STA 141A Final Project
Group 12 Project Proposal

Exploration of Relevant Parameters of Climate Change Using Multiple Linear Regression and Cross-Validation: A 25 year Study (1984-2008)

Code Lead and Primary Editor: Max Vo | maxvo@ucdavis.edu
Secondary Editor Josh Velazquez | jdvelazquez@ucdavis.edu

OVERVIEW

One of humanity's most pressing issues today is climate change, characterized by long-term shifts in global temperature and weather patterns. Its consequences include higher global temperatures, more frequent severe storms, and widespread poverty due to displacement, among others. Understanding the root causes of these shifts is crucial for mitigating climate change and reducing its impacts. Crucially this report asks: What are the common elements that affect temperature and weather patterns, and how do they influence global temperatures?

Informed by exploratory analysis, we created a model that utilized the variables ‘Year’, ‘MEI’, ‘CO2’, ‘CH4’, ‘N2O’, ‘CFC.11’, ‘CFC.12’, ‘Aerosols’ by determining a model that best fit Climate Data from 1984-2008. After testing various models, we utilized multiple linear regression which is a supervised statistical method. After conductive statistical analysis and testing, we identified that CH4 and N2O may not have a significant impact on the variable.

PREVIOUS LITERATURE AND CONTEXTUALIZATION

To better understand our objectives and data we have conducted a literature review on climate change and the factors which accelerate or accelerate its process. This review informs our progression with our data analysis.

Aerosols have a profound effect on Earth’s climate; however, since there is large variation in aerosols types, the overall effect of aerosols on climate is complex. Whether aerosols cool or warm the atmosphere is highly dependent on aerosol color: Darker aerosols, such as black carbon, absorb light and heat the atmosphere, while lighter aerosols, such as pure sulfates and nitrates, scatter light and cool the atmosphere. Between the two, the cooling impact of lighter aerosols is far more significant. For instance, the event of a volcanic eruption sends high volumes of sulfate dioxide, a reflective aerosol, into the atmosphere, causing a significant regional cooling effect. These once-in-a-while cooling effects highly outweigh the consistent heating effects of darker aerosols (Earth Observatory, 2010). It is likely that we observe a negative correlation between aerosols and global temperature.

Chlorofluorocarbons (CFCs) are a man-made greenhouse gas. CFCs absorb specific wavelengths of radiation that other greenhouse gasses cannot, trapping heat in the atmosphere that would otherwise not be present. CFCs are incredibly effective at trapping heat and have long lifespans. CFCs were banned in response to their threat to global temperatures in an international agreement, which is around the time our data collection starts. Therefore, we may see a diminishing effect of CFCs and global temperature with time. (Stone, 2023).

Total Solar Radiance is the total solar radiation that enters Earth’s atmosphere. This statistic indicates the total amount of solar energy within the climate system. TSI changes slowly with time, but has a large influence on global temperature. However, since TSI changes slowly with time, it may not be significant to include in our timespan of 24 years. (Ball et. al, 2022)

MEI describes the severity of El Niño/Southern Oscillation (ENSO), and contributes to variability in global climate. To discover patterns between global climate and ENSO, the Multivariate ENSO Index (MEI) serves as a basis. ENSO occurs due to a factor of multiple variables, six in total, meaning MEI can be difficult to accurately compute. (Wolter, 2011) However, we include the MEI variable in our study since although it oscillates there is a clear upward trend as time passes.

EXPLORATORY ANALYSIS

Data Collection

To find a relevant dataset for our report, we utilized Kaggle. Through Kaggle, we were able to view a quick description of the contents of the dataset. We settled on the dataset entitled “climate change” because it had the relevant variables needed to conduct an analysis of greenhouse gasses and global temperature. Furthermore, the dataset includes extra variables for us to create more complex models, such as ones that consider aerosol concentrations or El Niño/Southern Oscillation variability.

Dataset Link: ttps://www.kaggle.com/datasets/econdata/climate-change

Data Exploration

First, we explore the data to better understand how to proceed with our analysis. We first view the dataset, then use str() to understand the structure of our dataset and view each of our variables: Year, Month, MEI, CO2, CH4, N2O, CFC-11, CFC-12, TSI, Aerosols, and Temp.

SUMMARY OF INDEPENDENT VARIABLES

Variable	Description/Definition	Unit Measurement
CO2, N2O, CH4	The greenhouse gas with the highest concentration is Carbon dioxide (CO2), methane (CH4) is 2nd, nitrous oxide (N2O) is 3rd	ppmv (parts per million by atmospheric volume)
CFC.11, CFC.12	Chlorofluorocarbon-11 (CFC.11) and Chlorofluorocarbon-12 (CFC-12). A man-made greenhouse gas found in small concentrations with a high warming potential.	ppbv (parts per billion by atmospheric volume
Aerosols	Mean aerosol optical depth (effect aerosols have on scattering or absorbing solar radiation) per 550 nanometer.	N/A (dimensionless)
TSI	Total solar radiation entering Earth’s atmosphere.	W/m^2 (watts per meter squared)
MEI	The strength of current El Nino/Southern Oscillation.	N/A (dimensionless)
Temperature	Difference in global temperature and a single recorded temperature.	℃

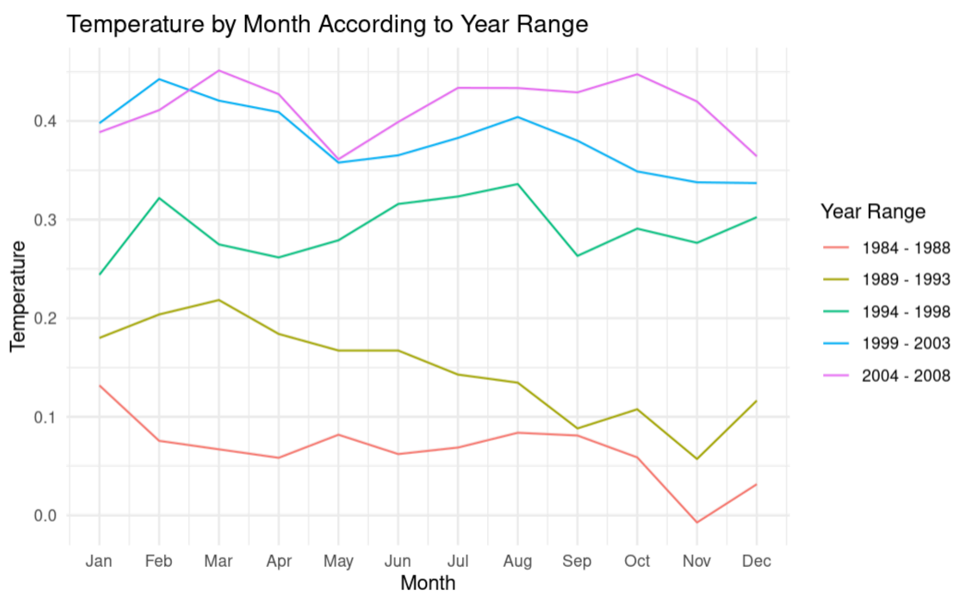
APPENDIX: Data Loading ~ First and Last 10 Entries of Dataset

Year	Month	MEI	CO2	CH4	N2O	`CFC-11`	`CFC-12`	TSI	Aerosols	Temp
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
2008	3	-1.64	386.	1793.	321.	245.	536.	1366.	0.0034	0.447
2008	4	-0.942	387.	1793.	321.	245.	536.	1366.	0.0033	0.278
2008	5	-0.355	388.	1796.	321.	245.	535.	1366.	0.0031	0.283
2008	6	0.128	388.	1792.	321.	245.	535.	1366.	0.0031	0.315
2008	7	0.003	386.	1783.	321.	244.	535.	1366.	0.0033	0.406
2008	8	-0.266	384.	1780.	321.	244.	535.	1366.	0.0036	0.407
2008	9	-0.643	383.	1795.	322.	244.	535.	1366.	0.0043	0.378
2008	10	-0.78	383.	1814.	322.	244.	535.	1366.	0.0046	0.44
2008	11	-0.621	384.	1812.	322.	244.	535.	1366.	0.0048	0.394
2008	12	-0.666	386.	1813.	322.	244.	535.	1366.	0.0046	0.33

	Year	Month	MEI	CO2	CH4	N2O	`CFC-11`	`CFC-12`	TSI	Aerosols	Temp
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1983	5	2.56	346.	1639.	304.	191.	350.	1366.	0.0863	0.109
2	1983	6	2.17	346.	1634.	304.	192.	352.	1366.	0.0794	0.118
3	1983	7	1.74	344.	1633.	304.	193.	354.	1366.	0.0731	0.137
4	1983	8	1.13	342.	1631.	304.	194.	356.	1366.	0.0673	0.176
5	1983	9	0.428	340.	1648.	304.	194.	357.	1366.	0.0619	0.149
6	1983	10	0.002	340.	1664.	304.	195.	359.	1366.	0.0569	0.093
7	1983	11	-0.176	342.	1658.	304.	196.	361.	1366.	0.0524	0.232
8	1983	12	-0.176	343.	1654.	304.	197.	362.	1366.	0.0486	0.078
9	1984	1	-0.339	344.	1659.	304.	197.	363.	1365.	0.0451	0.089
10	1984	2	-0.565	345.	1656.	304.	198.	364.	1366.	0.0416	0.013

By looking at the first and last 10 entries of the head and tail of the dataset we can identify certain immediate trends.The most identifiable difference is temperature. We can see that in May 1983, the recorded temperature differed from the global temperature by 0.109°C, while in December 2008, the recorded temperature differed by 0.33°C. Already, there is clear indication of warming. It is also clear that all aerosol concentrations, CO2, CH4, and N20 have increased from the first index from the last. (this hints at a correlation between those variables). The two variables that do not clearly increase with time are MEI and Aerosols, which indicates that we may have to take into account fluctuations with these variables in our model. We also notice how the data begins on month 5 of the year 1983. It may be beneficial to omit the first 8 rows of our data to make our data consistent(25 entries for each month). This preliminary analysis informs how we proceed with our data cleansing.

FIGURE 2: Exploration of How the Categorical Variable of Month Affects Temperature by Year



We then explore temperature by month according to a four year range. We clearly see an increasing trend in global temperature per 4 year period; the average difference in global temperature for each month is greater than that for the same month of the previous period, with an exception between the first 3 months of the 2004-2008 and 1999-2003 periods. Furthermore, every year follows similar monthly trends. For instance, temperature systemically dips throughout the year from January to December.

