

Do Sunspots Influence the Earth's Surface Temperatures & Precipitation?

Author: James Topor

Contents

Introduction	1
Data	3
Sunspot Data	3
Surface Temperature and Precipitation Data	3
Cases	4
Variables	5
Type of Study	5
Scope of inference	5
Exploratory Data Analysis	6
Data Refinement	6
Basic Summary Statistics + Visualization of Data	6
Inference	21
Hypothesis Statements	21
Theoretical Hypothesis Testing: Least Squares Line Fitting	22
Linear Regression Model Diagnostics	24
Summary of Hypothesis Testing and Inference	30
Conclusion	31
References:	31

Introduction

The earth's climate is affected by a variety of phenomena, many of which are external to the earth's atmosphere. These include the wobble of the earth's axis, gamma rays that continuously bombard the planet's atmosphere from outer space, the rate at which the earth's oceans absorb and release heat, natural variations in the earth's orbit around the sun, and variations in the sun's irradiance resulting from the sun's own internal stellar dynamics.

For example, scientists have long noted the possibility of a **relation between sunspot activity and variations in the earth's surface temperatures and precipitation**. Sunspots, which are basically areas of relatively high radial magnetic activity on the sun, manifest themselves as dark spots within the outermost layer of the sun. This layer, known as the photosphere, is the source of nearly all of the earth's solar radiation. The outer boundary of the photosphere typically has a temperature of approximately 4,200 degrees Kelvin.

However, the relatively high level of radial magnetic activity within a sunspot results in temperatures within a sunspot averaging 4,600 degrees Kelvin, or approximately 9.5% higher than the normal temperature of the photosphere. As such, when the number of observed sunspots is relatively high, the amount of energy being emitted by the sun is relatively higher than it would otherwise be. When this occurs, the amount of solar radiation being delivered to the earth's atmosphere increases, which might lead to changes in observable surface temperatures.

In fact, according to Geerts & Linacre (see Reference # 1), a lack of observable sunspots has been noted to coincide with colder periods in earth's climatic record. The number of observable sunspots varies over time as a result of the sun's own internal stellar dynamics. During what scientists consider "normal" solar conditions, the number of observable sunspots peaks roughly every 11 years or so. However, the sun also periodically experiences long periods of reduced irradiance, during which very few sunspots are observed, such as during the coldest intervals during earth's 1450 - 1820 "Little Ice Age". Furthermore, the length of the sunspot cycle can itself vary, with shorter intervals between peak sunspot activity (i.e., intervals of less than 11 years) appearing to coincide with higher average temperatures in earth's Northern Hemisphere and longer intervals (i.e., intervals > 11 years) appearing to coincide with lower average temperatures in earth's Northern Hemisphere.

Fortunately, regular periodic observations of sunspot counts, surface temperatures, and precipitation are readily available to us for the past 150 years or so. As such, we can conduct our own analysis into the relationship between sunspots and some of the earth's climatic variations.

Specifically, we will examine the following:

- Whether the average number of sunspots observed within a given calendar year might be predictive of the average surface temperature observed within the continental United States for that same calendar year;
- Whether the average number of sunspots observed within a given calendar year might be predictive of the average precipitation amount observed within the continental United States for that same calendar year.

Why is this research relevant?

The earth's climate has never remained static and is constantly in a state of flux. For example, in geologic terms, it wasn't that long ago (approximately 12,000 years) that the majority of the northern hemisphere was still buried under glaciers and sea levels were hundreds of feet lower than they are today. The rapid warming that coincided with the demise of that period of glaciation was the result of naturally occurring phenomena rather than due to any human activity. In fact, sea levels increased by approximately 400 feet during the period between 19,000 and 6,000 years ago, and since then have remained relatively steady (See References 2 & 3).

As the global community continues to debate the causes of climate change and what role humankind's activities may play in influencing earth's ever-changing climate, it is of crucial importance that we understand what role many naturally occurring phenomena play in shaping the weather we experience on an ongoing basis. Many such phenomena (e.g., the wobble of the earth's axis, cosmic rays that continually penetrate the earth's atmosphere, etc.) are widely recognized as having a meaningful impact on earth's climate even though their affects remain very hard to quantify. As such, humankind must be cognizant of naturally occurring factors that are beyond its control if we are to have realistic and practicable approaches to addressing climate change.

Data

To facilitate our analysis we will be drawing on time series data from two separate sources, one of which records sunspot observations on a regular basis, and another which tracks surface temperature and precipitation observations for the continental United States.

Sunspot Data

The Royal Observatory of Belgium has recorded daily sunspot observations since January, 1818. Access to the data repository containing those observations is available via the internet at the following web page:

<http://www.sidc.be/silso/datafiles>

The specific sunspot data set used for this analysis can be obtained from that page by scrolling down to the item labeled

“Yearly mean total sunspot number [1700 - now]”

and selecting the **CSV** option, which generates a .csv file containing the mean number daily sunspot observations for each calendar year for the period 1700 - 2015.

The .csv file contains 5 columns of data described as follows:

- **Column 1:** Gregorian calendar year (mid-year date)
- **Column 2:** Yearly mean total sunspot number.
- **Column 3:** Yearly mean standard deviation of the input sunspot numbers from individual stations.
- **Column 4:** Number of observations used to compute the yearly mean total sunspot number.
- **Column 5:** Definitive/provisional marker. ‘1’ indicates that the value is definitive. ‘0’ indicates that the value is still provisional.
(source: <http://www.sidc.be/silso/infosnytot>)

From this .csv file we will utilize data for the years 1900 - 2015 as part of our analysis.

Surface Temperature and Precipitation Data

The National Oceanic and Atmospheric Administration (NOAA) maintains a repository of climate-related observations that is accessible via the internet. Individual temperature and precipitation measurements are recorded at 1,218 different locations throughout the continental United States on a daily basis. The NOAA aggregates those measurements and computes average temperature and precipitation amounts for the continental United States and updates its online repository of that data on a regular basis.

Average surface temperature observations for the continental United States for the period 1900 - 2015 can be obtained via the following web page:

<http://www.ncdc.noaa.gov/cag/time-series/us/110/0/tavg/12/12/1900-2015>

The specific data set to be used in this analysis can be obtained as follows from that web page:

- For “**Parameter**” select “Average Temperature” from the drop down list;
- For “**Time Scale**” select “12-Month”;
- For “**Month**” select “December”;
- For “**Start Year**” select “1900”;
- For “**End Year**” select “2015”
- For “**State(Region)**” select “Contiguous US”
- For “**Climate Division/City**” select “All 48 States”

Then, click the “**Plot**” button. Once plotting has been completed, the data set is displayed on that page and can be downloaded as a .csv file by scrolling to the “**Download**” label and clicking the MS Excel icon. The .csv file will have a header that explains the two columns of data therein.

Precipitation observations for the continental United States for the period 1900 - 2015 can be obtained via the following web page:

<http://www.ncdc.noaa.gov/cag/time-series/us/110/0/pcp/12/12/1900-2015>

The specific data set to be used in this analysis can be obtained as follows from that web page:

- For “**Parameter**” select “Precipitation” from the drop down list;
- For “**Time Scale**” select “12-Month”;
- For “**Month**” select “December”;
- For “**Start Year**” select “1900”;
- For “**End Year**” select “2015”
- For “**State(Region)**” select “Contiguous US”
- For “**Climate Division/City**” select “All 48 States”

Then, click the “**Plot**” button. Once plotting has been completed, the data set is displayed on that page and can be downloaded as a .csv file by scrolling to the “**Download**” label and clicking the MS Excel icon. The .csv file will have a header that explains the two columns of data therein.

Cases

Each case is comprised of three distinct observations recorded for each calendar year for the years 1900 - 2015. The three observations are:

- 1) The average number of sunspots observed daily as recorded by The Royal Observatory of Belgium.
- 2) The average of recorded surface temperatures throughout the continental United States
- 3) The average of recorded precipitation amount throughout the continental United States

We have a total of 116 cases: one for each calendar year for the years 1900 - 2015.

Variables

Response Variables

For purposes of this analysis we will be examining two separate response variables:

- 1) **Average Temperature:** A continuous numeric variable whose value represents an average of the recorded surface temperatures in degrees Fahrenheit throughout the continental United States for a given calendar year for the years 1900 - 2015.
- 2) **Average Precipitation:** A continuous numeric variable whose value represents an average of the recorded precipitation in inches throughout the continental United States for a given calendar year for the years 1900 - 2015. All forms of surface precipitation, including rain, snow, sleet, and hail, are encompassed by this variable. (NOTE: The NOAA converts all non-rain precipitation measures to a rain-equivalent metric, thereby ensuring a standard unit of measurement for all types of precipitation).

Explanatory Variable

For purposes of this analysis we will utilize a single explanatory variable:

- **Average Sunspot Count:** A continuous numeric variable available to as the mean daily number of sunspots observed within a given calendar year for the years 1900 - 2015 as recorded by the Royal Observatory of Belgium.

Type of Study

Since we are relying upon observational data recorded each year for the period 1900 - 2015, this is strictly a retrospective observational study.

Scope of inference

Generalizability

The populations of interest for these analyses are:

- 1) Surface temperatures within the continental United States for the period 1900 - 2015;
- 2) Precipitation amounts within the continental United States for the period 1900 - 2015.

These analyses are generalizable to the populations in question. Since the annual mean temperature and annual mean precipitation amounts are the product of a very large number of individual temperature and precipitation measurements taken at 1,218 different locations throughout the continental United States, the samples are random.

However, as was reported in December, 2015 at the fall meeting of the American Geophysical Union (See Reference # 4), 808 of the 1,218 monitoring sites used to record the observations that serve as the basis for the NOAA data used in our analyses have been found to be overstating actual surface temperatures as a result of the monitoring equipment being either poorly sited or compromised by external heat sources. As a

result, the artificially inflated temperature readings from these 808 sites appear to skew the NOAA's average annual temperature readings upward by a statistically meaningful amount.

It is beyond the scope of our analyses here to attempt any adjustments to the NOAA data resulting from these biasing factors. Further work would need to be done to investigate whether the skewed temperature data from the 808 compromised NOAA monitoring sites invalidates our findings as well as our assumption that the samples are validly random.

Causality

These data definitely cannot be used to establish a causal link between sunspots and surface temperatures or precipitation since this is strictly an observational study. Observational studies can provide evidence of a naturally occurring association between variables, but they cannot by themselves show a causal connection. Causal connections can only be established through an experiment.

Exploratory Data Analysis

Data Refinement

As a first step in our exploratory data analysis we need to refine the raw data so that it conforms with our intended analytical approach. Each of the three data sets requires one or more refinements if we are to “harmonize” the format of the data so that we can perform direct comparisons between our explanatory and response variables. The R code for performing the required refinement was presented in the Project Proposal document and will not be repeated herein. Therefore, if you would like to examine the required refinement operations please either refer to that document or refer to the R Markdown version of this document.

Basic Summary Statistics + Visualization of Data

We'll use histograms, box plots, scatter plots, and R's **summary** function as part of our analysis. We start the analysis by generating some basic summary statistics for each of the relevant variables.

