Trends in Atmospheric Carbon Dioxide

Danny Ober-Reynolds, Jimin Nam, TJ Radigan, Katherine Wu

March 22, 2016

Abstract

This paper discusses the statistical approach in analyzing atmospheric carbon dioxde (CO2) trends, based on the annual and monthly average CO2 datasets provided by the Mauna Loa Observatory in Hawaii. We implement two of the fundamental models and outline the process of statistical testing and model adjustment. Even with the various transformations of both our models, there is always a increase in CO2 level with month and year. Thus, although there is seasonal variation, we have enough evidence to conclude that CO2 levels have a accelerating positive trend with each year, ultimately leading to higher levels with time.

1 Introduction

Modern climate scientists credit a large part of the recent trend in warmer global temperatures to increases in the global stock of greenhouse gases, especially CO2 resulting from human activities. CO2 can also vary naturally in time based on the season, and locally due to vegetation and human activities. The climate science hypothesis under consideration says that such temporally and spatially located CO2 will not cause global temperatures to vary significantly, while the long term global stock of CO2 can change both the rate at which the Earth warms and the long run equilibrium temperature.

A statistical approach is needed, one capable of separating the long term trends of CO2 from the seasonal fluctuations and one not affected by the localized effects of vegetation and economic production. A model able to predict the long run global stock of CO2 in the atmosphere that accounts for the natural fluctuations seen in the season could also be very useful in decision making regarding polluting production processes, which would be costly to give up. Our project is to build such a model, and use the results as evidence for (or against) the hypothesis that global stock values of CO2 have increased independent of seasonal fluctuations.

2 Data Source

In 1957, Dave Keeling was the first to capture accurate measures of CO2 in the atmosphere at a site high on the slopes of the Mauna Loa volcano. Since then, the goal of measuring CO2 in air masses that is representative of the globe. Presently, the CO2 measurements are collected at the Mauna Loa Observatory, in a location surrounded by miles of bare lava, and without vegetation or soil. In this particular setting, the measurements of CO2 will not be influenced by any absorbed or emitted locally by plants and soil, as well as human activities.

At Mauna Loa, a data selection criteria is strictly followed in order to maintain consistency and accuracy. First the standard deviation of minute averages should be less than 0.30 ppm per hour. Following that, it is regulated that the hourly average should differ by less than 0.25 ppm. Due to fluctuations of wind flow on Mauna Loa, it is noted that there is surface warming during the day, while it cools during the night. Based on the slope of the winds, particular hours are flagged, yet included as there the final requirement of data selection addresses the stability of background air. A general "outlier rejection" process is used, where a curve is fit to the preliminary daily means for each day calculated from the hours (post-elimination). From there, any hourly averages that reside outside of two standard deviations are "rejected" and marked from the sample collection. While no data is essentially thrown away, the hourly means are calculated based on the data that passes all the selection flags.

3 Data

While the Mauna Loa Observatory provides a wide variety of datasets, our analysis focused on both Mauna Loa CO2 monthly and annual mean data. The Mauna Loa CO2 monthly mean data are collected from March 1958 through February 2016. The variables included are as follows:

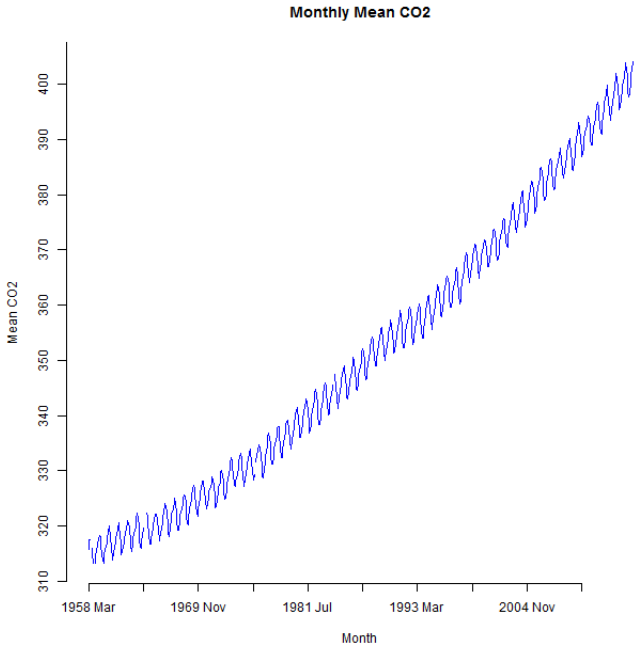
- Date (year and month) of when the data was collected in provided separately by month of each year from 1958 through 2016 (most recent). The variable of decimal date is combines the month and year to a single measurement in the format $year + \frac{x}{365}$, where x is the number of days that have passed since 1/1/1958 12:00 AM.

