ENVIRONMENTAL TENDENCIES OF SALT LAKE CITY

CORBIN APPLE, BRIDGET HYLAND, BEN STERLING

1. Introduction. It has been well established by The United States Environmental Protection Agency and other independent organizations that average temperature is increasing across the country. This study examines monthly temperature data from two stations: Rye Patch Dam, Nevada, and Salt Lake City International Airport, Utah. These stations were chosen because they are roughly at the same latitude (40.498°N and 40.790°N), longitude (118.316°W and 111.980°W), and elevation (1260.3m and 1287.8m, for Rye Patch Dam and Salt Lake City respectively). Our data is from the National Centers for Environmental Information [2], which contains monthly data about major meteorological parameters at many locations across the country. We chose Salt Lake City and Rye Patch Dam because of their similar geographical characteristics. Data for Rye Patch Dam begins in 1935, but data for Salt Lake City only reaches back to 1948, so we used only years from 1948 to 2020 in our analysis. The dataset includes many parameters, but of particular interest to us were average temperature, average precipitation, number of days with thunderstorms, and total minutes of sunshine.

The goal of our first hypothesis is to examine whether the effects of climate change statistically differ between these two locations; if they do, we may be able to infer that the climate change is predominantly man-made. We accomplished this by calculating the monthly temperature anomaly at each location and comparing the mean temperature anomalies of the two locations. The goal of our second hypothesis is to determine predictors of temperature. We use a multiple linear regression with temperature as the response variable and precipitation, days with thunderstorms, and minutes of sunshine as the dependent variables.

2. First Hypothesis. Climate change is a well-documented phenomenon: on average, the global temperature is increasing. In this section, we test the hypothesis that temperature increases in Salt Lake City and Rye Patch Dam are not equal.

Temperature anomaly is used to compare change in temperature over time. A mean temperature is calculated over a long period of time, and this long-term mean is subtracted from more recent observations. For example, it is known that the global average temperature was 13.9°C between 1901 and 2000. If the global average temperature in 2015 was 14.5°C, then the temperature anomaly for that year would be $14.5 - 13.9 = 0.6$°C. When analyzing temperature data over many years, it is advisable to use a temperature anomaly rather than absolute temperature because it eliminates seasonal variation within the year, allowing for a more significant result. This is an accepted and widely used method in climate analysis [3].

The GSOM records the monthly average temperature at each location, so we determined temperature anomaly monthly. We calculated a January anomaly by averaging the January temperatures of each year from 1948 to 1974. Then, we subtracted this long-term mean from each January temperature from 1975 to 2020. We did the same for the other eleven months and for both locations, yielding different anomalies for each. In our data files, this is column CG, labeled TAVG ADJ. We used this anomaly in our comparison in place of the absolute temperature. It served to reduce the variance of each sample to produce a meaningful result.

Because of the large number of observations in each sample, we invoked the central limit theorem to approximate the distribution as normal. The Shapiro-Wilk test for

normality gives a p-value of .01663 for the Salt Lake City sample and 0.0003392 for the Rye Patch Dam sample, which verifies that the normal approximation is appropriate ($P < 0.05$). The linear nature of the plots further supports our approximation.

Formally, our hypotheses are:

$$H_0 : \mu_{SLC} = \mu_{RPD} \ \ vs. \ \ H_a : \mu_{SLC} \neq \mu_{RPD}.$$

Because the two stations have similar geographical characteristics (i.e. latitude, longitude, and elevation), we determined that this is an observational matched pairs analysis and used the paired t-test at $\alpha = 0.05$ accordingly. As such, it was not necessary to determine equality of variance between the two samples. We determined the test statistic to be $t = -16.46$. This is less than the critical value $-t_{n-1,\alpha/2} = -1.648$ (where n = 487). Also, the p-value is 2.2×10^{-16}, which is less than the significance level $\alpha = 0.05$. Based on these results, we reject H_0 and conclude that the mean temperature anomaly at Salt Lake City is significantly greater than the mean temperature anomaly at Rye Patch Dam. In other words, since 1975, the temperature has increased more at Salt Lake City than at Rye Patch Dam. The reason for this is not known, but previous research suggests that Salt Lake City emits a relatively large amount of carbon pollution per capita, which hastens the effects of climate change there [5]. These results, and useful statistics, are summarized in Tables 1 and 2. We also present the Q-Q Plots in Figures 1 and 2 to visualize the data normality. The linear relationships suggest that both datasets are normal.

