

Utah State University

DigitalCommons@USU

All Graduate Plan B and other Reports

Graduate Studies

5-2015

Comparing Linear Mixed Models to Meta-Regression Analysis in the Greenville Air Quality Study

Lynsie M. Daley
Utah State University

Follow this and additional works at: <https://digitalcommons.usu.edu/gradreports>



Part of the [Mathematics Commons](#), and the [Statistics and Probability Commons](#)

Recommended Citation

Daley, Lynsie M., "Comparing Linear Mixed Models to Meta-Regression Analysis in the Greenville Air Quality Study" (2015). *All Graduate Plan B and other Reports*. 648.

<https://digitalcommons.usu.edu/gradreports/648>

This Report is brought to you for free and open access by the Graduate Studies at DigitalCommons@USU. It has been accepted for inclusion in All Graduate Plan B and other Reports by an authorized administrator of DigitalCommons@USU. For more information, please contact digitalcommons@usu.edu.



COMPARING LINEAR MIXED MODELS TO META-REGRESSION ANALYSIS IN THE
GREENVILLE AIR QUALITY STUDY

By

Lynsie M. Daley

A report submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Statistics

Approved:

Dr. John R. Stevens
Major Professor

Dr. Adele Cutler
Committee Member

Dr. Christopher D. Corcoran
Committee Member

Utah State University
Logan, Utah

2015

Copyright © Lysie M. Daley 2015
All Rights Reserved

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Stevens, for his continued support of me as a student both in my undergraduate years and the time I spent as a graduate student. He never gave up on me, and always gave me encouragement when things got tough. My family and friends deserve a huge expression of appreciation, all of them, but especially a few special people. From a young age, my mother told me often that she knew I was going to do great things in the world. She hasn't stopped telling me that and her belief in me as a student and as a person continues to inspire me. My husband, Jeremy, provided so much support to me at home while I was trying to complete my project and that made all the difference. I also want to thank Bill Francis, who was a great support through the whole process and went out of his way to provide some great resources to me. I dedicate this to my daughter, Hadleigh, who I hope I can always be a good example to.

ABSTRACT

Comparing Linear Mixed Models to Meta-Regression Analysis in the Greenville Air Quality Study

by

Lynsie M. Daley

Utah State University, 2015

Major Professor: Dr. John R. Stevens
Department: Mathematics and Statistics

The effect of air quality on public health is an important issue in need of better understanding. There are many stakeholders, especially in Utah and Cache Valley, where the poor air quality as measured by PM_{2.5} levels and consequent inversions can sometimes be the very worst in the nation. This project focuses on comparing two statistical methods used to analyze an important air quality data set from the Greenville Air Quality Study, focusing on a lung function response variable. A linear mixed model, with a random factor for subject, gives slope estimates and their significance for predictor variables of interest, especially PM_{2.5} levels. The method of meta-regression in this analysis is extended from looking at multiple studies to looking at multiple subjects from the single air quality study and the effect of PM_{2.5} on their lung function separately, finally combining the results using a model where the slope estimates for PM_{2.5} act as the response. With other predictors mean-centered, this meta-regression allows for interpretation of the model intercept as the overall mean effect size of PM_{2.5} on lung function. Both statistical methods were studied in depth in order to apply them appropriately to the data

set. The primary goal of applying both of these methods, aside from comparing their results, was to determine what significant role, if any, the PM 2.5 pollution levels played in the lung function of students after a 20 minute outdoor recess, therefore validating the results of previous analyses of the data.

CONTENTS

ACKNOWLEDGEMENTS	3
ABSTRACT.....	4
CONTENTS.....	6
LIST OF FIGURES	7
LIST OF TABLES	8
ABOUT THE GREENVILLE AIR QUALITY STUDY	9
LINEAR MIXED MODELS	13
META-REGRESSION ANALYSIS.....	18
RESULTS	21
DISCUSSION	35
REFERENCES.....	37
APPENDIX A – R SOURCE CODE.....	39

LIST OF FIGURES

Figure		Page
1	Histogram of n_i values	22
2	Histogram of FEF difference and normal Q-Q plot of FEF difference	23
3	Normal Q-Q plot of residuals and residuals vs. fitted values from mixed model	25
4	Histogram of residuals from mixed model	25
5	Boxplots of FEF difference by ethnicity	26
6	Meta-regression residuals vs. fitted and normal Q-Q plot of residuals	31
7	Meta-regression histogram of residuals	31
8	Residual and normal Q-Q plots of residuals of transformed Y	32
9	Histogram of residuals of transformed Y	32
10	Recess Guidelines for Utah Schools on Air Quality	36

LIST OF TABLES

Table		Page
1	Lung function measurements defined	10
2	AIC values for correlation structures in the mixed model	21
3	Variables included in full linear mixed model	23
4	Effect estimates for final linear mixed model	24
5	Results of PM 2.5 and asthma model.....	28
6	Results of PM 2.5 and indoor recess model.....	28
7	Covariates included in individual linear regression models	28
8	List of constant covariates included in meta-regression model.....	29
9	Partial matrix to set up meta-regression, with a row for each subject, and predictor columns mean-centered	30
10	Meta-regression results	33

1. ABOUT THE GREENVILLE AIR QUALITY STUDY

The Greenville Air Quality study, conducted at Greenville Elementary in North Logan, Utah, was carried out between January and March of 2007. The timing was important, since elevated PM 2.5 levels are a seasonal problem and typically peak during the winter months. The study was designed to answer an important public health question: Is the exposure to PM 2.5 during a 20 minute recess harmful enough to warrant keeping children inside when air quality is poor (Redd, 2015)? PM 2.5, or particulate matter less than 2.5 microns in width, are tiny particles or droplets that cause the air to appear hazy when levels become elevated. The concern they present is their ability to travel deep into the respiratory tract and reach the lungs (New York Department of Health, 2011). Exposure to these particles has been linked to eye, nose, and throat irritation, as well as worsening asthma, lung function, and heart disease (New York Department of Health, 2011). In Cache Valley, these particles are comprised mainly of ammonium nitrate; however, it is mainly the particle itself and not the chemical composition that causes damage (Redd, 2015).

Two years prior to the Greenville study, a pilot study was conducted at Hawthorne Elementary in Salt Lake City. This study was small and had little power for an analysis, but there they learned how to correctly measure for lung function and capacity (Packham, 2011). There were no other acute exposure studies in the literature, which prompted Dr. Edward Redd and the Bear River Health Department to design a more powerful study to capture both lung function and air quality data (Redd, 2015).

The Bear River Health Department partnered with Greenville Elementary and recruited over 100 student volunteers from grades 3-5. Each student's family submitted a demographic

Table 1 – Lung function measurements defined (Morgan Scientific, 2013)

Measurement:	Definition:	Example:
Forced Expiratory Flow (25-75%)	The average rate of airflow during the mid-portion of forced vital capacity.	
Peak Expiratory Flow Rate	The maximum flow at the outset of forced expiration. Measured in liters/second	
Forced Vital Capacity	The total volume of air expired after a full inspiration with maximum effort	
Forced Expiratory Volume, 1 Second	The volume of air expired in the first second during maximum expiratory effort.	

questionnaire, and each day before and after recess the students would check in with a table staffed by interns from Utah State University to have their lung function measured by a calibrated spirometer. Approximately half of these students were asthmatic, which provided an opportunity to determine whether this condition created a greater adverse effect from the bad air exposure the students had outdoors. The air quality inside the school library was also measured as a control for the outdoor measurements. In addition, A.M. and P.M. temperatures and humidity were captured each day.

Among the demographic data collected were information on the construction, heating type, and age of the students' home, whether or not they lived with smokers, asthma diagnosis by a physician, and whether the students took medications daily. Ethnicity, language spoken at home, and age were also captured. Information was collected each day on whether each student

was experiencing any respiratory symptoms such as coughing, wheezing, or upper respiratory infections (URIs).

Subject lung function was assessed using a spirometer, with students taking a deep breath and then blowing as hard and fast as they could into the spirometer tube. The lung function measurements collected were forced expiratory flow (FEF 25-75%), peak expiratory flow rate (PEFR), forced vital capacity (FVC), and forced expiratory volume in 1 second (FEV 1). See Table 1 for definitions of these measurements. It is understood that FEF 25-75% and the ratio between FEV 1 and FVC are the most sensitive to pollution particulates (Redd, 2015). Only FEF 25-75% measurements were used for the purposes of this report.

Forced expiratory flow (FEF 25-75%), used as the response variable in this report, is the mean of the flow during the interval of 25%-75% of remaining forced vital capacity, and is measured in liters per second. To understand FEF, forced vital capacity must be defined. Forced vital capacity is the total volume of air expired after a full inspiration with maximum effort. FEF is a flow rate measurement, so the more air can be expelled during the middle portion of forced vital capacity, the healthier the result is. Because this is the case, larger values are healthier than smaller values. Average normal values depend mostly on gender and age, height, mass, and ethnicity. Predicted normal values can be calculated based on those characteristics (Arnall 2015).

There were a couple of factors which affected the data collection in this study. 2007 was a generally mild year for poor air quality with fewer days where the pollution was considered to be at unhealthy levels. Also playing a role in the data collection was the fact that the Principal of Greenville Elementary had a policy of not allowing the children outdoors during recess if the PM 2.5 levels were above 50 micrograms per cubic meter. The primary investigators did not want to

change any protocol or habits at the school during the study, so this was continued, although children were still tested even if they had indoor recess (Redd, 2015).