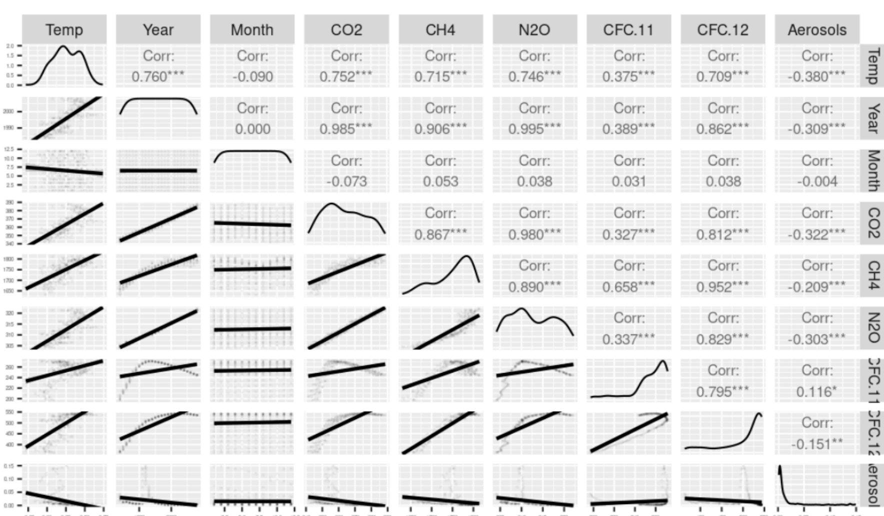
Descriptive Statistics (Summary and Correlations)

Next, we can take a look at a summary for each variable in the dataset:

```
      Year          Month                    CFC-11          CFC-12
 Min.   :1983   Min.   : 1.000          Min.   :191.3   Min.   :350.1
 1st Qu.:1989   1st Qu.: 4.000          1st Qu.:246.3   1st Qu.:472.4
 Median :1996   Median : 7.000          Median :258.3   Median :528.4
 Mean   :1996   Mean   : 6.552          Mean   :252.0   Mean   :497.5
 3rd Qu.:2002   3rd Qu.:10.000          3rd Qu.:267.0   3rd Qu.:540.5
 Max.   :2008   Max.   :12.000          Max.   :271.5   Max.   :543.8
      MEI              CO2                   TSI           Aerosols
 Min.   :-1.6350   Min.   :340.2        Min.   :1365   Min.   :0.00160
 1st Qu.:-0.3987   1st Qu.:353.0        1st Qu.:1366   1st Qu.:0.00280
 Median : 0.2375   Median :361.7        Median :1366   Median :0.00575
 Mean   : 0.2756   Mean   :363.2        Mean   :1366   Mean   :0.01666
 3rd Qu.: 0.8305   3rd Qu.:373.5        3rd Qu.:1366   3rd Qu.:0.01260
 Max.   : 3.0010   Max.   :388.5        Max.   :1367   Max.   :0.14940
      CH4           N2O                      Temp
 Min.   :1630   Min.   :303.7           Min.   :-0.2820
 1st Qu.:1722   1st Qu.:308.1           1st Qu.: 0.1217
 Median :1764   Median :311.5           Median : 0.2480
 Mean   :1750   Mean   :312.4           Mean   : 0.2568
 3rd Qu.:1787   3rd Qu.:317.0           3rd Qu.: 0.4073
 Max.   :1814   Max.   :322.2           Max.   : 0.7390
```

Above is the summary statistics for the dataset.. According to our temperature summary, it is clear that temperature increases on average, with its mean of 0.25°C hotter than the global average. We note the minimum, which is a negative number, meaning that there exists a year in the dataset the recorded temperature was less than the global average. A few other notable discoveries from our summary is that amongst the greenhouse gasses, CH4 has the widest range, with a minimum of 1630 and a maximum of 1814, not yet taking into account for outliers. N2O has the least spread, with a minimum of 303.7 and a maximum of 322.2, without taking into account outliers. Also, TSI barely varies, so we know the amount of solar radiation entering the atmosphere stays fairly constant; however, it may have a large impact on global temperature.

FIGURE 2: Scatter Plot Matrix (ANALYSIS OF CORRELATION)



We then conduct an analysis of correlation between the data set's variables. We are particularly interested in the first row and column, which shows us which variables positively correlate with temperature. Through the scatterplot matrix, we notice strong linear correlations between temperature and CO2, CH4, N2O, CFC.12, with correlation coefficients of 0.76, 0.752, 0.746 and 0.709 respectively. We also note strong correlations between some of our greenhouse gas variables, for instance, CO2 and N2O with a correlation coefficient of 0.980. This affirms our choice of using multiple regression analysis.

Data Cleansing

After exploring the data(ie. Looking at the data's structure, conducting exploratory plotting, and looking at specific data points), we have a better understanding of what adjustments should be made to provide a more rigorous analysis. The following adjustments help increase the quality and consistency of our data:

- We used na.omit() as we load in our data to remove missing values
- We omit the first 8 rows with our data to make each year consistent 12 month periods. Now, our data starts on month 1 of the year 1984, and ends on month 12 of the year 2008.
- We omit TSI from our analysis. As stated in the literature review, TSI changes too slowly to have a notable effect within 25 years
- We retain MEI as it trends upwards (even if it is noisy)
- We ensure that there are no duplicates, by comparing the number of rows in the dataset with the number of rows in the dataset with the unique() applied. There are no duplicate data points.

METHODOLOGY

Upon testing other modeling forms, we found them to be inclusive. Forms such as logistic regression, linear discriminant analysis (LDA), and cluster analysis did not yield relevant results. We found our best modeling form to be multiple linear regression. We assure there is no overfitting by comparing adjusted R2 and R2, then by selecting the best MSE. Check APPENDIX section, Selection Criteria as well as the Relevant Model Testing Sections, for more information.

Therefore we are using the Multiple Regression Model:

The multiple regression model is a supervised statistical learning method that will analyze various dependent variables and a single independent variable and make an estimation of the relationship between the two. In the context of our analysis, we will use multiple regression to find a relationship between global temperature increase and various climate predictors (CO2, N2O, etc.). The function in R utilizes OLS (ordinary least squares) to find the optimal coefficients for each parameter.

The general form of the multiple regression model will be:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta nXn + \epsilon$$

Where:

- T is the temperature change.
- X1,X2,...,Xn are the independent climate parameters.
- β 0 ,β 1 ,β 2 ,...,β n are the regression coefficients.
- ϵ is the error term. which is ~iid~ N(0, sigma^2)

Model Selection Using Leaps Package: https://cran.r-project.org/web/packages/leaps/leaps.pdf
Regsubset in the leaps package performs an exhaustive search over all possible subsets of the predictors in a linear regression model and selects the best subset according to some criterion, typically AIC, BIC, or adjusted R2. It identifies the best model at a given number of predictors(1-8 predictors for our purposes), where best is quantified using RSS. The syntax for this function is the same as lm().

After a List of Best predictors for each number of predictors is determined, we compare the MSE, R2 and adjR2 of each model to identify the model with the smallest MSE value that doesn't overfit the data. (adjR2 is much smaller than R2). More Information and clarity is provided in the APPENDIX sections 'Deciding on the Best Model' and 'Selection Criteria.'

Cross-Validation Approach

Cross-validation is a statistical method used to evaluate the performance of a model by partitioning the data into subsets, training the model on some subsets, and validating it on the remaining subsets. This process is repeated multiple(k) times to ensure that the model's performance is robust and not overly dependent on any particular partition of the data. With various climate predictors at our disposal, we must ensure that our model is robust.

1. **k-Fold Cross-Validation:** The dataset will be divided into k subsets (folds). The model will be trained on k-1 folds and tested on the remaining fold. This process will be repeated k times with each fold being used as the test set once.
2. **Leave-One-Out Cross-Validation (LOOCV):** A special case of k-fold cross-validation where k equals the number of observations (most information but computationally intensive)

Notable insights that can be drawn from using cross-validation: Cross-validation ensures we choose a model with high predictive power. This is done by curtailing the risk of choosing a model that overfits or under-fits the data by finding the most optimal balance between bias and variance.

RESULTS AND FURTHER ANALYSIS (multiple regression)

```
Call:
lm(formula = best_model_formula, data = climate)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.269500 -0.064557  0.001331  0.057791  0.307282 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 69.0353507 25.7736212   2.679  0.00782 ** 
Year        -0.0398327  0.0148996  -2.673  0.00793 ** 
MEI          0.0618198  0.0065293   9.468  < 2e-16 ***
CO2          0.0097461  0.0030543   3.191  0.00157 ** 
CH4         -0.0003409  0.0005341  -0.638  0.52381    
N2O          0.0217293  0.0136837   1.588  0.11338    
CFC.11      -0.0066150  0.0020318  -3.256  0.00126 ** 
CFC.12       0.0053424  0.0013216   4.042 6.78e-05 ***
Aerosols    -1.7626952  0.2339216  -7.535 6.20e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0963 on 291 degrees of freedom
Multiple R-squared:  0.7221,	Adjusted R-squared:  0.7144 
F-statistic: 94.51 on 8 and 291 DF,  p-value: < 2.2e-16
```

REGRESSION TABLE
from Appendix Section: Analyzing Our Model

For our model n = 300 and p(predictors) = 8

Model Specification

Y = 69.0354 − (0.0398)Year + (0.0618)MEI + (0.0097)CO2 − (0.0003)CH4 +(0.0217)N2O − 3(0.0066)CFC.11+ (0.0053)CFC.12 − (1.7627)Aerosols + ϵ

- Y is the dependent variable (Temperature)
- Year, MEI, CO2 CH4, N2O, CFC.11, CFC.12 and Aerosols are the included predictor variables.
- β0 = 69.0354 is the intercept; when all variables are at 0 units, temperature is 69 degrees; which lacks applicable meaning.
- $\beta 1$ = −0.0398, $\beta 2$ = 0.0618, $\beta 3$ = 0.0097, $\beta 4$ = − 0.0003, β 5 = 0.0217, β 6 = − 0.0066, β 7 = 0.0053, and $\beta 8$ = − 1.7627 are the coefficients for Year, MEI, CO2, CH4, N2O, CFC.11, CFC.12, and Aerosols respectively
    - Each slope coefficient represents the expected increase(positive β) or decrease(negative β) in the dependent variable for a one-unit change in the corresponding independent variable, holding all other variables constant.
- **Standard Errors**: Each coefficient estimate comes with a standard error, which measures the variability of the estimate. Smaller standard errors indicate more precise estimates of the coefficients.

Statistical Testing

<u>T-test for individual regression coefficients.</u>
Null Hypothesis (H_0): H0: βj = 0 for j = 1,..,8
- Implies that there is no linear relationship between the independent variable and the dependent variable.

Alternative Hypothesis (Ha): H1:βj ≠ 0 for j = 1,..,8
- Implies there is a significant linear relationship.

Compares the calculated $t_{statistic} = \frac{\beta j}{SE(\beta j)}$ to critical t-values with df(n-p-1). **From the regression output and conducting T-tests all variables except for CH4 and N2O are significant under the α = 0.05 significance level**

<u>EXAMPLE</u> (two-sided test):
$| t_{N20} |= 1.588 < t_{critical} = t(0.05/2, df = 300 - 8 - 1) = 1.968$

<u>F-statistics</u> APPENDIX Section: Analyzing Our Model and F-test Lack of Fit

1) **For overall model** (from regression output) tests the overall significance of the model:
Compares $F_{statistic} = \frac{Mean\ Sq\ (Model)}{Mean\ Sq\ (Residuals)}$ with $F_{critical}$ = F(α, df_{model} = total parameters - 1 = p, $df_{residual}$ = n−p−1)
Since $F_{statistic} = 94.51168 > F_{critical} = 1.970285$ the model concludes that the overall model is statistically significant at the α=0.05 level. This means that the predictors (independent variables) in your model collectively explain a significant amount of the variability in the dependent variable (response variable)