We then randomly selected 10% of the data from each station to remove from the dataset and treat as missing. Although our dataset contains ample missing values on its own, we removed more values in order to evalued the effects of missing data. Our analysis ignored months in which temperature data for Salt Lake City *or* Rye Patch Dam was missing. This resulted in a loss of total data points greater than 10%, leaving us with a new sample size (that is different for each run) for each station. Performing the Shapiro-Wilk test again, the distribution of the observations were still normal (from the p-value in Table 3 or Q-Q Plots in Figures 3 and 4). The new test statistic for a sample run was -15.03, which is less than the critical value $-t_{n-1,\alpha/2} = -1.648$ (where n = 390). The p-value was 2.2×10^{-16}, which is less than the significance level $\alpha = 0.05$ (note that the reported p-value is the same as before because the true value is below machine precision). These statistics are summarized in Tables 3 and 4. Therefore, we reject H_0 and conclude that the mean temperature anomaly at Salt Lake City is greater than the mean temperature anomaly at Rye Patch Dam. The addition of more missing values reduced the sample size and (theoretically) resulted in a higher p-value, but did not change the outcome of the hypothesis test.

3. Second Hypothesis. The second hypothesis analyzes which environmental factors are the best predicters of temperature for Salt Lake City. According to the United States Environmental Protection Agency (EPA), higher temperature results in either more or less precipation and higher frequency of storms[6]. This study performs a multiple linear regression of temperature against precipation, number of thunderstorms per month, and minutes of sunlight per month. In conjuction with the EPA article noted above, we also considered two key characteristics when deciding which parameters we would include in our study:

- Linear relationship with temperature
- Availability of data

For example, we did not include total monthly snowfall or number of days with fog because they have too much missing data despite having a decent linear relationship.

The parameter with the most significant linear relationship with temperature is, not suprisingly, minutes of sunlight. A major factor we had to consider here, however, was that sunlight data is only recorded from 1965 to 2004. Thus, in order to include this parameter in our analysis, we needed to restrict our study down to this range (roughly 479 total months, or observations). Additionally, we limited the data used in our study to months for which the data was available for all four parameters listed above. Therefore, when including temperature, precipation, and number of thunderstorms in our study, our sample size decreased from 479 months to 321 months because at least one of the these parameters was missing data in the 479 months of data avialable between 1965 and 2004.

For each of the parameters identified above, we examine its graph by temperature with a line of best fit. As expected, there is a strong positive correlation between minutes of sunlight and temperature, as seen in Figure 5. Similarly, albeit not as strong, there is also to be a positive correlation between number of thunderstorms and temperature, as seen in Figure 6. Lastly, the correlation between precipation and temperature is negative, as seen in Figure 7, which is expected in some regions as stated by the EPA [6].

As an additional preliminarly analysis, in order to obtain a strong multiple linear regression model to estimate temperature, we sought to use parameters that are not highly correlated with each other, as including additional variables that are highly correlated is more likely to simply increase model complexity as opposed to imporve the model fit. Figure 8, which is a correlation matrix of the three independent parameters, suggests that the independent variables of sunlight, days of thunderstorm, and precipation are not highly correlated, further supporting our decision to include them as parameters in our model.

At this point, we conducted the analysis on the selected paramters. Per Table 5, we can see the the model has an adjusted R^2 of $0.7969 \approx 0.8$, meaning that roughly 80.0% of the variance in temperature can be explained by the three predictors. Using the coefficients from the table, we can estimate temperature as $\hat{T} = \beta_0 + X_1\beta_1 + X_2\beta_2 + X_3\beta_3$, where indices 1, 2, and 3 reference minutes of sun, days of thunderstorms, and precipitation, respectively. Substituting β values from Table 5,

$$\hat{T} \approx -5.26 + 1.04 \times 10^{-3} X_1 + 0.84 X_2 - 4.80 \times 10^{-2} X_3.$$

Now that we have completed out multiple linear regression analysis on the three selected parameters, we turn to a final model comparison analysis. Here, we conduct a best subset analysis to locate the best model at each model size, including 1, 2, and 3 parameters. The best subset selection process considers all $\binom{3}{k}$ possibilities where k is the number of parameters (k = 1, 2, 3). For each k, the best subset method will select the model with the lowest SSE. As can been seen in Table 6, the best models for k = 1, 2, and 3 are minutes of sun, minutes of sun + days of thunderstorms, and minutes of sun + days of thunderstorms + precipation, respectively.