The data from the Greenville study has since been analyzed by several groups. Steve Packham, PhD found no statistically significant decrease in FEV1/FVC or PEFr function as a response of asthma, PM 2.5, or activity associated with recess. He did however find a significant increase in pre-recess FVC over the course of the study, as well as a significant decrease in FVC in students with URI symptoms (Packham 2011). It is possible that an increase in pre-recess FVC over the course of the study was an artifact of the students learning how to blow into the spirometers and improving over time. These findings in the very least provided evidence that the spirometry methods could detect effects of air quality if there were any. The results led to the Utah Department of Health abandoning previous guidelines and shifting to an hourly PM 2.5 monitoring guideline. Dr. Redd also stated that there was no significant decrease in lung function due to the air quality during the study in Cache Valley. This told him that short term exposure, 20 minutes as the students had, may not be very harmful to most people. It is still unclear how damaging long term exposure to PM 2.5 may be (Redd, 2015).

2. LINEAR MIXED MODELS

Mixed effects models are appropriate for use when there is a mix of different types of factors, both fixed and random. A random factor is defined as a factor which has many levels; all of interest, but only a representative sample of those levels is used in the analysis. Fixed effects include all levels available (Littell, 2002). These models work well for data that are correlated due to a grouping of subjects, repeated measurements over time, or multiple related outcome measurements at one point in time. Mixed models also handle uneven spacing of repeated measurements whether they are intentional or non-intentional. They allow for each level of a random factor to have its own intercept and slope randomly deviating from the mean intercept (Seltman, 2015). Linear mixed models are usually preferred with multiple profiles (subjects) of data and a smaller number of observations per subject (Jensen, 2006). The information from multiple subjects is pooled to improve estimates and inference (Seltman, 2015).

Variance-covariance structures are used in mixed models to describe the correlation present in the responses for a given subject. While the random effect for subject describes only the cause of correlation, a variance-covariance structure describes the type of correlation present (Seltman, 2015). The structures which were compared in the Greenville data analysis included compound symmetry (CS), autoregressive order 1 (AR(1)), and unstructured (SYMM).

Compound symmetry indicates the same correlation between all pairs of measurements. A random intercept for each level is used, which will remove the correlation between the other fixed effects (Seltman, 2015). An example of compound symmetry is given by:

$$\begin{bmatrix} \sigma^2 + \sigma_1 & \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma^2 + \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma^2 + \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 & \sigma^2 + \sigma_1 \end{bmatrix}$$

The autoregressive structure is one where the measurements are ordered, and adjacent measurements are more highly correlated than distant measurements (Seltman, 2015). An example of an AR(1) structure is given by:

$$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

The unstructured covariance structure imposes no restraints to the values. Each variance and covariance is estimated uniquely from the data. This structure usually results in the best fit, but at the cost of using up many degrees of freedom, which can make it difficult to justify (SAS, 2009). This structure is given by:

$$\begin{bmatrix} \sigma_1^2 & \sigma_{21} & \sigma_{31} & \sigma_{41} \\ \sigma_{21} & \sigma_2^2 & \sigma_{32} & \sigma_{42} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{43} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{bmatrix}$$

Either maximum likelihood (ML) estimation or restricted maximum likelihood (REML) estimation may be used when fitting a mixed model. There are advantages and disadvantages to each. In ML, the response variable is assumed to be normal with the mean depending on the model parameters. In REML, the principle of maximum likelihood can be applied to the residuals in order to remove the fixed variable effects. This means that the residual distribution does not depend at all on the fixed effects estimate, just on the variance components. Using ML, the deviance statistic measures the lack of fit the model has to the data, though it is hard to interpret. ML and REML are not robust to the assumptions of normality of the response variables. The deviance statistic given by REML can only be used to compare models that have the same fixed effects and differ only in their random component. Variances tend to be more realistic under REML than they are under ML estimation. When there is a need to compare

models using a variable selection, or model reduction technique, ML must be used because of the absence of the fixed effects in REML estimation (SSI Central, 2015). For this reason, ML was used in the Greenville data model.

As previously stated, the linear mixed model has fixed as well as random effects and is given by:

$$Y_i = \beta_0 + X_i\beta + Z_i\gamma + \epsilon_i, \text{ for } i = 1, 2 \dots m, m = \text{number of subjects}$$

Y_i is a length n_i vector of response measurements for subject i . β_0 is a scalar representing the Y-intercept of the regression model. β is a length p vector of fixed effect coefficients. X_i is a $n_i \times p$ covariate matrix of fixed effects. Z_i is an $n_i \times q$ matrix of predictor variables with random effects. γ is a length q vector of subject-specific random effects. ϵ_i is a length n_i vector of the errors.

The model can have two levels of correlation for the observations within a subject. The first comes from the random effects, γ , which cause all profile (or subject) measurements to be correlated. The second results from the within-profile variance-covariance matrix R_i (Jensen, 2006), due to correlations among error terms within subject. The model is flexible enough to allow errors to be independent or correlated. If correlated, R_i is usually a simple form such as compound symmetry or autoregressive order 1 to decrease the number of covariance parameters to be estimated. Autocorrelation is very common in time ordered data, as seen in the Greenville study.

The assumptions for linear mixed models are as follows:

- $\gamma \sim N_q(0, G)$, G $q \times q$ and positive definite
- $\epsilon_i := \begin{pmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{in_i} \end{pmatrix} \sim N_{n_i}(0, R_i)$, R_i $n_i \times n_i$ and positive definite

- γ independent of $\varepsilon_1, \dots, \varepsilon_m$
 - G = covariance matrix of random effects γ
 - R_i = covariance matrix of error vector ε_i in cluster (or subject) i
 - Z_i and X_i are matrices of known covariates (Spinka, 2007).

Issues with linear mixed models include zero estimates and non-convergence. When a negative estimate of a diagonal component of the variance-covariance matrix is given by the maximum likelihood algorithm, this is set to a zero estimate. It is set to zero, because it cannot be negative, thus causing the matrix to be singular. When no estimates are obtained due to the difficulty of maximizing the likelihood, non-convergence happens. This is more common when the data are unbalanced (Jensen, 2006).

In order to interpret a mixed model, the variance for each random factor should be looked at, as well as the residual variance. When looking at the variance for an individual random effect, it should be close to zero, with the variation falling into the residuals. Following this, the estimates and standard errors for the fixed effects should be examined. P-values are not as straight-forward in mixed models, and it is common to use the likelihood ratio test to obtain them. The `lme()` function in the `nlme` R package (Pinheiro et al., 2015) does this automatically and provides a p-value for each fixed effect in the model. These should be looked at for significance, and the slope estimates are interpreted just as they would be in a linear regression model. In order to interpret the overall fit of the model, AIC values between models with various subsets of the effects can be compared in order to select a final model. Once a final model is decided upon, Verbeke and Molenberghs (2000) proposed to assess goodness of fit by calculating an R^2 for each profile (Jensen, 2006). There are various ways to do this in the

literature, and there is not just one valid method. An overall goodness of fit can then be obtained by combining these values.

3. META-REGRESSION ANALYSIS

Typically, we analyze one group (study) of data as a whole and look at the effect of one or more covariates on a dependent response variable using regression methods. This practice is extended in meta-analysis where parameters are at the level of a study, or in the case of the Greenville data, a subject, and not at an individual observation level. Specifically, the effect of PM 2.5 on lung function can be quantified for each subject separately, due to repeated measurements. These subject-specific PM 2.5 effects can be combined across subjects while accounting for between-subject differences, to get a clearer picture of the true underlying effect size of interest. The dependent variable, such as the PM 2.5 effects in the Greenville study, becomes the effect size in the study. This field, meta-regression, is a branch of meta-analysis (Borenstein et al., 2009). Moving from regression to meta-regression, there are several differences that must be addressed and will be discussed in this section. They include weighting each study (subject), using the right type of model, and considering the ratio of studies (or subjects) to covariates.

A statistical problem that arises in meta-regression is that the estimated effect sizes from each study will have different variances, due in part to different sample sizes in the studies. To solve this problem, weighted least squares (WLS) is used in the meta-regression model, applying weights of $\frac{1}{SE^2}$, where SE is the standard error of the effect size estimate, to each study (Stanley, 2013). Becker and Wu (2007) refer to the meta-regression weight as the “reciprocal of the slope variance”.

The procedures of regression are used in meta-analysis to examine covariates and predicted effect size in subjects with certain characteristics. The fixed effect model assumes all

studies (subjects) in the meta-analysis are samples of a single, larger study and that all of their results reflect a single underlying true effect size. The random effect model assumes the studies are conducted independently, each reflects a separate true effect size, and the meta-regression is a combination of studies of the same matter.

Before building a meta-regression model, linear regression models are fitted to each subject in the study separately, using the response variable of interest. In the case of this analysis, the difference in FEF before and after recess is considered. The response variable, y_i , is regressed on the covariates of interest, with the exception of covariates whose values remain constant throughout subject i 's observations. These covariates do not provide any information and are omitted from the individual models, but introduced back into the aggregate meta-regression model. The model for subject i is given by the typical linear regression model form:

$$Y_i = \beta_0 + X_i\beta + \epsilon_i$$

Here, Y_i is the vector of response measurements for subject i , length n_i , β_0 is the model intercept, β is a length p vector of fixed effect coefficients, X_i is a $n_i \times p$ covariate matrix, and ϵ_i is a length n_i vector of the errors.

Once this model has been fit for each subject, the results are aggregated into a matrix for the meta-analysis. The results of the model included in this matrix are the slope estimates for the effect of interest from each model, in this case, the PM 2.5 effect, and their respective standard errors, the constant covariates are also introduced into the matrix for the meta regression model.

The typical model for a meta-regression is given by:

$$\theta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p X_{ip} + u_i + \epsilon_i$$

Here, θ_i is the effect size (PM 2.5 effect) estimated in study i (or subject i in the Greenville meta-regression), β_0 is the model intercept, β represents the effect for the j th covariate in the model,

X_{ij} is the value of covariate j for subject i , u_i represents a random effect explaining study-to-study variance (Viechtbaur, 2014), and ϵ_i are the sampling errors. The assumptions for this model are:

$$u_i \sim N(0, \tau^2)$$

$$\epsilon_i \sim N(0, v_i)$$

Here, v_i are the known sampling variances of the observed outcomes of effect size estimates, and τ^2 is a between subject (study) variance, with the square root of this estimate representing the estimated standard deviation of the underlying effect across subjects (studies) (Viechtbaur 2014). In this meta-regression model, observation i is weighted by $\frac{1}{SE_i^2}$, where SE_i is the standard error of the Y_i estimate.

