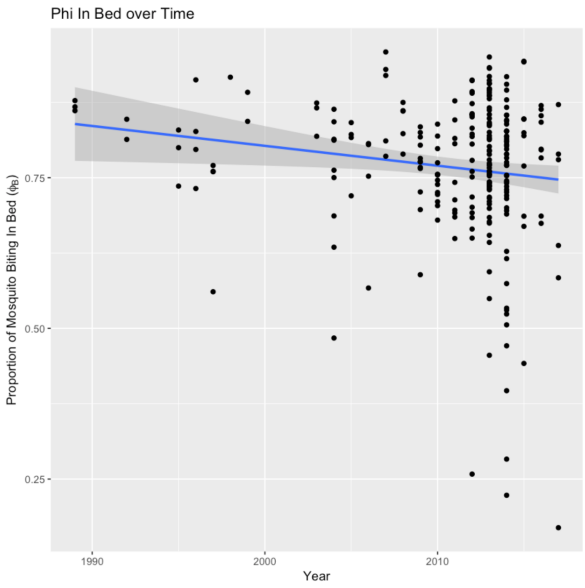
protected by LLINs and IRS in bed. While conducting a t-test may seem appealing for this conclusion, we can not use that test because both ϕ values are not independent.

After looking at the relationships between the two ϕ values, we decided to recreate two of the plots featured in the paper by Sherrard-Smith, et al.

The figure below is the recreation of a regression that was performed in the paper that showed that there was a slight decrease in indoor biting over time.



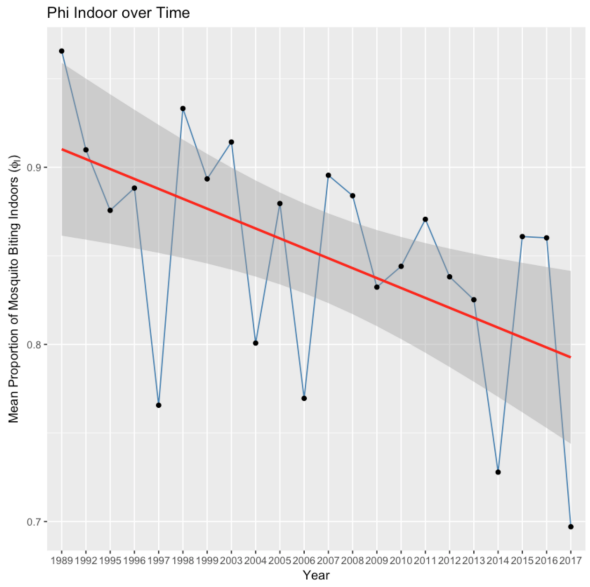
The second figure below is a similar regression that was performed that also demonstrated that there was a slight decrease in mosquito bed biting over time. In the first figure, on average, for every one year increase, the average change in ϕ_I is -0.0028 (p=0.036 with α=0.05 (R)). In the second figure, on average, for every one year increase, the average change in ϕ_B is -0.003 (p=0.018 with α=0.05 (R)). Therefore the decrease over time is significant.



What we can learn from these plots is that a simple linear regression does not fit the model well for a number of reasons. The first, is that the data is highly skewed to later years. The primary thing this plot is showing us, is that there was more data collected in years after 2010 as opposed to prior to 2010. These plots without the regression lines can be used to demonstrate that there were more data points in later years. Because we (and the authors of the paper aforementioned) do not have specific dates for when this data was collected but rather just have the year. If the temporal resolution was more granular, we could perform a time series decomposition to determine if there is a seasonal component, what the trend of the data is, and we could estimate the distribution of the noise. However, the analysis as above, may not be telling the whole story.

To potentially get a better idea of the true average change in the ϕ's that isn't influenced by the number of data points, we can look at the average points for each year. First,we started by grouping the data by country and year and averaging the ϕ values. The reason for this, is that the original dataset contained

a couple of "duplicate" values from various studies. In other words, if paper 1 has data for Benin in 2008 and paper 2 also has data for Benin in 2008, then we averaged the ϕ values for Benin in 2008 and kept just that. We then reran the same regressions on the averaged yearly data.



After trying the yearly average technique, we see that the regression line looks a bit better however, on average, for every one year increase, the average change in the average yearly ϕ_I value, is -0.0045 (p=0.007 with α=0.05). Therefore the relationship is significant.

For in bed (see Appendix), we found that on average, for every one year increase, the average change in the average yearly ϕ_B value, is -0.0044 (p=0.015 with α=0.05). So, though we would argue that the regression line looks better, we see that there is still a significant decrease over time.

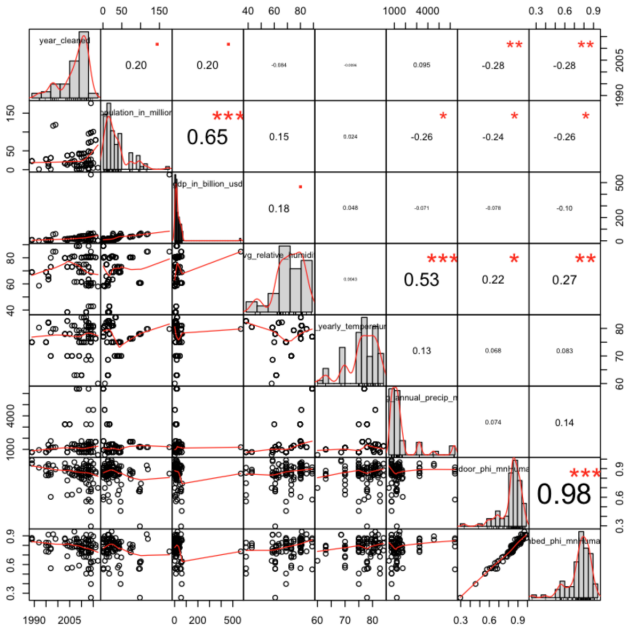
Given that these simple linear regressions of ϕ versus time potentially do not return accurate results, we decided to shift directions away from the paper and perform our own exploratory analysis. In particular, we used variables not included in the paper such as population, GDP, average humidity, average temperature, and average yearly precipitation. We conducted a multiple linear regression, corrected the assumptions, performed a sensitivity analysis, and checked for autocorrelation.

