Impact of the variability in latitude and minimum recorded temperature on the variability of Energy (AET) in the United States Ponderosa Pine: A Statistical Analysis

[Author: Leah Nicholson | Course: BA240 Statistics, Bellevue College | Date: August 10th, 2024]

Image source: Matt Strieby (2021), https://www.conifers.org/pi/Pinus_ponderosa.php

The latitude of Ponderosa Pines in the United States and their minimum recorded temperature can explain the variation in the organism's stored energy (AET).

AET (Energy, Y Var)	Latitude (X1 Var)	Min Temp (X2 Var)
45.14	34.82	(9.00)
28.71	34.52	(47.00)
29.75	43.25	(134.00)
37.53	37.35	(128.00)
38.14	48.07	(14.00)
46.12	39.77	16.00
13.11	35.50	10.00
27.37	42.07	(112.00)
54.62	37.07	38.00
30.02	45.17	(86.00)
36.42	39.03	(186.00)
33.18	46.40	(35.00)
31.60	45.45	(137.00)
32.87	43.88	(118.00)
35.11	36.42	(141.00)
25.88	47.72	(144.00)
20.62	33.28	(3.00)
28.66	41.82	(85.00)
45.59	38.72	8.00
33.29	47.92	(79.00)
19.47	34.48	(36.00)
30.94	35.78	(94.00)
21.55	36.75	(91.00)
51.23	39.38	(23.00)
44.12	41.27	(62.00)
47.20	37.35	21.00
24.13	35.20	(91.00)
27.87	34.12	(103.00)
25.58	45.70	(113.00)
41.11	39.92	(66.00)
30.37	44.40	(64.00)
23.91	45.87	(141.00)
25.70	44.63	(137.00)
36.83	40.60	(140.00)
45.83	39.20	(2.00)
27.65	34.22	(79.00)
30.82	46.60	(121.00)

(Data Continues)

(Data Continued)

AET (Energy) (Y Var)	Latitude (X1 Var)	Min Temp (X2 Var)
27.43	44.67	(47.00)
34.61	48.48	(80.00)
33.93	36.88	(143.00)
33.18	46.42	(35.00)
33.07	35.62	(103.00)
46.26	39.65	11.00
28.87	41.17	(107.00)
32.62	45.18	(97.00)
27.31	34.32	(53.00)
38.21	38.67	(83.00)
39.13	36.20	(123.00)
30.68	34.28	(28.00)
23.78	45.48	(118.00)
33.63	43.57	(80.00)

Data Source
https://www.kaggle.com/datasets/thedevastator/comprehensive-conifer-sampling-in-north-america?resource=download

Study Details
Date: Published November 19, 2021
Study: *Evaluating alternative study designs for optimal sampling of species' climatic niches*
Authors: Perret, Daniel and Sax, Dov
Site: [https://zenodo.org/records/5713338#.Y9Y3xtJBwUE]

Explanation of Variables and Theory

These data describe trees of the Ponderosa Pine species (*Pinus ponderosa*) in the United States. The variation in the dependent variable (AET or Energy) can be explained by the variation in the independent variables (Latitude and Minimum Recorded Temperature) due to the potential relationships between these variables, as outlined below:

Latitude (x^1):

Latitude can influence the amount of sunlight and thus the solar energy a location receives, which could directly impact the energy levels (AET) of pine trees. Locations closer to the equator (lower latitudes) receive more direct sunlight than locations closer to the poles (higher latitudes). This increased sunlight can lead to higher energy production in the pine trees through photosynthesis. Therefore, the variation in latitude could explain some of the variation in AET.

Minimum Recorded Temperature (x^2):

Temperature is a large factor for plant metabolism and growth. Lower temperatures can slow down metabolic processes and reduce the overall energy production (AET or Energy) in pine trees. Conversely, higher temperatures can increase these processes up to a certain limit (beyond which it may become detrimental to the plant). Therefore, the variation in the minimum recorded temperature could also account for some of the variation in AET.