Most of the time in meta-regression, the test for significance of covariate effects is based on the standard normal distribution. The Z test to test for significance of the slope is given as $\frac{\hat{\beta}}{SE(\hat{\beta})}$, with the null hypothesis that the coefficient, β , is equal to zero. The Z test can be used to test the significance of a single coefficient, β_j .

A consideration of the ratio of subjects to covariates is needed for valid interpretation. A rule of thumb here is that there should be at least 10 subjects for each covariate, though there are no hard and fast rules on this (Borenstein et al., 2009).

Interpretation of the model, once the effect sizes, intercept, and p-value are obtained focuses mainly on the intercept estimate. It is a good practice to center all covariates to their respective mean. This allows for interpretation of the intercept, β_0 , as the population mean effect size (Cooper and Hedges, 1994), which gives the overall picture of the effect size of interest.

4. RESULTS

4.1 Linear Mixed Model Results

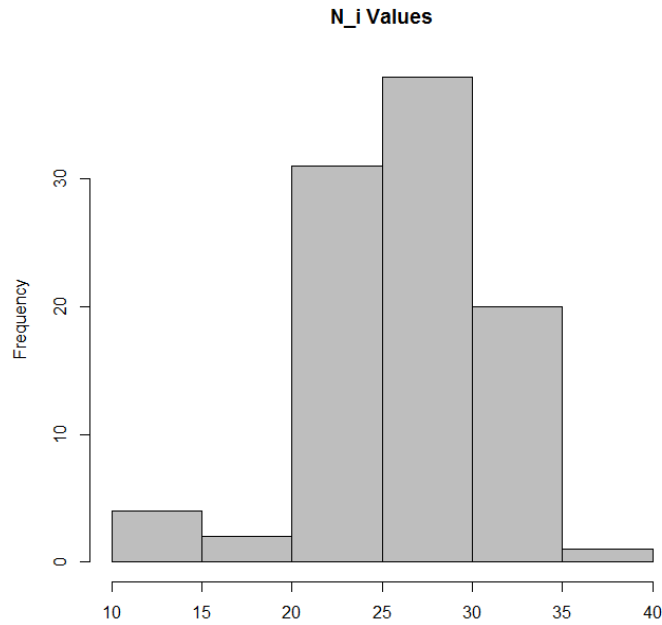
Table 2 – AIC values for correlation structures in the mixed model

Correlation Structure:	Akaike Information Criterion (AIC)
AR(1)	-267.7231
US (SYMM)	Failed to Converge
Compound Symmetry	-266.9819

Using a linear mixed model, the Greenville data were analyzed taking into account a random effect for subject (student), and an AR(1) correlation structure. The correlation structure of the model was selected by comparing AIC values using the various structures (see Table 2). The AIC values were very similar, aside from compound symmetry, which would not allow the model to converge, leading to AR(1) being chosen as it made the most sense for the repeated measures over time in these data.

The original data included 101 students with a total of 4563 observations. 5 students were eliminated from the analysis due to all of the outcomes data missing, leaving 96 distinct students in the model. Approximately 1/3 of the rows also had to be eliminated for missing values which rendered them useless in the analysis. This resulted in 2837 remaining rows. n_i , referred to in the models on pages 15 and 19 as the number of observations for subject i , ranges from 13 to 37. This distribution is shown in Figure 1, which is slightly skewed toward the lower range of subject observations.

Figure 1: Histogram of n_i values



The response, or dependent variable in the model, was the difference between the post-recess and pre-recess forced expiratory flow (FEF 25-75%). This was calculated from the data as: [post-recess FEF – pre-recess FEF]. Pre-Recess FEF values ranged from .33 to 4.64 liters/second, while post-recess FEF values ranged from .45 to 3.79 liters/second. Recall that higher values of FEF 25-75% are healthier, therefore negative values of FEF difference would mean that FEF values and lung function decreased over the period of recess, while positive values would mean that lung function increased during recess. The more negative or positive the value is, the more lung function decreased or increased. FEF difference ranged from -1.82 to 2.44 liters/second in difference. This variable appears symmetric, but long-tailed, without the use of a transformation (Figure 2).

Figure 2 – Histogram of FEF difference and normal Q-Q plot of FEF difference

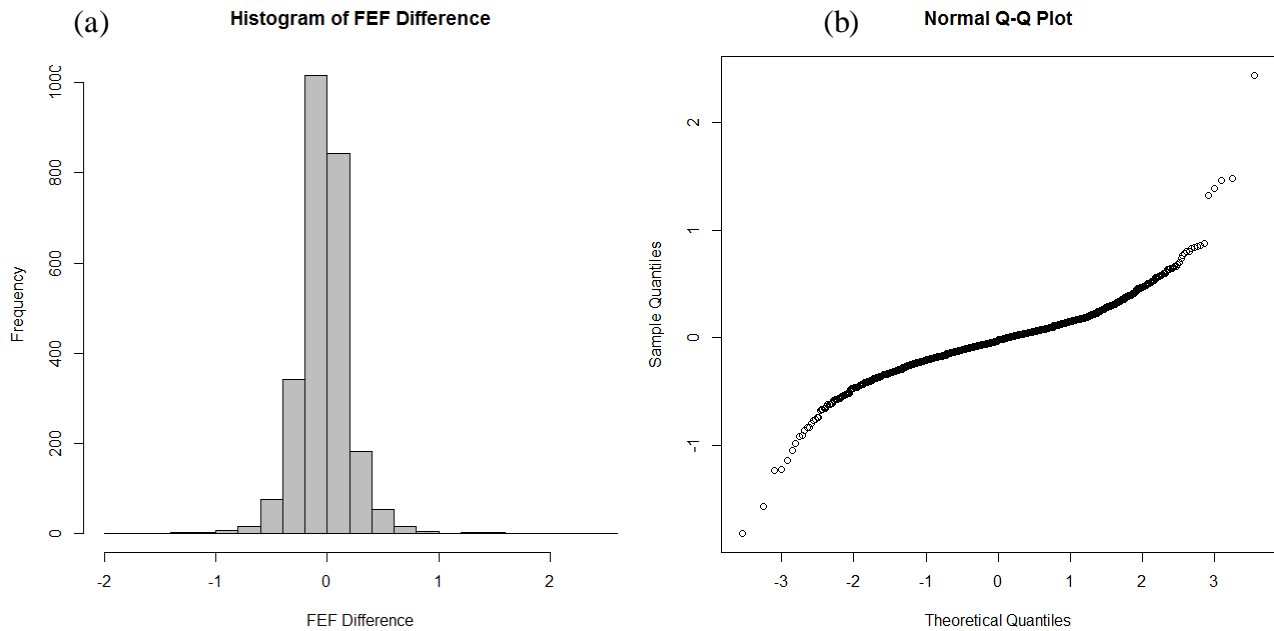


Table 3 – Variables included in full linear mixed model

Fixed Effect or Factor	Definition
Date	Number from 0 to 33, with 0 being the first day of the study
Spirometer ID	Number from 1-7 denoting which spirometer was used
Grade	Number from 3-5 denoting grade level
Age	Number denoting age of subject in years
Gender	Sex of the subject (coded 1=female, 0=male)
Sum of Criteria	A sum of 3 scores denoting respiratory health
Race	Race of the subject
Home Type	Type of home lived in by the subject: single-family, apt., etc.
Numbr of Smokers in Home	Number of smokers living in the home with the subject
URI Flag	Flag denoting whether the subject had URI symptoms
Indoor Recess Flag	Flag denoting whether the subject spent recess indoors
Passive Recess Flag	Flag denoting whether the subject did not engage in activity
Lingering Cough	A score from 0-3 denoting severity of lingering cough
Wheezing with Cold	A score from 0-3 denoting severity of wheezing with cold
Wheezing with no Cold	A score from 0-3 denoting severity of wheezing and no cold
Wheezing and Hard to Breathe	A score from 0-3 denoting severity of wheezing and breathing
Wheezing with Exercise	A score from 0-3 denoting severity of wheezing with exercise
Cough with Exercise	A score from 0-3 denoting severity of coughing with exercise
Chest Tightness	A score from 0-3 denoting chest tightness
Dr. Diagnosis	A flag denoting dr. diagnosis of asthma
Daily Medication	A flag denoting whether daily medication is taken
Asthma Medication	A flag denoting whether asthma medication is taken
PM PM 2.5	Average PM 2.5 measurement from 11a-2pm
AM Temperature	Average outdoor temperature (F°) from 7-10am
PM Temperature	Average outdoor temperature (F°) from 11a-2pm
PM Humidity	Average outdoor humidity from 11a-2pm