ANOVA TABLE from Appendix Section: Analyzing Our Model

```
Analysis of Variance Table

Response: Temp
           Df Sum Sq Mean Sq  F value    Pr(>F)    
Year        1 5.6137  5.6137 605.3399 < 2.2e-16 ***
MEI         1 0.5917  0.5917  63.8094 3.220e-14 ***
CO2         1 0.0151  0.0151   1.6299   0.20274    
CH4         1 0.0260  0.0260   2.8026   0.09519 .  
N2O         1 0.0184  0.0184   1.9817   0.16028    
CFC.11      1 0.0030  0.0030   0.3268   0.56797    
CFC.12      1 0.2172  0.2172  23.4208 2.115e-06 ***
Aerosols    1 0.5266  0.5266  56.7824 6.195e-13 ***
Residuals 291 2.6986  0.0093                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2) **F-stat from ANOVA** compares the variability explained by the predictors with the variability not explained by the model (error/residuals):

Compares $F_{\text{statistic}} = \frac{\textit{Sum of Mean Sq (Predictors)}}{\textit{Mean Sq (Residuals)}}$ with $F_{critical} = F(\alpha, df_{\text{numerator}} = p, df_{\text{residual}} = n-p-1)$

Since $F_{\text{statistic}} = 756.0935 > F_{\text{critical}} = 1.970285$ the variability explained by the predictors (independent variables) in your model is statistically significantly different from the variability not explained by the model (residuals)

3) **Lack-of-Fit F-test**: This test specifically examines whether the chosen model adequately fits the data. It compares the lack-of-fit mean square with the residual mean square. If the model does not fit well, the lack-of-fit F-test would indicate a significant result:

Compares $F_{\text{Lack of Fit}} = \frac{\textit{Mean Sq Lack of Fit}}{\textit{Mean Sq (Residuals)}}$ with $F_{critical} = F(\alpha, df_{\text{Lack of Fit}} = n\text{-}p, df_{\text{residual}} = n-p-1)$

Since $F_{\text{Lack of Fit}} = 1 < F_{critical} = 1.213079$ We fail to reject the Ho of no lack of fit and so we conclude that model adequately fits the data

Cross Validation

Conducting cross-validation is important for determining providing an unbiased estimate of model performance. It is a robust technique for assessing how a model will generalize to an independent dataset. The Mean Squared Error (MSE) calculated from cross-validation is the average squared difference between the observed actual outcomes and the predictions of our model. The MSE values provide a quantitative measure of how closely the model's predictions match the actual outcomes. Therefore, lower MSE values indicate a better fit of the model to the data.

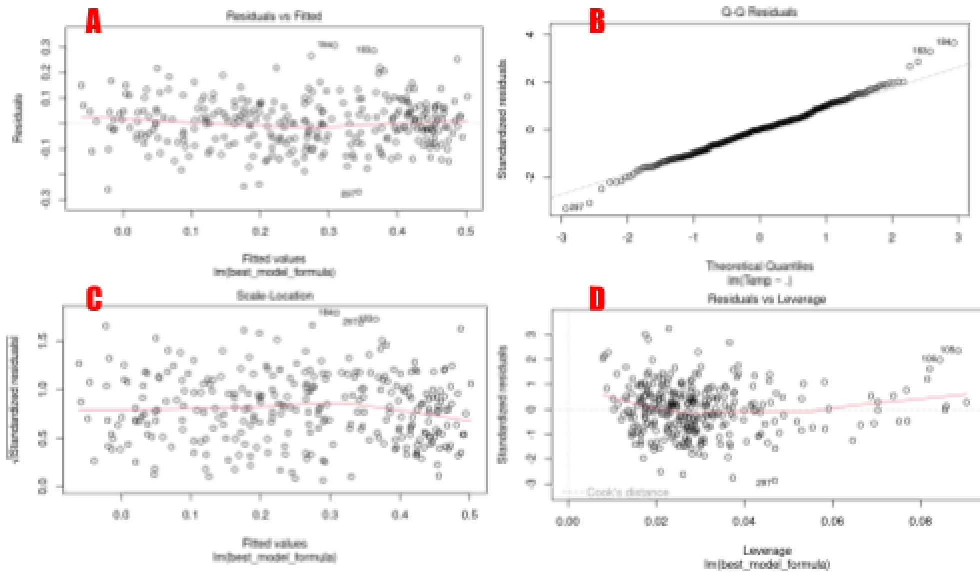
Using the Cross-Validation Techniques Defined in The Methodology our MSE values are:

- **LOOCV MSE:** 0.009562674
- **K-fold CV MSE:** 0.009526567

Both MSE values are sufficiently small, indicating that the model predictions are very close to the actual values. This implies that the model performs well and has a good fit to the data. The small difference between the LOOCV and K-fold CV MSE suggests that the model generalizes well across different subsets of the data, demonstrating robustness and reliability in its predictions

Residual Analysis

FIGURES 3 (A-D)



A) **Residuals vs Fitted**:

- **Purpose**: To check the linearity assumption and detect any non-linear patterns.
- **Interpretation**:
  - The residuals should be randomly scattered around the horizontal line (y = 0) without forming any specific patterns.
  - There seems to be no clear patterns which suggest a linear relationship between our predictors and independent variables.
  - **Does not violate assumption of homoscedasticity**

B) Normal Q-Q (Quantile-Quantile) Plot:

- **Purpose**: To assess whether the residuals are normally distributed.
- **Interpretation**:
  - The points should fall approximately along a straight diagonal line.
  - The deviations near the end of the QQplot suggest that the residuals may not be normally distributed.
  - As the QQplot is heavily tailed (upper end of the QQplot curve upward and the bottom end of the QQplot curves downward) this suggests that there are more extreme values (outliers) than expected under normality.

C) Scale-Location Plot (also called Spread-Location Plot):

- **Purpose**: To check the homoscedasticity (constant variance) of the residuals.
- **Interpretation**:
  - The plot shows the square root of standardized residuals against the fitted values.
  - For the most part, the the variance in square root of standardized residuals are constant, but there is a slight down curve towards the end indicates that the spread (variance) of residuals may changes systematically as the fitted values increase; This could be solvable with logistic transformations of the model
  - As there is no funnel shape (heteroscedasticity) the variance of the residuals appears constant

D) Residuals vs Leverage:

- **Purpose**: To identify influential observations and data points that might disproportionately affect the model fit.
- **Interpretation**:
  - We Look for points that have high leverage (i.e., they are far from the mean of the predictor variables) and high residuals (large vertical distances from the horizontal line y = 0).
  - Points outside of the Cook's distance(D) lines (the dashed lines which are not in the plot frame in the case of our model) may warrant investigation as they may need to be accounted for in the model

CONCLUSIONS

This study provides insight into the most critical factors driving climate change over a 25-year period and the effectiveness of multiple regression models in predicting future climate trends. The use of cross-validation ensures the robustness of the findings, offering a reliable tool for policymakers and researchers. The parameters we defined as significant (CO_2, MEI, CFC.11, CFC.12, Aerosols) grant insight into what causes of climate change policymakers should focus on when looking to diminish its effects on global temperatures. It is also important to note that manmade greenhouse gasses (CFC.11, CFC.12) prevail in significance compared to other naturally occurring gasses (N_2O, CH_4). In a broader context, this analysis reveals that other man-made factors should be scrutinized when analyzing climate change, as a small concentration of CFCs were found to have the most significant effects.

Future Directions:

As hinted previously, our research provides multiple avenues for advancing our research questions. We could analyze larger datasets with more variables over a longer timespan. Longitudinal Data would provide us with opportunities to create clearer parameters. We could also utilize a more robust model, by addressing issues such as non-normality in the residual by utilizing data transformations (box-cox, logistic) or nonparametric statical learning forms.

REFERENCES

Ball, W., & Haigh, J., & et. al. (2022, September 9). The climate data guide: Total Solar Irradiance (TSI) datasets: An overview. *National Center for Atmospheric Research.* https://climatedataguide.ucar.edu/climate-data/total-solar-irradiance-tsi-datasets-overview

Earth Observatory. (2010, November 2). Aerosols and incoming sunlight (direct effects). https://earthobservatory.nasa.gov/features/Aerosols/page3.php

Stone, K. (2023). What is the concentration of CFCs in the atmosphere, and how much do they contribute to global warming? *Climate Portal.* https://climate.mit.edu/ask-mit/what-concentration-cfcs-atmosphere-and-how-much-do-they-contribute-global-warming

United Nations. Key findings. https://www.un.org/en/climatechange/science/key-findings

Wolter, K. (2011). El Niño/southern oscillation behaviour since 1871 as diagnosed in an extended multivariate ENSO index (MEI.ext) *International Journal of Climatology, 31(7)*, 1074-1087. https://doi.org/10.1002/joc.2336

APPENDIX ~ CLIMATE ANALYSIS: STA141A Final Project

April 12

Data Loading and Preliminary Analysis

```
data = na.omit(read.csv("/cloud/project/climate_change.csv"))

# For Consistency in Year Based/Month Based Analyses, we remove the first 8 months
# of the dataset (as the data randomly begins from the 5th month of 1983)

# Therefore, our dataset starts from 1984
climate <- data[-(1:8),]

#brief summary; head()/taiL()
head(climate,10)
```

```
##    Year Month    MEI    CO2     CH4     N2O  CFC.11  CFC.12      TSI Aerosols
## 9  1984     1 -0.339 344.05 1658.98 304.130 197.219 363.359 1365.426   0.0451
## 10 1984     2 -0.565 344.77 1656.48 304.194 197.759 364.296 1365.662   0.0416
## 11 1984     3  0.131 345.46 1655.77 304.285 198.249 365.044 1366.170   0.0383
## 12 1984     4  0.331 346.77 1657.68 304.389 198.723 365.692 1365.566   0.0352
## 13 1984     5  0.121 347.55 1649.33 304.489 199.233 366.317 1365.778   0.0324
## 14 1984     6 -0.142 346.98 1634.13 304.593 199.858 367.029 1366.096   0.0302
## 15 1984     7 -0.138 345.55 1629.89 304.722 200.671 367.893 1366.114   0.0282
## 16 1984     8 -0.179 343.20 1643.67 304.871 201.710 368.843 1365.978   0.0260
## 17 1984     9 -0.082 341.35 1663.60 305.021 202.972 369.800 1365.867   0.0239
## 18 1984    10  0.016 341.68 1674.65 305.158 204.407 370.782 1365.787   0.0220
##      Temp
## 9   0.089
## 10  0.013
## 11  0.049
## 12 -0.019
## 13  0.065
## 14 -0.016
## 15 -0.024
## 16  0.034
## 17  0.025
## 18 -0.035
```

```
tail(climate,10)
```

```
##     Year Month    MEI    CO2     CH4     N2O  CFC.11  CFC.12      TSI Aerosols
## 299 2008     3 -1.635 385.97 1792.84 321.295 245.430 535.979 1365.673   0.0034
## 300 2008     4 -0.942 387.16 1792.57 321.354 245.086 535.648 1365.715   0.0033
## 301 2008     5 -0.355 388.50 1796.43 321.420 244.914 535.399 1365.717   0.0031
## 302 2008     6  0.128 387.88 1791.80 321.447 244.676 535.128 1365.673   0.0031
## 303 2008     7  0.003 386.42 1782.93 321.372 244.434 535.026 1365.672   0.0033
## 304 2008     8 -0.266 384.15 1779.88 321.405 244.200 535.072 1365.657   0.0036
```

```
## 305 2008     9 -0.643 383.09 1795.08 321.529 244.083 535.048 1365.665   0.0043
## 306 2008    10 -0.780 382.99 1814.18 321.796 244.080 534.927 1365.676   0.0046
## 307 2008    11 -0.621 384.13 1812.37 322.013 244.225 534.906 1365.707   0.0048
## 308 2008    12 -0.666 385.56 1812.88 322.182 244.204 535.005 1365.693   0.0046
##      Temp
## 299 0.447
## 300 0.278
## 301 0.283
## 302 0.315
## 303 0.406
## 304 0.407
## 305 0.378
## 306 0.440
## 307 0.394
## 308 0.330
```

