Using Statistics to Better Characterize the Residential Wood Combustion Emissions Project

Andrew S. Moore
Maya C. Thompson
Bryan T. Stines

Faculty Advisors: Mr. William F. Hunt, Jr.
Dr. Roger Woodard

Client

- Mid-Atlantic Regional Air Management Association
  - Susan Wierman, Executive Director
  - Megan Schuster, Project Manager

Problems Created by Residential Wood Combustion (RWC)

- Health Effects:
  - Cardiovascular Disease
  - Respiratory Illness
    - Asthma, Bronchitis, Emphysema
  - Release of cancer-causing chemicals
- Accounts for approximately 8% of PM2.5 emissions in MANE-VU region

Problems Created by RWC

- Quality of Life Effects

  - Contributes to Regional Haze

    * Reduced Visibility

http://www.manevu.org/air_pollution.asp

Equipment used for RWC

- Indoor Equipment
  - Fireplaces
  - Woodstoves
  - Furnace/Boilers
  - Pellet Stoves
- Outdoor Equipment
  - Barbeques
  - Fireplaces
  - Fire Pits
  - Chimneys

Background

- 1999: EPA implemented a program to decrease haze and improve visibility in natural parks
- Mid-Atlantic/Northeast Visibility Union (MANE-VU)

  - Created to coordinate regional haze planning activities for the region

Background

- E.H. Pechan & Associates, Inc.
  - Contracted by MARAMA to develop an emissions inventory for 2002 in the MANE-VU region
- Population Research Systems
  - Conducted telephone survey of MANE-VU residents in 2003

States Included in the Survey

- Connecticut
- Delaware
- District of Columbia
- Maine
- Maryland
- Massachusetts
- New Hampshire
- New Jersey
- New York
- Pennsylvania
- Rhode Island
- Vermont

Initial Approach Used by E.H. Pechan & Associates, Inc.

- 3 Stratification Variables used to group survey respondents
  - Geographic Location
  - Household Type
  - Heating Degree Day Zone
- General Linear Model created to analyze significant effects on household emissions of PM 2.5
  - Model consisted of three-way interactions from the original stratification variables

Creation of “Super Cells”

- Based on the mean household emissions for each combination of the 3 stratification variables, certain groups were combined because they were deemed to not be significantly different
  - Other, RNF/S/U, High/Med/Low
  - Other, RF, High/Med/Low
  - Single, Urban, High/Med/Low
- This process resulted in 12 “super cells” and a new general linear model was created

Results of General Linear Model with Collapsed Categories

	DF	SS	MSE	F Value	Pr > F
Model	12	2419383.57	201615.3	23.69	< 0.0001
Error	1880	16000571.14	8510.94		
Corrected Total	1892	18419954.71			

R-Square	CV	Root MSE	Mean
0.08	393.71	92.25	23.43

Parameter	Estimate	Std. Error	T value	Pr > |t|	Significant
Sing, Urb, All	1.88	6.41	0.29	0.7693	
Oth, RNF/S/U, All	1.91	3.78	0.51	0.6129	
Oth, RF, All	6.01	6.8	0.88	0.3768	
Sing, Sub, Low	18.14	10.58	1.71	0.0867	
Sing, RNF, Low	19.45	8.53	2.28	0.0227	****
Sing, RF, Low	46.68	7.53	6.2	< 0.0001	****
Sing, Sub, Med	39.7	11.27	3.52	0.0004	****
Sing, RNF, Med	24.81	9.67	2.57	0.0104	****
Sing, RF, Med	55.89	9.89	5.65	< 0.0001	****
Sing, Sub, High	21.4	12.01	1.78	0.0749	
Sing, RNF, High	72.67	9.95	7.31	< 0.0001	****
Sing, RF, High	79.94	7.03	11.36	< 0.0001	****

Analysis of Collapsed Categories: Are Cells Significantly Different?

i/j	1	2	3	4	5	6	7	8	9	10	11	12
1		1.0000	<0.0001	0.0037	0.0020	<0.0001	0.9864	0.9130	0.9941	0.9984	0.3042	1.0000
2	1.0000		<0.0001	<0.0001	<0.0001	<0.0001	0.7716	0.5462	0.9269	0.9546	0.0662	1.0000
3	<0.0001	<0.0001		0.0572	0.7059	1.0000	<0.0001	0.0003	0.0016	<0.0001	0.1009	<0.0001
4	0.0037	<0.0001	0.0572		0.9999	0.6350	0.4116	0.8268	0.8275	0.5521	1.0000	0.0004
5	0.0020	<0.0001	0.7059	0.9999		0.9896	0.1851	0.5154	0.5377	0.2769	0.9955	0.0003
6	<0.0001	<0.0001	1.0000	0.6350	0.9893		0.0030	0.0284	0.0477	0.0097	0.5545	<0.0001
7	0.9864	0.7716	<0.0001	0.4116	0.1851	0.0030		1.0000	1.0000	1.0000	0.9572	0.8914
8	0.9130	0.5462	0.0003	0.8268	0.5154	0.0284	1.0000		1.0000	1.0000	0.9977	0.7101
9	0.9941	0.9269	0.0016	0.8275	0.5377	0.0477	1.0000	1.0000		1.0000	0.9943	0.9569
10	0.9984	0.9546	<0.0001	0.5521	0.2769	0.0097	1.0000	1.0000	1.0000		0.9648	0.9774
11	0.3042	0.0662	0.1009	1.0000	0.9955	0.5545	0.9572	0.9977	0.9943	0.9648		0.1359
12	1.0000	1.0000	<0.0001	0.0004	0.0003	<0.0001	0.8914	0.7101	0.9569	0.9774	0.1359	