Basic Summary Statistics: Average Number of Sunspots Observed

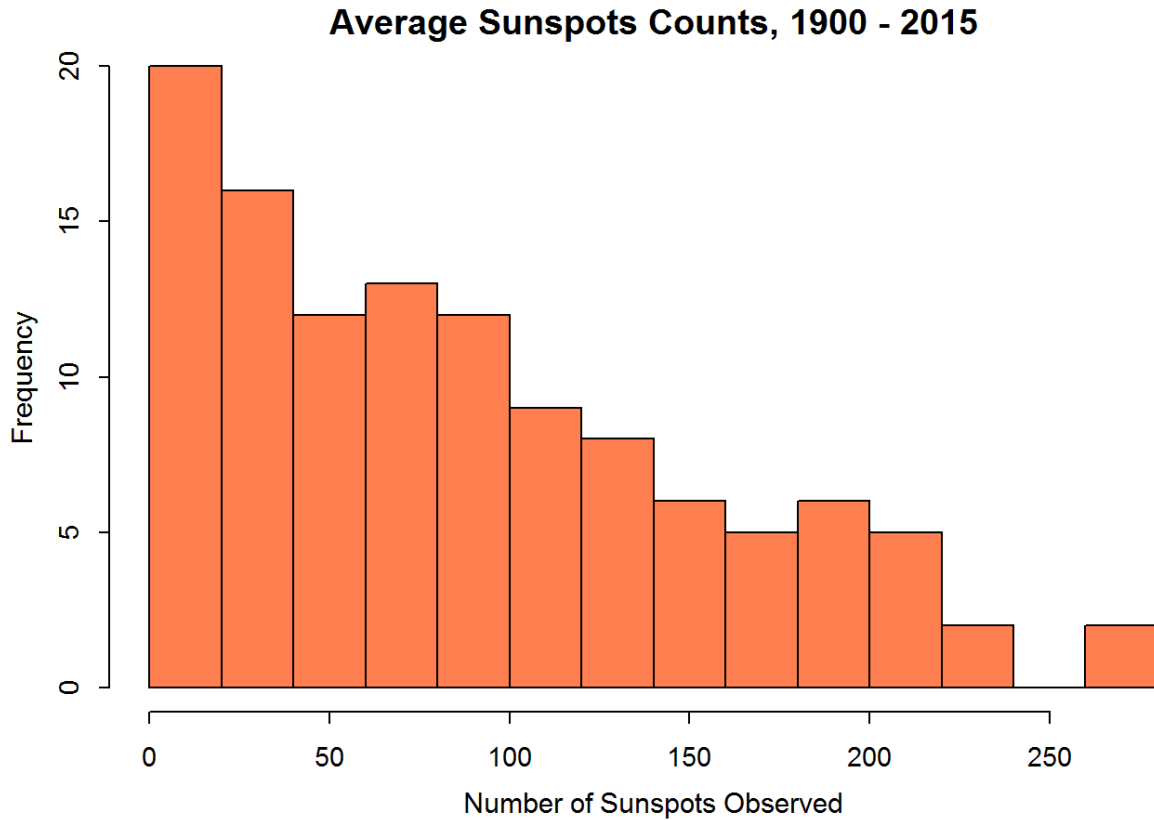
```
summary(ssdf$AvgSpots)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      \n##      2.40  27.70   76.25   88.24 132.90  269.30
```

```
sd(ssdf$AvgSpots)
```

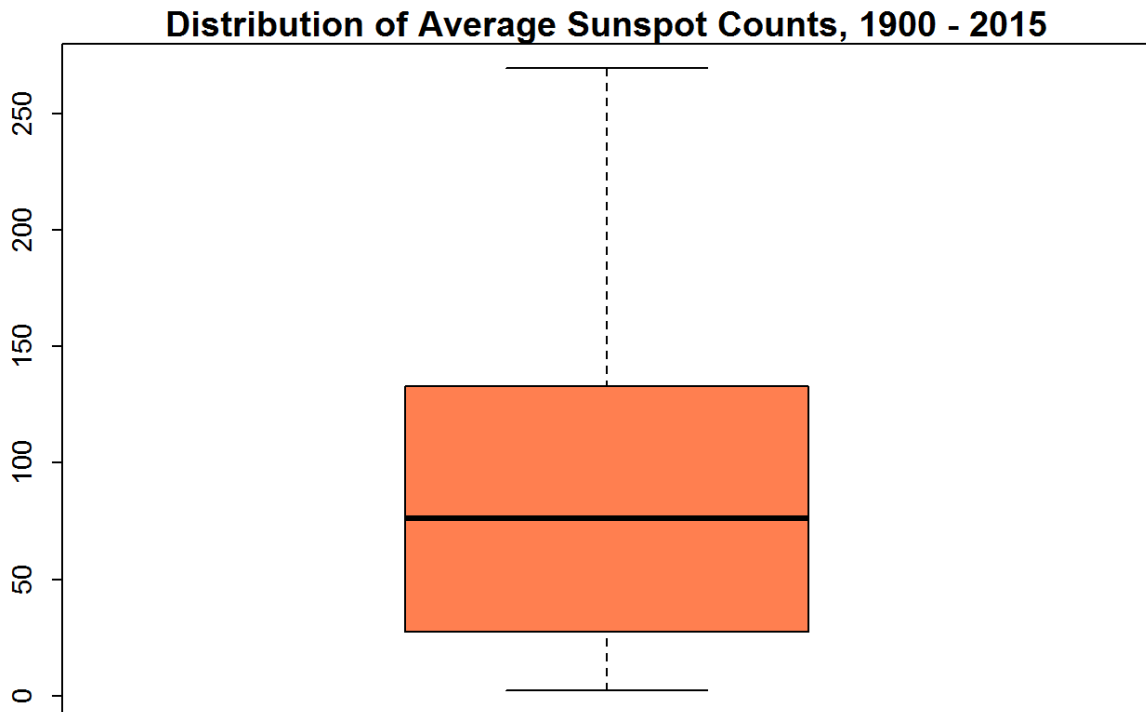
```
## [1] 66.94521
```

As can be seen in the summary statistics shown above, the average number of sunspots observed each year for the period 1900 - 2015 was 88.24 with a standard deviation of 66.94, while the median was 76.25. Since the mean is so much larger than the median, it appears that we have a right-skewed distribution. A plot of a histogram of the individual values confirms this:

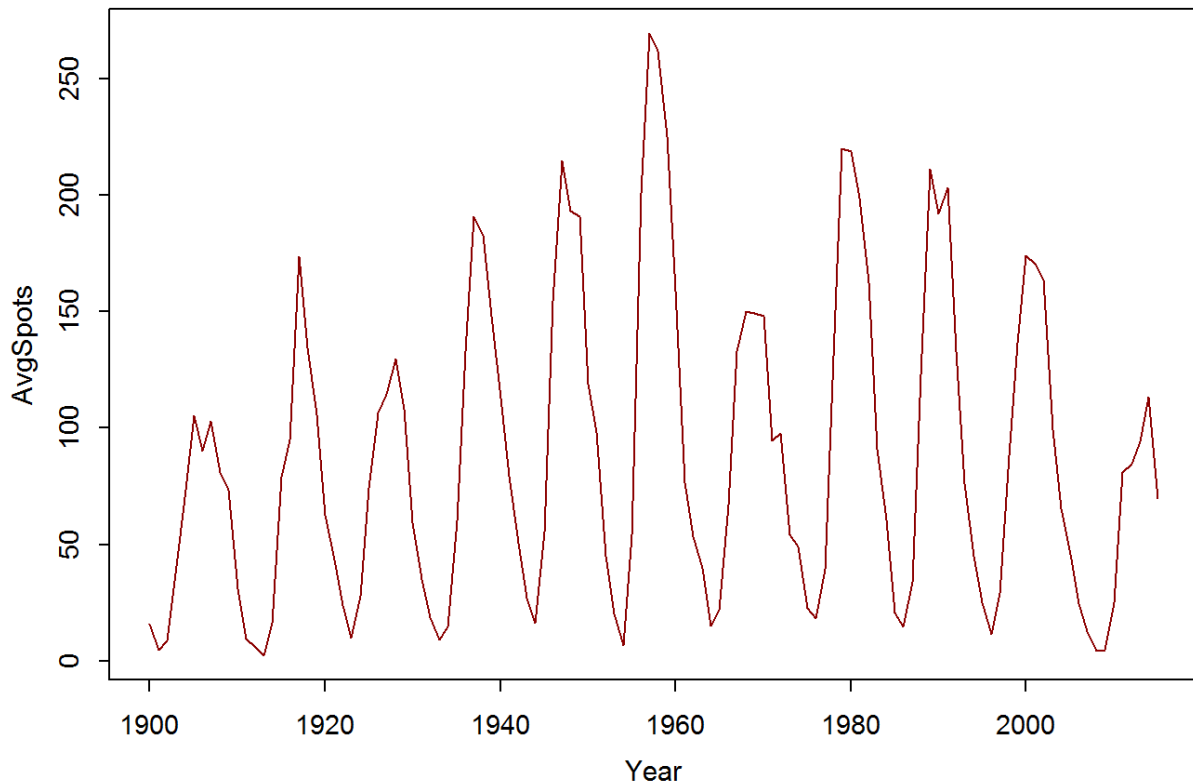


Indeed, we have a strongly right-skewed distribution that is zero-bounded on the lefthand side. This makes sense intuitively since we can't observe a negative number of sunspots in any given year.

Examining a boxplot of our sunspot data provides further confirmation of its skew:



Plotting the average sunspot counts according to their time series reveals a rather striking confirmation of the periodicity of sunspot activity. Sunspot counts do indeed appear to ebb and flow in roughly 11 year cycles.



As was discussed earlier, this periodicity is a result of the sun's internal solar dynamics, with the sun's internal magnetic activity varying in a seemingly predictable fashion. When sunspot counts are relatively high, the amount of energy released by the sun is higher than it would otherwise be due to the corresponding increase in the sun's internal magnetic activity.

Basic Summary Statistics: Average Temperatures Within the Continental USA

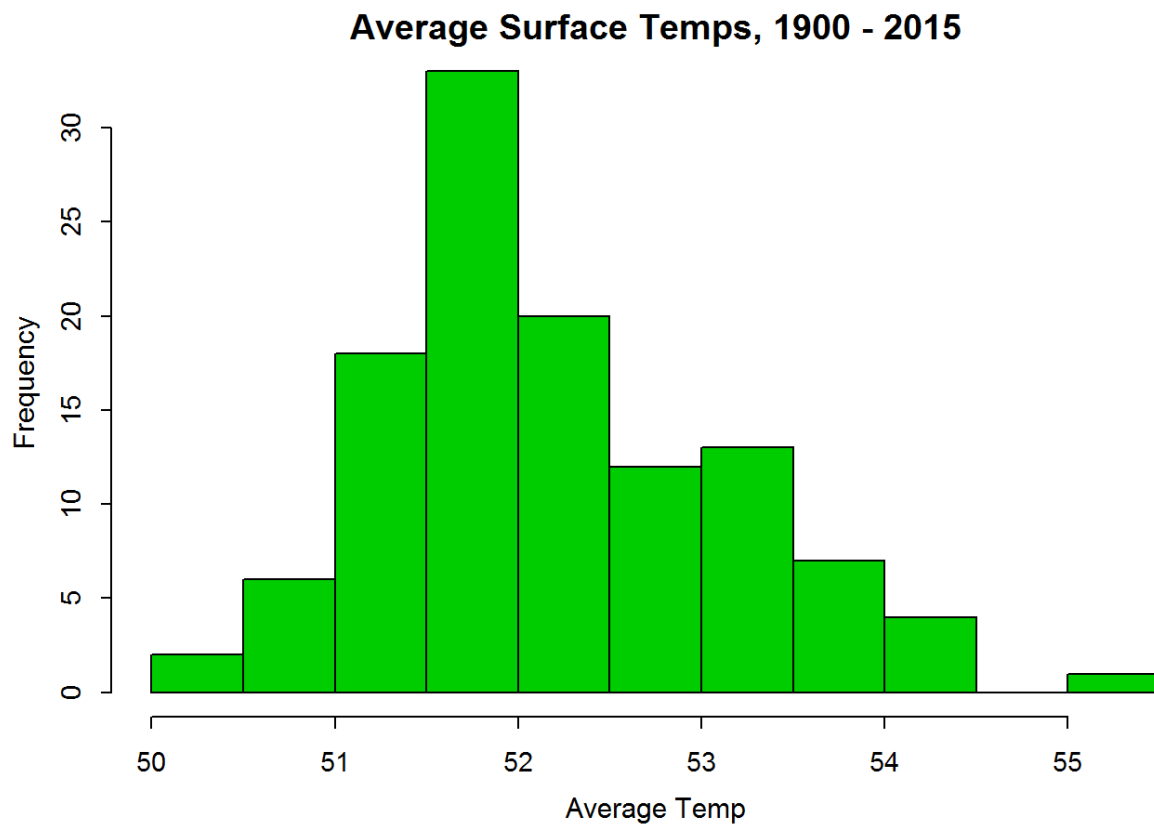
```
summary(tempdf$Value)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  50.06  51.58   51.98   52.20  52.80   55.28
```