```r
summary(climate)
```

```
##       Year          Month             MEI               CO2
##  Min.   :1984   Min.   : 1.00   Min.   :-1.6350   Min.   :341.4
##  1st Qu.:1990   1st Qu.: 3.75   1st Qu.:-0.4125   1st Qu.:353.8
##  Median :1996   Median : 6.50   Median : 0.2250   Median :362.3
##  Mean   :1996   Mean   : 6.50   Mean   : 0.2573   Mean   :363.8
##  3rd Qu.:2002   3rd Qu.: 9.25   3rd Qu.: 0.8197   3rd Qu.:373.8
##  Max.   :2008   Max.   :12.00   Max.   : 3.0010   Max.   :388.5
##       CH4            N2O            CFC.11          CFC.12            TSI
##  Min.   :1630   Min.   :304.1   Min.   :197.2   Min.   :363.4   Min.   :1365
##  1st Qu.:1726   1st Qu.:308.6   1st Qu.:247.5   1st Qu.:478.0   1st Qu.:1366
##  Median :1766   Median :311.7   Median :259.0   Median :528.9   Median :1366
##  Mean   :1753   Mean   :312.6   Mean   :253.5   Mean   :501.3   Mean   :1366
##  3rd Qu.:1787   3rd Qu.:317.0   3rd Qu.:267.3   3rd Qu.:540.7   3rd Qu.:1366
##  Max.   :1814   Max.   :322.2   Max.   :271.5   Max.   :543.8   Max.   :1367
##     Aerosols            Temp
##  Min.   :0.00160   Min.   :-0.2820
##  1st Qu.:0.00280   1st Qu.: 0.1288
##  Median :0.00550   Median : 0.2510
##  Mean   :0.01535   Mean   : 0.2600
##  3rd Qu.:0.01200   3rd Qu.: 0.4100
##  Max.   :0.14940   Max.   : 0.7390
```

```r
# structure of dataframe
str(climate)
```

```
## 'data.frame':    300 obs. of  11 variables:
##  $ Year    : int  1984 1984 1984 1984 1984 1984 1984 1984 1984 1984 ...
##  $ Month   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ MEI     : num  -0.339 -0.565 0.131 0.331 0.121 -0.142 -0.138 -0.179 -0.082 0.016 ...
##  $ CO2     : num  344 345 345 347 348 ...
##  $ CH4     : num  1659 1656 1656 1658 1649 ...
##  $ N2O     : num  304 304 304 304 304 ...
##  $ CFC.11  : num  197 198 198 199 199 ...
##  $ CFC.12  : num  363 364 365 366 366 ...
##  $ TSI     : num  1365 1366 1366 1366 1366 ...
##  $ Aerosols: num  0.0451 0.0416 0.0383 0.0352 0.0324 0.0302 0.0282 0.026 0.0239 0.022 ...
##  $ Temp    : num  0.089 0.013 0.049 -0.019 0.065 -0.016 -0.024 0.034 0.025 -0.035 ...
```

```r
# checking for any duplicate data (there is none)
nrow(climate)==nrow(unique(climate))
```

```
## [1] TRUE
```

Exploratory Analysis

FIGURE 2: Testing Correlation Between Variables

```r
library(GGally)
```

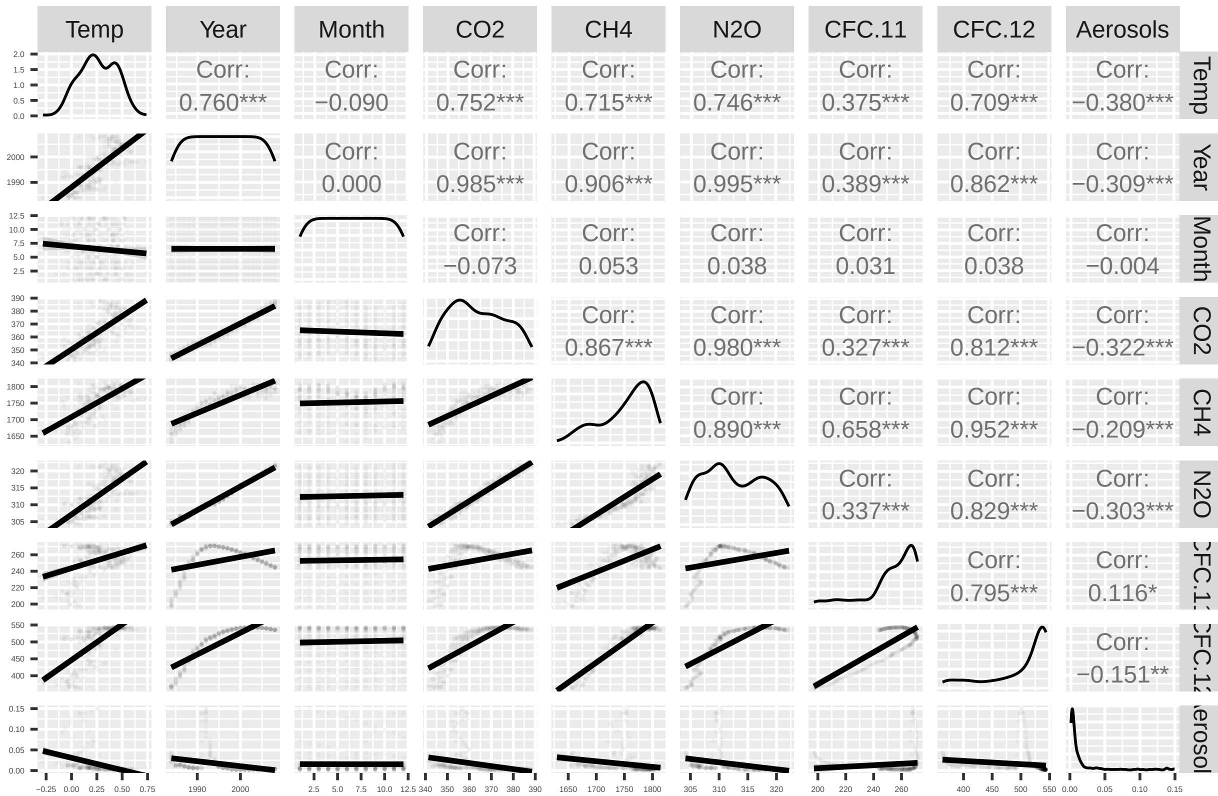
```
## Loading required package: ggplot2

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
library(dplyr)
```

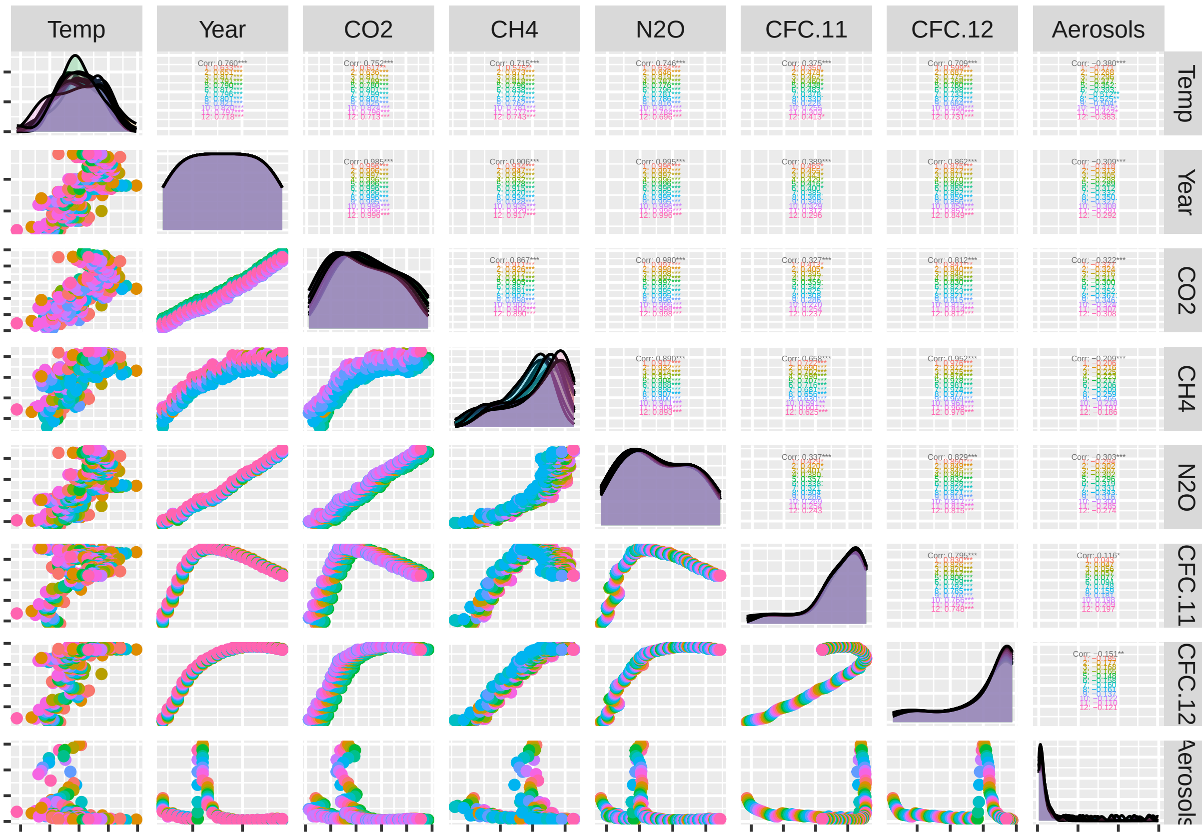
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)

### GROUPINGS OF THESE VECTORS EXPLAINED IN THE PAPER ###
variables <- c("Temp","Year","Month","MEI","CO2","CH4","N2O","CFC.11","CFC.12", "TSI", "Aerosols")
variables_of_interest <- c("Temp","Year","Month","CO2","CH4","N2O","CFC.11","CFC.12", "Aerosols")
variables_of_interest2 <- c("Temp","Year","CO2","CH4","N2O","CFC.11","CFC.12", "Aerosols")

# General Scatter Plot Matrix
climate %>%
  ggpairs(columns = variables_of_interest,
          upper = list(continuous = wrap('cor', size = 3)),
          lower = list(continuous = wrap('smooth',size = .1, alpha = 0.03))) +
  theme_grey() +
  theme(axis.text = element_text(size = 3)) +
  labs(title = "FIGURE 2: Scatter Plot Matrix")
```

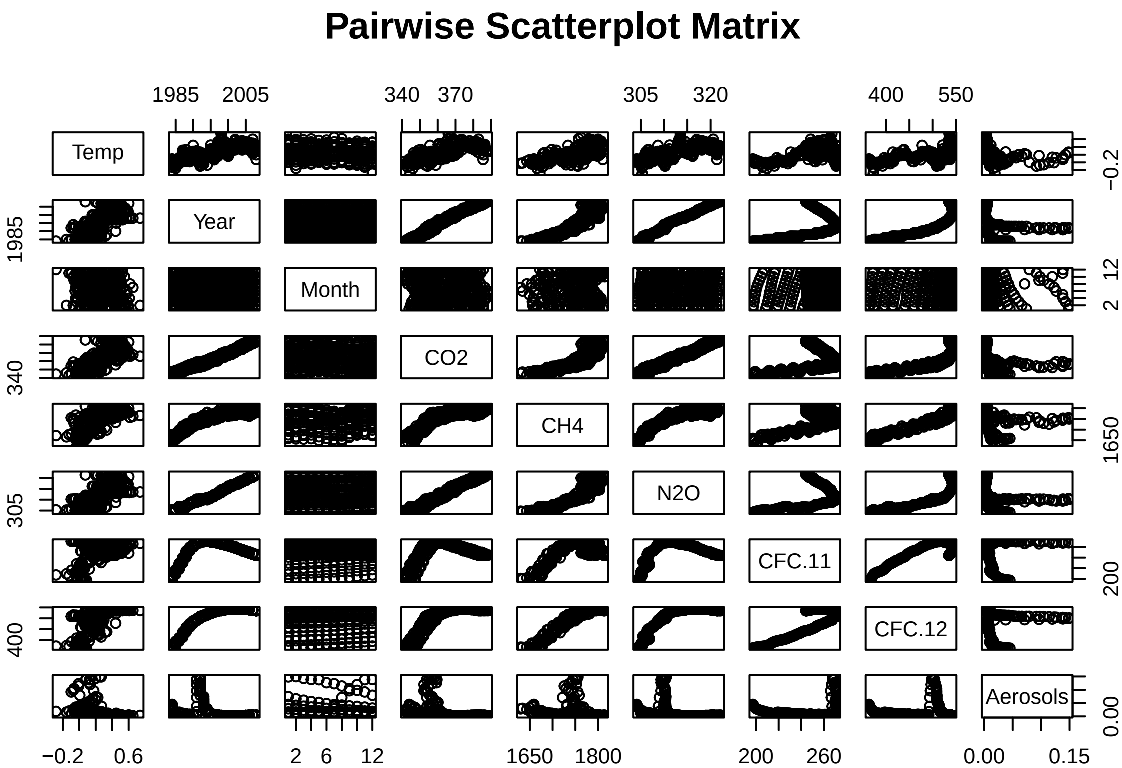
FIGURE 2: Scatter Plot Matrix




```r
# Scatter Plot Matrix Accounting for Categorical Variable
climate %>%
  ggpairs(columns = variables_of_interest2,
          aes(color = factor(Month)),
          upper = list(continuous = wrap('cor', size = 1)),
          lower = list(combo = list(continuous = "smooth", discrete = "boxplot"),
                       size = 0.1, alpha = 0.1),
          diag = list(continuous = wrap('densityDiag', alpha = 0.2))) +
  theme_grey() +
  theme(axis.text = element_text(size = 1))
```




```r
# Lower Half is Correlations; Diagonal Is Density Functions; Upper Half is Corr Values

pairs(climate[variables_of_interest], main = "Pairwise Scatterplot Matrix")
```