1: O, RF, All
2: O, RNF/S/U, All
3: S, RF, High
4: S, RF, Low
5: S, RF, Med
6: S, RNF, High
7: S, RNF, Low
8: S, RNF, Med
9: S, Sub, High
10: S, Sub, Low
11: S, Sub, Med
12: S, Urb, All

Pr > |t| for Ho: LSMean(i)=LSMean(j)

Means are significantly different (alpha=0.05)

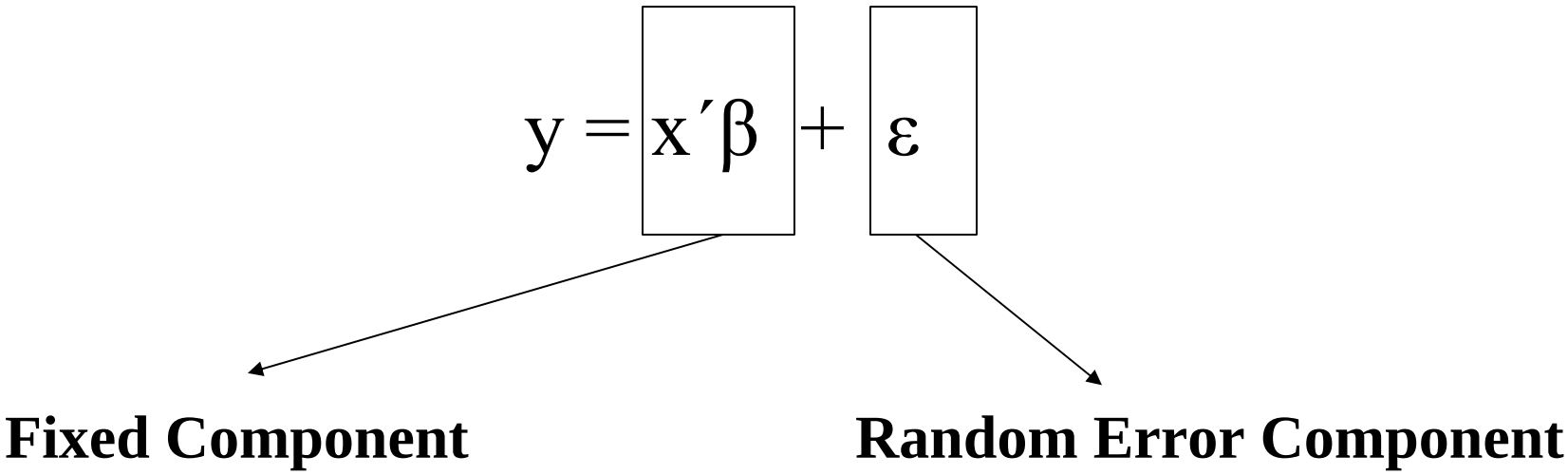
Means not significantly different (alpha=0.05)

Analysis of Collapsed Categories

- Results of this analysis indicate many categories are still not significantly different from each other
- The general linear model parameter estimates for some classifications are not significantly different from zero
  - We are unable to remove these parameters from the model because they represent the three-way interaction between geographic location, household type, and heating degree day zone
  - If these parameters were removed, there would be no estimate for a large part of the population

Assumptions of the General Linear Model

- The observed values of the dependent variable can be written as the sum of a fixed component which is a linear function of the independent variables and a random error component:



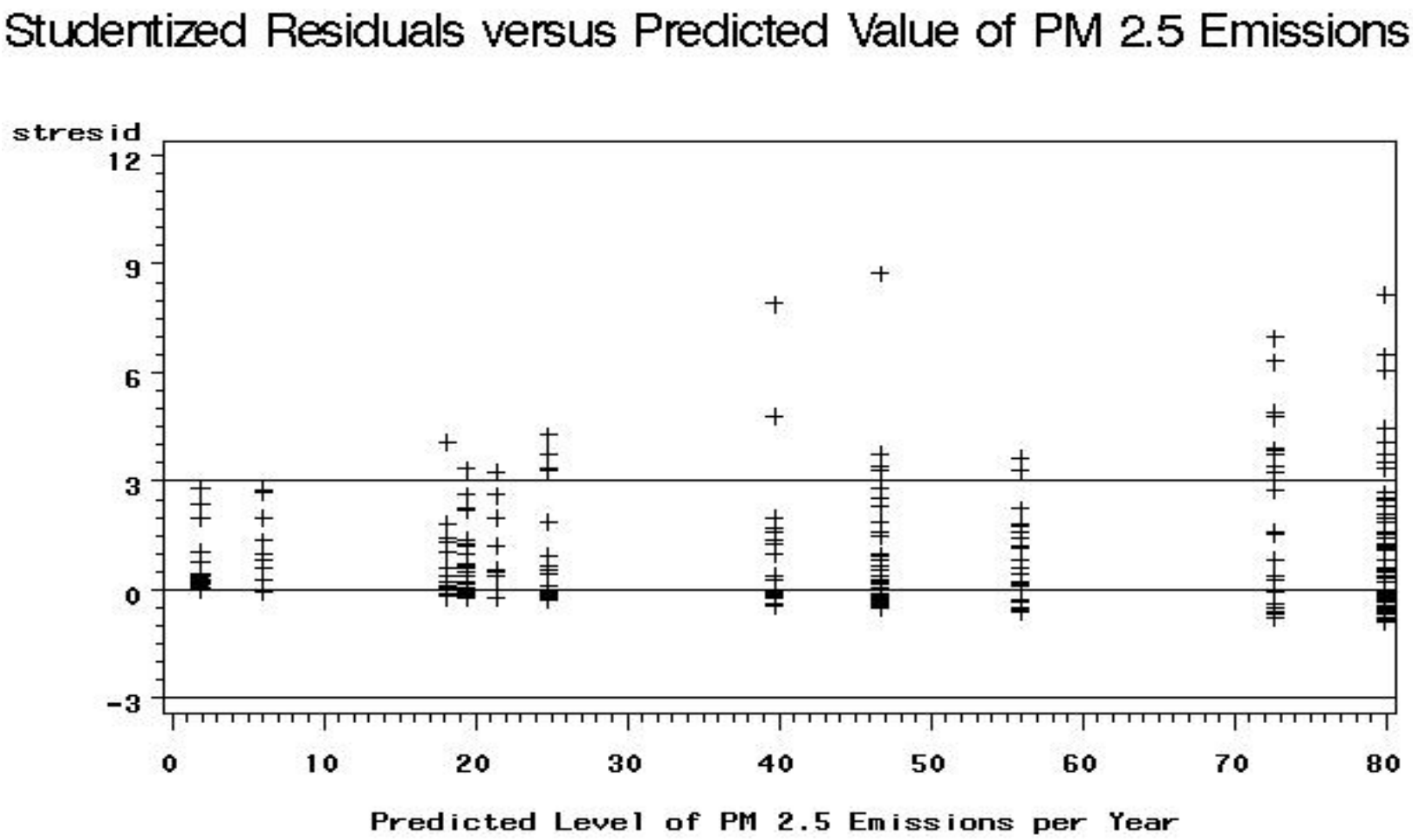
Assumptions about the Random Error Components

1. Random error components are independent
2. Random error components have a constant variance
3. Random error components are identically, normally distributed with $E(\varepsilon)=0$