- Average is the variable for the monthly mean CO2 mole fraction based on daily averages, where the mole fraction of CO2 is expressed as parts per million (ppm). Missing data is marked as -99.99, and later replaced as NA in R.

- Interpolated represents the "average" column, where the missing data points are replaced by interpolated values. These values are computed through a process where the average seasonal cycle in a 7-year window around each monthly value is found for each month, in order to represent the slow change of the seasonal cycle.

- Trend (season corr) represents values found to be linearly interpolated for missing months. The interpolated monthly mean is then the sum of the average seasonal cycle value and the trend value for the missing month.
- The number of days (# days) is self-explanatory, and is only counted if there is data for the daily means of the month. In the case that there is no data for the daily means of the month, a -1 is used; however, that also is replaced as NA in R.

Using the monthly mean CO2 dataset, the following graphs show the time trend of monthly CO2:



As shown on the graph, there is suspicion that there is an increase in mean CO2, as we observe seasonal effects and the general increase with time. We further exploit this in the discussion section.

The Mauna Loa CO2 annual mean dataset is also collected from 1958 through 2015. The variables included in this set are as followed:

- Year is included from 1958 through 2015, and there is no missing year in the annual mean dataset.
- Mean is the variable for the annual mean CO2, where CO2 is expressed as a mole fraction in dry air, micromol/mol (ppm).
- Uncertainty (unc) represents the estimated uncertainty in the annual mean is found through the standard deviation of the differences in annual mean values, which are determined independently.

4 Methods

4.1 Simple Regression Model

In order to answer the question of whether the mean CO2 level is significantly changing each year, a regression model based on the data described above was created. To determine the incremental effect of one year on mean CO2 levels, our first model is a simple linear regression model predicting mean CO2 levels with year as the explanatory variable from the annual data. Since the data was collected annually, 57 years of data (at one year intervals) were observed with their mean CO2 levels to determine whether there was a relationship. The model estimated is:

$$mean = \beta_0 + \beta_1 year + \varepsilon \tag{1}$$

This model has only 57 data points, a relatively small sample, and yearly mean CO2 levels, a relatively large time step. The model could be improved by accounting for more information that is available–namely, monthly data on CO2 levels.

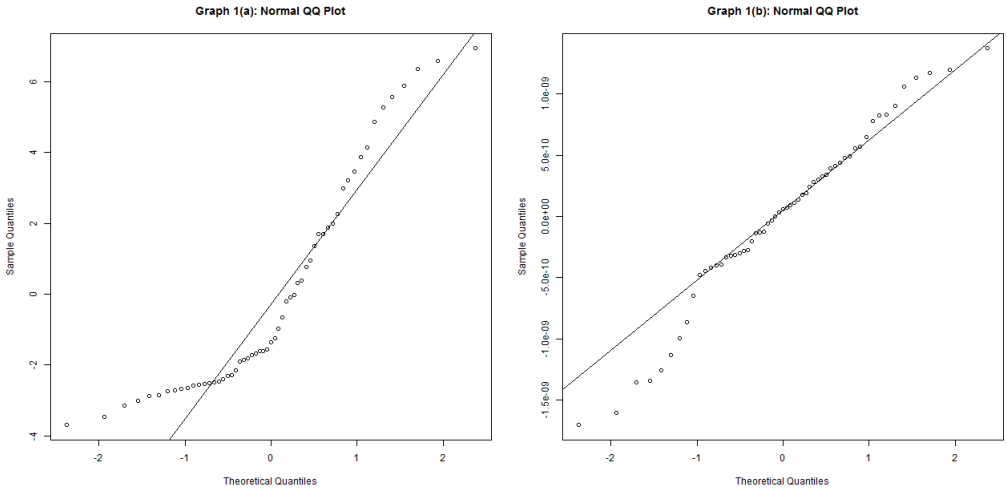
4.2 Multiple Regression Model

Our second multiple regression model is able to control for seasonality associated with CO2 levels in different months. Since CO2 levels are recorded for each month of the year, evaluating changes in CO2 levels for a fixed month between two years controls for seasonality and gives a better indication of the effect of the increasing year. It also serves to dramatically increase the number of data points, and the model benefits from the greater sample size. The multiple regression model accounts for months as factor variables, or dummy variables. January, month 1, is taken to be the base case. Each other month is a binary factor variable which equals one in that month and zero in other months. In this way, the effects of individual months can be controlled out of the yearly effect, and the relative CO2 levels in individual months can be assessed. This allows for inference on seasonality effects as well as isolated yearly effects. It is worthy of note that seven records were omitted in the monthly data series due to missing pieces of data.