Table 4 – Effect estimates for final linear mixed model

	Value	Std.Error	DF	t-value	p-value
(Intercept)	0.00403497	0.04397381	2468	0.091759	0.9269
spiro.f2	-0.06190396	0.01995089	81	-3.102817	0.0026
spiro.f3	0.00561935	0.01999515	81	0.281036	0.7794
spiro.f4	-0.01763338	0.02011807	81	-0.876495	0.3834
spiro.f5	-0.03593843	0.01966795	81	-1.827258	0.0713
spiro.f6	-0.06018398	0.02014909	81	-2.986933	0.0037
spiro.f7	-0.02884340	0.02080600	81	-1.386302	0.1695
gender.f2	-0.02082011	0.01092515	81	-1.905704	0.0602
race.f2	-0.01460347	0.05638529	81	-0.258994	0.7963
race.f3	0.30148454	0.07306460	81	4.126274	0.0001
race.f4	0.09370347	0.07279648	81	1.287198	0.2017
race.f5	0.01601010	0.01994912	81	0.802547	0.4246
URI	0.03584515	0.01246625	2468	2.875375	0.0041
Indoor_Recess	0.08263983	0.01843690	2468	4.482306	0.0000
PM25OUT	0.00084334	0.00044337	2468	1.902131	0.0573
Wheeze_Hard_Breathe	-0.01162160	0.00706161	81	-1.645745	0.1037
meddaily.f.L	0.00732257	0.01127500	81	0.649452	0.5179
meddaily.f.Q	0.07790942	0.02131243	81	3.655586	0.0005
PMTEMP	0.00281774	0.00082195	2468	3.428115	0.0006
PMHUMID	-0.00099436	0.00057520	2468	-1.728706	0.0840
Indoor_Recess:PM25OUT	-0.00079861	0.00062346	2468	-1.280940	0.2003

The random effect in the model was the subject ID denoting each student. See Table 3 for the list of fixed effects and factors included in the full model. The model was fit in R using the lme() function from the nlme package (Pinheiro et al., 2015). Once the full model was fit, a backward selection using AIC was implemented to reduce the model to the most important factors. The final model gives an AIC of -297.294, and the parameter estimates are shown in Table 4. An R squared for the mixed model can be calculated as:

$$R^2 = 1 - \frac{RSS_{Model}}{RSS_{Intercept\ Only\ Model}}$$

Here, RSS_{Model} are the residual sum of squares for the full model and $RSS_{Intercept\ Only\ Model}$ are the residual sum of squares for the null, or intercept only model. The R squared for the mixed model was .033, meaning that the full model explains approximately 3% of the variation in the data. The residuals (Figures 3 and 4) look symmetric, but long-tailed. Mixed model assumptions

include normality of error terms, but it is the symmetry that is essential as the mixed model methods are robust against this kind of departure from normality.

Figure 3 - Normal Q-Q plot of residuals and residuals vs. fitted values from mixed model

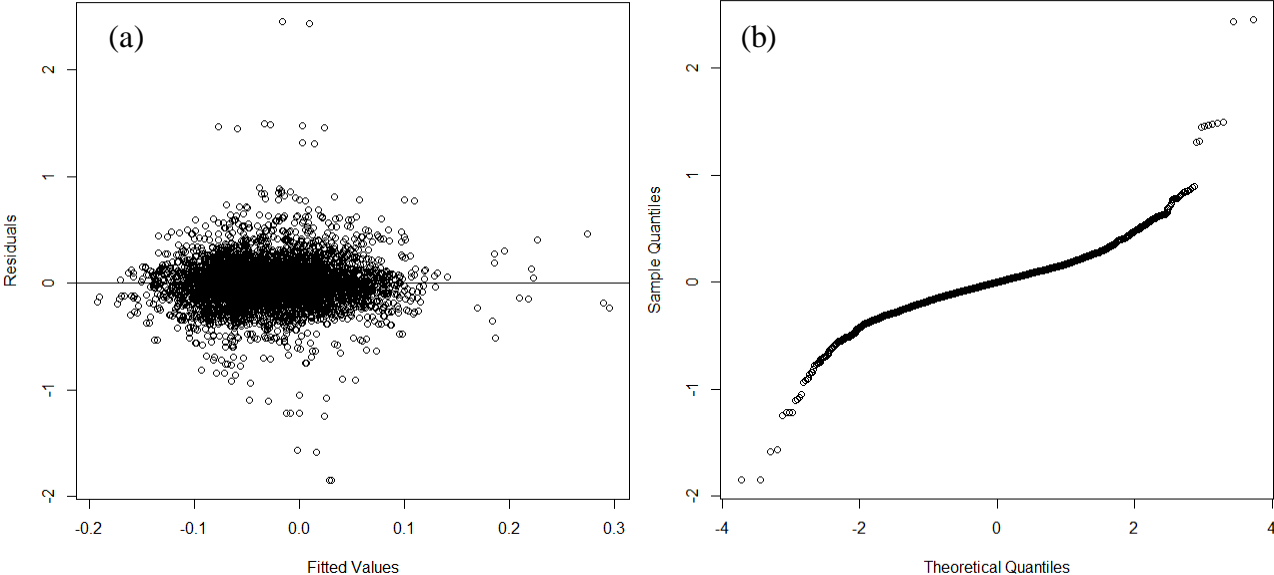


Figure 4 – Histogram of residuals from mixed model

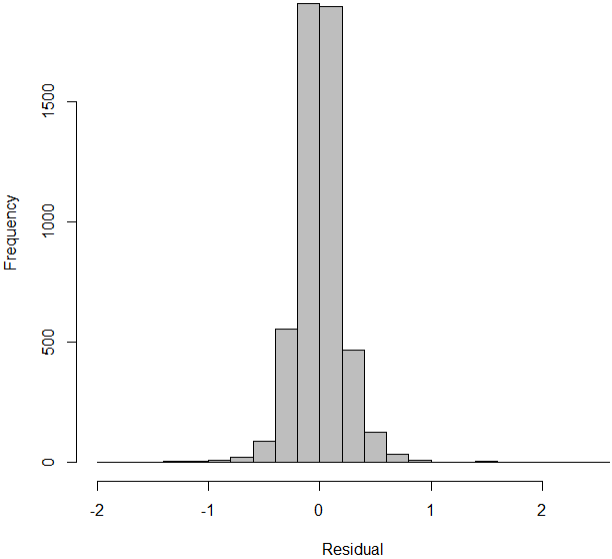
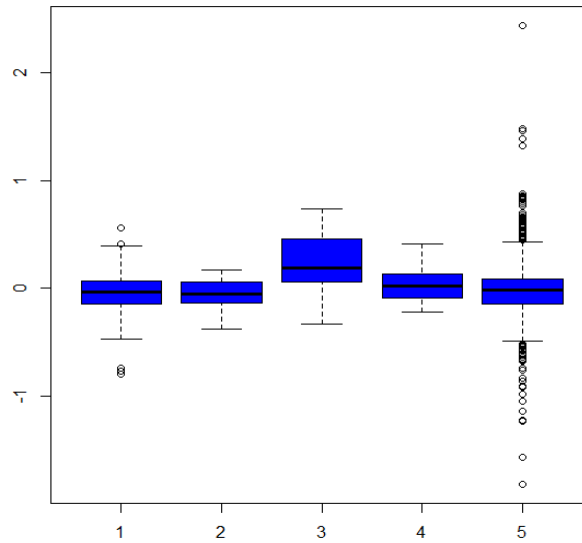


Figure 5 – Boxplots of FEF difference by ethnicity



The following conclusions can be made from this model, based on the summary in Table 4:

- Several spirometers used seem to have a significant effect on the difference in FEF measurements taken before and after recess. Using a post-hoc means comparison, spirometer 1 differs significantly from spirometer 2, and spirometer 2 differs significantly from spirometer 3.
- Students of Japanese (race level 3) descent appear to differ significantly in their lung function, all other factors constant (see Figure 5). They have significantly greater lung function, based on the difference in FEF before and after recess, compared to other races in a post hoc comparison. It is worth noting that only 1 of the 96 students, or 1.5% of them, were of Japanese descent.
- The presence of URI (upper respiratory infection) symptoms seemed to have a positive effect (.036) on FEF difference. One reason for this could be that a period of activity,

such as recess, may have helped lung function for students who were having those symptoms upon going outside.

- An indoor recess seems to have the largest positive impact on FEF Difference by a factor of .083, all other factors held constant. This is significant with a p-value $< .0001$
- As outdoor PM 2.5 increases, FEF difference actually increases by a factor of .001, which is marginally significant (p-value .0573).
- Taking daily medication for asthma has a positive effect on the FEF difference. Daily medication increases that difference by a factor of .078
- As temperature outdoors increases, FEF increases by a factor of .003

In order to look separately at how an indoor recess and asthma together with air quality affected the FEF difference, a couple of additional models were considered, both in the linear model framework, still with a random effect for subject, and an AR(1) correlation structure. The first model looks at the asthma variable interacting with the PM 2.5 measurements. The results of this model are shown in Table 5. Looking at asthma and PM 2.5, we can see that their interaction is non-significant, and neither of these covariates are significant alone. We cannot show that asthma has any effect on the FEF difference, while also controlling for the effect of PM 2.5, and there is also no evidence that the effect of PM 2.5 depends on asthma status.

The second additional model looks at the effect of indoor recess on the FEF difference, controlling for PM 2.5 levels. The results of this model are shown in Table 6. These results show that there is no interaction present between indoor recess and the PM 2.5 levels, so the effects of PM 2.5 do not depend on having recess indoors versus outdoors. There is however, a substantial effect of indoor recess alone on FEF difference. The results show that controlling for PM 2.5 levels, the FEF difference increases by a factor of .073 for students that stay inside for recess.

Table 5 – Results of PM 2.5 and asthma model

	Value	Std.Error	DF	t-value	p-value
(Intercept)	-0.03700669	0.011753413	2472	-3.1485912	0.0017
Asthma	0.00706422	0.016456665	94	0.4292620	0.6687
PM25OUT	0.00067139	0.000394237	2472	1.7030161	0.0887
Asthma:PM25OUT	-0.00031599	0.000549463	2472	-0.5750935	0.5653

Table 6 – Results of PM 2.5 and indoor recess model

	Value	Std.Error	DF	t-value	p-value
(Intercept)	-0.03979907	0.008883373	2471	-4.480176	0.0000
Indoor_Recess	0.07252212	0.018568725	2471	3.905606	0.0001
PM25OUT	-0.00006033	0.000398618	2471	-0.151351	0.8797
Indoor_Recess:PM25OUT	-0.00037128	0.000610118	2471	-0.608541	0.5429

Overall, the linear mixed model does not seem to show an adverse effect from PM 2.5 exposure during recess by looking at the FEF levels before and after, all other measured factors held constant. This is in line with what has been found in previous analyses of these data.