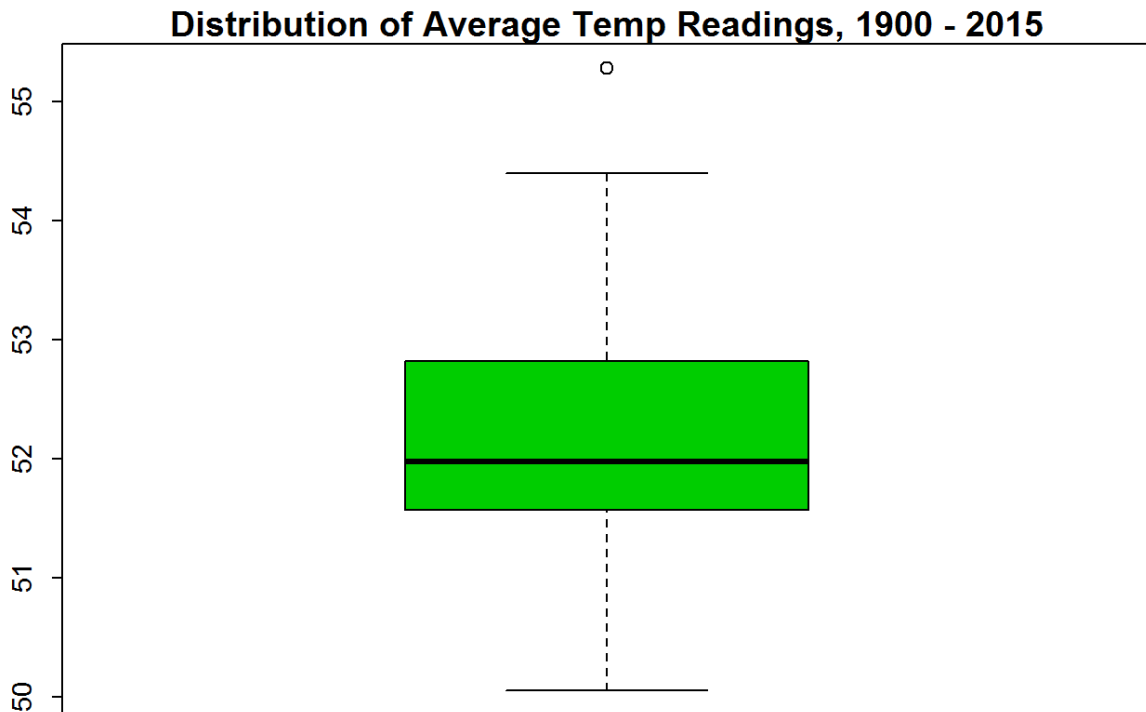
```
sd(tempdf$Value)
```

```
## [1] 0.9327599
```

The summary data for our average surface temperature within the continental United States reveals a mean average temperature of 52.20 degrees Fahrenheit, a standard deviation of 0.9327, and a median of 51.98 degrees. The mean being greater than the median suggests that we once again may be dealing with a right-skewed distribution. A plot of a histogram for the temperature data confirms this:



Examining a boxplot of our temperature data provides further confirmation of its skew:



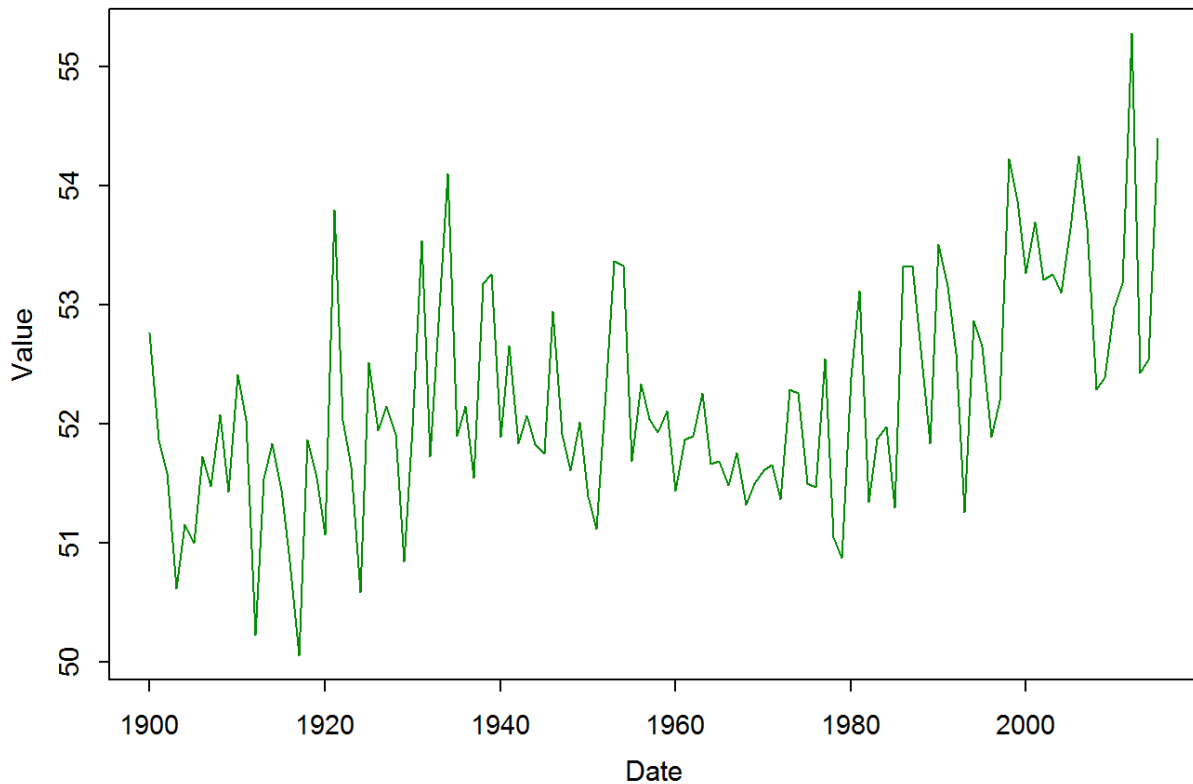
The boxplot also reveals the presence of an outlier with a value greater than 55. We can find the year corresponding to this outlier using R's native **subset** function:

```
subset(tempdf, Value == max(tempdf$Value))
```

```
##      Date Value  
## 113 2012 55.28
```

Our outlier average temperature reading is for the year 2012.

Plotting our temperature data according to its time series reveals what appears to be an upward trend toward the end of the period:



However, we cannot draw any conclusions regarding possible trends from such a plot since what we are observing could simply be the result of randomness. Furthermore, our data only span a very brief 116 year period of the earth's history: the earth has been both much warmer and much colder at different times during its evolution, and both it and its surrounding astrophysical environment continue to evolve on an ongoing basis. As such, attempting to attribute variations in such a graph to a "trend" is not recommended.

Basic Summary Statistics: Average Precipitation Amounts Within the Continental USA

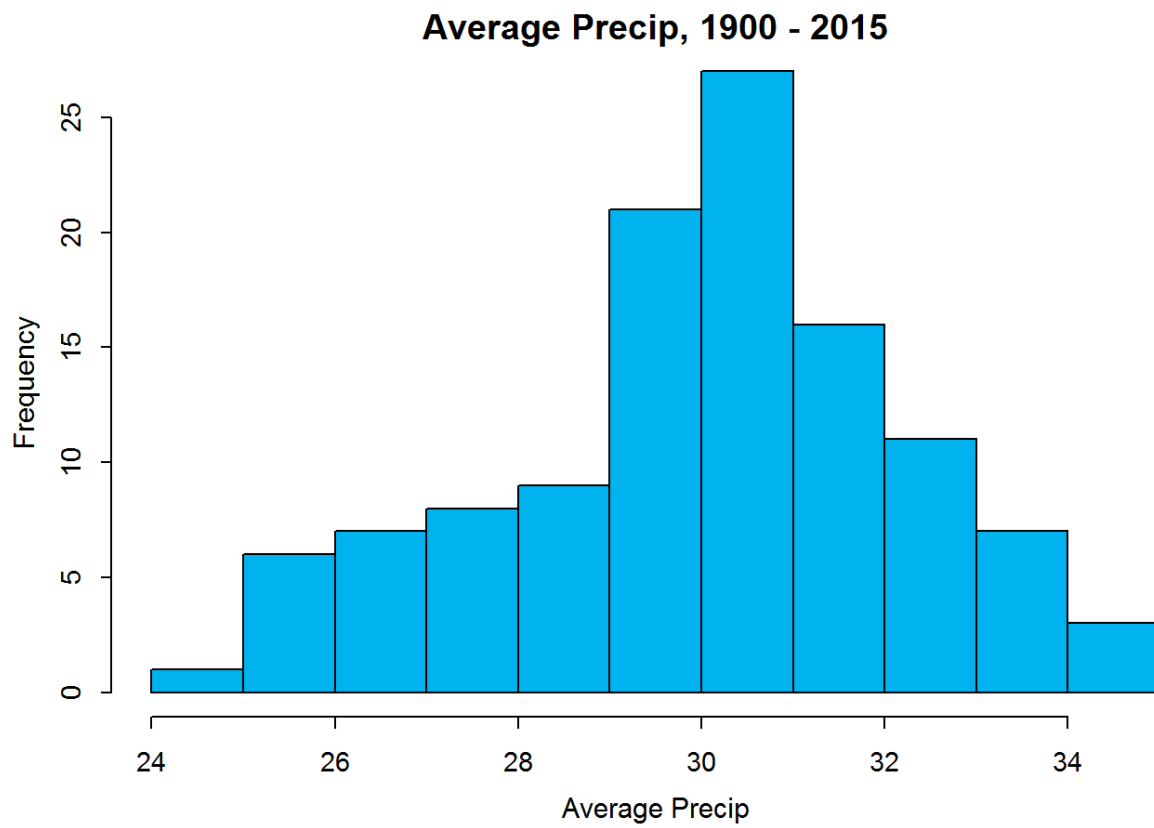
```
summary(precipdf$Value)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  24.91   28.78   30.28   30.04   31.31   34.96
```