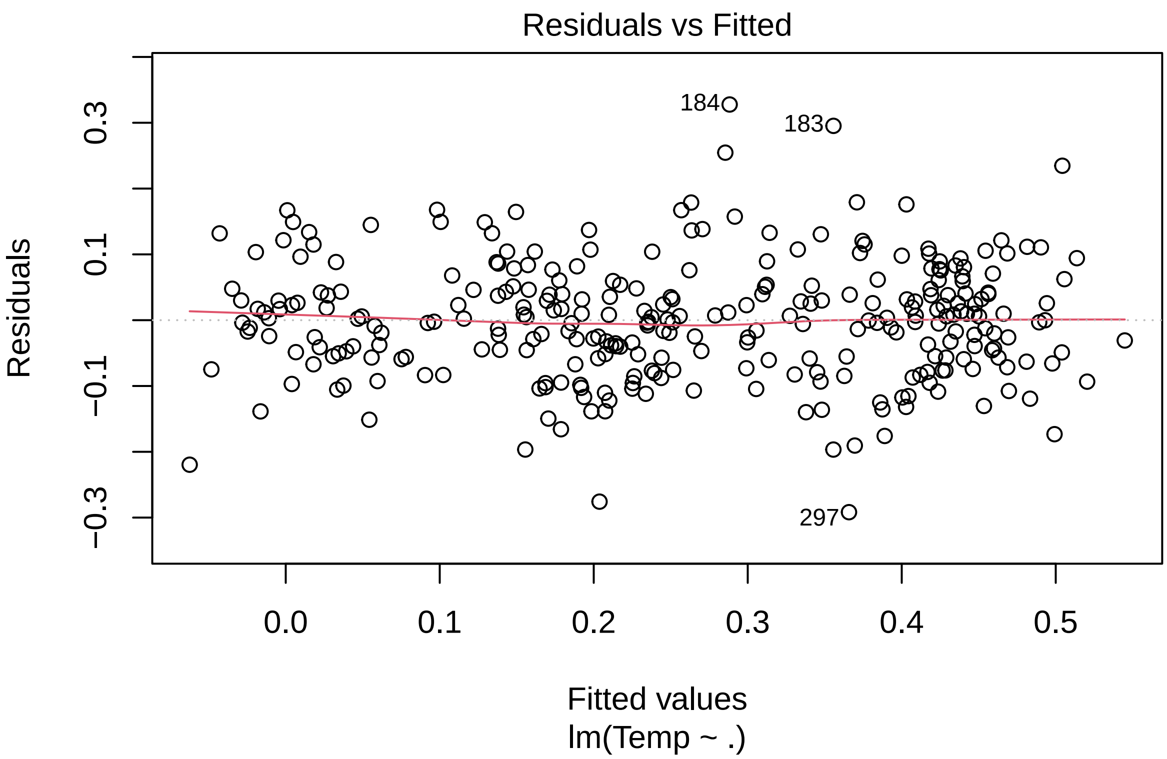
Testing Linear and Logistic ~ the plots are not actually used, but inform decisions

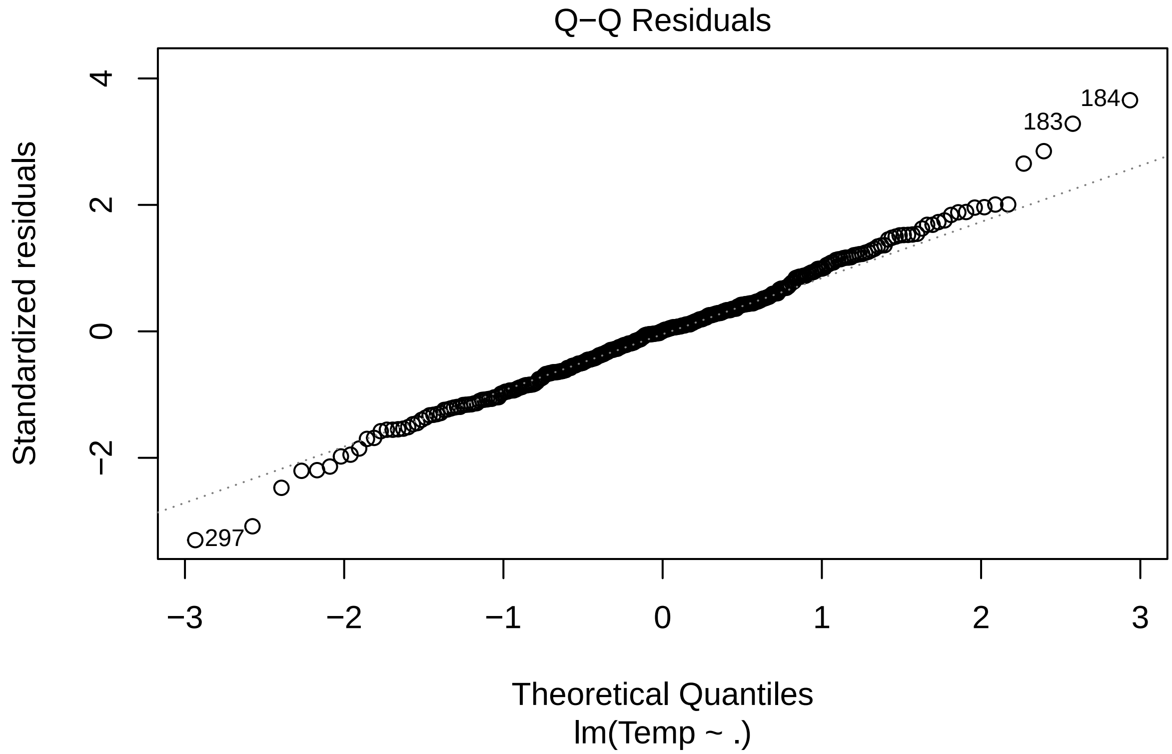
```
linear_climate_model = lm(Temp ~ ., data = climate)
summary(linear_climate_model)
```

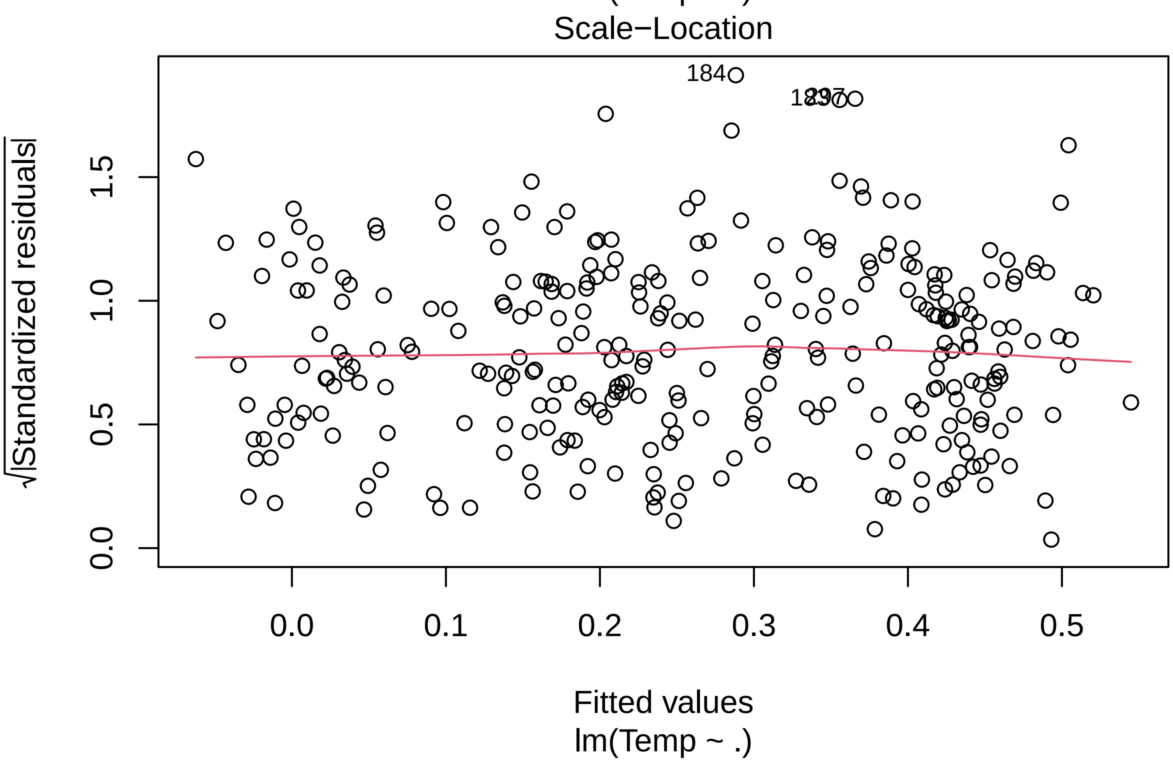
```
##
## Call:
## lm(formula = Temp ~ ., data = climate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29175 -0.05693  0.00049  0.04889  0.32774
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.170e+02  5.334e+01  -2.193  0.02912 *
## Year         6.073e-05  1.927e-02   0.003  0.99749
## Month       -4.667e-03  2.064e-03  -2.262  0.02447 *
## MEI          6.716e-02  6.224e-03  10.790  < 2e-16 ***
## CO2          1.962e-03  3.146e-03   0.624  0.53329
## CH4         -3.832e-05  5.123e-04  -0.075  0.94041
## N2O         -3.048e-03  1.859e-02  -0.164  0.86988
## CFC.11      -5.325e-03  2.001e-03  -2.661  0.00823 **
## CFC.12       3.414e-03  1.416e-03   2.410  0.01656 *
## TSI          8.571e-02  1.897e-02   4.519 9.06e-06 ***
## Aerosols    -1.762e+00  2.239e-01  -7.869 7.20e-14 ***
## ---
```