4.2 Meta-Regression Results

Table 7- Covariates included in individual linear regression models

Covariates:
P.M. measurements of PM 2.5
Indoor Recess Flag
P.M. Temperature
P.M. Humidity

The results of a meta-regression modeling the effect size of the PM 2.5 measurement taken in the afternoon will now be looked at. In order to perform a meta-regression on these data,

a linear regression was first performed for each student individually using the `lmList()` function in R from the package `lme4` (Bates et al., 2015). This generated a PM 2.5 effect estimate for each student, as well as the standard error of that estimate. See Table 7 for a list of predictor variables included in the individual linear regressions.

Categorical variables which were constant throughout each day of the study were excluded from the individual models, but their values are later included in the meta-regression. See Table 8 for a list of these constant covariates. The matrix created to implement the meta-regression includes an individual slope estimate for the effect of interest (outdoor PM 2.5), their standard errors, and the various constant covariates to be included in the analysis. A portion of this matrix is shown in Table 9. Once the matrix from the individual regressions was defined, a weighted multiple regression was applied using the effect sizes of the PM 2.5 measurement as a response, and weights given by the inverse of the squared standard error of the PM 2.5 effect.

Table 8: List of constant covariates included in meta-regression model

Constant Covariates:
Spirometer ID (6 flag variables for 7 spirometers)
Gender
Asthma Flag
Race (4 flag variables for 5 levels)
Upper Respiratory Infection (URI) Flag
Wheezing with Trouble Breathing Flag
Physician Diagnosis of Asthma Flag
Taking Medication Daily Flag (2 flag variables for 3 levels)

Table 9 – Partial matrix to set up meta-regression, with a row for each subject, and predictor columns mean-centered

coefs25OUTIN	SE25OUTIN	spiro.f2	spiro.f3	spiro.f4	spiro.f5	spiro.f6	spiro.f7
-0.0023649905	0.002753786	-0.1458333	-0.1354167	-0.1458333	-0.1458333	0.84375	-0.1354167
-0.0017993886	0.002678871	0.8541667	-0.1354167	-0.1458333	-0.1458333	-0.15625	-0.1354167
-0.0016116741	0.002660327	0.8541667	-0.1354167	-0.1458333	-0.1458333	-0.15625	-0.1354167
0.0001656486	0.003331656	-0.1458333	0.8645833	-0.1458333	-0.1458333	-0.15625	-0.1354167
-0.0026583252	0.003061057	-0.1458333	-0.1354167	-0.1458333	-0.1458333	-0.15625	0.8645833
0.0062414672	0.004113231	-0.1458333	-0.1354167	-0.1458333	-0.1458333	0.84375	-0.1354167
gender	asthma	race.f2	race.f3	race.f4	race.f5	URI	whb
-0.4583333	-0.5208333	-0.01041667	-0.01041667	-0.01041667	0.1145833	-0.8854167	-0.4479167
-0.4583333	0.4791667	-0.01041667	-0.01041667	-0.01041667	0.1145833	-0.8854167	0.5520833
-0.4583333	-0.5208333	-0.01041667	-0.01041667	-0.01041667	0.1145833	0.1145833	-0.4479167
0.5416667	0.4791667	-0.01041667	-0.01041667	-0.01041667	0.1145833	-0.8854167	-0.4479167
-0.4583333	0.4791667	-0.01041667	-0.01041667	-0.01041667	0.1145833	0.1145833	-0.4479167
-0.4583333	-0.5208333	-0.01041667	-0.01041667	-0.01041667	0.1145833	0.1145833	-0.4479167
dx	med.f.L	med.f.Q					
-0.46875	-0.05208333	-0.1666667					
0.53125	-0.05208333	-0.1666667					
-0.46875	-0.05208333	-0.1666667					
0.53125	0.94791667	-0.1666667					
0.53125	-0.05208333	-0.1666667					
-0.46875	-0.05208333	-0.1666667					

This meta-regression was carried out using the `lm()` function in R with the weights option. The covariates were centered around their means so that the intercept could be interpreted as the overall population mean effect size, denoting whether there is a significant effect (not equal to 0) of PM 2.5 on lung function, specifically the difference of FEF before and after recess.

The possible right-skewness of the residuals (Figures 6 and 7) suggests the need to consider a transformation as follows. Let $Y = \text{PM 2.5 effect (slope for subject)}$ and $\text{Var}[Y] \approx (\text{SE of PM 2.5 effect})^2$. Consider the transformation $Y' = (c + Y)^\lambda = f(Y)$ for some constants c and λ . By the Delta method (Rice 1995), $\text{Var}[f(Y)] \approx [f'(\mu_y)]^2 * \text{Var}(Y)$. Therefore, $\text{Var}[Y'] \approx [\lambda(c + \mu_y)^{\lambda-1}]^2 * \text{Var}(Y) \approx [\lambda(c + Y)^{\lambda-1}]^2 * \text{Var}(Y)$.