5 Results

The results for the simple regression model predict an increase in 1.522 ppm of mean CO2 levels for each passing year, with a standard error of 0.025. This yields a p-value of (2 * 10^-16), offering an extremely strong case for significance at traditional significance levels of 0.05, 0.01, and even 0.001. The model's adjusted R-squared value of 0.9856 indicates that roughly 98.56% of sample variation in CO2 levels can be explained by the passage of time. For following graphs display the Normal Q-Q plot for the original simple regression model (Graph 1a) and the transformed model (Graph 1b) that best fit our assumptions.

Unfortunately, the original simple linear regression model (1) displayed significant evidence of violating the normality assumption of simple linear regression.



A Shapiro-Wilk normality test of the residuals resulted in a p-value of $2.723 * 10^{-5}$, confirming our suspicion that this assumption was violated. After trying several transformations the mean and year variable, we found that the simple linear regression model

$$mean^{-2.75} = \beta_0 + \beta_1 \cdot year + \varepsilon \tag{2}$$

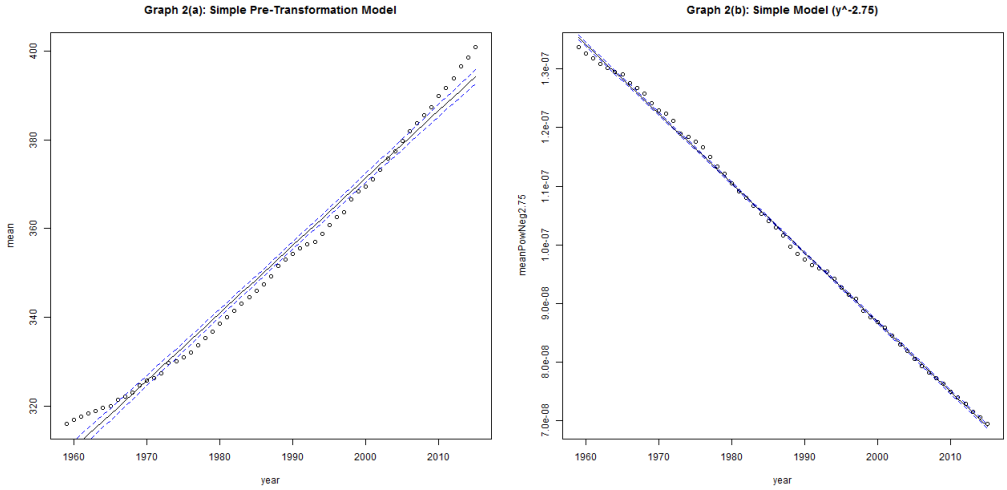
fit this assumption much better, as displayed in the Graph (1b): Normal Q-Q Plot.

A Shapiro-Wilk normality test resulted in a p-value of 0.2475. This updated simple linear regression model still found a very statistically significant negative coefficient on year, indicating that as years pass the mean does indeed increase. Going forward, we used this transformed mean as our outcome variable. We

include both the regression results of our original simple regression model, and our newly transformed model in our appendix.

Least squares regression of model (2) found an estimate of the coefficient on year of $-1.184 * 10^{-9}$ with a standard error of $5.768 * 10^{-12}$. In a two-tailed hypothesis test with $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$, this yields a p-value of less than $2 * 10^{-16}$, indicating strong evidence that this effect is statistically significant at traditional significance levels of 0.05, 0.01, and even 0.001. The model's coefficient of determination value of 0.9987 indicates that roughly 99.87% of sample variation in CO2 levels can be explained by the passage of time.

The following graphs include our simple regression model on the annual mean dataset as a scatter plot of the annual mean CO2 data, with the predicted line, and the 95% confidence band.



Our multiple regression model

$$mean^{-2.75} = \beta_0 + \beta_1 year + \beta_2 February + ... + \beta_{12} December + \varepsilon$$

indicates that seasonality does play a role in monthly mean CO2 levels, but the coefficeint year ($\hat{\beta_1} = -1.182 * 10^{-9}$, $sd(\hat{\beta_1}) = 1.959 * 10^{-12}$) did not change signficantly and remained statistically signficant with a two-tailed hypothesis test (as in our simple linear regression model) resulting in a p-value less than $2 * 10^{-16}$. This indicates that that there is strong statistical evidence that as year increases, the mean CO2 level increases as well. The adjusted R-squared value is actually slightly lower for the multiple regression model at 0.9981, indicating that roughly 99.81% of sample variation in CO2 levels can be explained by the passage of time after controlling for seasonality.