4 Results

As mentioned above, our primary analysis was a multiple linear regression.

4.1 Correlation

To determine which variables we should include in our analysis, we began by looking at a correlation plot of all of our variables:



Diagonally, the variables are: year (1989-2017), population (in millions of people), GDP (in billions of USD), average relative humidity (as a percentage), average yearly temperature (in Fahrenheit), average yearly precipitation (in millimeters), the proportion of mosquito bites indoors, and the proportion of mosquito bites in bed.

The most important things to note about the correlation plot above is the correlation and the significance. First, we notice that the correlation between ϕ_B and ϕ_I is very high at 0.98. This relationship was elaborated on in the Methods section.

For another, we see that population and GDP are highly correlated (0.65). Note that GDP is defined as

$$GDP = C + G + I + NX$$

where C is total consumption, G is total government expenditure, I is the sum of investments, and NX are net exports. The definition of GDP makes it clear why population is correlated with it. Though population is not explicitly a part of the formula, it is implicitly a component. In fact, if we look at the scatter plot of population and GDP, we see that the values are nearly perfectly correlated except for at least one outlier.

Another value of note is the correlation between relative humidity and precipitation. This relationship seems logical as when it is more humid, it rains more or vise versa.

Now, the relationships between the other predictors and the ϕ values are what we are most interested in looking at. First, note that there is a slightly negative relationship between year and both ϕ values. It is interesting to note that there appears to be a weakly negative relationship between year and the proportion of biting indoors/in bed (i.e. we see lower ϕ values for later years).

We can also see that population has a weakly negative relationship with ϕ. It's possible that countries with more people have better access to protection against mosquitoes such as nets and insecticide. Although, looking at the scatter plot of the relationship between the two, we see that the relationship is likely not significant and/or non-linear.

Finally, if we look at average relative humidity, we see a weak positive relationship with the ϕ values. Once again, this is reasonable as we would expect to see more mosquitoes in more humid regions.

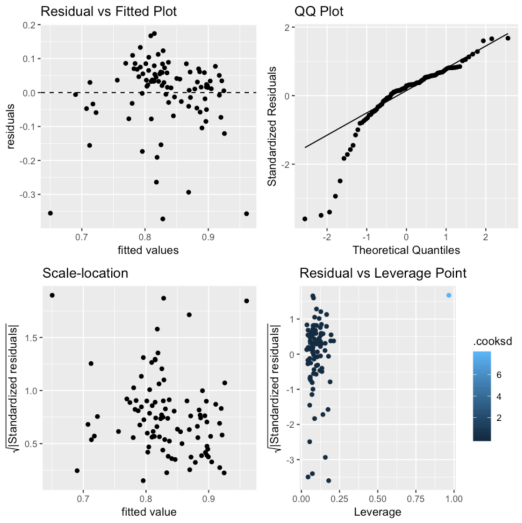
4.2 Multiple Linear Regression

We then focused on two multiple linear regressions:

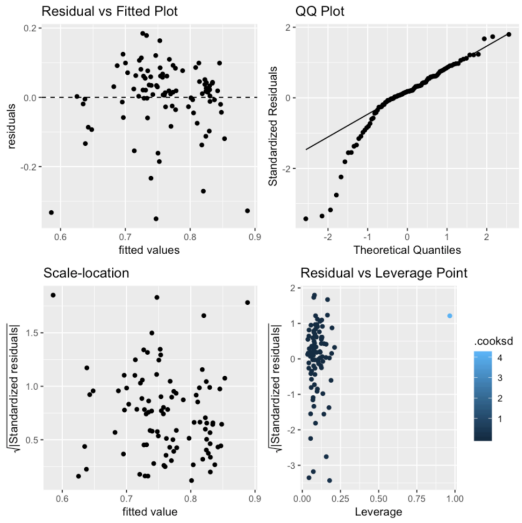
$$\begin{aligned}\phi_I &= \beta_0 + \beta_1 Y + \beta_2 S + \beta_3 P + \beta_4 G + \\ &\quad + \beta_5 H + \beta_6 T + \beta_7 Pr + \varepsilon \\ \phi_B &= \beta_0 + \beta_1 Y + \beta_2 S + \beta_3 P + \beta_4 G + \\ &\quad + \beta_5 H + \beta_6 T + \beta_7 Pr + \varepsilon\end{aligned}$$

where Y= year, S = species (a factor variable), P = population, G = GDP, H = humidity, T = temperature, and Pr = precipitation. Note that we excluded country and the other ϕ from each regression. The reason for excluding country is that H, T and Pr are all the same for each country and year combo. We excluded the other ϕ value from each regression because the ϕ values were highly correlated and were both derived values.