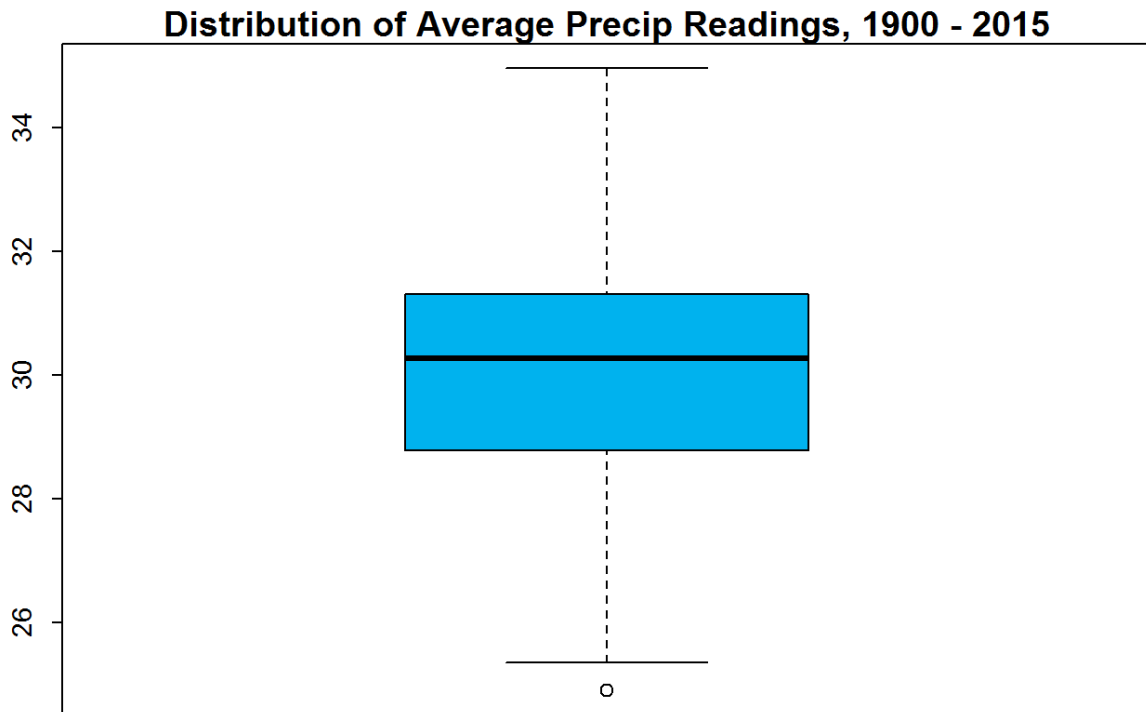
```
sd(precipdf$Value)
```

```
## [1] 2.22302
```

The summary data for our average precipitation within the continental United States reveals a mean amount of 30.04 inches, a standard deviation of 2.223 inches, and a median of 30.28 inches. The mean being nearly identical to the median suggests that we may have a nearly normal distribution of average precipitation amounts. A plot of a histogram for the precipitation data confirms this:



Examining a boxplot of our temperature data provides further confirmation of the near normal distribution:



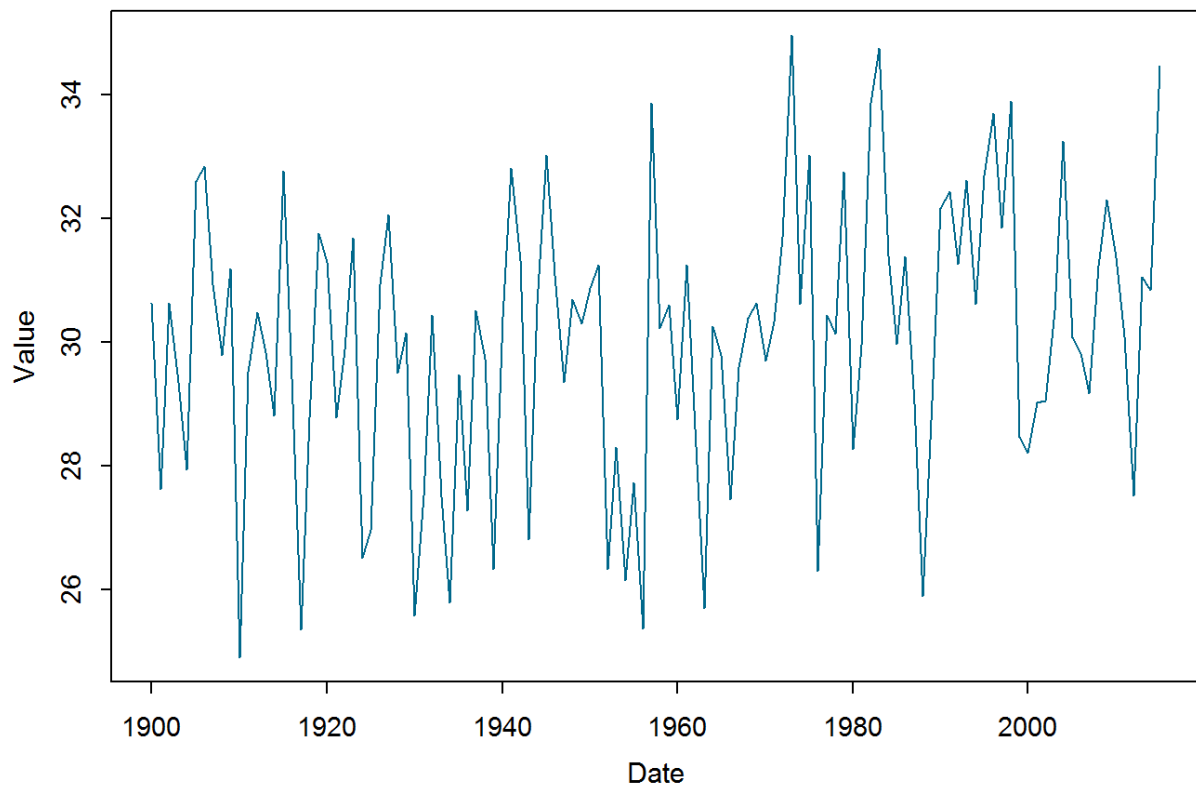
The boxplot also reveals the presence of an outlier with a value less than 25. We can find the year corresponding to this outlier using R's native **subset** function:

```
subset(precipdf, Value == min(precipdf$Value))
```

```
##      Date Value  
## 11 1910 24.91
```

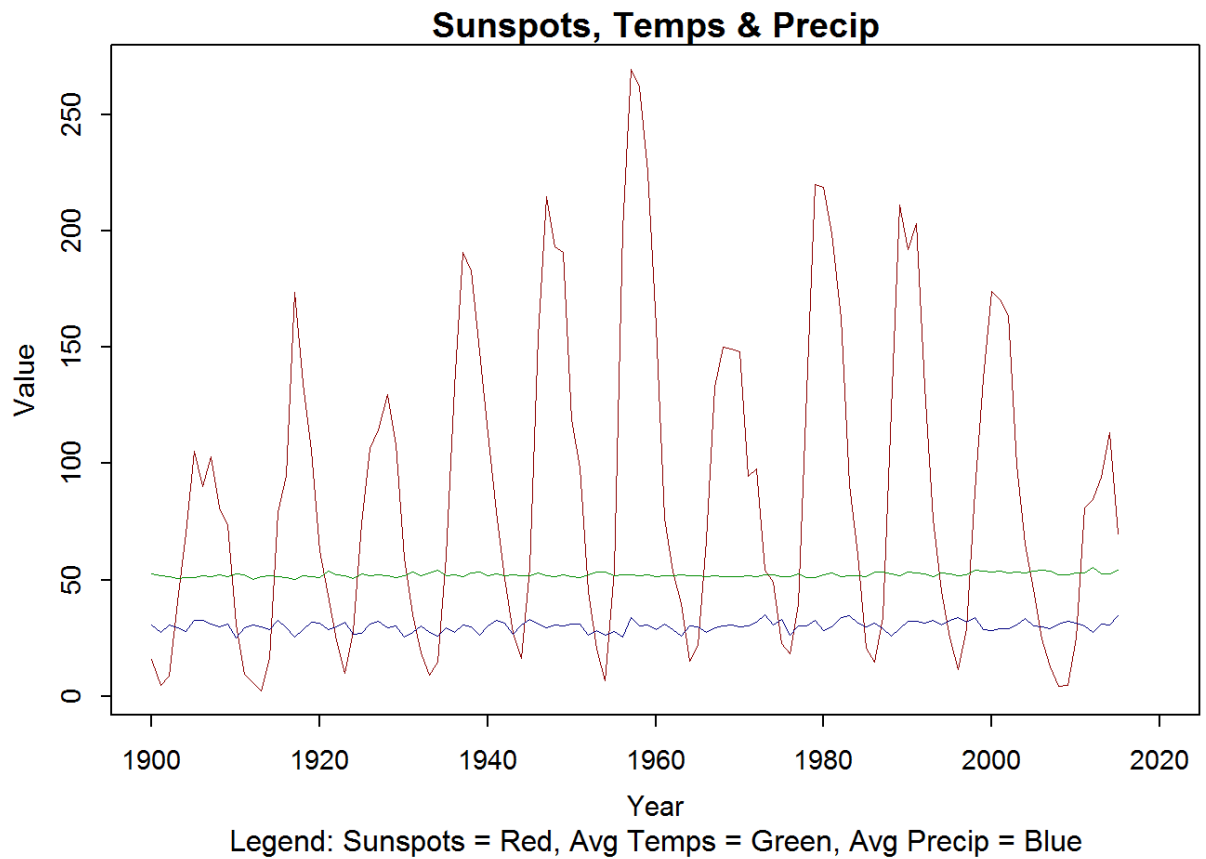
Our outlier average precipitation reading is for the year 1910.

Plotting our precipitation data according to its time series reveals no obvious trend or clear periodicity:



Co-Plotting of Explanatory and Response Variables

Plotting the raw values of all three variables together yields no obviously useful information since they are all quantifying very different metrics of very different magnitudes:



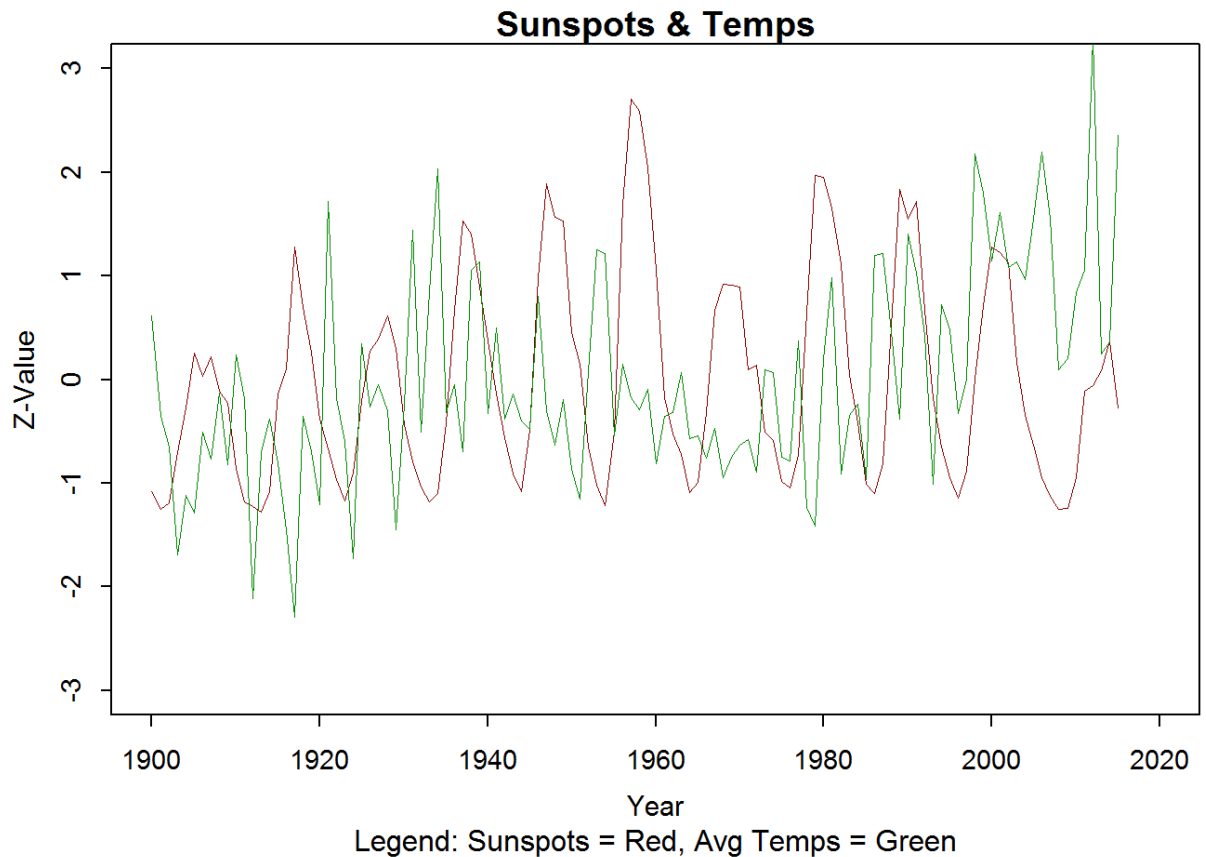
As we can clearly see, the sunspot data dominates the plot since it varies over a vastly wider range of possible values than does either the temperature or precipitation data.