In the multiple regression model, every month has significantly different mean CO2 concentrations from January, the base case. For example, the coefficient on May was estimated to be $-2.796 * 10^{-9}$ (standard deviation $1.593 * 10^{-10}$) with a p-value less than $2 * 10^{-16}$, indicating that May has a significantly higher CO2 concentration at any reasonable significance level. The regression results of our post-transformation is included in our appendix.

6 Discussion

In order to determine whether CO2 levels increased over the course of years and months, both a simple linear regression model on the annual mean dataset with annual mean CO2 and a multiple regression model on the monthly mean dataset with monthly mean CO2 were created. Once the results were tabulated from the following simple regression model and multiple regression model, the data explicitly shows apparent correlations with various data points over a time series.

6.1 Simple Regression Model: Annual Mean CO2 Dataset

Based on a scatter plot of the simple linear regression model, it does suggest that there is an increasing linear relationship between the annual mean data with annual mean CO2 over the course of several years. It also suggests that there are no unusual data points in the data set, and it illustrates that variation around the estimated regression line from 1958 to 2016 is constant, suggesting that the assumption of equal error variances is reasonable. Because no data plots are extremely far apart from the predicted line, overall error in the data is not expected to be extreme. In addition, many of the plotted points fall within or extremely close to the 95% confidence band around the predicted line. Thus, estimations provided in the data are reasonably reliable to make overall assumptions from data collected at the Mauna Loa Observatory, Hawaii site.

While the simple linear regression model appeared to have a positively linear trend with annual mean CO2 and year(s), it appears that our model was flawed after testing for nonlinearity, unequal error variances, or outliers through a residuals versus fitted plot, and normality through the Shapiro-Wilk normality test. The residuals versus fitted plot of the annual data depicted a "U" shaped figure, suggesting that the model is not a good fit and may require a transformation of the independent variable. Furthermore, this model also violated the normality assumption, as discussed in Results, with a very small p-value on the Shapiro-Wilk normality test. With evidence of an inaccurate model, we made structural changes to the form of the model to find a better fit; however, we note that numerous transformation of both the independent and dependent variables failed to better the model. We did settle for a regression model of

$$mean^{-2.75} = \beta_0 + \beta_1 \cdot year + \varepsilon,$$

which does satisfy the Shapiro-Wilk normality test; however, it make the coefficient very difficult to interpret. Nonetheless, even with the transformed simple regression model, it appears that the long term trends of CO2 separated from seasonal fluctuations have increased over time.

6.2 Multiple Regression Model: Monthly Mean CO2 Dataset

To expand on this study and to provide more validity behind the relationship between CO2 levels and time, we created a multiple regression model from the monthly mean dataset with monthly mean CO2.

With observations that are on a monthly basis, we have more data points to provide a more comprehensive understanding of the trends between the two variables we are looking at. In our multiple regression model, it appears that there is still a positive linear relationship between the two variables even after accounting for seasonality effects throughout a single year. While the predictive power regarding the effect of yearly CO2 levels does not differ drastically from the annual data, it gives more bases and evidence behind the analysis of trends between time and CO2 levels. This suggests that there is still strong evidence for a trend in continuously increasing CO2 levels, which can lead to issues such warmer global temperatures at an accelerating rate in the long-term future. Furthermore, by being able to look at the different levels ranging across consecutive months, not only can we see that the peak level for CO2 becomes high and higher with each passing year, but it does so in large increments at an accelerating rate over time.

Similar to our simple linear regression model, the residuals versus fitted values graphs of our multiple regression model also reflect a pattern, indicating potential nonlinearity. In order to find a line of better fit, we went through various transformations and settled with a multiple regression model of:

$$mean^{-2.75} = \beta_0 + \beta_1 year + \beta_2 February + ... + \beta_{12} December + \varepsilon.$$

Even with a transformed model, we find that the trend is still positive, giving evidence towards increasing CO2 levels over longer lengths of time. Thus, it is possible to assume that even after adjusting for seasonality, the overall CO2 levels will continue to grow in size, but according to the trend, at a more linear rate.

Conclusively, a statistical analysis of annual and monthly averages of concentration of atmospheric carbon dioxide establish a long term increasing trend of CO2 levels over time. While the annual mean data itself does not account for changes due to seasons, the monthly data supports the trend shown in the annual model with further detail about potential changes throughout a single year over a series of time. Additionally, the low standard error and high adjusted R-square provide validity to relationships between the two variables even after controlling for seasonality and adjusting for nonlinearity in the initial model.