Figure 6- Meta-regression residuals vs. fitted and normal Q-Q plot of residuals

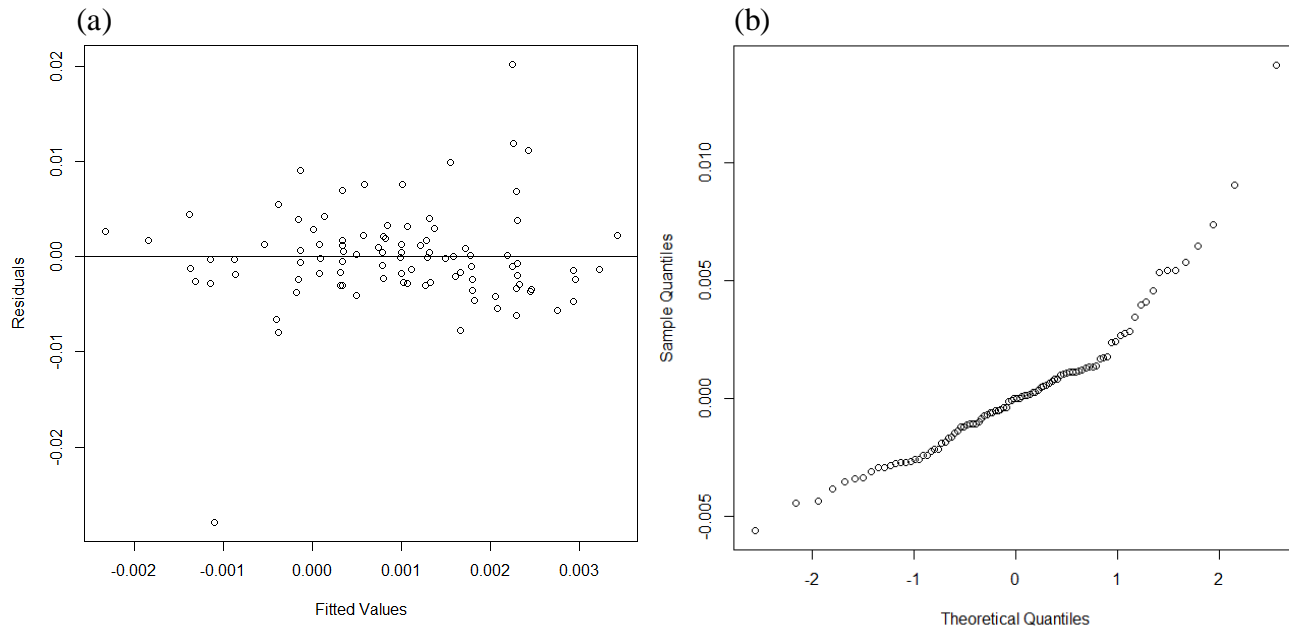


Figure 7- Meta-regression histogram of residuals

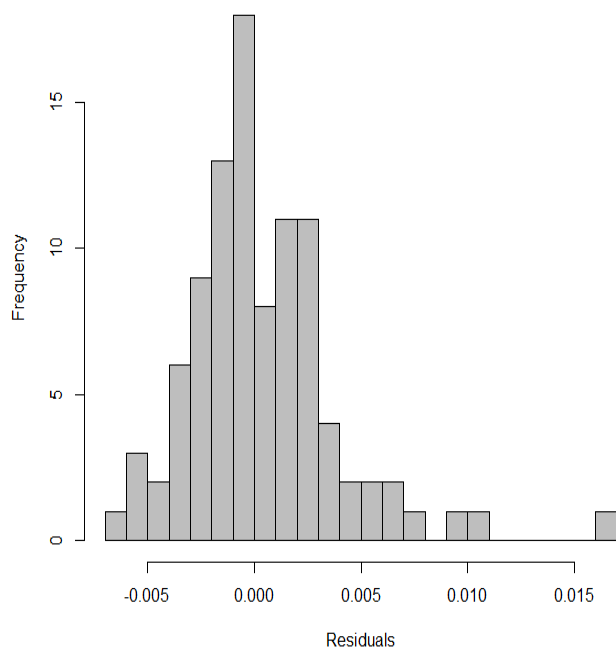


Figure 8 – Residual and normal Q-Q plots of residuals of transformed Y

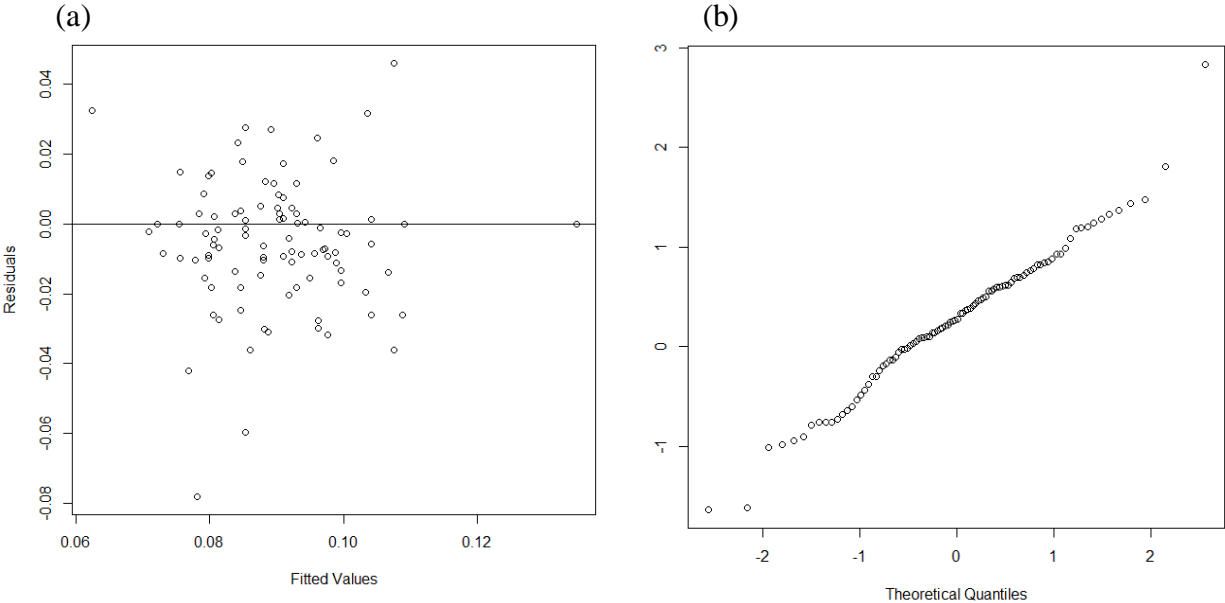
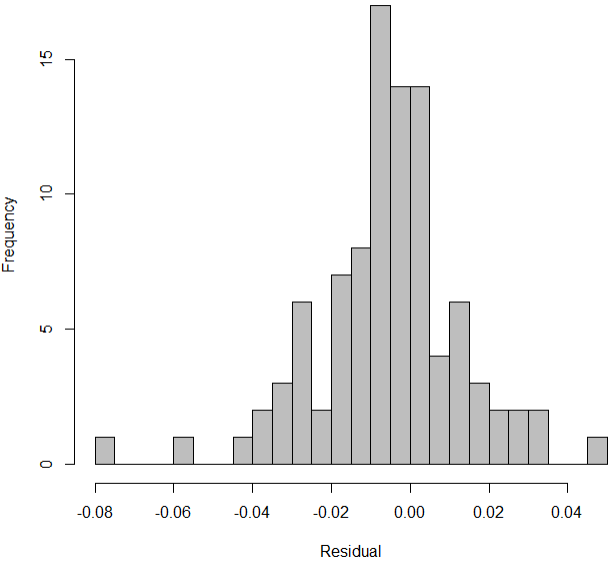


Figure 9 – Histogram of residuals of transformed Y



When running the meta-regression again, Y was transformed as follows: $Y' = (c + PM\ 2.5\ Slope)^\lambda$, and the weights were redefined as $weight = [\lambda(c + PM\ 2.5\ Slope)^{\lambda-1} * (SE\ PM\ 2.5\ Slope)]^{-2}$, where $c = -\min(\text{PM 2.5 slopes for all subjects}) = .00651$, and $\lambda = .5$. This transformation produced generally normal residuals, centered around 0 with no evidence of heteroscedasticity (see Figures 8 and 9).

Table 10 – Meta-regression results (with transformed Y)

```

Weighted Residuals:
  Min      1Q  Median      3Q      Max
-1.3948 -0.5819 -0.1958  0.1213  3.9465

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0896564  0.0020275  44.221  <2e-16 ***
spiro.f2     -0.0139220  0.0074234  -1.875  0.0645 .
spiro.f3      0.0103787  0.0072402   1.433  0.1558
spiro.f4     -0.0027280  0.0071360  -0.382  0.7033
spiro.f5     -0.0054142  0.0075572  -0.716  0.4759
spiro.f6     -0.0084218  0.0076058  -1.107  0.2716
spiro.f7     -0.0182839  0.0090213  -2.027  0.0462 *
gender       -0.0008056  0.0040653  -0.198  0.8434
asthma       -0.0006812  0.0069476  -0.098  0.9222
race.f2      -0.0216608  0.0260145  -0.833  0.4076
race.f3       0.0338588  0.0218041   1.553  0.1246
race.f4       0.0282413  0.0263850   1.070  0.2878
race.f5      -0.0058855  0.0080378  -0.732  0.4663
URI           0.0123830  0.0066059   1.875  0.0646 .
whb          -0.0013219  0.0031046  -0.426  0.6714
dx            0.0054279  0.0073639   0.737  0.4633
med.f.L      -0.0076688  0.0099572  -0.770  0.4435
med.f.Q       0.0058640  0.0057772   1.015  0.3133
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9048 on 77 degrees of freedom
Multiple R-squared:  0.2835,    Adjusted R-squared:  0.1253
F-statistic: 1.792 on 17 and 77 DF,  p-value: 0.04436

```

The results from the meta-regression model in R are shown in Table 10. The important result from the model is the effect size of the intercept. For the Greenville data, after back transforming, the intercept estimate was .0015 with a 95% confidence interval of (.00082,

.00227). Because this interval does not contain 0, we conclude from this meta-regression analysis that the mean effect of the PM 2.5 level on the difference in FEF lung function before and after recess is, controlling for other factors, significant. The positive effect size suggests that for each increase of 1 ppm in the PM 2.5 measurements, lung function increases by .0015. Note also, that there are no other significant effects in the meta-regression except for spirometer, which suggests a difference in FEF difference depending on which spirometer was used. All of the other covariate p-values are well above the threshold of .05.

4.3 Comparison

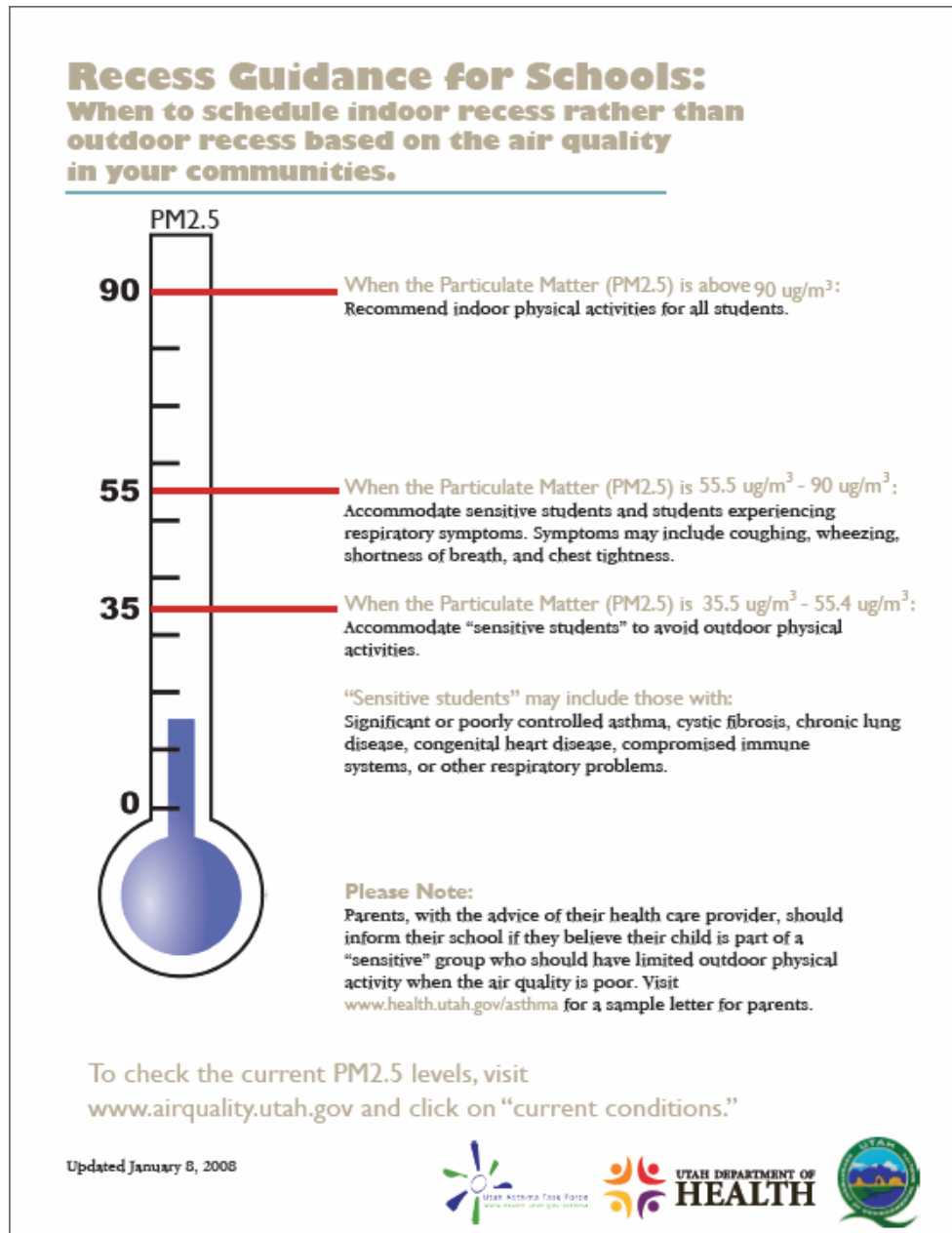
Comparing meta-regression to a mixed model with the Greenville Air Quality Study data, there are not a lot of differences in the results. The 95% confidence interval for the effect of PM 2.5 on lung function in the linear mixed model was (-.00002, .0017). This confidence interval in the meta-regression was (.00082, .00227), which is a similar interval but not including 0. Because the meta-regression confidence interval does not include 0, this method concluded that there is a significant effect on FEF difference by PM 2.5 measurements, by a factor of .0015. In the mixed model, we treated subject as a random factor with repeated measurements. In the meta-regression, we considered each subject separately and aggregated the PM 2.5 effect size estimates across subjects. In both methods, each subject had its own intercept estimate. In the linear mixed model, this is due to the fact that subject was treated as a random factor. Considering each subject separately may have allowed for a clearer picture of how PM 2.5 really affected FEF difference before and after recess. Although the difference in the PM 2.5 slope estimates from the two methods was small, it was a large enough difference that a significant result was found using meta-regression.