This inherent variance in the raw data can be addressed by standardizing the values of each of our variables and then replotting them. We can standardize the values of each of our variables using the following formula:

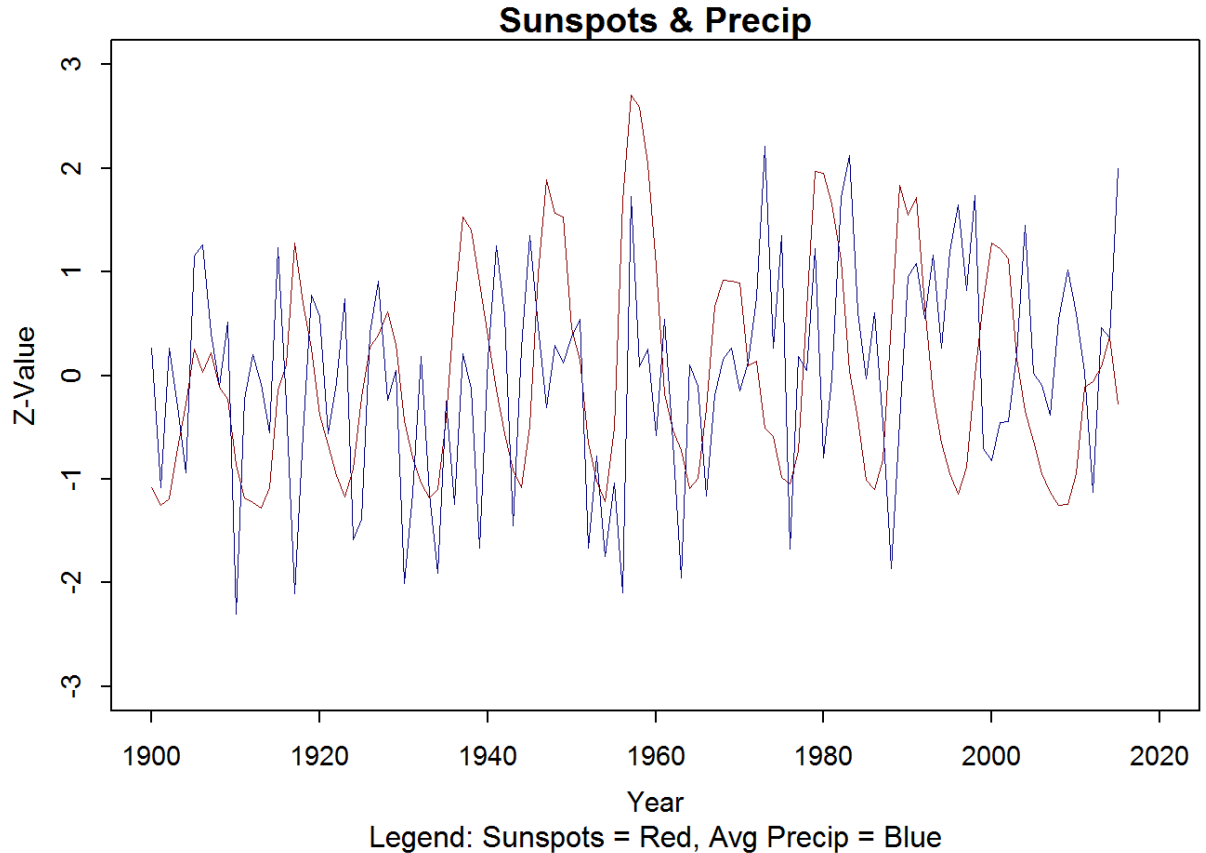
$$(value - mu)/stddev$$

where 'mu' is the mean value for our variable and 'stddev' is the standard deviation for the variable. Applying this formula converts each of our data values to a Z-score, and once we have Z-score equivalents of our data we can more readily co-plot the respective time series data for our three variables.

For our standardized plots, instead of plotting all three variables together we will create two separate plots, each of which will plot our explanatory variable (sunspot count) against one of the response variables (temperature, precipitation).



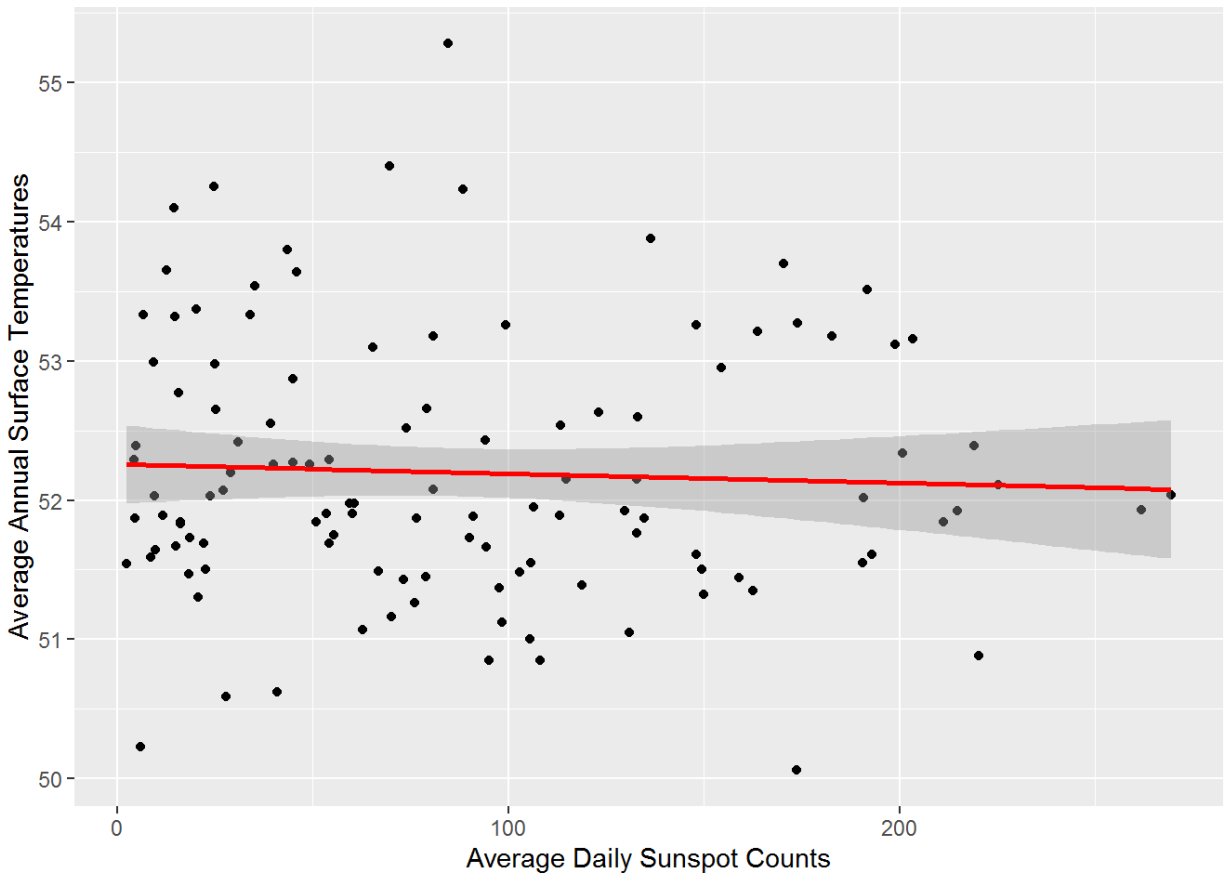
Our plot of the standardized sunspot and temperature time series data do not allow us to draw any definitive conclusion as to whether average sunspot counts might be predictive of average surface temperatures within the continental USA. In some instances, spike or dips in sunspot counts appear to coincide with spikes or dips in average temperatures, and in some instances they do not.



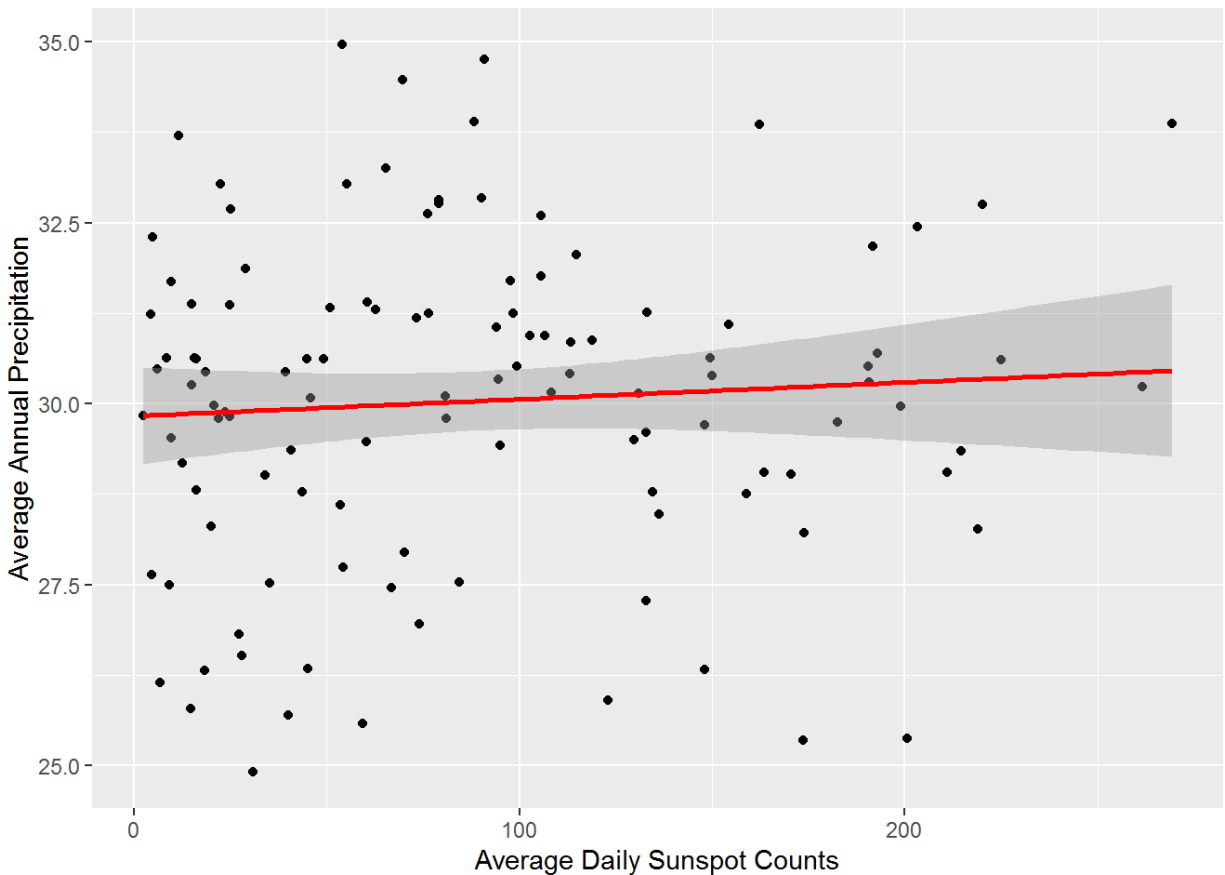
Similarly, our plot of the standardized sunspot and precipitation time series data does not allow us to draw any definitive conclusion as to whether average sunspot counts might be predictive of average precipitation amounts within the continental USA. As with the temperature data, in some instances spikes or dips in sunspot counts appear to coincide with spikes or dips in precipitation amounts, and in some instances they do not.