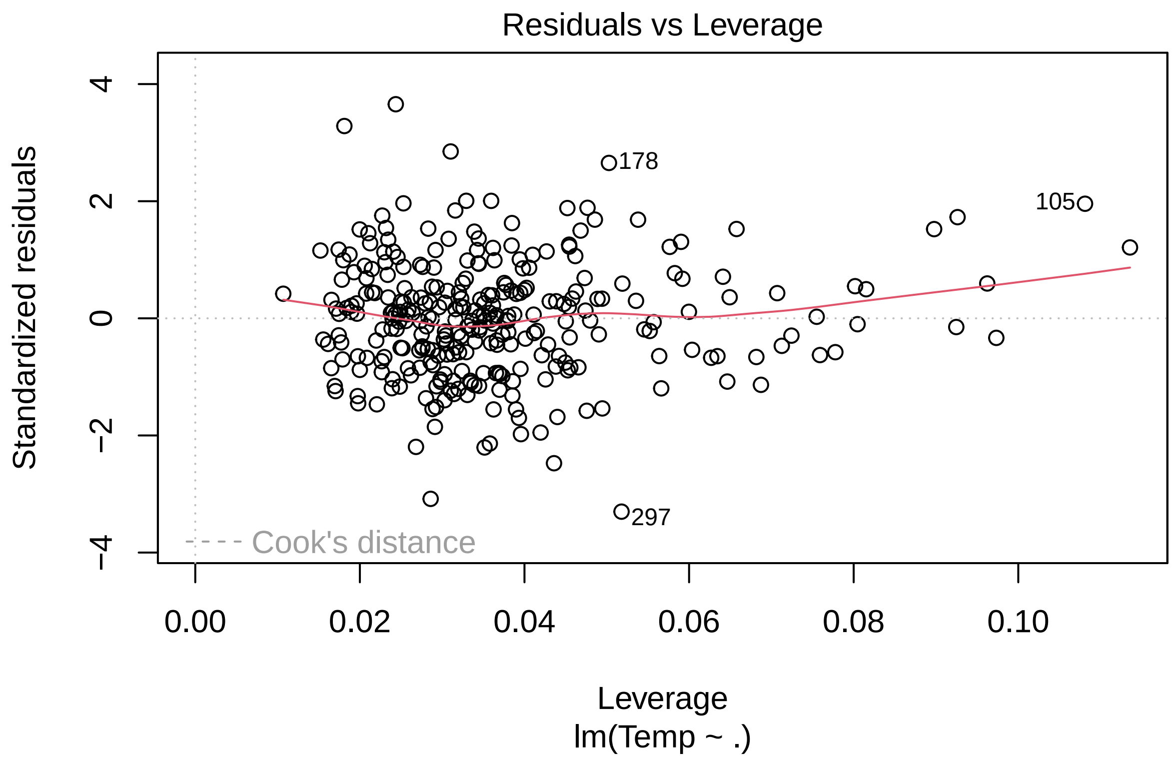
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09076 on 289 degrees of freedom
## Multiple R-squared:  0.7548, Adjusted R-squared:  0.7464
## F-statistic: 88.98 on 10 and 289 DF,  p-value: < 2.2e-16
```

```r
# Exploratory Visualization of the Response Variable ~ FIGURES 2a-2d
plot(linear_climate_model)
```




Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(Temp ~ .)
184
183
297
−3
−2
−1
0
1
2
3
4


Scale−Location
√|Standardized residuals|
Fitted values
lm(Temp ~ .)
184
297
0.0
0.1
0.2
0.3
0.4
0.5
1.0
1.5

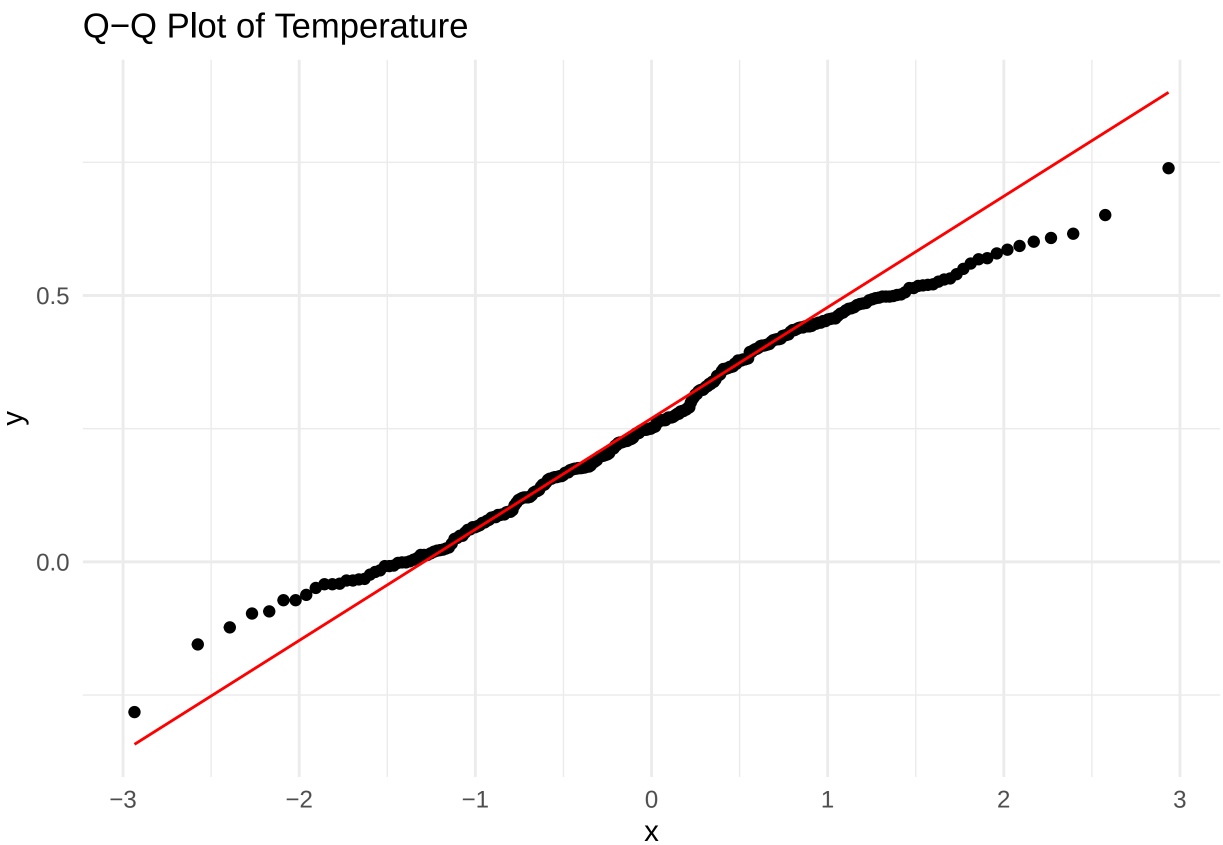



```
############## PROB NEEDS FIXING ################
binomial_climate_model <- climate %>%
  mutate(HighTemp = ifelse(Temp < mean(Temp), 1, 0)) %>%
  glm(HighTemp ~ . - Temp, data = ., family = "binomial") # deal with multi-collinearity

# balancing sensitivity and specificity of response variable
summary(binomial_climate_model)
```

```
##
## Call:
## glm(formula = HighTemp ~ . - Temp, family = "binomial", data = .)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.565e+02  2.172e+03   0.164  0.86961
## Year         7.452e-01  7.711e-01   0.966  0.33384
## Month        1.364e-01  7.870e-02   1.733  0.08311 .
## MEI         -1.469e+00  3.199e-01  -4.591  4.4e-06 ***
## CO2         -1.206e-02  1.224e-01  -0.099  0.92149
## CH4          8.272e-03  2.024e-02   0.409  0.68275
## N2O         -9.541e-01  7.335e-01  -1.301  0.19336
## CFC.11       1.356e-01  9.458e-02   1.434  0.15167
## CFC.12      -9.209e-02  6.096e-02  -1.511  0.13088
## TSI         -1.131e+00  7.638e-01  -1.481  0.13869
## Aerosols     3.811e+01  1.222e+01   3.118  0.00182 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 415.77  on 299  degrees of freedom
## Residual deviance: 168.15  on 289  degrees of freedom
## AIC: 190.15
##
## Number of Fisher Scoring iterations: 6
```

```r
# NEED TO USE ROC CURVE OR OTHER METHOD TO DETERMINE THRESHOLDS OR
# CLUSTERING/HIERARCHICAL METHODS INSTEAD

##### BRIEF VISUALIZATION W/ GGPLOT #####
ggplot(climate, aes(sample = Temp)) +
  geom_qq() +
  geom_qq_line(col = "red") +
  ggtitle("Q-Q Plot of Temperature") +
  theme_minimal()
```



Testing LDA ~ Confusion Table

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```r
lda_climate_model = lda(Temp ~ ., data = climate)
summary(lda_climate_model)
```

```
##         Length Class  Mode
## prior    238   -none- numeric
## counts   238   -none- numeric
## means   2380   -none- numeric
## scaling  100   -none- numeric
## lev      238   -none- character
## svd       10   -none- numeric
## N          1   -none- numeric
## call       3   -none- call
## terms      3   terms  call
## xlevels    0   -none- list
```

```r
# Create a confusion matrix
predictions <- predict(lda_climate_model)
confusion_matrix <- table(Actual = climate$Temp, Predicted = predictions$class)
### WE DID NOT PRINT THE OUTPUTS AS IT IS INCOLCUSIVE AND PRINTS TOO MUCH DATA
```

Deciding on The Best Model

```r
library(leaps)


#### OLD CODE THAT USES ALL VARIABLES ####
predictor_names <- names(climate)[-which(names(climate) == "Temp")]
all_subsets <- regsubsets(Temp ~ ., data = climate,
                        nvmax = length(predictor_names), method = "exhaustive")
# Get the list of all subsets
all_subsets_list <- summary(all_subsets)$which; all_subsets_list
```

```
##    (Intercept)  Year Month   MEI   CO2   CH4   N2O CFC.11 CFC.12   TSI Aerosols
## 1         TRUE  TRUE FALSE FALSE FALSE FALSE FALSE  FALSE  FALSE FALSE    FALSE
## 2         TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE  FALSE  FALSE FALSE    FALSE
## 3         TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE  FALSE  FALSE FALSE     TRUE
## 4         TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE  FALSE  FALSE  TRUE     TRUE
## 5         TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE  FALSE  FALSE  TRUE     TRUE
## 6         TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE   TRUE   TRUE  TRUE     TRUE
## 7         TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE   TRUE   TRUE  TRUE     TRUE
## 8         TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE   TRUE   TRUE  TRUE     TRUE
## 9         TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE   TRUE   TRUE  TRUE     TRUE
## 10        TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE   TRUE   TRUE  TRUE     TRUE
```

```r
#### ~ you would use predictor_names in the for loop instead

included_vars <- c("Year", "MEI", "CO2", "CH4", "N2O", "CFC.11", "CFC.12", "Aerosols")
formula_str <- paste("Temp ~", paste(included_vars, collapse = " + "))
formula <- as.formula(formula_str)
all_subsets <- regsubsets(formula, data = climate,
                          nvmax = length(included_vars), method = "exhaustive")

####### WE CAN CHANGE THE METHOD OF SEARCH but since the algorithm returns the best
# model of each size (number of parameters 2-10 or number of predictors 1-9),
```

```r
# so the results do not depend on a penalty model for model size: it doesn't make
# any difference whether you want to use AIC, BIC, CIC, DIC #######

# Initialize a dataframe to store model specifications, MSE, and adjR2
model_info <- data.frame(
  model = character(),
  MSE = numeric(),
  R2 = numeric(),
  adjR2 = numeric(),
  stringsAsFactors = FALSE
)

# Extract Values for Every Model Size (1-9 predictors)
for (i in 1:length(included_vars)) {
  # Extract the coefficients for the best model of size i
  model_coef <- coef(all_subsets, id = i)
  model_formula <- as.formula(paste("Temp ~",
                    paste(names(model_coef)[-1], collapse = "+"))) #create formulas
  # Fit Models
  fit <- lm(model_formula, data = climate)

  # Calculate MSE
  predictions <- predict(fit, newdata = climate)
  mse <- mean((climate$Temp - predictions)^2)
  # Get R2 Values
  r2 <- summary(fit)$r.squared
  adj_r2 <- summary(fit)$adj.r.squared

  # Store model information
  model_info <- rbind(model_info, data.frame(
    model = deparse(model_formula),
    MSE = mse,
    R2 = r2,
    adjR2 = adj_r2,
    stringsAsFactors = FALSE
  ))
}

model_info
```

```
##                                                               model         MSE
## 1                                                       Temp ~ Year 0.013655576
## 2                                                 Temp ~ Year + MEI 0.011683088
## 3                                      Temp ~ Year + MEI + Aerosols 0.009825178
## 4                           Temp ~ MEI + CFC.11 + CFC.12 + Aerosols 0.009338815
## 5                     Temp ~ MEI + CO2 + CFC.11 + CFC.12 + Aerosols 0.009240142
## 6              Temp ~ Year + MEI + CO2 + CFC.11 + CFC.12 + Aerosols 0.009074121
## 7        Temp ~ Year + MEI + CO2 + N2O + CFC.11 + CFC.12 + Aerosols 0.009008043
## 8 Temp ~ Year + MEI + CO2 + CH4 + N2O + CFC.11 + CFC.12 + Aerosols 0.008995450
##          R2     adjR2
## 1 0.5781145 0.5766987
## 2 0.6390540 0.6366234
## 3 0.6964536 0.6933771
## 4 0.7114797 0.7075675
```

```
## 5 0.7145282 0.7096732
## 6 0.7196573 0.7139165
## 7 0.7216988 0.7150272
## 8 0.7220878 0.7144476
```