5. DISCUSSION

Considering the results of the various analyses performed on these data, it is possible that when we look at the difference in FEF lung function after a 20 minute recess with exposure to the outside air, which may be polluted by an inversion, there is no evidence of a significant, adverse effect to the children playing outside for that period of time. It is interesting that the more traditional method of using a mixed model produced a non-significant result, while meta-regression showed a slight positive significant difference in lung function after an exposure to the PM 2.5 outdoors. Whether or not a child had asthma did not change the effect of PM 2.5 on lung function. This was shown by the non-significant effect of the asthma covariate when a model was run looking just at asthma and PM 2.5 levels and their effect on the lung function difference. Speaking with Dr. Redd who was instrumental in this study, he and the staff of Greenville felt the results of the study show that for a healthy child, playing outside may actually be safer than an indoor recess on those poor air days. This was due to observation during the course of the study on indoor recess days that more students were prone to injuries as a result of being active in such a confined space; for example, running into walls, other students, etc. (Redd 2015).

Ultimately, each school or school district in the state of Utah will choose their own course of action, with this depending on the resources they have available to them. Schools in Cache, Davis, Weber, and Salt Lake counties all have access to hourly air quality data, which allows them to make the decision each time they are going to send kids to recess. Some schools have created a PM 2.5 cutoff, while others are still considering all of the factors involved before making a decision. The recommendation from the State of Utah is shown in Figure 10 (Packham 2011). The analysis of this data supports the notion of continuing to leave the decision of how to

Figure 10 – Recess Guidelines for Utah Schools on Air Quality (Utah Dept of Health 2008)



handle poor air quality and when to restrict outdoor recess up to individual schools and districts in the state of Utah. Future study of this issue is being considered by Utah State University student health, and Dr. Redd recommends future study of long term exposure to poor air, something we don't know the effect of just yet (Redd 2015).

REFERENCES

- Arnall, David. "Pulmonary Function Tests". <http://www2.nau.edu/~daa/lecture/pft.htm>. 2006. Accessed 23 November, 2015.
- Bates D, Maechler M, Bolker B, Walker, S. "Fitting Linear Mixed-Effects Models Using lme4". *Journal of Statistical Software*, 67(1), 1-48. 2015.
- Becker B. and Wu M. "The Synthesis of Regression Slopes in Meta-Analysis". *Statistical Science*, 22(3), 414-429. 2007.
- Borenstein M., Hedges L.V., Higgins J.R.T., and Rothstein H.R. *Introduction to Meta-Analysis*. John Wiley and Sons. New York, NY. 2009.
- Cooper, H. and Hedges, L.V. *The Handbook of Research Synthesis*. Russell Sage Foundation, New York, NY. 1994.
- Jensen, Willis Aaron. "Profile Monitoring for Mixed Model Data". PhD, dissertation. Virginia Tech, 2006.
- Littell, Ramon C. "SAS for Linear Models". 2002.
- Morgan Scientific. "What is a Pulmonary Function Test?" <http://www.morgansci.com/solutions/resource/what-is-a-test-pulmonary-function-test/>. 2013. Accessed 30 October, 2015.
- New York State Department of Health. "Fine Particles (PM 2.5) Questions and Answers". http://www.health.ny.gov/environmental/indoors/air/pmq_a.htm. 2011. Accessed 30 October, 2015.
- Packham, Steve. "Utah's Recess Guidance: Based on Air Quality". <http://www3.epa.gov/airnow/2011conference/communications/Packham.pptx>. 2011. Accessed 30 October, 2015.
- Pinheiro J, Bates D, Debhoy S, Sarkor D, and R Core Team. "nlme: Linear and Nonlinear Mixed Effects Models". R Package Version 3.1-120. 2015.
- Redd, Dr. Edward. "Telephone Interview with Dr. Redd" 18 September 2015.
- Rice, John A. *Mathematical Statistics and Data Analysis*. Duxbury Press, Belmont, CA, 2nd edition. 1995.
- SAS. SAS/STAT(R) 9.2 User's Guide, Second Edition. 2009. "REPEATED Statement". http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_mixed_sect019.htm Accessed 30 October, 2015.

- Seltman, Howard J. *Experimental Design and Analysis*. Carnegie Mellon University. Ebook. <http://www.stat.cmu.edu/~hseltman/309/Book/chapter15.pdf> . 2012. Accessed 30 October, 2015.
- Spinka, Christie. “Data Analysis II – Linear Mixed Models” <http://www.stat.missouri.edu/~spinkac/stat8320/LinearMixedModels.pdf> 2007. Accessed 30 October, 2015.
- SSI Central. “Full ML vs Restricted ML”. www.ssicentral.com/hlm/help6/faq/Full ML vs Restricted ML.pdf Accessed 30 October, 2015.
- Stanley, T.D and Doucouliagos, Hristos. “Meta-Regression Approximations to Reduce Publication Selection Bias”. *Research Synthesis Methods* 5(1): 60-78. 2014.
- Utah Department of Health. “Frequently Asked Questions on Air Quality and Recess”. <http://Health.utah.gov/asthma/pdfs/airquality/faqrecess.pdf>. 2008. Accessed 30 October, 2015.
- Verbeke, G. and Molenberghs, G. *Linear Mixed Models for Longitudinal Data*. Springer-Verlag. New York, NY. 2000.
- Viechtbaur, Wolfgang. “Linear Regression and the Mixed-Effects Meta-Regression Model”. The metafor Package. http://www.metaforproject.org/doku.php/tips:regression_with_rma. 2014. Accessed 30 October, 2015.

APPENDIX A: R CODE

```
##Read in dataset
data = read.csv("asthma.csv",header=TRUE)

##Pre-analysis data exploration
hist(data$Spiro,xlab="Spiro ID",main="Spirometer Use",col="black")
hist(data$Grade,xlab="Grade",main="Grade Distribution",col="black")
hist(data$Age,xlab="Age",main="Age Distribution",col="black")
summary(data$Gender)
summary(data$Race)
summary(data$Language)
table(data$Home_Type)
table(data$Built)
table(data$Construct)
hist(data$Num_Smoke,xlab="Number of Smokers in Home",main="Num Smokers
Distribution",col="black")
table(data$Stove)
table(data$Heat)
hist(data$Cough_Stay,xlab="Cough Won't Go
Away",main="Distribution",col="black")
hist(data$Wheeze_Cold,xlab="Wheezing with
Cold",main="Distribution",col="black")
hist(data$Wheeze_No_Cold,xlab="Wheezing with No
Cold",main="Distribution",col="black")
hist(data$Wheeze_Hard_Breath,xlab="Wheezing Hard to
Breathe",main="Distribution",col="black")
hist(data$Wheeze_Exercise,xlab="Wheezing with
Exercise",main="Distribution",col="black")
hist(data$Cough_Exercise,xlab="Cough with
Exercise",main="Distribution",col="black")
hist(data$Chest_Tight,xlab="Chest Tight",main="Distribution",col="black")
hist(data$Sleep_Prob,xlab="Sleep Problems",main="Distribution",col="black")
table(data$Dr_Dx)
table(data$Asthma1)
table(data$Bronchitis)
table(data$Rad)
table(data$Pneumonia)
table(data$Asth_Bronch)
table(data$Med_Rx)
table(data$Med_Daily)
table(data$Asthma)
hist(data$Crit13,xlab="Criteria 1-3",main="Distribution",col="black")
hist(data$Crit22,xlab="Criteria 2-2",main="Distribution",col="black")
hist(data$Crit32,xlab="Criteria 3-2",main="Distribution",col="black")
hist(data$Sum_Crit,xlab="Sum of Criteria",main="Distribution",col="black")
table(data$Color)
hist(data$Pre_FEF,xlab="Pre Recess FEF",main="Distribution",col="grey")
hist(data$Pre_PEFR,xlab="Pre Recess PEFR",main="Distribution",col="grey")
hist(data$Post_FEF,xlab="Post Recess FEF",main="Distribution",col="grey")
hist(data$Post_PEFR,xlab="Post Recess PEFR",main="Distribution",col="grey")
table(data$URI)
table(data$Indoor_Recess)
table(data$Passive_Recess)
table(data$Asthma_Meds)
```



```

hist(data$Pre_FVC,xlab="Pre Recess FVC",main="Distribution",col="grey")
hist(data$Pre_FEV,xlab="Pre Recess FEV",main="Distribution",col="grey")
hist(data$Pre_FEVFVC,xlab="Pre Recess FEV FVC
Ratio",main="Distribution",col="grey")
hist(data$Post_FVC,xlab="Post Recess FVC",main="Distribution",col="grey")
hist(data$Post_FEV,xlab="Post Recess FEV",main="Distribution",col="grey")
hist(data$Post_FEVFVC,xlab="Post Recess FEV FVC
Ratio",main="Distribution",col="grey")
hist(data$AM25OUT,xlab="Avg Outdoor PM2.5
AM",main="Distribution",col="black")
hist(data$PM25OUT,xlab="Avg Outdoor PM2.5
PM",main="Distribution",col="black")
hist(data$INPM25,xlab="Avg Indoor PM2.5 PM",main="Distribution",col="black")
hist(data$AMTEMP,xlab="Avg AM Outdoor
Temperature",main="Distribution",col="black")
hist(data$PMTEMP,xlab="Avg PM Outdoor
Temperature",main="Distribution",col="black")
hist(data$PMHUMID,xlab="Avg PM Humidity
Outdoors",main="Distribution",col="black")

##create lung function differences
data$FEF_Diff = data$Post_FEF - data$Pre_FEF
data$FVC_Diff = data$Pre_FVC - data$Post_FVC
data$FEV_Diff = data$Post_FEV - data$Pre_FEV
data$FEVFVC_Diff = data$Post_FEVFVC - data$Pre_FEVFVC

##check response variable for normality
hist(data$FEF_Diff,xlim = c(-.5, .75))
qqnorm(data$FEF_Diff)

##check for interactions among continuous variables
pairs(data$AM25OUT,data$PM25OUT,data$AMTEMP,data$PMTEMP,data$PMHUMID)

##define nominal and ordinal factors in the data
data$race.f = factor(data$Race)
data$spiro.f = factor(data$Spiro)
data$grade.f = ordered(data$Grade)
data$gender.f = factor(data$Gender)
data$lang.f = factor(data$Language)
data$hometype.f = factor(data$Home_Type)
data$built.f = factor(data$Built)
data$construct.f = factor(data$Construct)
data$stove.f = factor(data$Stove)
data$heat.f = factor(data$Heat)
data$meddaily.f = ordered(data$Med_Daily)

hist(data$FEF_Diff,xlab="FEF Difference",col="grey",main="Histogram of FEF
Difference",breaks=24)

##Perform a linear mixed effects regression
library(nlme)

aqmod = lme(FEF_Diff ~ Date + spiro.f + grade.f + Age + gender.f + Sum_Crit
+ race.f + hometype.f + Num_Smoke + URI + Indoor_Recess*PM25OUT +
Passive_Recess
+ Cough_Stay + Wheeze_Cold + Wheeze_No_Cold + Wheeze_Hard_Breathe +
Wheeze_Exercise