Scatterplots With a Least Squares Line Added

As an alternative, we can fit a model to each of our explanatory / response variable pairs, generate a scatter plot for each pair, and add a line to those plots based on the slope and intercept of the linear model that predicts values for our response variable for any given value of the explanatory variable.



As we can see, fitting a line to our temperature and sunspot data shows a slight negative correlation between the two variables. The raw scatterplot without the line shows no obvious correlation between the two variables.



Fitting a line to our precipitation and sunspot data shows a slight positive correlation between the two variables. The raw scatterplot without the line shows no obvious correlation between the two variables.

We can use R's native **cor.test** function to calculate the correlation coefficients for our two pairs of explanatory / response variables:

Correlation Coefficient - Average Sunspots vs. Average Temperatures

```
##
## Pearson's product-moment correlation
##
## data:  ssdf$AvgSpots and ssdf$AvgTemp
## t = -0.51352, df = 114, p-value = 0.6086
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2283567  0.1354629
## sample estimates:
##           cor
## -0.0480402
```

Correlation Coefficient - Average Sunspots vs. Average Precipitation

```
##
## Pearson's product-moment correlation
##
```

```
## data:  ssdf$AvgSpots and ssdf$AvgPrecip
## t = 0.74973, df = 114, p-value = 0.455
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1137231  0.2491803
## sample estimates:
##           cor
## 0.07004583
```

The correlation coefficients for both are close to zero, indicating that very little, if any, of the variation we see in our average temperature and average precipitation data is explained by the average number of sunspots observed each year. We will explore these hypotheses in greater detail in the following section.

Inference

During our exploratory data analysis our scatter plots suggested that very little of the variability in either the average temperature readings or the average precipitation amounts within any given year can be explained by the average number of sunspots observed during that same year. Making use of that insight we can propose two hypotheses:

Hypothesis Statements

Sunspot Counts and Average Temperature Readings

A hypothesis test for whether average sunspot counts are a valid predictor of average temperature readings can be stated as:

H0: $\beta = 0$: The true linear model of the relationship between average sunspot counts and average temperature readings has a slope of zero.

HA: $\beta \neq 0$: The true linear model of the relationship between average sunspot counts and average temperature readings has a slope of other than zero.

Sunspot Counts and Average Precipitation Amounts

A hypothesis test for whether average sunspot counts are a valid predictor of average precipitation amounts can be stated as:

H0: $\beta = 0$: The true linear model of the relationship between average sunspot counts and average precipitation amounts has a slope of zero.

HA: $\beta \neq 0$: The true linear model of the relationship between average sunspot counts and average precipitation amounts has a slope other than zero.

Theoretical Hypothesis Testing: Least Squares Line Fitting

We can attempt to test each of these hypotheses using linear regression. In fact, we've already accumulated the results of least squares line fitting for both of these hypotheses via our use of R's native **lm** function for purposes of plotting our "best fit" lines within the scatter plots we produced in Part 3. Therefore, we'll review those results and then examine the residuals to see whether the conditions for least squares line fitting have in fact been satisfied. If the conditions for the use of least squares line fitting have not been met, the results of our least squares line fitting cannot be relied upon.

Sunspot Counts vs. Average Temperature Readings

The results of our least square line fitting for sunspots vs. average temperatures are:

```
summary(fit1)

##
## Call:
## lm(formula = AvgTemp ~ AvgSpots, data = ssdf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.0820 -0.6181 -0.2171  0.5522  3.0784
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  52.2582049   0.1441493  362.528   <2e-16 ***
## AvgSpots     -0.0006694   0.0013035   -0.514    0.609
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9358 on 114 degrees of freedom
## Multiple R-squared:  0.002308,    Adjusted R-squared:  -0.006444
## F-statistic: 0.2637 on 1 and 114 DF,  p-value: 0.6086
```

From the above, the characteristic equation for the estimated value of average temperature readings given an average sunspot count is:

$$\hat{y} = 52.2582 + -0.0006694 * \text{AverageSunspotCount}$$

and, given 114 degrees of freedom, our 95% confidence interval is:

$$-0.0006694 + / - (1.981 * 0.0013035)$$

The 1.981 coefficient given here was obtained by looking up the critical value of the t-distribution for a two-tailed test at a .05 confidence level for 114 degrees of freedom.

We can calculate the bounds for the confidence interval as follows:

```
# NOTE: See slide 54 of Chap 7 slide deck for this formula
lower <- -0.0006694 - 1.981 * 0.0013035
upper <- -0.0006694 + 1.981 * 0.0013035
c(lower, upper)
```

```
## [1] -0.003251634 0.001912833
```

As we can see, the confidence interval includes the test value for our null hypothesis, which is zero. Furthermore, the very small average sunspot coefficient coupled with the relatively large p-value of 0.609 indicates that our least squares modeling appears to have found very little relationship between average sunspot counts and average temperature readings. The very small R-Squared value of 0.002308 provides confirmation of this.

In fact, assuming the 95% confidence interval stated above and based on the calculated p-Value of 0.609, we fail to reject the null hypothesis and conclude that we do not have sufficient evidence to show that the true linear model representing the relationship between average sunspot counts and average temperature readings has a non-zero slope.

Sunspot Counts vs. Average Precipitation Amounts

The results of our least square line fitting for sunspots vs. average precipitation amounts are:

```
summary(fit2)
```

```
##
## Call:
## lm(formula = AvgPrecip ~ AvgSpots, data = ssdf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.9919 -1.2918  0.1773  1.3864  5.0042
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.830003   0.343099   86.94  <2e-16 ***
## AvgSpots     0.002326   0.003102    0.75   0.455
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.227 on 114 degrees of freedom
## Multiple R-squared:  0.004906, Adjusted R-squared: -0.003822
## F-statistic: 0.5621 on 1 and 114 DF, p-value: 0.455
```

From the above, the characteristic equation for the estimated value of average precipitation amounts given an average sunspot count is:

$$\hat{y} = 29.830003 + 0.002326 * \text{AverageSunspotCount}$$

and, given 114 degrees of freedom, our 95% confidence interval is:

$$0.002326 + / - (1.981 * 0.003102)$$

The 1.981 coefficient given here was obtained by looking up the critical value of the t-distribution for a two-tailed test at a .05 confidence level for 114 degrees of freedom.

We can calculate the bounds for the confidence interval as follows:

```
# NOTE: See slide 54 of Chap 7 slide deck for this formula
lower <- 0.002326 - 1.981 * 0.003102
upper <- 0.002326 + 1.981 * 0.003102
c(lower, upper)
```

```
## [1] -0.003819062 0.008471062
```

As we can see, the confidence interval includes the test value for our null hypothesis, which is zero. Furthermore, the very small average sunspot coefficient coupled with the relatively large p-value of 0.455 indicates that our least squares modeling appears to have found very little relationship between average sunspot counts and average temperature readings. The very small R-Squared value of 0.004906 provides confirmation of this.

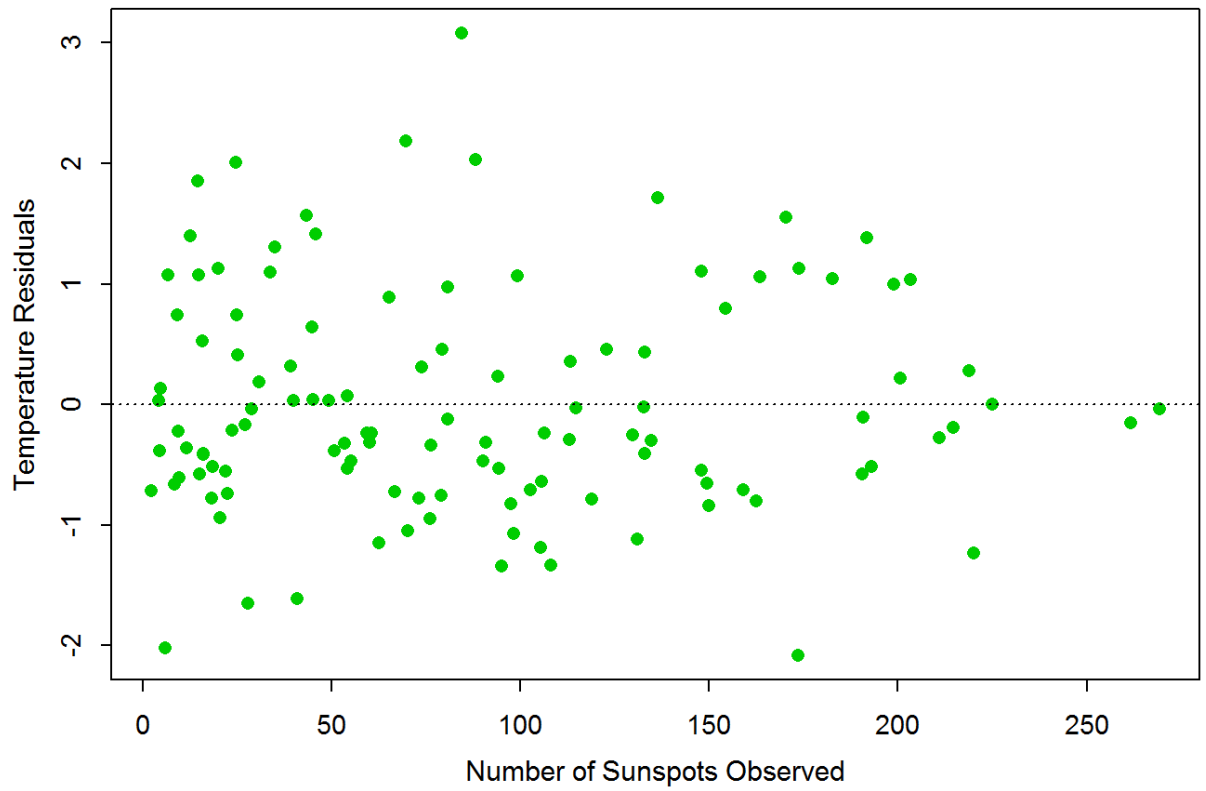
In fact, assuming the 95% confidence interval stated above and based on the calculated p-Value of 0.455, we fail to reject the null hypothesis and conclude that we do not have sufficient evidence to show that the true linear model representing the relationship between average sunspot counts and average precipitation readings has a non-zero slope.