Now that we have our three best models at each k, we will calculate the following metrics for each model:

- Adjusted R^2
- AIC
- BIC

The adjusted R^2 will measure what percentage of the variance in the model is explained by the parameters; this measure penalizes model complexity relative to R^2. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are both penalized-likelihood criteria, where the BIC penalizes model complexity a

bit more than AIC. Overall, the best model of the three from Table 6 will have the greatest adjusted R^2 and the lowest AIC and BIC values. Define:

- Model 1: Performs regression against minutes of sunlight
- Model 2: Performs regression against minutes of sunlight + days of thunderstorm
- Model 3: Performs regression against minutes of sunlight + days of thunderstorm + precipitation

Table 7 contains the results. The full model, model 3, is the best overall model to choose to estimate temperature because it has the highest Adjusted R^2 and lowest AIC and BIC values.

To demonstrate our results, let us consider the following example to predict the average monthly temperature of a new observation Y^*. Notationally, the following formula enables the calculation of an $(1-\alpha)$-level prediction interval for Y^*.

$$Y^* \pm \hat{Y}^* + t_{n-(k+1),\alpha/2} s \sqrt{1 + \mathbf{x}^{*T}\mathbf{V}\mathbf{x}^*}$$

where $\hat{Y}^*$ is the estimate of Y^*, $t_{n-(k+1),\alpha/2}$ is the critical value, n is the number of observations (n = 321), k is the number of predictors (k = 3), s is the estimate for σ, equivalent to $\sqrt{MSE}$, $\mathbf{x}^*$ is the predictor variables for the three parameters, and $\mathbf{V} = (\mathbf{X}^T\mathbf{X})^{-1}$, with $\mathbf{X}$ representing the matrix of the values of the predictor variables [4].

Now, given a new month with 15,000 minutes of sunlight, 4 days with thunderstorms, and 35 millimeters of total precipation, we can calculcate a 95% prediction interval, as calculated below. To avoid extrapolation, the parameter figures above are all within the ranges of values observed for each parameter in the 321 oberservations analyzed. Utilizing the prediction interval with 95% confidence, the average monthly temperature of this new month would be between 4.48°C and 19.63°C (with $\hat{Y}^* = 12.06$°C).

To investigate the effects of missing data on the multiple linear regression, we removed 5% of temperature observations, 5% of sunlight observations, 5% of thunderstorm observations, and 5% of preceipitation observations. We then ignored any data point for which at least one of these four observations was missing, resulting in a smaller sample size. Because of the random nature of the removal of data, the new sample size is approximately 80% of the original sample size for each run. The new adjusted R^2 is $0.797 \approx 0.8$, indicating that roughly 80% of the variation in temperature can be explained by the three predictors. This is not appreciably different from the adjusted R^2 for the full data set from 1965 to 2004. The new estimate for temperature is

$$\hat{T} \approx -4.79 + 1.03 \times 10^{-3} X_1 + 0.85 X_2 - 6.30 \times 10^{-2} X_3.$$

See Table 8 for the β values and p-values for all three predictor variables in the missing data case.

4. Missing Data. When data are missing at random, the value of the missing data is not related to the reason for its missingness. In this case, the complete data points are essentially a random sample of the entire dataset, so missing points can be ignored in the analysis. We evaluated the effects of data missing at random on our hypothesis test and multiple linear regression in sections 2 and 3 respectively.

Missing data is non-ignorable if the value of the missing data is related to the reason for its missingness. For example, a storm could cause data collection equipment

to become inoperable, and the storm data would not be collected. In this case, it is not possible to ignore the missing data in the analysis because the complete data is no longer a random sample, and therefore does not represent the whole of the data. The Heckman correction is a method that can correct bias from this non-random sample[1]. In the context of regression, suppose we want to fit the linear model

$$Y_i = X_i\beta + \epsilon$$

to our data. Heckman's model lets us obtain unbiased estimators of β. First, we introduce the selection equation $P(R_{yi} = 1|X_i^s) = \Phi(X_i^s\beta^s)$, where R_{yi} is an indicator variable equal to 1 if Y_i is observed and 0 if Y_i is missing. We use maximum likelihood estimation, a method of estimating parameters by maximizing the likelihood function, to find estimates for $\hat{\beta}^s$. Using these estimates, we can compute the inverse Mills ratio $\hat{\lambda}_i = \frac{\phi(X_i^s\hat{\beta}^s)}{\Phi(X_i^s\hat{\beta}^s)}$ for each observation in the sample, where ϕ is the standard normal density and Φ is the standard normal distribution function. Finally, using the linear equation $Y_i = X_i\beta + \hat{\lambda}_i\beta_\lambda + \eta_i$, where $\eta \sim \mathcal{N}(0, \sigma_\eta^2)$, we can find estimators for $\hat{\beta}$ and $\hat{\beta}_\lambda$. There also exists a one-step procedure for estimating β which generates a lower standard error, but the two-step procedure is used for faster computation.