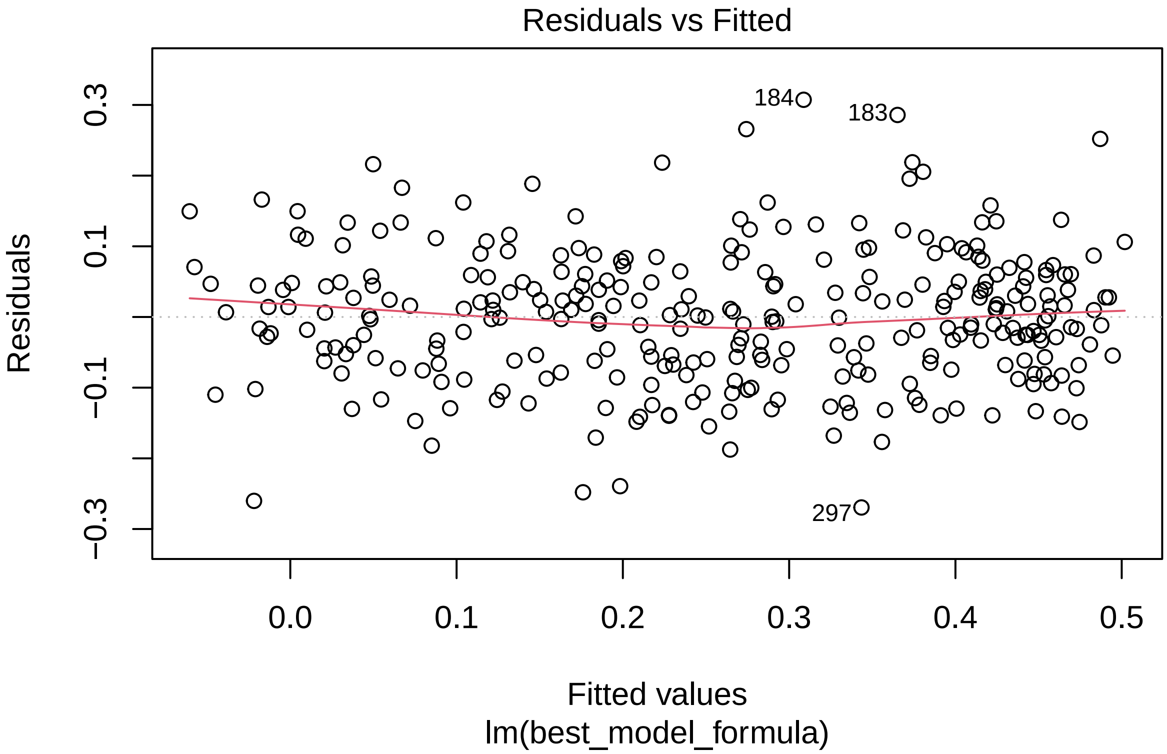
Selection Criteria

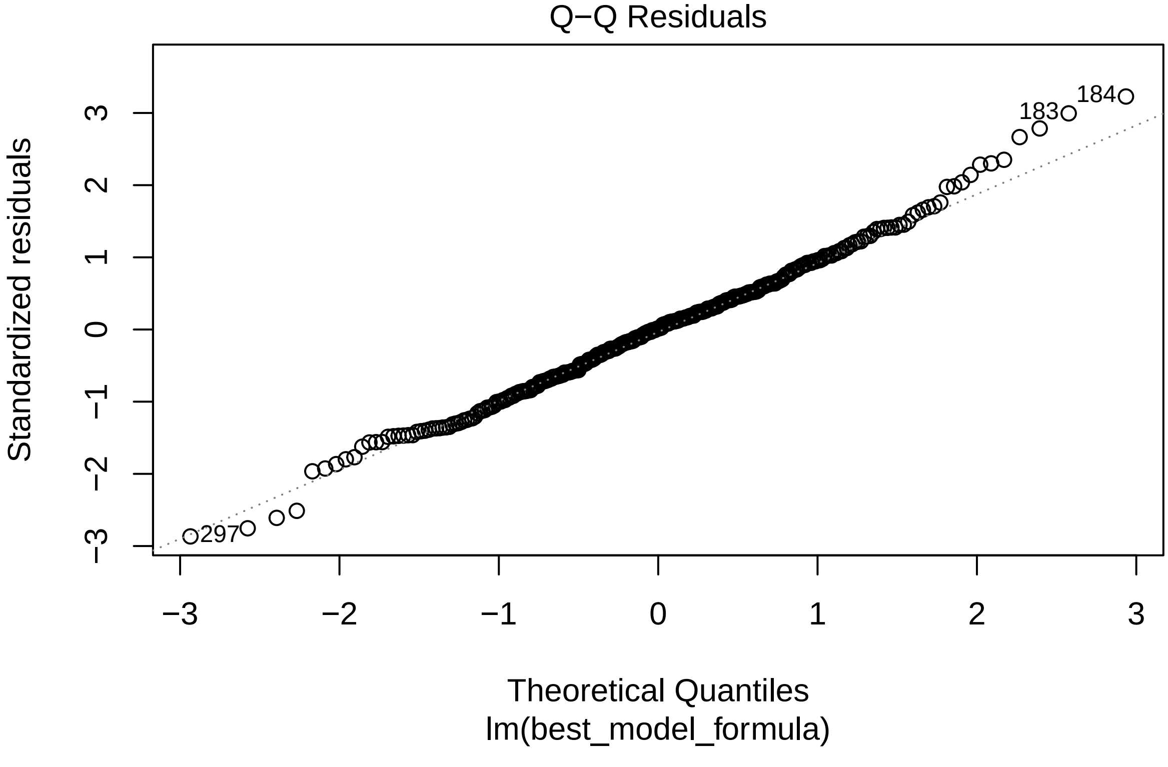
```r
# Regsubsets Identifies the Best Model at each # of predictors (1-9);
# We decide on the best model by considering the model with the lowest MSE that
# does not overfit the data (the adjusted R2 is not significantly smaller than the R2)

filter <- model_info[model_info$adjR2 >= (0.95 * model_info$R2), ]

# Choose the model with the lowest MSE from the filtered models
best_model <- filter[which.min(filter$MSE), ]

print(best_model)
```

```
##                                                              model        MSE
## 8 Temp ~ Year + MEI + CO2 + CH4 + N2O + CFC.11 + CFC.12 + Aerosols 0.00899545
##          R2     adjR2
## 8 0.7220878 0.7144476
```

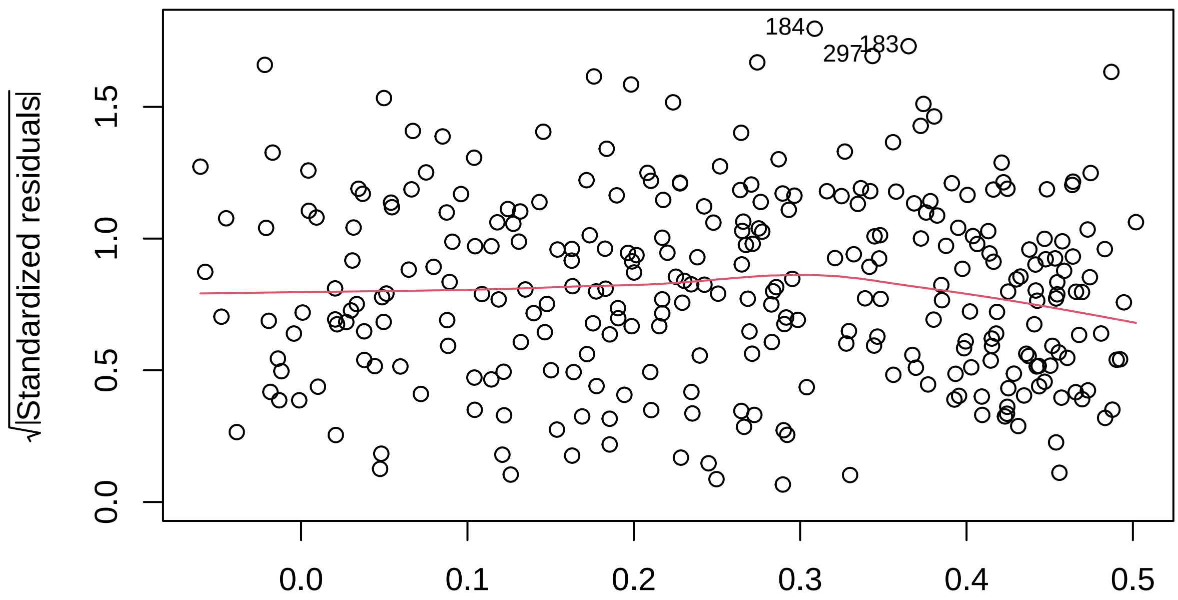
Analyzing Our Model: Figure 3

```r
best_model_formula <- as.formula(best_model$model)


# Fit the best model
best_fit <- lm(best_model_formula, data = climate)
# FIGURE 3 (a-d)
plot(best_fit)
```