The plots that follow are the standard 4 plots for checking regression assumptions.



In the figure above, looking at the residuals versus fitted values, we see that we do not appear to satisfy linearity because there appears to be a slight parabolic curve. We mostly satisfy constant variance (some clustering but not too bad). We also note from the QQ plot that the residuals and data are very clearly not normal. The final assumption is that the residuals are independent. The figures above do not provide any insight into this but later in this section, we'll dive into checking and correcting this assumption.



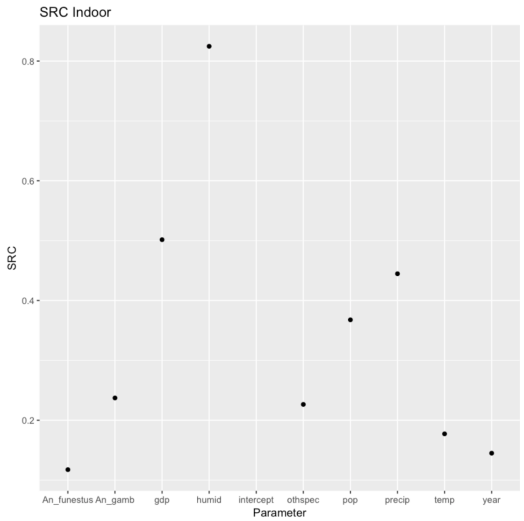
Note, the figures demonstrate the ϕ_I and ϕ_B assumptions respectively. Because both ϕ values are highly correlated, we see that there isn't a significant difference between the plots above and the plots for ϕ_I. We again, do satisfy homoskedasticity but don't satisfy linearity and do not satisfy normality.

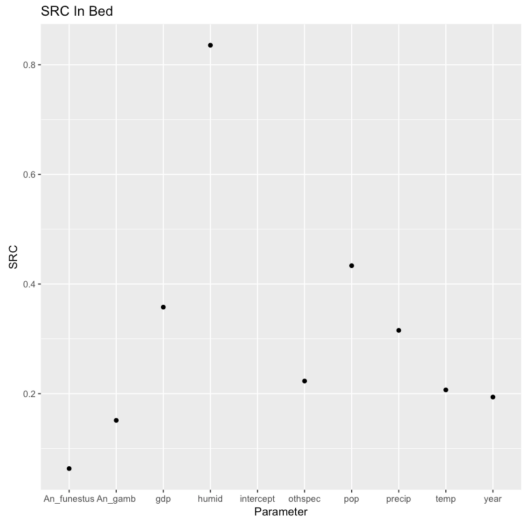
Note, also that in both figures, by studying the residual vs leverage point plot, we see one blue point in the upper right corner. That value corresponds to Nigeria in 2014. Because it was so influential (the Cook's Distance ¿ 1), we chose to exclude that value.

In the following sections, we determine what the best parameters are and whether we should be using a reduced or full model.

4.3 Sensitivity

Here, we conducted a sensitivity analysis on the parameters to determine how the model output changes with varying parameter inputs. If a small change in the parameter value results in a big model output change, then the output is sensitive to that parameter [5]. Our sensitivity results can be seen in the following SRC plots.





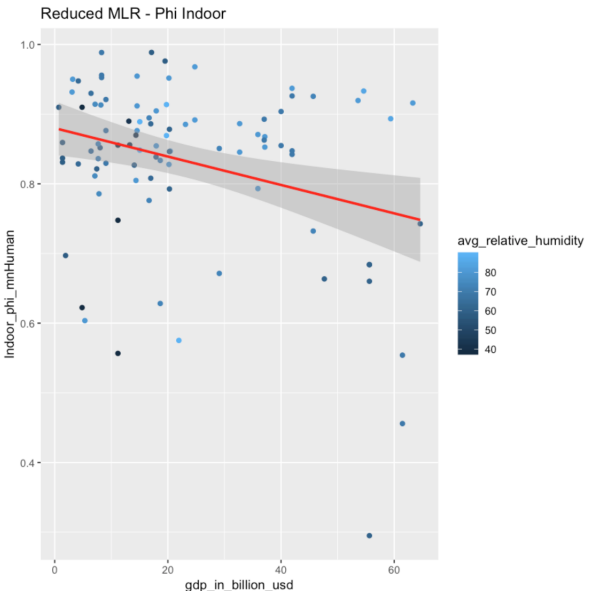
From these plots, it is clear that humidity is the most sensitive parameter in both models. This is not surprising because it is the only significant parameter in our new model that does not contain the influential point. Another interesting thing to take note is that population (p=0.250) is less sensitive than GDP (p=0.152) in the indoor model while population (p=0.004) is more sensitive than GDP (p=0.477) in the in bed model. These align with the p-values for the parameter values that we obtained in our multiple regression model because the lower the p-value of the parameter the higher the sensitivity of the parameter. This indicates that as the parameter becomes more significant, its sensitivity increases.