Based on the multiple regression model with our monthly mean CO2 level dataset, we can predict the monthly average CO2 level for each month in 2016, as listed in the following table:

Month	Estimated mean $(\hat{y})^{-\frac{1}{2.75}}$
January	403.05
February	404.34
March	405.97
April	408.19
May	405.64
June	408.14
July	405.64
August	402.09
September	399.34
October	399.25
November	401.52
December	403.81

6.3 Future Studies

For future studies, there are a few potential weaknesses in the analysis and data that must be referenced when creating a judgment on the overall analysis. Local influence on carbon dioxide at the Mauna Loa Observatory such as weather and air may have slightly influenced the data collection process, which cannot be expunged from the model. There are also potential equipment failures that may affect the quality of the data collected in a short period of time. Because the experiment is over a long period of time, we have enough data to overlook these potential equipment failures when paying attention to the broad trend. Furthermore, there is potential to create a model that does adjust for the normality issue in the original model in this study that can create a stronger line of best fit that better represents the data without any bias.

7 Conclusion

In our analysis on the change of CO2 levels in Mauna Loa, we developed two separate models based on the monthly and yearly datasets that is made available by the Earth Systems Research Laboratory. After creating and adjusting both our simple and multiple regression models, we have enough evidence to say that CO2 levels have a accelerating positive trend with each year, ultimately leading to higher levels with time. We do note that CO2 levels do vary based on seasonality; nonetheless, the overall broad trend is consistent in both month and annual dataset. Even with the various transformations of both our models, there is always a increase in CO2 level with month and year. It appears that with several years, we can observe that fossil-fuel emissions are building up in the atmosphere, which can lead to large implications about global warming and environmental initiatives on a global scale.

8 Appendix

8.1 Regression Results: Summary

Simple Regression Model: Original

```
Residuals:
   Min     1Q Median     3Q    Max
-3.697 -2.492 -1.361  1.881  6.941

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.672e+03  4.883e+01  -54.72   <2e-16 ***
year         1.522e+00  2.457e-02   61.93   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.052 on 55 degrees of freedom
Multiple R-squared:  0.9859,    Adjusted R-squared:  0.9856
F-statistic:  3835 on 1 and 55 DF,  p-value: < 2.2e-16
```

Simple Regression Model: Post-Transformation

```
Residuals:
       Min         1Q     Median         3Q        Max
-1.706e-09 -3.322e-10  5.988e-11  4.390e-10  1.372e-09

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.455e-06  1.146e-08   214.2   <2e-16 ***
year        -1.184e-09  5.768e-12  -205.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.164e-10 on 55 degrees of freedom
Multiple R-squared:  0.9987,    Adjusted R-squared:  0.9987
F-statistic: 4.216e+04 on 1 and 55 DF,  p-value: < 2.2e-16
```

Multiple Regression Model: Post-Transformation

```
Residuals:
       Min         1Q     Median         3Q        Max
-1.706e-09 -3.322e-10  5.988e-11  4.390e-10  1.372e-09

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.455e-06  1.146e-08   214.2   <2e-16 ***
year        -1.184e-09  5.768e-12  -205.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.164e-10 on 55 degrees of freedom
Multiple R-squared:  0.9987,    Adjusted R-squared:  0.9987
F-statistic: 4.216e+04 on 1 and 55 DF,  p-value: < 2.2e-16
```

We note that even after the transformation of our multiple regression model, the normality assumption is still violated according to the Shapiro-Wilks Test.