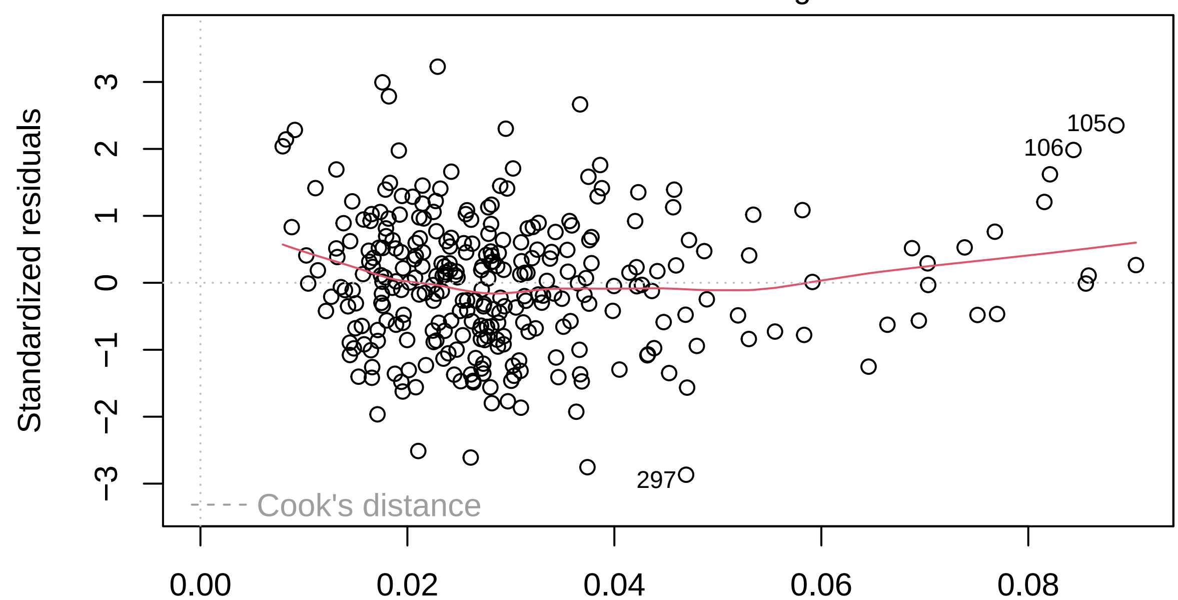

Residuals vs Fitted
Residuals
0.3
0.1
−0.1
−0.3
184
183
297
0.0
0.1
0.2
0.3
0.4
0.5
Fitted values
lm(best_model_formula)


Q−Q Residuals
Standardized residuals
3
2
1
0
−1
−2
−3
184
183
297
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(best_model_formula)

Scale−Location



Fitted values
lm(best_model_formula)
Residuals vs Leverage



Leverage
lm(best_model_formula)

```r
# Analysis ~ t-test
summary(best_fit)
```

```
##
## Call:
## lm(formula = best_model_formula, data = climate)
##
```

```
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.269500 -0.064557  0.001331  0.057791  0.307282
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 69.0353507 25.7736212   2.679  0.00782 **
## Year        -0.0398327  0.0148996  -2.673  0.00793 **
## MEI          0.0618198  0.0065293   9.468  < 2e-16 ***
## CO2          0.0097461  0.0030543   3.191  0.00157 **
## CH4         -0.0003409  0.0005341  -0.638  0.52381
## N2O          0.0217293  0.0136837   1.588  0.11338
## CFC.11      -0.0066150  0.0020318  -3.256  0.00126 **
## CFC.12       0.0053424  0.0013216   4.042 6.78e-05 ***
## Aerosols    -1.7626952  0.2339216  -7.535 6.20e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0963 on 291 degrees of freedom
## Multiple R-squared:  0.7221, Adjusted R-squared:  0.7144
## F-statistic: 94.51 on 8 and 291 DF,  p-value: < 2.2e-16
```

```r
qt(1-0.05/2, 300-8-1)
```

```
## [1] 1.96815
```

```r
# F-stat overall significance of the model
summary(best_fit)$fstatistic[1]
```

```
##    value
## 94.51168
```

```r
qf(1-0.05,df1 = 8, df2 = 300-8-1)
```

```
## [1] 1.970285
```

```r
anova_table <- anova(best_fit)
print(anova_table)
```

```
## Analysis of Variance Table
##
## Response: Temp
##            Df Sum Sq Mean Sq  F value    Pr(>F)
## Year        1 5.6137  5.6137 605.3399 < 2.2e-16 ***
## MEI         1 0.5917  0.5917  63.8094 3.220e-14 ***
## CO2         1 0.0151  0.0151   1.6299   0.20274
## CH4         1 0.0260  0.0260   2.8026   0.09519 .
## N2O         1 0.0184  0.0184   1.9817   0.16028
## CFC.11      1 0.0030  0.0030   0.3268   0.56797
## CFC.12      1 0.2172  0.2172  23.4208 2.115e-06 ***
## Aerosols    1 0.5266  0.5266  56.7824 6.195e-13 ***
## Residuals 291 2.6986  0.0093
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Calculating F-val
mean_sq_model <- sum(anova_table$`Mean Sq`[1:8])
```

```
F_value <- mean_sq_model /anova_table$`Mean Sq`[9]
# F-stat compares the variability explained by the predictors (Mean Sq Model)
# with the variability not explained by the model (Mean Sq Residuals)
F_value
```

```
## [1] 756.0935
```

F-test Lack of Fit

```
residuals <- residuals(best_fit)
fitted_values <- fitted(best_fit)

# Calculate lack-of-fit sum of squares
n <- length(residuals)
mean_residuals <- mean(residuals)
lack_of_fit_ss <- sum((residuals - mean_residuals)^2)

# Calculate residual sum of squares
residual_ss <- sum(residuals^2)

# Degrees of freedom for the lack-of-fit test
df_lack_of_fit <- n - length(coefficients(best_fit)) #Adjust for number of coefficients

# Degrees of freedom for residuals
df_residuals <- df.residual(best_fit)

# Calculate F-value for lack of fit
F_lack_of_fit <- (lack_of_fit_ss / df_lack_of_fit) / (residual_ss / df_residuals)
F_critical <- qf(1 - 0.05, df_lack_of_fit, df_residuals)

# Calculate p-value for lack of fit
p_value_lack_of_fit <- pf(F_lack_of_fit, df_lack_of_fit, df_residuals, lower.tail = FALSE)

# Print results
cat("F-critical for LOF:", F_critical, "\n")
```

```
## F-critical for LOF: 1.213079
```

```
cat("F-value for lack of fit:", F_lack_of_fit, "\n")
```

```
## F-value for lack of fit: 1
```

Cross Validation

```
library(caret)
```

```
## Loading required package: lattice
```

```
# Leave-One-Out-Cross-Validation
train_control_loocv <- trainControl(method = "LOOCV")
loocv <- train(best_model_formula, data = climate, method = "lm", trControl = train_control_loocv)

# Calculate MSE for LOOCV
```

```
mse1 <- loocv$results$RMSE^2
cat("LOOCV MSE:", mse1, "\n")
```

```
## LOOCV MSE: 0.009562674
```

```
# K-fold cross-validation
train_control_kfold <- trainControl(method = "cv", number = 10)
kfold <- train(best_model_formula, data = climate, method = "lm", trControl = train_control_kfold)

# Calculate MSE for K-fold CV
mse2 <- kfold$results$RMSE^2
cat("K-fold MSE:", mse2, "\n")
```

```
## K-fold MSE: 0.009185652
```

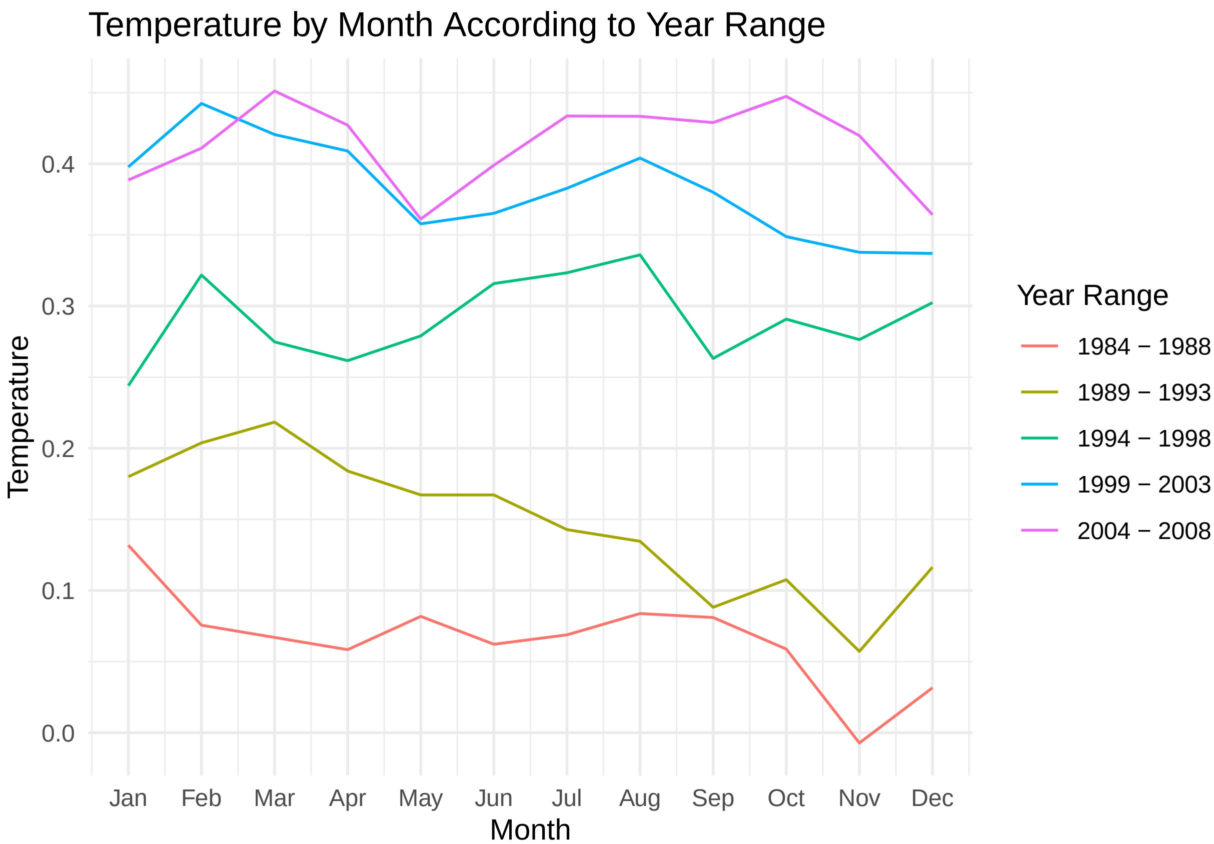
Additional EXPLORATORY Analysis

```
split_month <- climate %>%
  split(.$Month) %>%
  lapply(function(.) {
    .[order(.$Year, decreasing = FALSE), ]
  })

# split_month
### NOT PRINTED AS THE OUTPUT IS TOO LONG
```

```
aggregate_temperature <- function(df) {

  # Create a new column that groups years into sets of 5
  df <- df %>%
    mutate(YearGroup = (row_number() - 1) %/% 5 + 1)

  # Calculate the average temperature for each group and rename the year group
  result <- df %>%
    group_by(YearGroup) %>%
    summarise(
      YearRange = paste(min(Year), max(Year), sep = " - "),
      Temperature = mean(Temp),
      Month = first(Month) # unchanged
    ) %>%
    ungroup()

  return(result)
}

# Apply the function to each split_month dataframe
bymonth <- lapply(split_month, aggregate_temperature)
# bymonth
### NOT PRINTED AS THE OUTPUT IS TOO LONG
```

FIGURE 1 Exploration of How the Categorical Variable of Month Affects Temperature by Year

```r
by_yeargroup <- list()

for (i in 1:5) {
  new_df <- do.call(rbind, lapply(bymonth, function(df) df[i, ]))
  by_yeargroup[[i]] <- new_df
}
# by_yeargroup
### NOT PRINTED AS THE OUTPUT IS TOO LONG

plot <- ggplot() +
  labs(title = "Temperature by Month According to Year Range",
       x = "Month",
       y = "Temperature") +
  scale_color_discrete(name = "Year Range") +
  scale_x_continuous(breaks = 1:12, labels = month.abb) +  # Set x-axis breaks and labels
  theme_minimal()

# Iterate through each data frame in the by_yeargroup list and add a geom_line() for each
for (i in 1:length(by_yeargroup)) {
  plot <- plot + geom_line(data = by_yeargroup[[i]], aes(x = Month,
                                    y = Temperature, color = factor(YearRange)))
}

plot
```


Temperature by Month According to Year Range
Temperature
0.4
0.3
0.2
0.1
0.0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month
Year Range
1984 − 1988
1989 − 1993
1994 − 1998
1999 − 2003
2004 − 2008