Model Validity and Assumptions about the Random Error Components

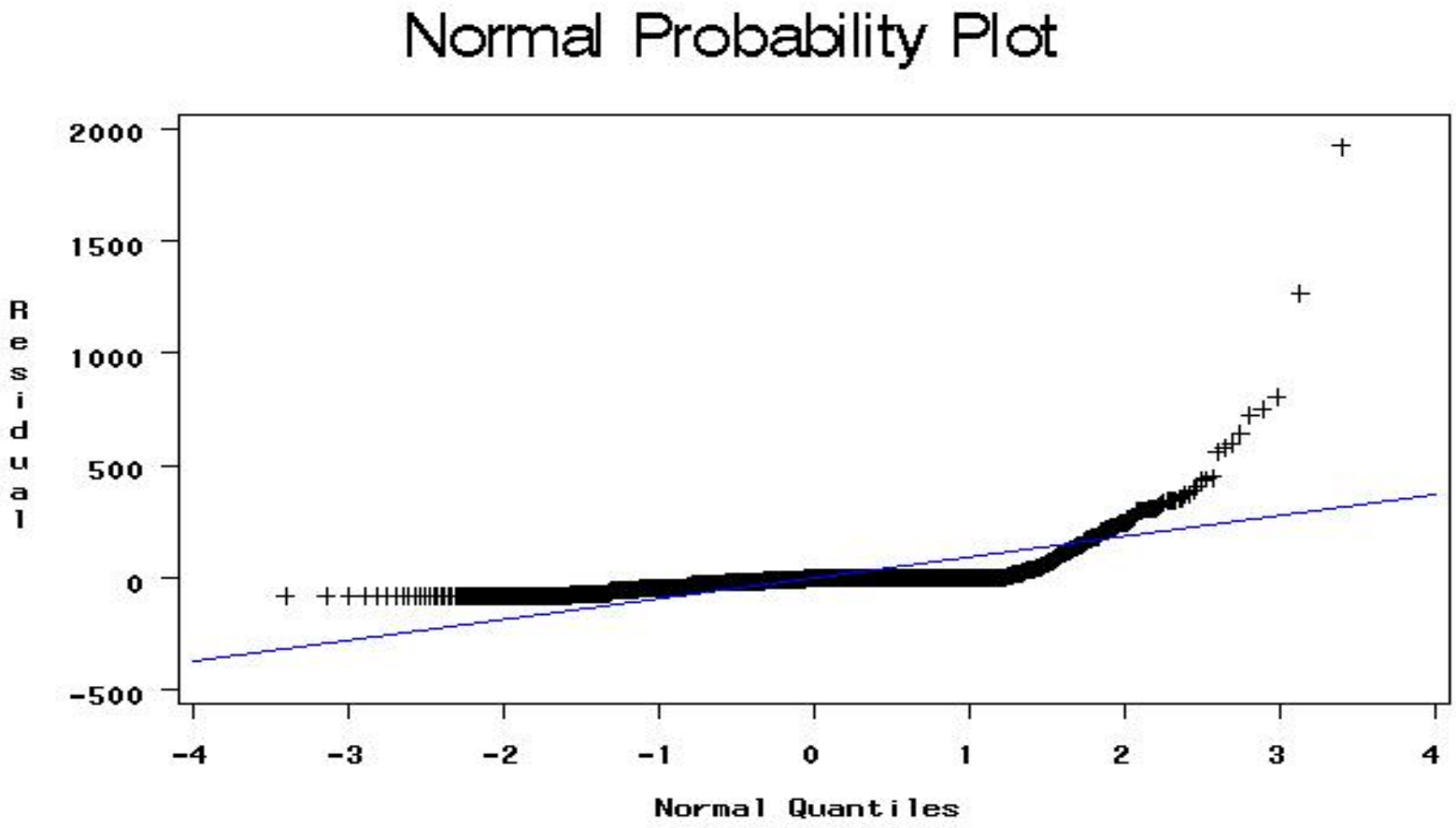
- Assumptions 1 and 2 ensure that the linear parameter estimates from the general linear model are the MVUE among linear estimators
- The normality assumption ensures that the linear parameter estimates from the general linear model are the maximum likelihood estimates with a known distribution
  - Required for confidence limits and significance levels to be exactly valid
  - Non-normality can reduce the power of statistical tests

Assumption of Constant Variance



Random error components do not have a constant variance

Assumption of Normally Distributed Random Errors



Plot of residuals should fall along the straight line if they are approximately normal

Primary Source of Problems for Current Analysis

- Majority of respondents do not participate in residential wood combustion
  - Distribution of PM 2.5 Emissions is Highly Skewed
  - Original stratification variables only explain approximately 8% of the variability in individual household emissions of PM 2.5

Objectives of Student Analysis

- Focus on PM 2.5 Emissions of Wood Burners
  - Distribution
  - Central Tendency
  - Variability
- Analyze the effect of the stratification variables on household emissions of PM 2.5