Linear Regression Model Diagnostics

Sunspot Counts vs. Average Temperature Readings

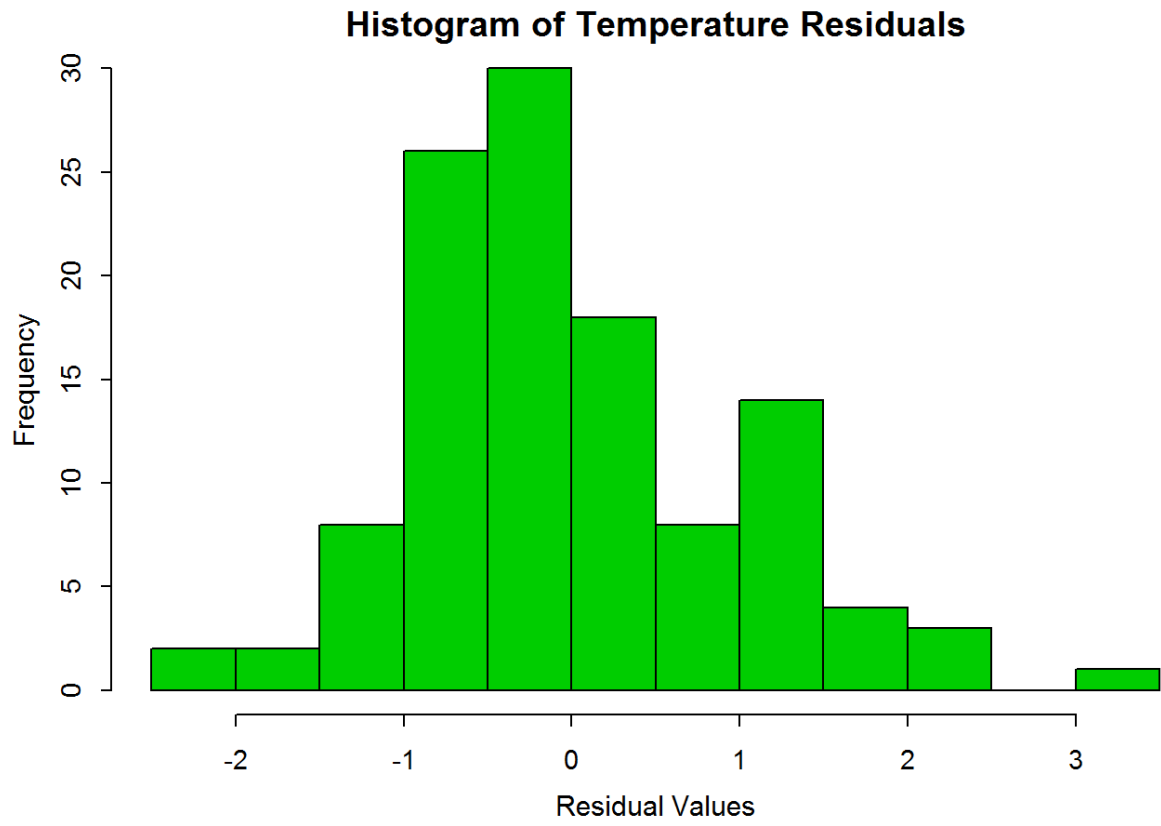
The conditions required for least squares line fitting are:

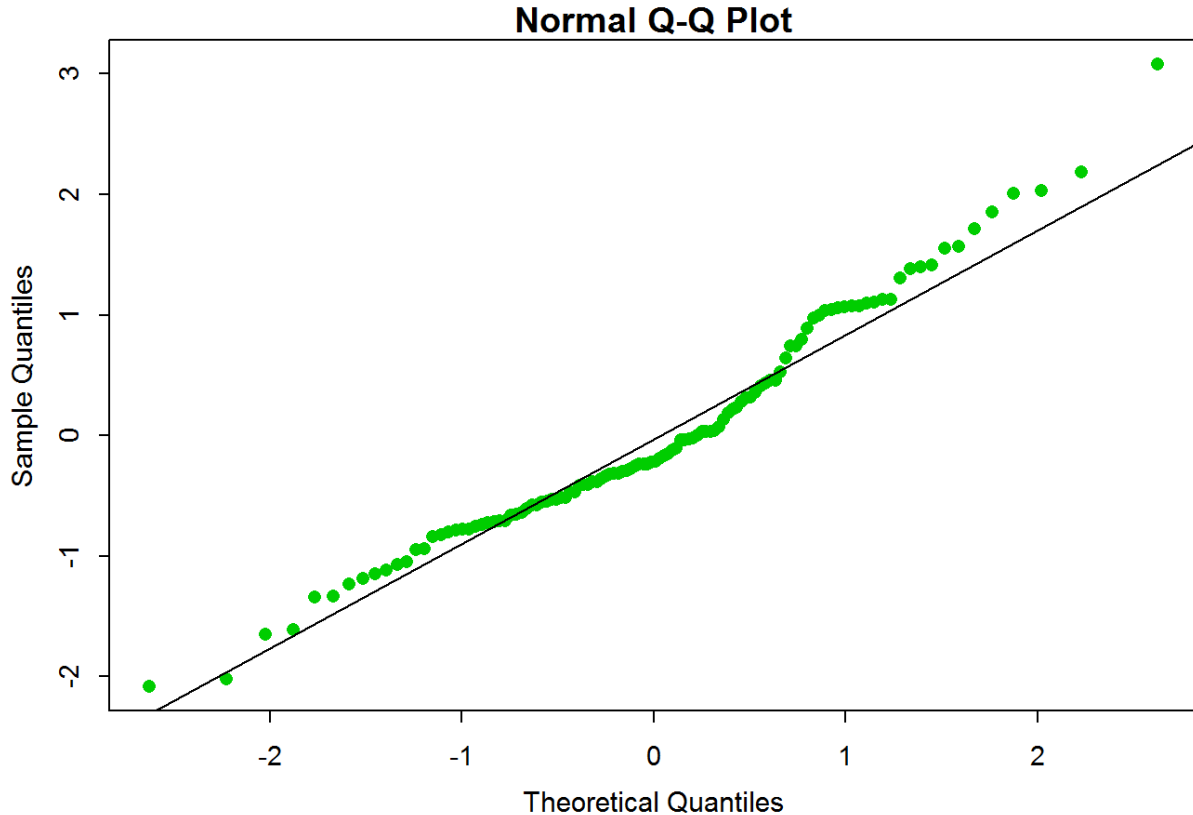
1) Linearity: To check for linearity We can plot the residuals of our least squares line fitting model for the temperature data against our average sunspot counts:



The plot shows no obvious patterns, which suggests that it is reasonable to try to fit a linear model to the data.

2) Nearly normal residuals: To check for the normality of the residuals of the temperature data we can examine a both a histogram of the residuals and a quantile plot:





The histogram shows that the distribution of the residuals is skewed to the right. The quantile plot indicates the presence of an outlier in the upper righthand corner as well as a fair amount of variation from the normal line beyond the first theoretical quantile. Each of these factors prevents the residual data from having a nearly normal distribution. Without a nearly normal distribution of the residuals, it is not reasonable to try to fit a linear model to the sunspot/temperature data.

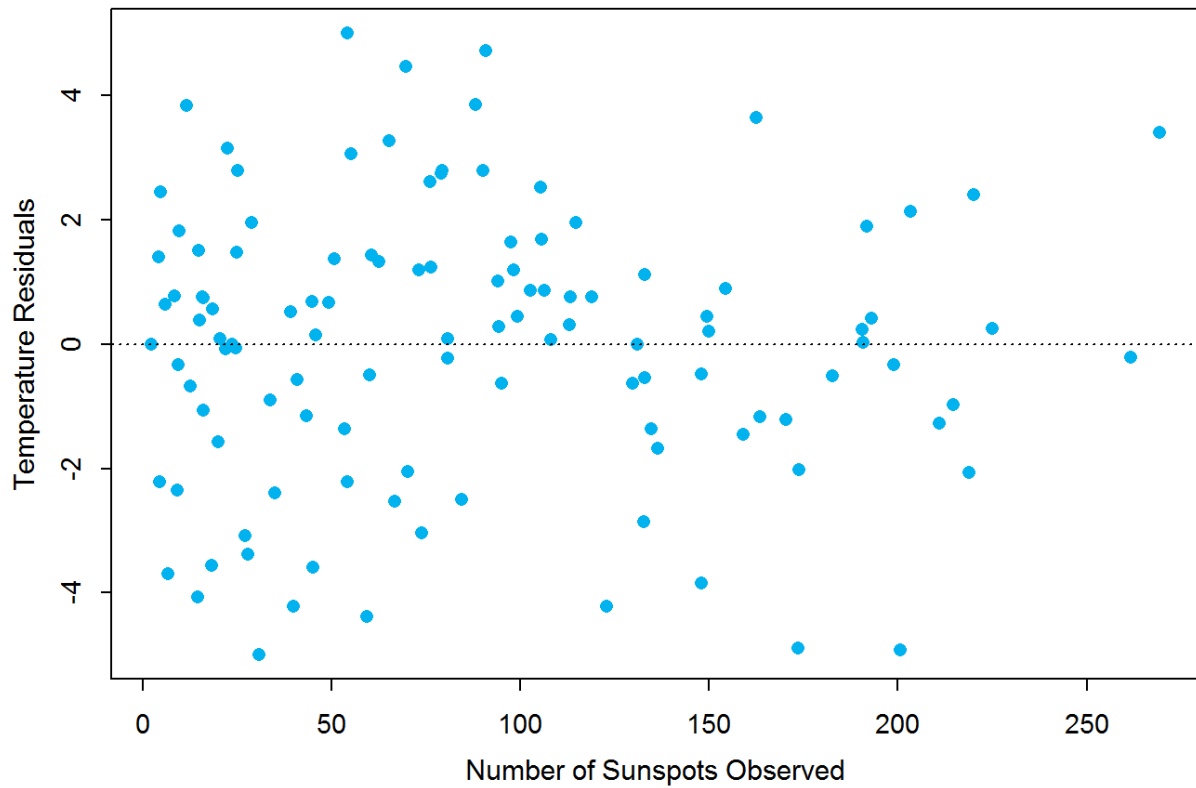
3) Constant variability: The residual plot of the temperature data shows that we do not have constant variability since larger average sunspot counts appear to yield less variability in average temperature readings than do smaller average sunspot counts. However, the lack of variability at higher average sunspot counts may simply be an artifact of the relative paucity of such high average counts within our data. In other words, there are relatively few years where average sunspot counts exceeded 200 within our 116 years of data. As such, we should expect to see less variability in our residual plots simply due to the fact that there are relatively few such data points within our data set. Nevertheless, this lack of variability relative to high average sunspot counts fails to satisfy the need for constant variability.