```

```

+ Cough_Exercise + Chest_Tight + Dr_Dx + meddaily.f + Asthma_Meds
+ AMTEMP + PMTEMP + PMHUMID, data=data, random=~1|Subject, method="ML")

#introduce correlation structure into the model
Aqmodcor = update(aqmod, correlation = corAR1())

#create a null model for comparison
nmod = lme(FEF_Diff ~ 1, data=data, random=~1|Subject, method="ML")
naqmod <- update(nmod, correlation = corAR1())

smrnull = summary(naqmod)
smraqmod = summary(Aqmodcor)

#Look at other correlation structures and compare AIC values
aqmod2 <- update(aqmod, correlation = corCAR1())
a2 = summary(aqmod2)
aqmod3 <- update(aqmod, correlation = corCompSymm())
a3 = summary(aqmod3)
aqmod9 <- update(aqmod, correlation = corSymm())
a4 = summary(aqmod9)

#backward selection for model reduction
library(MASS)
aqmodr = stepAIC(Aqmodcor, direction="backward")
summary(aqmodr)

aqmodr = update(aqmodr, ~.+Indoor_Recess*PM25OUT) ##add back in
#check the residuals of the model
Summary(aqmodr)
plot(aqmodr)

## calculate a pseudo-rsquared
rssfull = sum(aqmodr$residuals^2)
rssnull = sum(naqmod$residuals^2)
rsq = 1 - (rssfull/rssnull)
rsq

hist(aqmodr$residuals, main= "Residuals: Linear Mixed Model",
col="grey", breaks = 24)

#post hoc means comparisons on significant factors
pairwise.t.test(data$FEF_Diff, data$spiro.f, p.adj="none")
pairwise.t.test(data$FEF_Diff, data$race.f, p.adj="none")
pairwise.t.test(data$FEF_Diff, data$meddaily.f, p.adj="none")

#looking at boxplots of FEF difference by factor
boxplot(data$FEF_Diff~data$spiro.f, col="blue")
boxplot(data$FEF_Diff~data$race.f, col="blue")
boxplot(data$FEF_Diff~data$gender.f, col="blue")
boxplot(data$FEF_Diff~data$Asthma, col="blue")
boxplot(data$FEF_Diff~data$URI, col="blue")
boxplot(data$FEF_Diff~data$Indoor_Recess, col="blue")
boxplot(data$FEF_Diff~data$Wheeze_Hard_Breathe, col="blue")
boxplot(data$FEF_Diff~data$Dr_Dx, col="blue")
boxplot(data$FEF_Diff~data$meddaily.f, col="blue")

```

```

##additional models to look at effect of asthma, indoor recess with PM 2.5 on
FEF_Difference
asth= lme(FEF_Diff ~ Asthma*PM25OUT,data=data,random=~1|Subject,method="ML")
asth = update(asth, correlation = corAR1())
summary(asth)
plot(asth)

asth2 = lme(FEF_Diff ~ Dr_Dx*PM25OUT,data=data,random=~1|Subject,method="ML")
asth2 = update(asth2, correlation = corAR1())
summary(asth2)
plot(asth2)

recess = lme(FEF_Diff ~ Indoor_Recess*PM25OUT,
data=data,random=~1|Subject,method="ML")
recess = update(recess, correlation = corAR1())
summary(recess)
plot(recess)

#performing individual linear regressions on each subject
library(lme4)
fits = lmList(FEF_Diff~PM25OUT + Indoor_Recess + PMTEMP + PMHUMID | Subject,
data=data)
summ = summary(fits)

##creating histogram of ni values
ni = tapply(data$Subject,data$Subject,length)
ni = as.vector(ni)
hist(ni)

##extracting needed coefficients and standard errors into a matrix for meta
regression
coefs25OUT = summ$coefficients[,1,2]
SE25OUT = summ$coefficients[,2,2]
coefs25OUTIN = as.vector(coefs25OUT)
SE25OUTIN = as.vector(SE25OUT)

spiro = tapply(data$Spiro,data$Subject,max)
spiro.f2 = as.numeric(spiro==2)
spiro.f3 = as.numeric(spiro==3)
spiro.f4 = as.numeric(spiro==4)
spiro.f5 = as.numeric(spiro==5)
spiro.f6 = as.numeric(spiro==6)
spiro.f7 = as.numeric(spiro==7)
spiro.f2 = as.vector(spiro.f2-mean(spiro.f2))
spiro.f3 = as.vector(spiro.f3-mean(spiro.f3))
spiro.f4 = as.vector(spiro.f4-mean(spiro.f4))
spiro.f5 = as.vector(spiro.f5-mean(spiro.f5))
spiro.f6 = as.vector(spiro.f6-mean(spiro.f6))
spiro.f7 = as.vector(spiro.f7-mean(spiro.f7))
gender = tapply(data$Gender,data$Subject,max)
gender = as.vector(gender-mean(gender))
asthma = tapply(data$Asthma,data$Subject,max)
asthma = as.vector(asthma-mean(asthma))
race = tapply(data$Race,data$Subject,max)
race.f2 = as.numeric(race==2)
race.f3 = as.numeric(race==3)
race.f4 = as.numeric(race==4)

```

```

race.f5 = as.numeric (race==5)
race.f2 = as.vector (race.f2-mean (race.f2))
race.f3 = as.vector (race.f3-mean (race.f3))
race.f4 = as.vector (race.f4-mean (race.f4))
race.f5 = as.vector (race.f5-mean (race.f5))
URI = tapply (data$URI, data$Subject, max)
URI = as.vector (URI-mean (URI))
whb = tapply (data$Wheeze_Hard_Breathe, data$Subject, max)
whb = as.vector (whb-mean (whb))
dx = tapply (data$Dr_Dx, data$Subject, max)
dx = as.vector (dx-mean (dx))
med = tapply (data$Med_Daily, data$Subject, max)
med.f.L = as.numeric (med==1)
med.f.Q = as.numeric (med==2)
med.f.L = as.vector (med.f.L-mean (med.f.L))
med.f.Q = as.vector (med.f.Q-mean (med.f.Q))
amat =
as.data.frame (cbind (coefs25OUTIN, SE25OUTIN, spiro.f2, spiro.f3, spiro.f4, spiro.f
5, spiro.f6, spiro.f7, gender, asthma, race.f2, race.f3, race.f4, race.f5, URI, whb, dx,
med.f.L, med.f.Q))

w = 1/ (amat$SE25OUTIN^2)

#perform metaregression
metareg = lm (coefs25OUT ~ spiro.f2 + spiro.f3 + spiro.f4 + spiro.f5 +
spiro.f6 + spiro.f7 + gender + asthma + race.f2 + race.f3 + race.f4 + race.f5
+ URI + whb + dx + med.f.L + med.f.Q, data=amat, weights=w)
summeta = summary (metareg)
plot (metareg)
hist (metareg$residuals)

##transform coefs25OUT
amat$ytr = (-min (amat$coefs25OUTIN)+amat$coefs25OUTIN)^.5

#define weights for meta regression transformed
wt1 = .5*(-min (amat$coefs25OUTIN)+amat$coefs25OUTIN)^-.5
wt2 = wt1*amat$SE25OUTIN
wtt = wt2^-2

#perform metaregression transformed
metareg = lm (ytr ~ spiro.f2 + spiro.f3 + spiro.f4 + spiro.f5 + spiro.f6 +
spiro.f7 + gender + asthma + race.f2 + race.f3 + race.f4 + race.f5 + URI +
whb + dx + med.f.L + med.f.Q, data=amat, weights=wtt)
summeta = summary (metareg)

#check residuals of metaregression
plot (metareg)
hist (metareg$residuals, xlab="Residuals", main="Histogram of Meta-Regression
Residuals", col="grey", breaks=24)

#95% confidence interval for intercept effect size
confint (metareg, ' (Intercept) ', level=.95)
intervals (aquad, level=.95)

```