8.2 R Code

```
# Group Project I: Trends in Atmospheric Carbon Dioxide (#5) # STP 429:  Dassanayake - T/TH 1:30-2:45 # Team:  Danny Ober-Reynolds, Jimin Nam, Katherine Wu, TJ Radigan
      # Part I: The Data
      # Set Directory (different for everyone) # setwd("C:/Users/dober_000/Documents/School 2016 Spring/STP 429/Project1")
      # Import Data # Source:  http://www.esrl.noaa.gov/gmd/ccgg/trends/ # Sets:  Mauna Loa CO2 monthly mean data and Mauna Loa CO2 annual mean data monthlydta = read.table("co2_mm_mlo.txt")
annualdta = read.table("co2_annmean_mlo.txt")
      # Rename Columns names(monthlydta) <- c("year", "month", "decimaldate", "average", "interpolated", "trend", "ndays") names(annualdta) <- c("year", "mean", "unc")
      # Replace missing observations with "NA" # -99.99 and -1 denote missing monthlydta[monthlydta == -99.99] <- NA monthlydta[monthlydta == -1] <- NA
      # Part II: The Model
      # 1) simple linear regression model on the annual mean dataset with annual mean CO2 simplereg = lm(mean ~ year, data=annualdta) # residuals for model 1 plot(fitted(simplereg),
residuals(simplereg), main = "SLR: mean ~ year, Fitted vs.  Residuals", xlab = "Fitted", ylab = "Residuals")
      # Diagnostic graphs (saves to working directory) png('SimpleRegDiagnosticGraphs.png', width = 640, height = 640, units = "px") par(mfrow=c(2,2)) plot(simplereg) dev.off()
      # Normal QQ plot (saves to working direectory) png('Graph1a_SimpleRegQQPlot.png', width = 640, height = 640, units = "px") qqnorm(residuals(simplereg), main = "Graph 1(a):  Normal
QQ Plot") qqline(residuals(simplereg)) dev.off() # Test normality assumption shapiro.test(residuals(simplereg))
      # We see a pattern in residual vs fitted plot.  We'll try transforming the independent variable
      # standardize year annualdta$stdYear = (annualdta$year - mean(annualdta$year))/sd(annualdta$year) # square annualdta$yearSqrd = annualdta$year * annualdta$year # other year
variables (staring from 1958 = 0) annualdta$yearFrom1958 = annualdta$year - 1958 annualdta$yearFrom1958Sqrd = annualdta$yearFrom1958 * annualdta$yearFrom1958 annualdta$yearFrom1958Sqrt
= sqrt(annualdta$yearFrom1958) annualdta$yearFrom1958log = log(annualdta$yearFrom1958) annualdta$stYearFrom1958 = (annualdta$yearFrom1958 - mean(annualdta$yearFrom1958))/sd(annualdta$yearFrom1958)
annualdta$invYearFrom1958 = 1/annualdta$yearFrom1958 #Tranformed mean (dependent variable) annualdta$sqrtMean = sqrt(annualdta$mean) annualdta$sqrdMean = annualdta$mean * annualdta$mean
annualdta$logMean = log(annualdta$mean) annualdta$meanPowNeg2.75 = annualdta$mean^(-2.75)
      # Alternative SLR models # Standardized year simpleregDeMeanedYr <- lm(mean ~ stdYear, data = annualdta) par(mfrow=c(2,2)) plot(simpleregDeMeanedYr) shapiro.test(residuals(simpleregDeMeanedYr))
      # Year squared simpleregSqrdYr <- lm(mean ~ yearSqrd, data = annualdta) par(mfrow=c(2,2)) plot(simpleregSqrdYr) shapiro.test(residuals(simpleregSqrdYr))
      # Year from 1958 squared (THIS MODEL HAS THE BEST NORMALITY ASSUMPTION) simpleregYrFm1958Sqrd <- lm(mean ~ yearFrom1958Sqrd, data = annualdta) par(mfrow=c(2,2)) plot(simpleregYrFm1958Sqrd)
shapiro.test(residuals(simpleregYrFm1958Sqrd))
      # Year from 1959 sqrt simpleregYrFm1958Sqrt <- lm(mean ~ yearFrom1958Sqrt, data = annualdta) par(mfrow=c(2,2)) plot(simpleregYrFm1958Sqrt) shapiro.test(residuals(simpleregYrFm1958Sqrt))
      # Log year from 1958 simpleregLogYrFrom1958 <- lm(mean ~ yearFrom19581og, data = annualdta) par(mfrow=c(2,2)) plot(simpleregLogYrFrom1958) shapiro.test(residuals(simpleregLogYrFrom1958))
      # Standardized year from 1958 simpleregStYrFrom1958 <- lm(mean ~ stYearFrom1958, data = annualdta) par(mfrow=c(2,2)) plot(simpleregStYrFrom1958) shapiro.test(residuals(simpleregStYrFrom1958))
      # Inverse year from 1958 simpleregInvYearFrom1958 <- lm(mean ~ invYearFrom1958, data = annualdta) par(mfrow=c(2,2)) plot(simpleregInvYearFrom1958) shapiro.test(residuals(simpleregInvYearFrom1958))