After obtaining our estimates for β, we can draw σ_η^{2*}, β^*, and β_λ^* from the linear equation. Also, we can draw η^* from $\mathcal{N}(0, \sigma_\eta^{2*})$. Finally, for each missing Y, we can impute

$$Y_i^* = X_i\beta^* + \frac{-\phi(\widehat{X_i^s\beta^s})}{1 - \Phi(\widehat{X_i^s\beta^s})}\beta_{\lambda i}^* + \eta^*.$$

5. R Code Compilation Instructions. To run the R scripts, it is imperative to do so from the /data subdirectory as both reference data from a relative path. The team ran the scripts from Terminal with the "Rscript" command instead of RStudio; they could possibly be run from RStudio, but we did not test it with this.

If not installed already, one can install the libraries used by the scripts by entering the following commands in Terminal:

- R
- >install.packages("ggpubr")
- >install.packages("ggplot2")
- >install.packages("latex2exp")
- >install.packages("car")
- >install.packages("corrplot")
- >install.packages("leaps")

One can run the scripts by executing the following commands in Terminal:

- cd <Base Directory>/AMS_572_Project/data
- Rscript Hypothesis_1.R
- Rscript Hypothesis_2.R

REFERENCES

[1] J. E. Galimard, S. Chevret, C. Protopopescu, and M. Resche-Rigon, *A multiple imputation approach for MNAR mechanisms compatible with Heckman's model*, Statistics in Medicine, 35(2016), pp. 2907-2920.

[2] J. H. Lawrimore, R. Ray, S. Applequist, B. Korzeniewski, and M. J. Menne, *Global summary of the month (gsom) dataset, version 1*, 2016, https://doi.org/10.7289/V5QV3JJ5 (accessed 2020-11-14).

[3] NASA, *The exploring the environment project*, http://ete.cet.edu/gcc/?/globaltemp_anomalies/.
[4] A. C. Tamhane and D. D. Dunlop, *Statistics and Data Analysis from Elementary to Intermediate*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2000.
[5] The White House, *The threat of carbon pollution: Utah*, https://obamawhitehouse.archives.gov/sites/default/files/docs/state-reports/climate/Utah%20Fact%20Sheet.pdf.
[6] United States Environmental Protection Agency, *What climate change means for utah*, https://19january2017snapshot.epa.gov/sites/production/files/2016-09/documents/climate-change-ut.pdf.

6. Appendix: Tables and Figures.

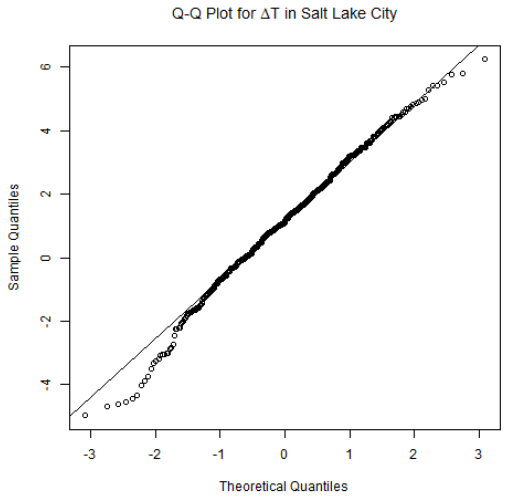

FIG. 1. *Q-Q Plot in Salt Lake City*

	$\Delta Temp$ Salt Lake City	$\Delta Temp$ Rye Patch
μ	1.1229	0.1317
σ	1.9765	1.9226
n	487	487
$P_{Shapiro}$	0.0166	0.0003

TABLE 1
Temperature Difference Statistics

Parameter	Value
t	-16.46
df	486
P	2.2×10^{-16}
Conf. Interval	$(-1.1095, -0.8728)$