Summary Statistics: Wood Burners

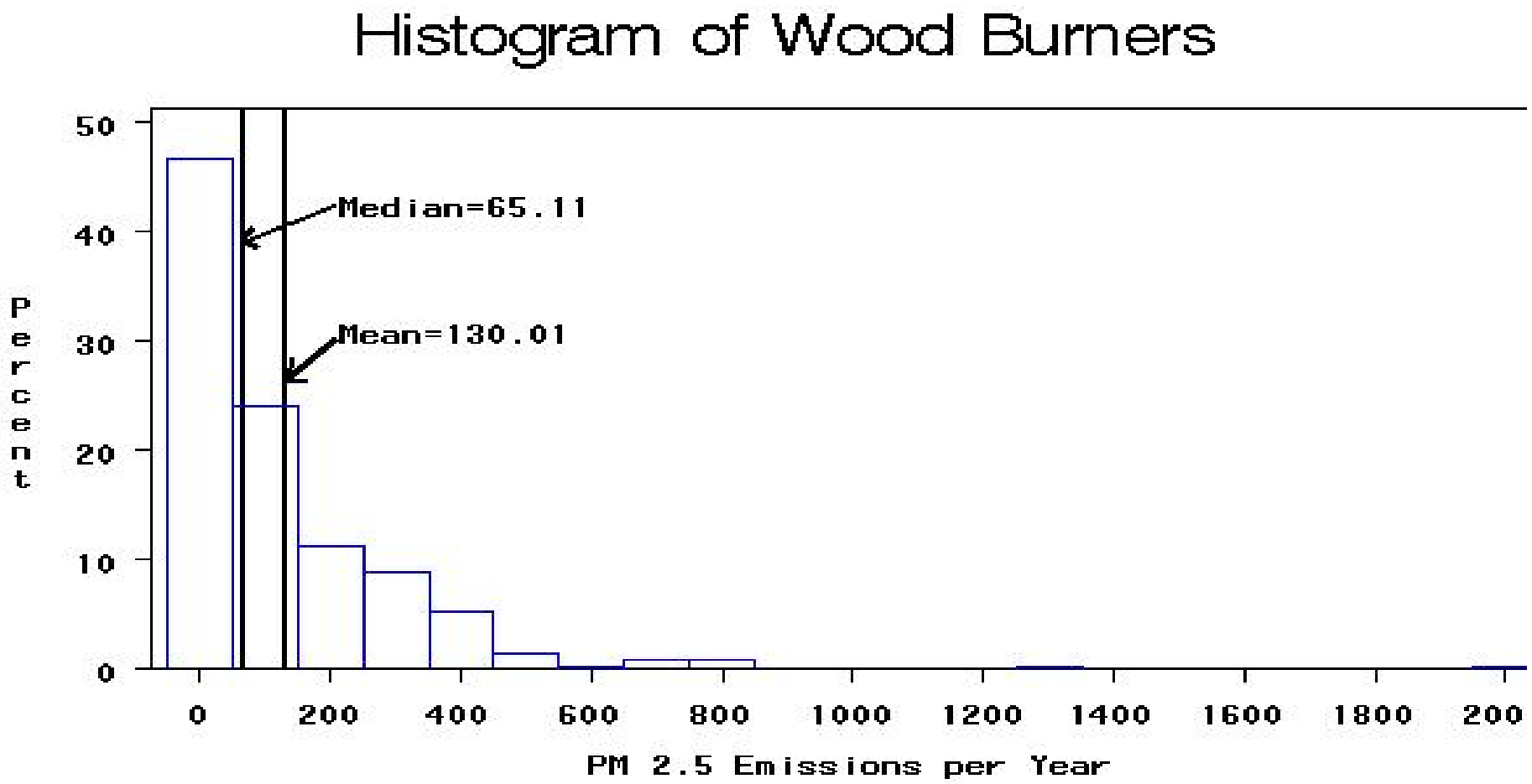
Central Tendency			
N	**Mean**	**Median**	**Mode**
341	130.01	65.11	130.21

Variability				
Variance	**Std. Dev.**	**Std. Error**	**Range**	**IQR**
37224	192.94	10.45	2001	172.61

Quantiles	**Estimates**
100%	2001.06
99%	831.18
95%	416.17
90%	325.53
75%	188.99
50%	65.11
25%	16.38
10%	2.94
5%	1.37
1%	0.51
0%	0.38

Smallest Observation	**Largest Observation**
0.37927	762.979
0.40630	831.179
0.43621	849.919
0.50577	1324.187
0.60549	2001.06