4.4 Partial F-test

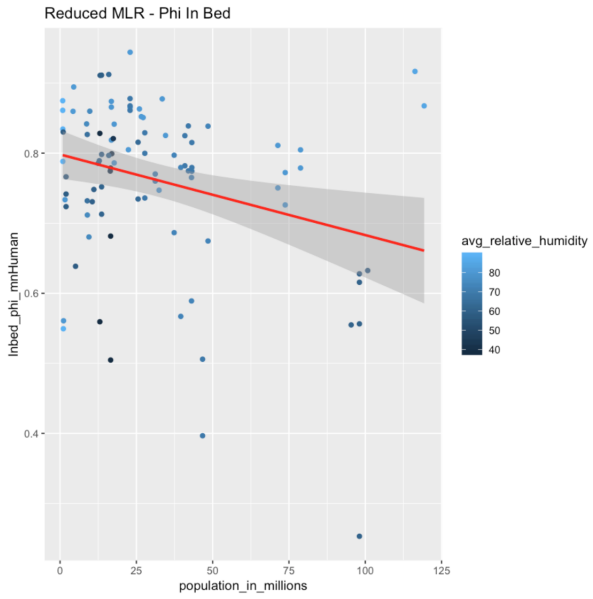
In this section, we decided to explore the possibility of reducing our model. Since humidity was the only significant parameter in our model, we decided that correlation between parameters may be causing parameters to not be significant in the model. With that said, we decided to remove all of the parameters except for GDP and humidity in the indoor model and to remove all of the parameters except for population and humidity in the in bed model. So the reduced models are:

$$\phi_I = \beta_0 + \beta_1 G + \beta_2 H + \varepsilon$$
$$\phi_B = \beta_0 + \beta_1 P + \beta_2 H + \varepsilon$$

These were the respective parameters that we decided to use because they had the highest sensitivity measures. This test went well and both reduced models were significant based on the partial F-test conducted which means that we should be using these models instead of the full models. Below, we see the results of the multiple linear regression.



For the figure above, we can see that a reduced multiple linear regression doesn't quite fit the data well, however we find that this is the best result given the parameters we included in our analysis.



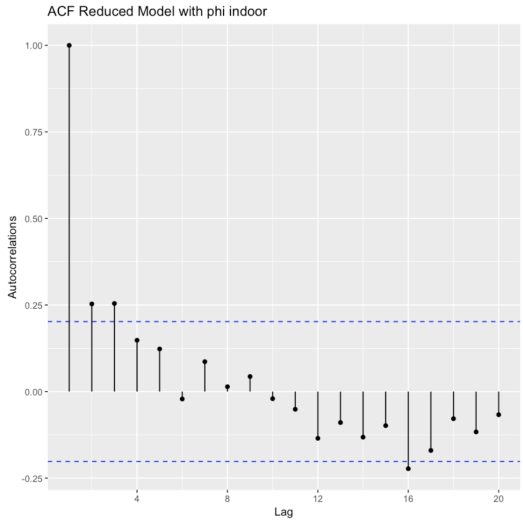
Similarly here, we see that our model doesn't fit the data perfectly but it is reasonable. In this figure in particular, it's possible that the data is skewed by countries that have larger populations.

4.5 Normality and Linearity Assumptions

Unfortunately, reducing our model did not correct the normality and linearity assumptions that are being violated. To correct these violations, we decided to explore transformations of specific parameters and attempted to utilize the box-cox function in R to see how the response could be transformed. These transformations did not provide any better results so we attempted to see if an interaction term was present in the model. We decided to look at the interaction of the mosquitoes species and humidity to see if different species lived in more humid areas which in turn would result in different proportions. This attempt at finding an interaction term was also unsuccessful and insignificant so we ultimately decided that the error terms were correlated because of the temporal component at the level of years and may be causing the normality and linearity assumptions to not be upheld.

4.6 Auto Correlation Function

As mentioned in the Multiple Linear Regression subsection, the final assumption that needed to be checked is that the error terms are uncorrelated. Because the data included a temporal component at the level of years, we investigated whether the residuals were correlated with residuals at previous time steps (lag(h)) when looking at the two reduced models.

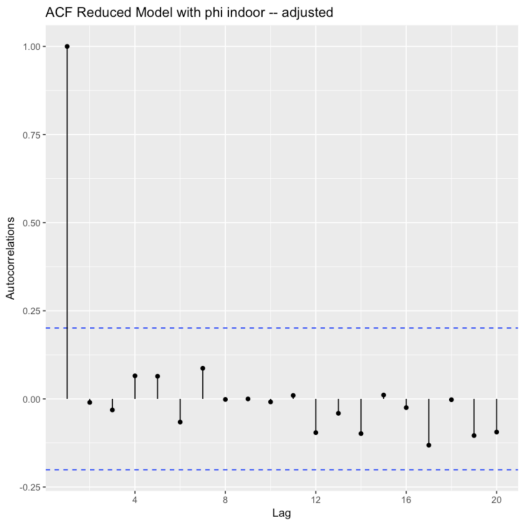


The figure above indicates that we are seeing autocorrelation at up to lag(2). This means that the residuals are correlated between time t and $t-2$. The autoregressive model of order 2 (AR(2)) is:

$$X_t = \mu + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \varepsilon_t$$

where X_t is the time series (in our case, the residuals of the MLR of the proportion of peo-

ple being bit inside) and φ_1 and φ_2 are constants. After adjusting for the errors having autocorrelated terms, we see the new ACF looks much better:



The figure above tells us that we do not have any significant correlation between time steps t and $t - h$. For similar plots of in bed mosquito biting residual ACFs, please refer to the Appendix.

5 Conclusion

Overall, this paper provided insight into how indoor and in bed mosquito bites can be attributed to a country's attributes, but we unfortunately cannot make any definite conclusions. This is because of the fact that all of our data was not only all observational, but assumptions were made about the responses and predictors such as how ϕ is calculated and how respective values for years are grouped together. Furthermore, we were not able to completely correct the implied non-linearity relationship given in the residuals versus fitted values plot and the normality assumption was still not upheld after creating our new model.