      # Sqrt mean (dependent variable) simpleregSqrtMean <- lm(sqrtMean ~ year, data = annualdta) par(mfrow=c(2,2)) plot(simpleregSqrtMean) shapiro.test(residuals(simpleregSqrtMean))
      # Sqrd mean (dependent variable) simpleregSqrdMean <- lm(sqrdMean ~ year, data = annualdta) par(mfrow=c(2,2)) plot(simpleregSqrdMean) shapiro.test(residuals(simpleregSqrdMean))
      # Log mean (dependent variable) simpleregLogMean <- lm(logMean ~ year, data = annualdta) par(mfrow=c(2,2)) plot(simpleregLogMean) shapiro.test(residuals(simpleregLogMean))
      # Double log model simpleregDoubleLog <- lm(logMean ~ yearFrom1958log, data = annualdta) par(mfrow=c(2,2)) plot(simpleregDoubleLog) shapiro.test(residuals(simpleregDoubleLog))
      # sqrt mean, log year from 1958 simpleregSqrtMeanLogYr <- lm(sqrtMean ~ yearFrom1958log, data = annualdta) par(mfrow=c(2,2)) plot(simpleregSqrtMeanLogYr) shapiro.test(residuals(simpleregSqrtMeanLogYr
      # mean^-2.75, year simpleregMeanPowNeg275 <- lm(meanPowNeg2.75 ~ year, data = annualdta) par(mfrow=c(2,2)) plot(simpleregMeanPowNeg275) shapiro.test(residuals(simpleregMeanPowNeg275))
# results:  p-value 0.2475
      # Save graphs from mean^-2.75 model png('SimpleRegPowModelDiagnosticGraphs.png', width = 640, height = 640, units = "px") par(mfrow=c(2,2)) plot(simpleregMeanPowNeg275) dev.off()
      # Normal QQ plot (saves to working direectory) png('Graph1b_SimpleRegPowModelQQPlot.png', width = 640, height = 640, units = "px") qqnorm(residuals(simpleregMeanPowNeg275),
main = "Graph 1(b):  Normal QQ Plot") qqline(residuals(simpleregMeanPowNeg275)) dev.off()
      # The last model, simpleregMeanPowNeg275, appears to have the least issues with the normality assumption.  However, there still appears to be a sort of pattern in the residuals
vs.  fitted, indicating heteroskedasticity.  Perhaps including the months as a factor variable will help.
      # 2) multiple regression model on the monthly mean dataset with monthly mean CO2 multiplereg = lm(average ~ year + factor(month), data=monthlydta) # check residuals par(mfrow=c(2,2))
plot(multiplereg) # get residuals - use qqnorm
      # We still see the same issue as before, using mean rather than mean^-2.75.
      monthlydta$averagePowNeg2.75 = monthlydta$average^(-2.75)
      # 2a) multiple regression model on the monthly mean dataset with monthly mean CO2 and mean^-2.75 as response variable multipleregMeanPowNeg275 <- lm(averagePowNeg2.75 ~ year
+ factor(month), data = monthlydta) par(mfrow=c(2,2)) plot(multipleregMeanPowNeg275) shapiro.test(residuals(multipleregMeanPowNeg275)) # results:  p-value 0.01407 summary(multipleregMeanPowNeg275)
      # Multiple Reg QQ plot png('Graph3a_MultipleRegPowModelQQPlot.png', width = 640, height = 640, units = "px") qqnorm(residuals(multipleregMeanPowNeg275), main = "Graph 3(a):
Normal QQ Plot") qqline(residuals(multipleregMeanPowNeg275)) dev.off()
      # Part III: Graphs # 1) For the simple linear regression model:  scatter plot of the annual mean CO2 data, with the predicted line and its 95% confidence band attach(annualdta)
predictlineSimpleReg <- predict(simplereg, interval="confidence") png('SimpleRegScatterAndPrediction.png', width = 640, height = 640, units = "px") plot (mean ~ year, data=annualdta,
main = "Graph 2(a):  Simple Pre-Transformation Model") lines(year[order(year)], predictlineSimpleReg[order(year), 1]) lines(year[order(year)], predictlineSimpleReg[order(year), 2],
col="blue", lty=2) lines(year[order(year)], predictlineSimpleReg[order(year), 3], col="blue", lty=2) dev.off() # save graph (export)
      # 1a) For the fixed simple linear regression model (simpleregMeanPowNeg275):  scatter plot of the annual mean CO2 data, with the predicted line and its 95% confidence bands