TABLE 2
Paired t Test Results for Temperature Difference

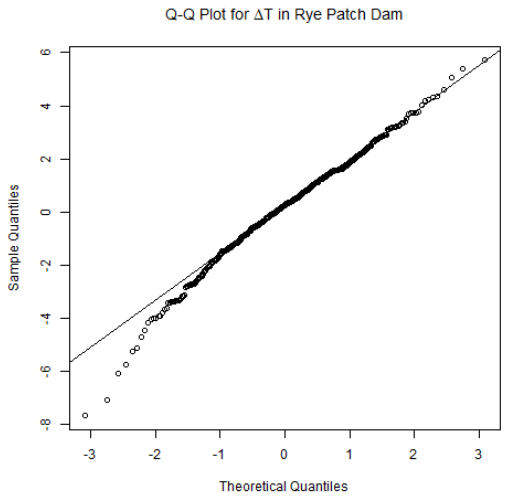

FIG. 2. *Q-Q Plot in Rye Patch*

	$\Delta Temp$ Salt Lake City	$\Delta Temp$ Rye Patch
μ	1.1282	0.1457
σ	1.9850	1.9051
n	390	390
$P_{Shapiro}$	0.0238	0.0005

TABLE 3
Temperature Difference Statistics for Missing Values

Parameter	Value
t	-15.03
df	389
P	2.2×10^{-16}
Conf. Interval	$(-1.1109, -0.8539)$

TABLE 4
Paired t Test Results for Temperature Difference for Missing Values

	Temperature	Intercept	Minutes of Sun	Days of Thunderstorms	Precipitation
$P_{Shapiro}$	2.662×10^{-8}	–	6.845×10^{-9}	9.438×10^{-15}	2.817×10^{-12}
β	–	-5.261	1.041×10^{-3}	0.8471	-4.800×10^{-2}
$P(>\|t\|)$	–	2.30×10^{-8}	$< 2.00 \times 10^{-16}$	$< 2.00 \times 10^{-16}$	$< 2.91 \times 10^{-7}$

TABLE 5
Multilinear Regression for Predicting Temperature (Adjusted $R^2 = 0.7969$, $n = 321$)