Even with all of these caveats, we still learned a lot from this project. We learned how to conduct an effective exploratory data analysis, how to check for linear regression assumptions in others' work, how to conduct a sensitivity analysis, how to utilize autocorrelation, and most importantly we were able to see how material that we learned in class can be applicable to real world research questions. Additionally, we ultimately learned that in order to model ϕ for in bed and indoors the most significant model is the reduced model that only has humidity and GDP or population as predictors. This is quite shocking considering neither were in the original dataset. Overall, we believe that we corrected for the error made in the original paper and were able to create our own model that more accurately and correctly portrayed the data.

To improve the design of this study, we would correct how the study used many data points for a given year composed of different sources. By doing this, they introduced additional error by not keeping the data collection consistent and they skewed their data towards years with more data points. We would also request that data points were measured by not only year, but month or week so that seasonality could be accounted for. Finally, we would try to find more data points from earlier years so that a more accurate analysis can be made for prior years.

Finally, if time permitted or if we decide to return to this research, we would would look into how a non-parametric methodology could potentially provide a better result and if it could correct the error violations that were found. In addition to this, we would also investigate how error terms may be spatially correlated.

References

[1] "Malaria." https://www.who.int/news-room/fact-sheets/detail/malaria, January 2020.

[2] E. Sherrard-Smith, J. E. Skarp, A. D. Beale, C. Fornadel, L. C. Norris, S. J. Moore, S. Mihreteab, J. D. Charlwood, S. Bhatt, P. Winskill, *et al.*, "Mosquito feeding behavior and how it influences residual malaria transmission across africa," *Proceedings of the National Academy of Sciences*, vol. 116, no. 30, pp. 15086–15095, 2019.

[3] D. B. Larremore, A. Clauset, and C. O. Buckee, "A network approach to analyzing highly recombinant malaria parasite genes," *PLoS computational biology*, vol. 9, no. 10, 2013.

[4] A. J. MacDonald and E. A. Mordecai, "Amazon deforestation drives malaria transmission, and malaria burden reduces forest clearing: a retrospective study," *The Lancet Planetary Health*, vol. 3, p. S13, 2019.

[5] X. Zhou and H. Lin, *Sensitivity Analysis*, pp. 1046–1048. Boston, MA: Springer US, 2008.

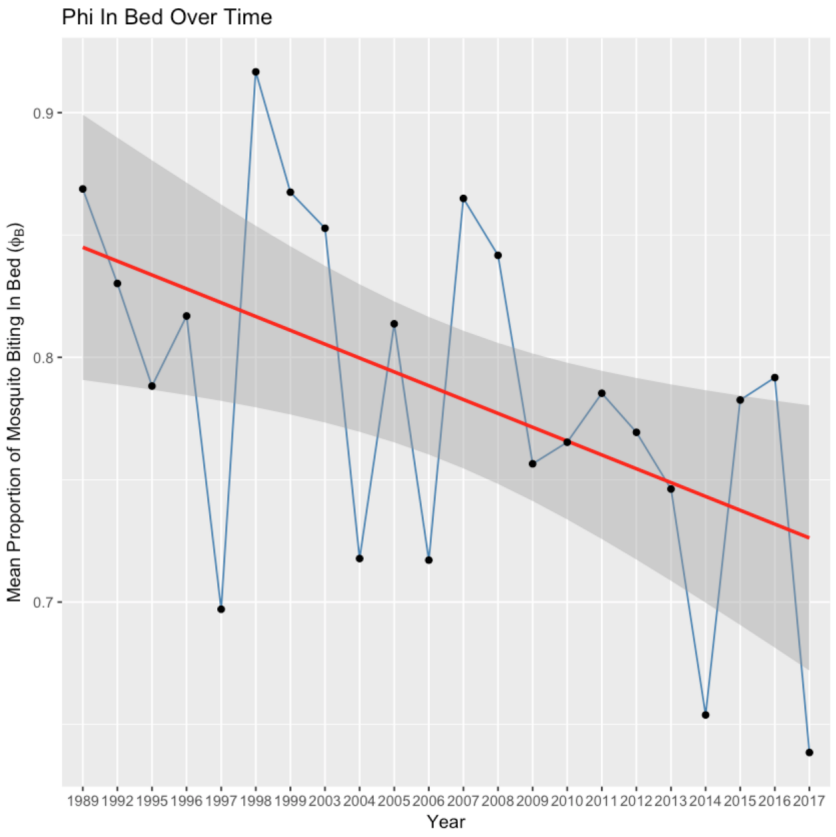
[6] "Interpreting residual plots to improve your regression."

[7] D. A. Pfeffer, T. C. Lucas, D. May, J. Harris, J. Rozier, K. A. Twohig, U. Dalrymple, C. A. Guerra, C. L. Moyes, M. Thorn, *et al.*, "malariaatlas: an r interface to global malariometric data hosted by the malaria atlas project," *Malaria journal*, vol. 17, no. 1, pp. 1–10, 2018.