predictlineSimpleRegPowModel <- predict(simpleregMeanPowNeg275, interval = "confidence") png('SimpleRegPowModelScatterAndPrediction.png', width = 640, height = 640, units = "px")
plot (meanPowNeg2.75 ~ year, main = "Graph 2(b):  Simple Model (y^-2.75)") lines(year[order(year)], predictlineSimpleRegPowModel[order(year), 1]) lines(year[order(year)], predictlineSimpleRegPowModel[orde
2], col="blue", lty=2) lines(year[order(year)], predictlineSimpleRegPowModel[order(year), 3], col="blue", lty=2) dev.off() detach(annualdta)
      # 2) For the multiple regression model:  connected line plot of the monthly mean CO2 data to show the trend fixYearMonthWrong <- function(yearMonthWrong) { if (nchar(yearMonthWrong)
< 6){ year <- substr(yearMonthWrong,0,4) month <- substr(yearMonthWrong,5,5) result <- as.integer(paste(year,"0",month, sep = "")) } else { result <- as.integer(yearMonthWrong) }
return (result) }
      monthlydta$yearMonthWrong = paste(monthlydta$year, monthlydta$month, sep = "") monthlydta$yearMonth = unlist(lapply(monthlydta$yearMonthWrong, fixYearMonthWrong)) monthlydta
<- monthlydta[order(monthlydta$yearMonth),]
      # Plot Monthly Mean CO2 png('MontlyMeanCO2.png', width = 640, height = 640, units = "px") plot(0,0,type = "n", xlim = c(0,696), ylim = c(min(monthlydta$average, na.rm = TRUE),
max(monthlydta$average, na.rm = TRUE)), axes = FALSE, ann = FALSE) lines(monthlydta$average, type = "l", col = "blue") title(main = "Monthly Mean CO2", xlab = "Month", ylab = "Mean
CO2") axis(1, at = c(seq(from = 1, to = 700, by = 70)), labels = c("1958 Mar","1964 Jan","1969 Nov","1975 Sep","1981 Jul","1987 May","1993 Mar","1999 Jan","2004 Nov","2010 Sep"))
axis(2, at = c(seq(310,410,by=10))) dev.off()
      ## Bin and histogram CO2 values annually binsAnnually <- 10 cutpointsAnnually <- quantile(annualdta$mean, (0:binsAnnually)/binsAnnually, na.rm = TRUE) binnedAnnually <- cut(annualdta$mean,
cutpointsAnnually, include.lowest = TRUE, na.rm = TRUE) summary(binnedAnnually) png('AnnualCO2Distribution.png', width = 640, height = 640, units = "px") plot(binnedAnnually, main
= "Annual CO2 Distribution") dev.off()
      ## Bin and histogram CO2 values monthly binsMonthly <- 10 cutpointsMonthly <- quantile(monthlydta$average, (0:binsMonthly)/binsMonthly, na.rm = TRUE) binnedMonthly <- cut(monthlydta$average,
cutpointsMonthly, include.lowest =TRUE, na.rm = TRUE) summary(binnedMonthly) png('MonthlyCO2Distribution.png', width = 640, height = 640, units = "px") plot(binnedMonthly, main =
"Monthly CO2 Distribution") dev.off()

      # Calculate predicution values from the monthly model.  Coefficients:  # (Intercept) year factor(month)2 factor(month)3 factor(month)4 factor(month)5 factor(month)6 #2.451319e-06

-1.181988e-09 -5.999201e-10 -1.342449e-09 -2.340740e-09 -2.795695e-09 -2.318805e-09 #factor(month)7 factor(month)8 factor(month)9 factor(month)10 factor(month)11 factor(month)12

#-1.193374e-09 4.507518e-10 1.763915e-09 1.805585e-09 7.221231e-10 -3.499770e-10 jan = (2.451319e-06 + -1.181988e-09*2016)^(-1/2.75) feb = (2.451319e-06 + -1.181988e-09*2016 - 5.999201e-10)^(-1/2.75)

mar = (2.451319e-06 + -1.181988e-09*2016 - 1.342449e-09)^(-1/2.75) apr = (2.451319e-06 + -1.181988e-09*2016 - 2.340740e-09)^(-1/2.75) may = (2.451319e-06 + -1.181988e-09*2016 - 2.795695e-09)^(-1/2.75)

jun = (2.451319e-06 + -1.181988e-09*2016 - 2.318805e-09)^(-1/2.75) jul = (2.451319e-06 + -1.181988e-09*2016 - 1.193374e-09)^(-1/2.75) aug = (2.451319e-06 + -1.181988e-09*2016 + 4.507518e-10)^(-1/2.75)

sep = (2.451319e-06 + -1.181988e-09*2016 + 1.763915e-09)^(-1/2.75) oct = (2.451319e-06 + -1.181988e-09*2016 + 1.805585e-09)^(-1/2.75) nov = (2.451319e-06 + -1.181988e-09*2016 + 7.221231e-10)^(-1/2.75)

dec = (2.451319e-06 + -1.181988e-09*2016 - 3.499770e-10)^(-1/2.75)
```

References

[1] Dr. Pieter Tans, NOAA/ESRL (www.esrl.noaa.gov/gmd/ccgg/trends/) and Dr. Ralph Keeling, Scripps Institution of Oceanography (scrippsco2.ucsd.edu/).