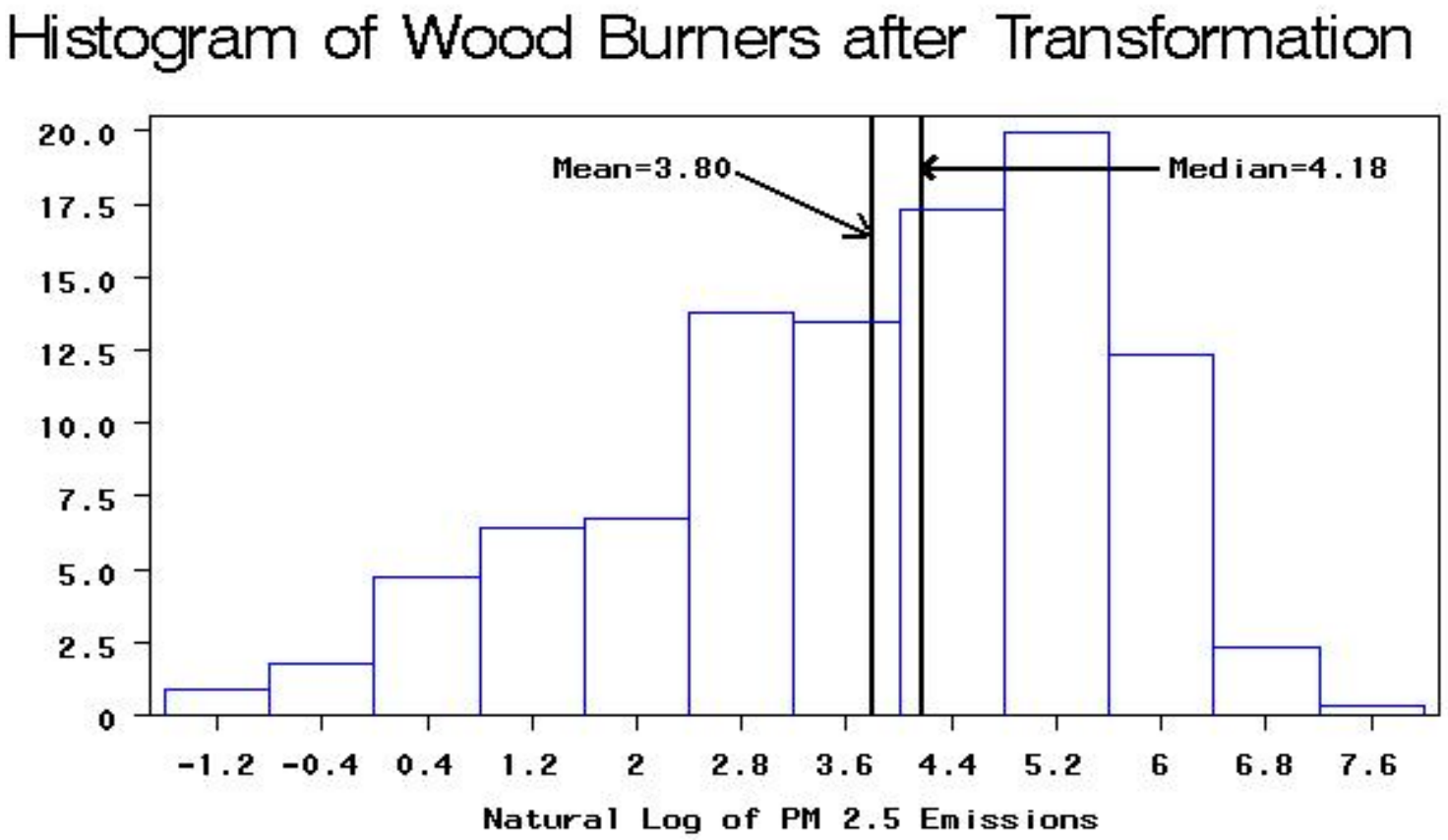
Distribution of Wood Burners



Data Transformation

- Since the distribution of household emissions was so highly skewed, a linear transformation of the response variable was considered
  - $\sqrt{PM2.5Emissions}$
  - $\ln(PM2.5Emissions)$
- Natural Log transformation produced the best results with respect to normalizing the data

Distribution of Natural Log of PM 2.5 Emissions



Analysis of Wood Burners

- Once the response variable was transformed to more closely resemble a normal distribution, a model was generated to determine the effects of geographic location, household type, and heating degree days on household emissions of PM 2.5

Model Building

- The first model to be analyzed contained all main effects, two-way interactions, and three way interactions

Source	DF	SS	MS	F	Pr > F
Model	21	178.87	8.52	3.09	<0.0001
Error	319	879.48	2.76		
Corrected Total	340	1058.35			

R Square	CV	Root MSE	Mean
0.1690	43.70	1.66	3.80

Source	MS	F Value	Pr > F
Geog Loc	11.31	4.10	0.0071
HDD Zone	7.88	2.86	0.0588
HH Type	0.01	0.00	0.9506
GL*HDD	2.14	0.78	0.5886
GL*HH	4.92	1.78	0.1500
HDD*HH	0.30	0.11	0.8956
GL*HH*HDD	2.76	1.00	0.4077

Model Building: 2nd Attempt

- Three-way interactions in previous model were not significant (p-value=0.4077) so a new model was generated with main effects and two-way interactions

Source	DF	SS	MS	F	Pr > F
Model	17	167.84	9.87	3.58	<0.0001
Error	323	890.5	2.76		
Corrected Total	340	1058.35			

R Square	CV	Root MSE	Mean
0.1586	43.70	1.66	3.80

Source	MS	F Value	Pr > F
Geog Loc	6.58	2.39	0.0691
HDD Zone	9.86	3.58	0.0291
HH Type	0.01	0.01	0.9414
GL*HDD	1.14	0.41	0.8691
GL*HH	4.23	1.53	0.2053
HDD*HH	0.93	0.34	0.7147

Model Building: 2nd Attempt Continued

- Two-way interactions were not significant when the main effects for the stratification variables were also included

Model Building: 3rd Attempt

- A model with only main effects for geographic location, household type, and HDD zone was created

Source	DF	SS	MS	F	Pr > F
Model	6	145.89	24.31	8.9	<0.0001
Error	334	912.46	2.73		
Corrected Total	340	1058.35			