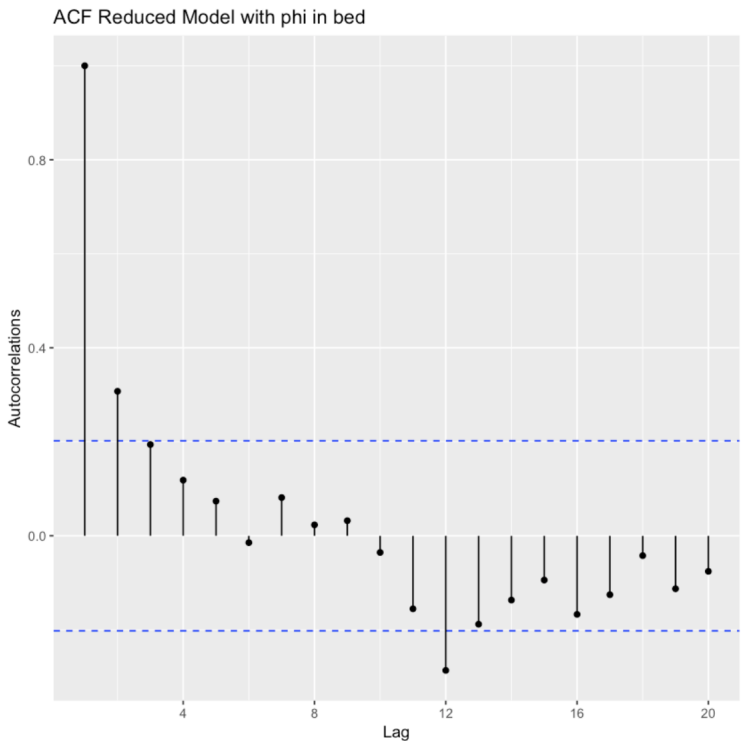
6 Appendix

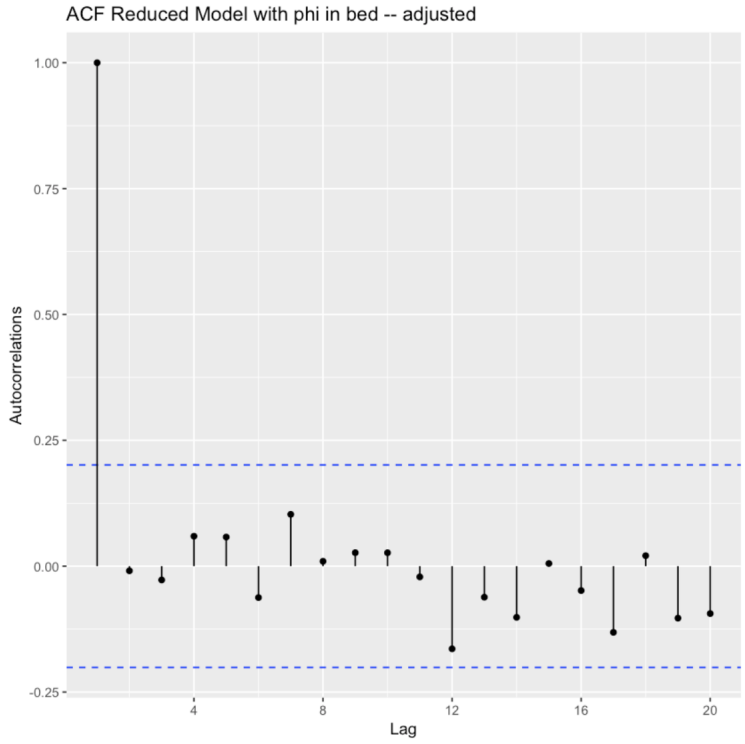
6.1 Plots

Below is the figure of the regression on the averaged yearly in bed ϕ values.



The ACF for in bed is very similar to the ACF of indoor. The two following plots indicate that we also see that the residuals can be modeled with AR(2) process and when we account for that, the ACF looks much better.


ACF Reduced Model with phi in bed
Autocorrelations
0.8
0.4
0.0
4
8
12
16
20
Lag


ACF Reduced Model with phi in bed -- adjusted
Autocorrelations
1.00
0.75
0.50
0.25
0.00
-0.25
4
8
12
16
20
Lag