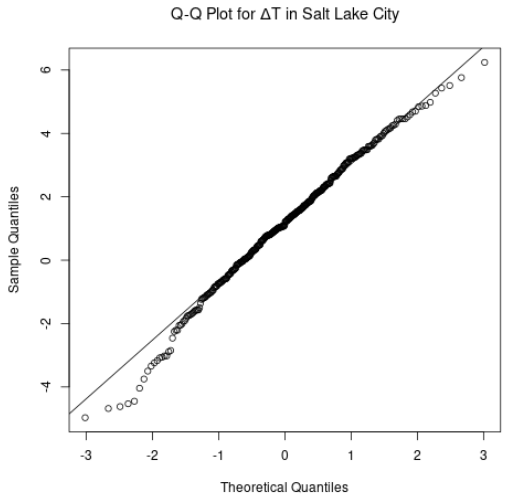

FIG. 3. *Q-Q Plot in Salt Lake City with 10% missing data*

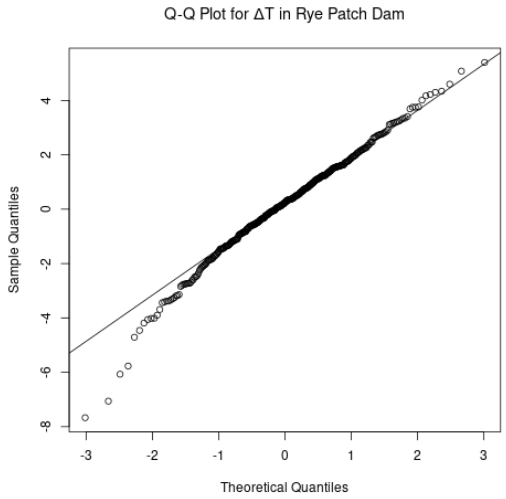

FIG. 4. *Q-Q Plot in Rye Patch with 10% missing data*

# Variables	Minutes of Sun	Days of Thunderstorms	Precipitation
1	*		
2	*	*	
3	*	*	*

TABLE 6
Optimal Subset selection for Multilinear Regression

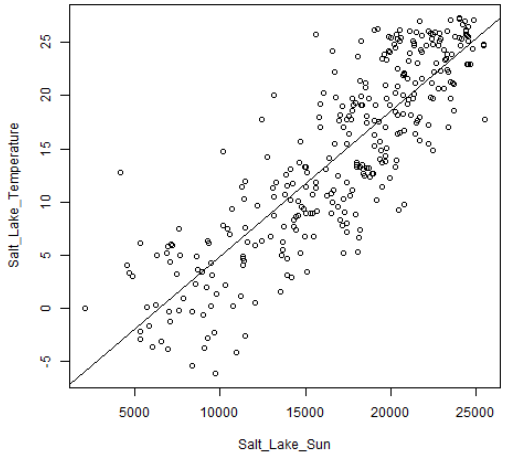

FIG. 5. *Salt Lake City Temperature versus Sun*

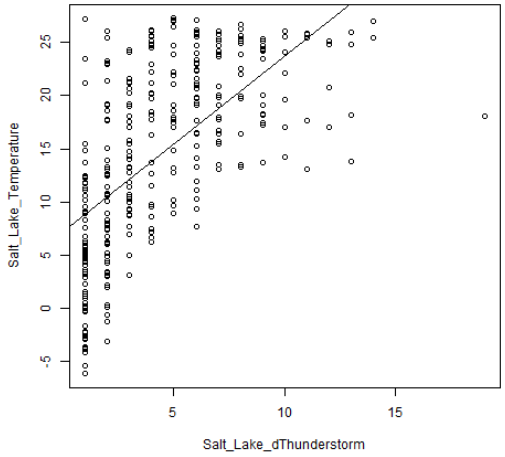

FIG. 6. *Salt Lake City Temperature versus Days of Thunderstorms*

	Model 1	Model 2	Model 3
AIC	1875.498	1805.823	1781.139
BIC	1886.812	1820.909	1799.996
Adjusted R^2	0.7258	0.78	0.7969

TABLE 7
Model Information for Multilinear Regression

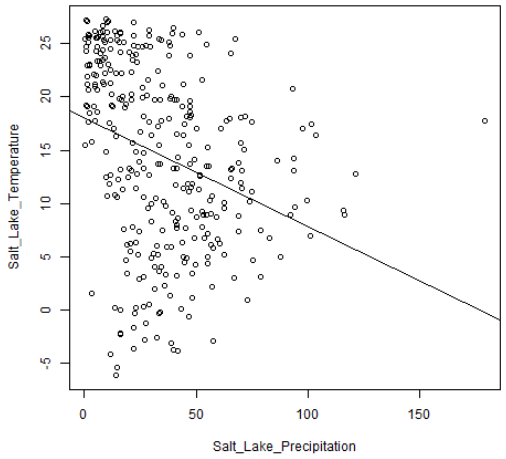

FIG. 7. *Salt Lake City Temperature versus Precipation*

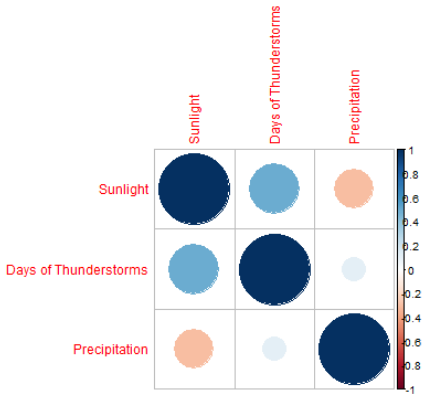

FIG. 8. *Correlation plot of chosen dependent variables*

	Temperature	Intercept	Minutes of Sun	Days of Thunderstorms	Precipitation
$P_{Shapiro}$	3.259×10^{-7}	–	2.592×10^{-7}	2.011×10^{-13}	1.88×10^{-8}
β	–	-4.795	1.037×10^{-3}	0.8541	-6.308×10^{-2}
$P(>\|t\|)$	–	4.85×10^{-6}	$< 2.00 \times 10^{-16}$	$< 2.00 \times 10^{-16}$	$< 1.94 \times 10^{-8}$

TABLE 8

Multilinear Regression for Predicting Temperature with Missing Data (Adjusted $R^2 = 0.797$, $n = 260$*)*

	Model 1	Model 2	Model 3
AIC	1530.985	1481.879	1451.769
BIC	1541.667	1496.121	1469.572
Adjusted R^2	0.7226	0.7713	0.797

TABLE 9
Model Information for Multilinear Regression with Missing Data