R Square	CV	Root MSE	Mean
0.1378	43.50	1.65	3.80

Source	MS	F Value	Pr > F
Geog Loc	30.08	11.01	<0.0001
HDD Zone	19.81	7.25	0.0008
HH Type	0.98	0.36	0.5494

Model Building: 3rd Attempt Continued

- The main effect for household type was not significant in the previous model so it was removed and a new model was generated

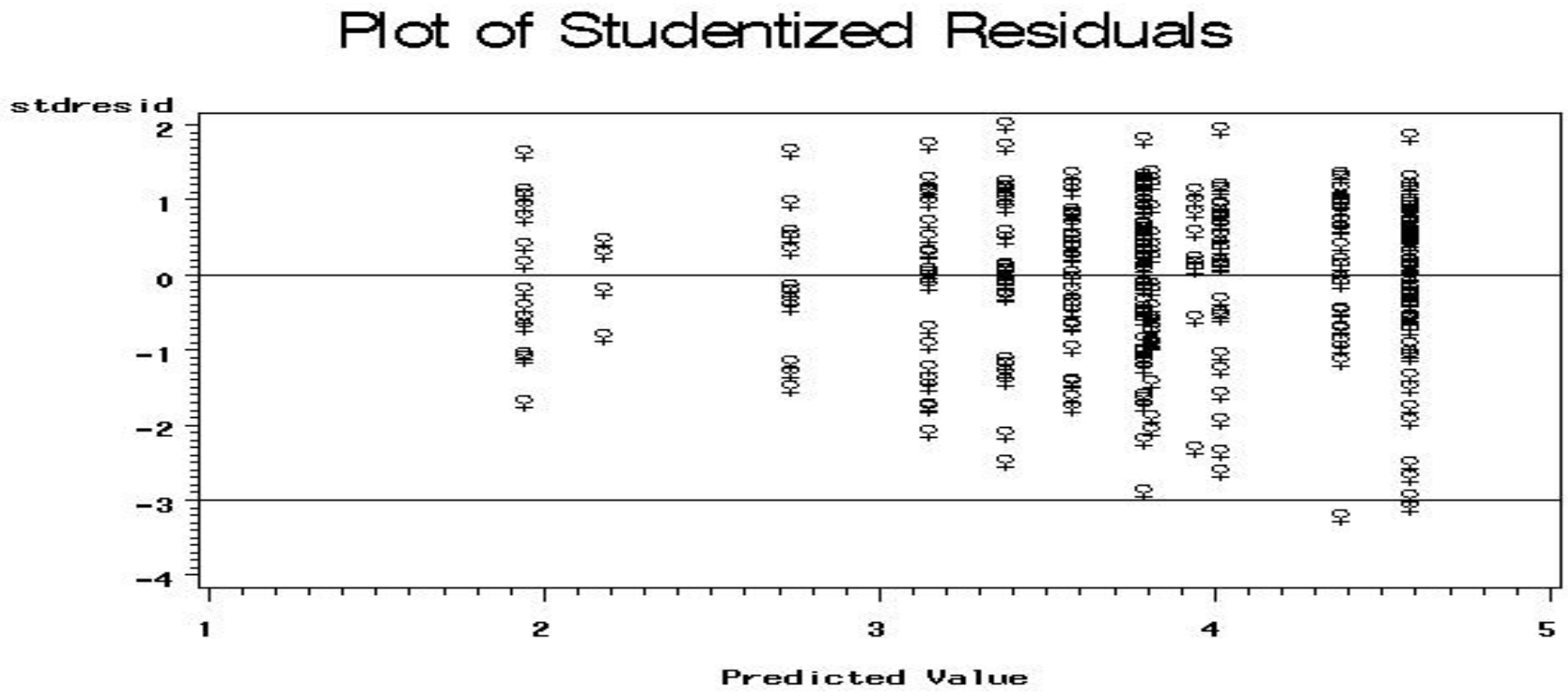
Source	DF	SS	MS	F	Pr > F
Model	5	144.91	28.98	10.63	<0.0001
Error	335	913.44	2.73		
Corrected Total	340	1058.35			