6.2 Code

```
install.packages("PerformanceAnalytics")
library(ggplot2)
library(dplyr)
library(latex2exp)
library(PerformanceAnalytics)
data = read.table("data/mean_phi_by_country_and_year.csv",
sep =',', header = TRUE)
data = data[ , !(names(data) %in% c('X'))]
head(data,5)
p <- ggplot(data, aes(x=year_cleaned, y=Indoor_phi_mnHuman)) +
    geom_smooth(method = "lm") +
    geom_point() + xlab('Year')+
    ylab(TeX("Mean Proportion of Mosquito Biting Indoors Across
    Studies($\\phi_I$)")) +
    ggtitle('Phi Indoor over Time')
p
summary(lm(data$Indoor_phi_mnHuman~data$year_cleaned))
p <- ggplot(data, aes(x=year_cleaned, y=Inbed_phi_mnHuman)) +
    geom_smooth(method = "lm") +
    geom_point() + xlab('Year')+
    ylab(TeX("Mean Proportion of Mosquito Biting In Bed Across
    Studies($\\phi_I$)")) +
    ggtitle('Phi In Bed over Time')
p
summary(lm(data$Indoor_phi_mnHuman~data$year_cleaned))
#Phi Indoor
years=c("1989","1992","1995","1996","1997","1998","1999",
        "2003","2004","2005","2006","2007","2008","2009",
        "2010","2011","2012","2013","2014","2015","2016",
        "2017")
j=1
en = c()
for(k in years){
    en[j]=with(data,mean(Indoor_phi_mnHuman[year_cleaned==k]))
    j=j+1
}
df=data.frame(years,en)
p <- ggplot(df, aes(x=years, y=en,group=1)) +
 geom_line( color="steelblue") +
 geom_smooth(method = "lm", color='red') +
  geom_point() +
  xlab('Year')+ ylab(TeX("Mean Proportion of Mosquito Biting
```

```
  Indoors ($\\phi_I$)")) +
    ggtitle('Phi Indoor over Time')
p
#Phi Inbed
years=c("1989","1992","1995","1996","1997","1998","1999",
        "2003","2004","2005","2006","2007","2008","2009",
        "2010","2011","2012","2013","2014","2015","2016",
        "2017")
j=1
en = c()
for(k in years){
    en[j]=with(data,mean(Inbed_phi_mnHuman[year_cleaned==k]))
    j=j+1
}
df=data.frame(years,en)
p <- ggplot(df, aes(x=years, y=en,group=1)) +
 geom_line( color="steelblue") +
 geom_smooth(method = "lm", color='red') +
  geom_point() +
  xlab('Year')+ ylab(TeX("Mean Proportion of
  Mosquito Biting In Bed ($\\phi_B$)")) +
    ggtitle('Phi In Bed Over Time')
p
chart.Correlation(data[, c('year_cleaned',
                        'population_in_millions',
                        'gdp_in_billion_usd',
                         'avg_relative_humidity',
                         'avg_yearly_temperature_F',
                         'avg_annual_precip_mm',
                         'Indoor_phi_mnHuman',
                        'Inbed_phi_mnHuman')], histogram=TRUE, pch=19)
                        lmod = lm(Indoor_phi_mnHuman ~ . - Country_clean
          - Inbed_phi_mnHuman, data = data)
summary(lmod)

tmp<- melt(data[,c('Indoor_phi_mnHuman', 'Inbed_phi_mnHuman')])
ggplot(tmp,aes(x=value, fill=variable)) + geom_density(alpha=0.25)

#without point 67

reminf=data[-c(67),]
lmod = lm(Indoor_phi_mnHuman ~ . - Country_clean
          - Inbed_phi_mnHuman, data = reminf)
summary(lmod)
par(mfrow=c(2,2))
```

```
plot(lmod)
othlmod=lm(Inbed_phi_mnHuman ~ . - Country_clean
          - Indoor_phi_mnHuman, data = data)
summary(othlmod)
par(mfrow=c(2,2))
plot(othlmod)
othlmod=lm(Inbed_phi_mnHuman ~ . - Country_clean
          - Indoor_phi_mnHuman, data = reminf)
summary(othlmod)
par(mfrow=c(2,2))
plot(othlmod)
library(caret)
install.packages("ggthemes")
library(ggthemes)
install.packages("sensitivity")
library(sensitivity)
library(boot)
intercept=c(rep(3.041e+00 ,94))
year=reminf$year_cleaned*-1.274e-03
newdata=reminf
newdata <- newdata %>%
      mutate(species = ifelse(species == "An_funestus",1,0))
An_funestus=newdata$species*1.614e-02

gamb=reminf
gamb <- gamb %>%
      mutate(species = ifelse(species == "An_gambiae_sl",1,0))
An_gamb=gamb$species*2.886e-02

other=reminf
other <- other %>%
      mutate(species = ifelse(species == "other",1,0))
othspec=other$species*-4.119e-02

pop=reminf$population_in_millions*-7.770e-04
gdp=reminf$gdp_in_billion_usd*-1.736e-03
humid=reminf$avg_relative_humidity*4.044e-03
temp=reminf$avg_yearly_temperature_F*2.019e-03
precip=reminf$avg_annual_precip_mm*-1.949e-05


X <- data.frame(intercept,year,An_funestus,An_gamb,othspec,pop,gdp,humid,temp,precip)
head(X)

# linear model : Y = X1 + X2 + X3
```

```
y <- with(X, intercept + year + An_funestus + An_gamb +
othspec + pop + gdp + humid + temp + precip)


# # sensitivity analysis

x <- src(X, y, rank=FALSE,  nboot = 0, conf=0.95)

print(x)
plot(x,main="SRC Indoor Phi")
intercept=c(rep( 3.791e+00,94))
year=reminf$year_cleaned*-1.703e-03
othnewdata=reminf
othnewdata <- othnewdata %>%
      mutate(species = ifelse(species == "An_funestus",1,0))
An_funestus=othnewdata$species*8.702e-03

othgamb=reminf
othgamb <- othgamb %>%