Since two of the three conditions required for linear regression have not been met, we should not use a linear model for modeling of the relationship between average sunspot counts and average temperature readings. **Therefore, we will not be able to rely upon the output of this least squares line fitting for purposes of inferring a relationship between the sunspot and temperature data.**

Sunspot Counts vs. Average Precipitation Amounts

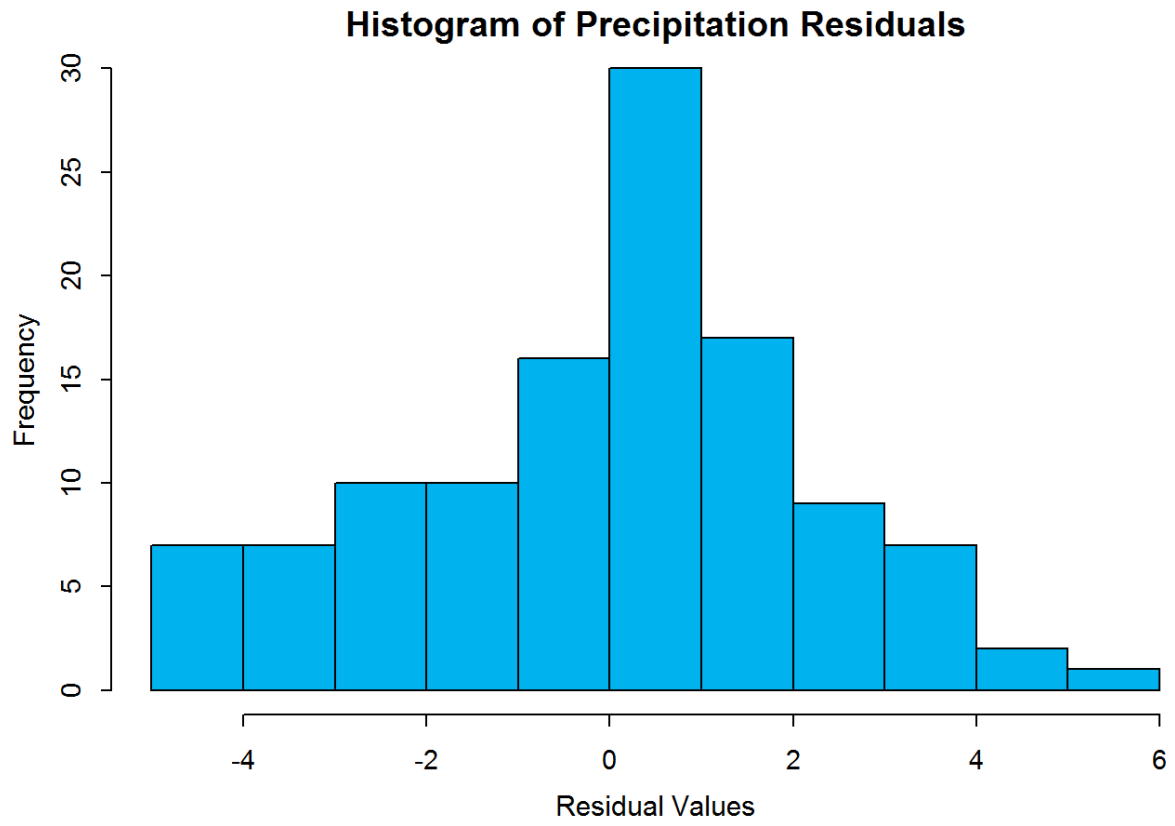
The conditions required for least squares line fitting are:

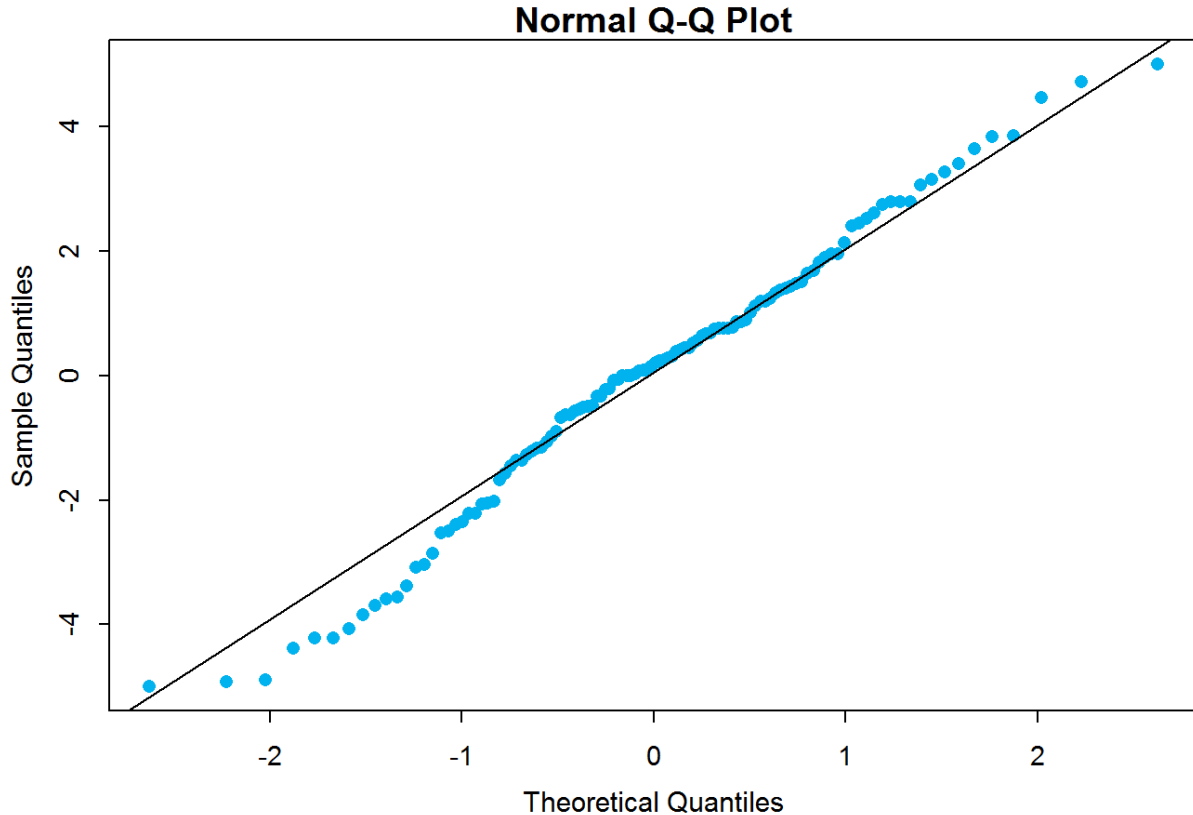
1) Linearity: To check for linearity We can plot the residuals of our least squares line fitting model for the precipitation data against our average sunspot counts:



The plot shows no obvious patterns, which suggests that it is reasonable to try to fit a linear model to the data.

2) Nearly normal residuals: To check for the normality of the residuals of the precipitation data we can examine both a histogram of the residuals and a quantile plot:





The histogram shows that the distribution of the residuals is slightly skewed to the left, but overall the distribution appears to be nearly normal. The quantile plot also shows evidence of that left skew from the normal line below the (-1) theoretical quantile. Despite the slight left skew the distribution does appear to be nearly normal. That suggests that it should be reasonable to fit a linear model to the data.

3) Constant variability: The residual plot of the precipitation data shows that we do appear to have constant variability since larger average sunspot counts appear to yield roughly the same amount of variability in average precipitation amounts as do smaller average sunspot counts.

Since each of the three conditions required for linear regression appear to have been met, use of a linear model for modeling of the relationship between average sunspot counts and average precipitation readings appears to be reasonable. **Therefore, we should be able to rely upon the output of this least squares line fitting for purposes of inferring a relationship between the sunspot and precipitation data.**

Summary of Hypothesis Testing and Inference

Our hypothesis tests have confirmed the observations we made during our exploratory data analysis: there appears to be very little, if any, relationship between average sunspot counts for a given calendar year and the average surface temperatures and precipitation amounts within the continental United States for the years 1900 - 2015. Additionally, since we have a lack of normality and a lack of constant variability in the residuals of our average temperature data relative to the average sunspot counts, we should not rely on linear modeling to assess the relationship between average sunspot counts and average surface temperature readings within the continental United States.

However, for the average precipitation data we do appear to have satisfied each of the three conditions required for linear least squares modeling. As such, we should be able to rely upon linear modeling for assessing the relationship between average sunspot counts and average precipitation measurements within the continental United States.

Conclusion

Throughout this study we've tried to determine whether or not average sunspot counts for the years 1900 - 2015 were predictive of average surface temperatures and average precipitation amounts within the continental United States for that same period using sunspot data from the Royal Observatory of Belgium and climate data from the National Oceanic and Atmospheric Administration (NOAA). Both our exploratory data analysis and our hypothesis testing yielded no evidence of a tangible predictive aspect of average sunspot counts relative to either the surface temperature data or the precipitation data for the continental United States for the years 1900 - 2015.

A primary reason for undertaking this study was to examine one of the many possible natural influencers of earth's climate. One of the most significant challenges we encountered after starting this study was the inherent limitations of the time series data we have available to us for such research. Humankind has regularly recorded surface temperature and precipitation measurement for a very brief period of geologic time (roughly 150 years or so), and sunspots for only a slightly longer period. There are, of course, estimates of earlier surface temperatures derived from scientific studies of things such as tree rings, archaeological samples, and glacial ice. However, such estimates are obviously not the results of firsthand observations and as such may be influenced by many hard-to-quantify potential confounding variables.

Furthermore, observable sunspot counts ebb and wane based on the sun's own internal stellar dynamics, and some of these dynamics appear to follow cycles that are between hundreds of years and even tens of thousands of years in duration, vastly longer than our 116 years worth of data. The same can be said of other variables that are known to influence earth's climate, such as the wobble of the earth's axis and the cycle of natural variations in the earth's orbit around the sun. As such, it seems highly unlikely from an intuitive standpoint that any results we've achieved herein might be applicable beyond our admittedly limited data set.

Our data collection efforts for this study also brought to our attention recently uncovered evidence of significant issues regarding the validity of the NOAA's surface temperature data for the continental United States (See Reference # 4). It is possible that revising the NOAA data to remove the biases it has accumulated as a result of the large number of compromised data collection sites would change the results of our analysis, and this may warrant further investigation.

Other possible avenues for future research related to our work here would include:

- Revisiting our existing data to see if there is any predictive aspect of average sunspot counts as it relates to the magnitude of change in either average surface temperatures or average precipitation amounts;
- Revisiting our existing data to see if there is any latency between increases in average sunspot counts and changes (either increases or decreases) in average surface temperatures or average precipitation amounts.

References:

General Sunspot Facts

1. Geerts & Linacre, <http://www-das.uwyo.edu/~geerts/cwx/notes/chap02/sunspots.html>

Glacial Meltoff

2. <http://noc.ac.uk/news/global-sea-level-rise-end-last-ice-age>
3. http://www.giss.nasa.gov/research/briefs/gornitz_09/

Inaccuracy of 808 out of 1,218 NOAA monitoring sites:

4. <https://fallmeeting.agu.org/2015/files/2015/12/Press-Release-NEW-STUDY-OF-NOAA-USHCN.pdf>
5. <http://dailycaller.com/2015/12/17/exclusive-noaa-relies-on-compromised-thermometers-that-inflate-u-s-warming-trend/#ixzz3uft0kzy4>
6. <http://wattsupwiththat.com/2015/12/17/press-release-agu15-the-quality-of-temperature-station-siting-matters-for-tempera>

Sunspot Data

As noted above, the sunspot data was provided by The Royal Observatory of Belgium. Their website is:

<http://www.sidc.be/silso/home>

The proper citation for the Royal Observatory of Belgium and the data contained therein is:

Source: WDC-SILSO, Royal Observatory of Belgium, Brussels.

Surface Temperature and Precipitation Data

As noted above, surface temperature and precipitation data have been obtained via the National Oceanic and Atmospheric Administration, which is a branch of the United States Department of Commerce. Their website is:

<http://www.ncdc.noaa.gov/>

The data used in this analysis is accessible via this specific web page:

<http://www.ncdc.noaa.gov/cag/>