R Square	CV	Root MSE	Mean
0.1369	43.45	1.65	3.80

Source	MS	F Value	Pr > F
Geog Loc	31.85	11.68	<0.0001
HDD Zone	20.05	7.35	0.0007

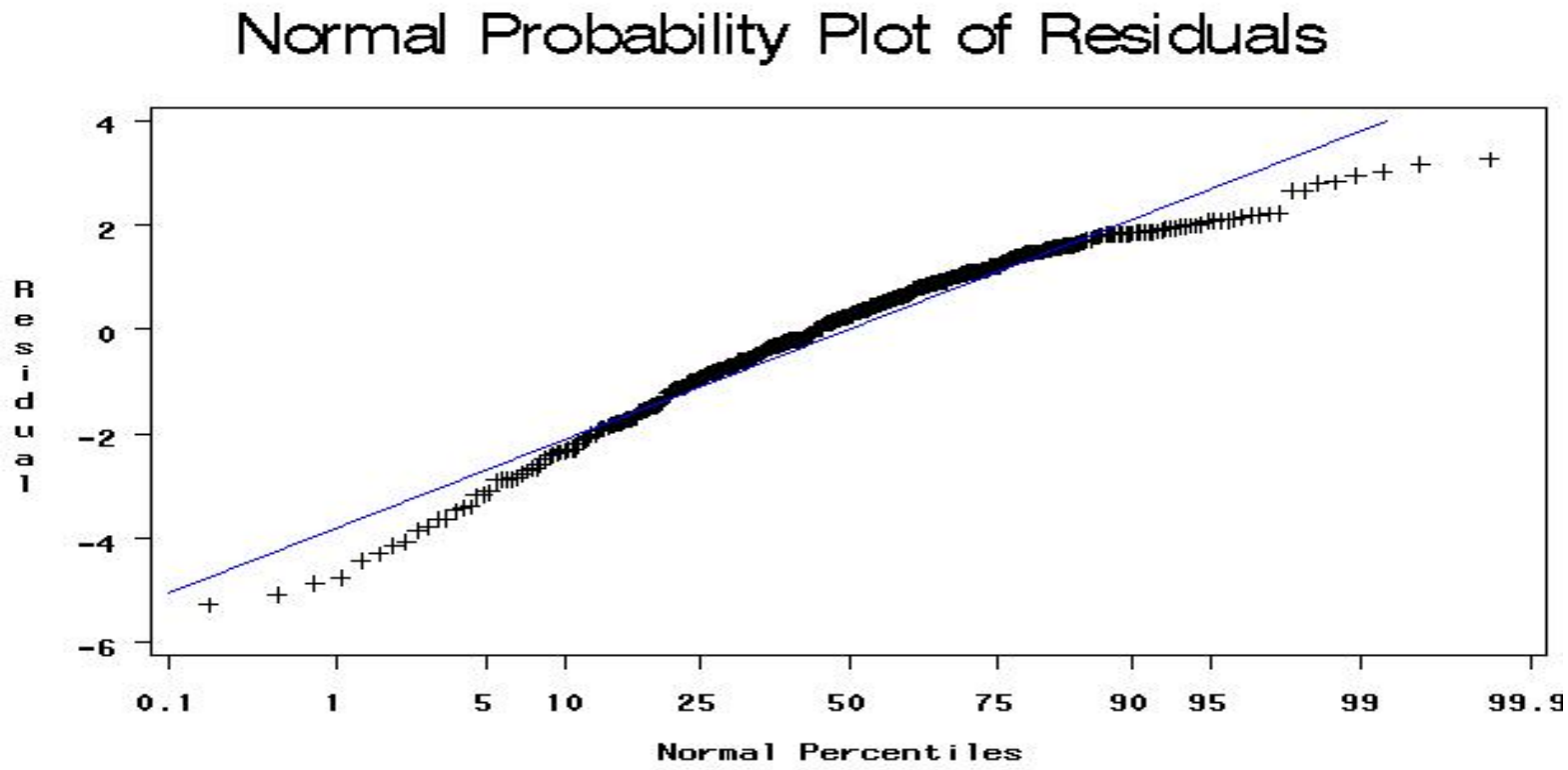
- Once the final model was generated, an analysis of the residuals was performed to determine if model assumptions were satisfied

Assumptions of Random Error Components: Constant Variance



- Majority of residuals within 3 standard deviation limits
- Variance approximately constant

Assumptions of Random Error Components: Normality



• Slight downward concavity indicates some left skewness but no serious departures from normality

Using Model to Predict Emissions

- Preliminary estimates from the final proposed model indicate that emissions for the MANE-VU region may be significantly less than the estimate provided in the original analysis

  - However, further analysis needs to be conducted, specifically with regard to the standard error of the estimate and the possibility of a slightly biased estimator

Conclusions

- Household emissions of PM 2.5 are strongly positively skewed (mean >> median)

  - Natural log transformation of response variable alters the distribution so that analysis of variance can be performed and linear models may be created

Conclusions

- Geographic Location and Heating Degree Day Zone have a significant effect on the amount of PM 2.5 emissions produced by people who burn wood
- Household Type does not significantly affect the amount of emissions at the household level for wood burners
- These variables do not interact in a significant way to affect the amount of emissions produced by wood burners

Recommendations

- Attempt to incorporate more information into the model to increase the amount of variability explained by the model:
  - Equipment Type
  - Type of wood used
- Use the model for predicting household emissions and creating an emission inventory for the region
- Investigate the possibility of using additional information to improve the estimate of total emissions through ratio estimation