      mutate(species = ifelse(species == "An_gambiae_sl",1,0))
An_gamb=othgamb$species*1.840e-02

othother=reminf
othother <- othother %>%
      mutate(species = ifelse(species == "other",1,0))
othspec=othother$species*-4.056e-02

pop=reminf$population_in_millions*-9.155e-04
gdp=reminf$gdp_in_billion_usd*-1.238e-03
humid=reminf$avg_relative_humidity*4.096e-03
temp=reminf$avg_yearly_temperature_F*2.355e-03
precip=reminf$avg_annual_precip_mm*-1.382e-05


X <- data.frame(intercept,year,An_funestus,An_gamb,othspec,pop,gdp,humid,temp,precip)
head(X)

# linear model : Y = X1 + X2 + X3

y <- with(X, intercept + year + An_funestus + An_gamb +
othspec + pop + gdp + humid + temp + precip)


# # sensitivity analysis
```

```
x <- src(X, y, rank=FALSE,  nboot = 0, conf=0.95)

print(x)

plot(x)
#indoor
mse1=mean(lmod$residuals^2)

mse1

#inbed
mse2=mean(othlmod$residuals^2)

mse2
redlmod = lm(Indoor_phi_mnHuman ~ . - Country_clean
          - Inbed_phi_mnHuman - year_cleaned -
          species - population_in_millions - avg_yearly_temperature_F
             - avg_annual_precip_mm, data = reminf)
par(mfrow=c(2,2))
plot(redlmod)
summary(redlmod)

redothlmod=lm(Inbed_phi_mnHuman ~ . - Country_clean
          - Indoor_phi_mnHuman - year_cleaned -
          species - gdp_in_billion_usd - avg_yearly_temperature_F
             - avg_annual_precip_mm, data = reminf)
summary(redothlmod)
plot(redothlmod)
#inbed
anova(redlmod,lmod)

#indoor
anova(redothlmod,othlmod)
#indoor
newmse1=mean(redlmod$residuals^2)

newmse1

#inbed
newmse2=mean(redothlmod$residuals^2)

newmse2
lmodinteract = lm(Indoor_phi_mnHuman ~ . - Country_clean
          - Inbed_phi_mnHuman + species:avg_relative_humidity, data = reminf)
```

```
summary(lmodinteract)
anova(lmod, lmodinteract)
othlmodinteract = lm(Inbed_phi_mnHuman ~ . - Country_clean
          - Indoor_phi_mnHuman + species:avg_relative_humidity, data = reminf)
summary(othlmodinteract)
anova(othlmod, othlmodinteract)
par(mfrow=c(2,2))
plot(lmod)
bacf = acf(lmod$residuals, plot=FALSE)
alpha <- 0.95
conf.lims <- c(-1,1)*qnorm((1 + alpha)/2)/sqrt(bacf$n.used)

bacf$acf %>%
  as_tibble() %>% mutate(lags = 1:n()) %>%
  ggplot(aes(x=lags, y = V1)) +
  scale_x_continuous(breaks=seq(0,41,4)) +
  geom_hline(yintercept=conf.lims, lty=2, col='blue') +
  labs(y="Autocorrelations", x="Lag", title= "ACF") +
  geom_segment(aes(xend=lags, yend=0)) +geom_point()
  #+ theme_setting
  arima_fit <- with(data, arima(Indoor_phi_mnHuman, order = c(1, 0, 0),
                xreg = cbind(year_cleaned, population_in_millions,
                gdp_in_billion_usd,
                             avg_relative_humidity,
                             avg_yearly_temperature_F,
                             avg_annual_precip_mm
                            )))
                            bacf = acf(arima_fit$residuals, plot=FALSE)
alpha <- 0.95
conf.lims <- c(-1,1)*qnorm((1 + alpha)/2)/sqrt(bacf$n.used)

bacf$acf %>%
  as_tibble() %>% mutate(lags = 1:n()) %>%
  ggplot(aes(x=lags, y = V1)) + scale_x_continuous(breaks=seq(0,41,4)) +
  geom_hline(yintercept=conf.lims, lty=2, col='blue') +
  labs(y="Autocorrelations", x="Lag", title= "ACF") +
  geom_segment(aes(xend=lags, yend=0)) +geom_point() #+ theme_setting
  par(mfrow=c(2,2))
arima_fit
library(nlme)
genls = gls(Indoor_phi_mnHuman ~ . - Country_clean - Inbed_phi_mnHuman,
            data = data,
            correlation=corARMA(p=1,q=0))
summary(genls)
par(mfrow=c(2,2))
```

```
plot(genls$fitted, genls$residuals)
abline(h=0)
qqnorm(genls$residuals)
qqline(genls$residuals)
lmod2 = lm(Indoor_phi_mnHuman ~ log(population_in_millions)+
gdp_in_billion_usd+
avg_relative_humidity+avg_yearly_temperature_F+
avg_annual_precip_mm+Indoor_phi_mnHuman - Country_clean
          - Inbed_phi_mnHuman, data = data)
head(data)
summary(lmod2)
par(mfrow=c(2,2))
plot(lmod2)
data[67,]
moddata=data[-c(67),]
lnewmod2 = lm(Indoor_phi_mnHuman ~ population_in_millions-gdp_in_billion_usd+
avg_relative_humidity-avg_yearly_temperature_F-
avg_annual_precip_mm - Country_clean
          - Inbed_phi_mnHuman, data = moddata)

summary(lnewmod2)
par(mfrow=c(2,2))
plot(lnewmod2)
library(MASS)
boxcox(lnewmod2, lambda = seq(-2, 2, 1/10),
plotit = TRUE, eps = 1/50, xlab = expression(lambda),
       ylab = "log-Likelihood")

## Plot MLR
p = ggplot(reminf, aes(y=Indoor_phi_mnHuman,x=gdp_in_billion_usd,
    color=avg_relative_humidity))+
    geom_point()+
    geom_smooth(method = "lm", color='red') +
    ggtitle('Reduced MLR - Phi Indoor')
p
p = ggplot(reminf, aes(y=Inbed_phi_mnHuman,x=population_in_millions,
    color=avg_relative_humidity))+
    geom_point()+
    geom_smooth(method = "lm", color='red') +
    ggtitle('Reduced MLR - Phi In Bed')
p
```