Descriptive Statistics

AET (Energy, Y Var)	
Mean	33.15
Standard Error	1.20
Median	32.62
Mode	33.18
Standard Deviation	8.57
Sample Variance	73.40
Kurtosis	0.10
Skewness	0.40
Range	41.51
Minimum	13.11
Maximum	54.62
Sum	1,690.77
Count	51.00

Latitude (X1 Var)	
Mean	40.48
Standard Error	0.65
Median	39.77
Mode	37.35
Standard Deviation	4.66
Sample Variance	21.74
Kurtosis	(1.37)
Skewness	0.13
Range	15.20
Minimum	33.28
Maximum	48.48
Sum	2,064.25
Count	51.00

Min Temp (X2 Var)	
Mean	(74.78)
Standard Error	7.57
Median	(83.00)
Mode	(47.00)
Standard Deviation	54.09
Sample Variance	2,925.77
Kurtosis	(0.81)
Skewness	0.30
Range	224.00
Minimum	(186.00)
Maximum	38.00
Sum	(3,814.00)
Count	51.00

Variables Represent (Key):

AET (Energy, Y Var)
Amount of stored energy in the conifer species.
This will be the dependent variable (y) in the regression model.

Latitude (X^1 Var)
Latitude location of Ponderosa Pine in the USA.
This will be the first independent variable (x^1) in the regression model.

Min Temp (X^2 Var)
Minimum recorded temperature for the individual conifer tree.
This will be the second independent variable (x^2) in the regression model.

Summary Output:
Regression - AET (Energy, Y Var) with Latitude (X1 Var)

Regression Statistics	
Multiple R	0.04
R Square	0.00
Adjusted R Square	(0.02)
Standard Error	8.65
Observations	51.00

ANOVA

	df	SS	MS	F	Significance F
Regression	1.00	4.90	4.90	0.07	0.80
Residual	49.00	3,665.21	74.80		
Total	50.00	3,670.11			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	35.87	10.69	3.36	0.00	14.40	57.35
Latitude (X1 Var)	(0.07)	0.26	(0.26)	0.80	(0.59)	0.46

Regression Equation Prediction: $\hat{y} = 35.87 - 0.07(x^1)$

Explanation:
This equation predicts the AET (Energy) of Ponderosa Pines in the USA based on their Latitude (x^1).

Y-int: The value of 35.87 is the y-intercept and represents the predicted AET (Energy) when latitude (x^1) is zero. (However, after looking up and discovering that the lowest latitude is in Florida (at 24° 32′ 28.19" N), it may be that this latitude is the appopriate boundary for latitude in this equation).

Slope: The value of - 0.07 represents the slope with respect to latitude. For each unit increase in latitude, the AET (Energy) decreases by 0.07 units on average. This suggests that as we move away from the equator (up in latitude), the energy levels (AET) of the pine trees decrease.

x^1 : This variable represents the latitude. Each value of x^1 can be used to get a predicted value of the AET (Energy).

$\hat{y}$: This value is the predicted AET (Energy) for each tree at a specific latitude.

Summary Output:
Regression - AET (Energy, Y Var) with Min Temp (X2 Var)

Best Model

Regression Statistics	
Multiple R	0.38
R Square	0.15
Adjusted R Square	0.13
Standard Error	7.99
Observations	51.00

ANOVA

	df	SS	MS	F	Significance F
Regression	1.00	542.47	542.47	8.50	0.01
Residual	49.00	3,127.64	63.83		
Total	50.00	3,670.11			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	37.71	1.92	19.62	0.00	33.85	41.57
Min Temp (X2 Var)	0.06	0.02	2.92	0.01	0.02	0.10

Regression Equation Prediction: $\hat{y} = 37.71 + 0.06(x^2)$

Explanation:
This equation predicts the AET (Energy) of Ponderosa Pines in the USA based on their Minimum Recorded Temperature (x^2).

Y-int: The value of 37.71 is the y-intercept and represents the predicted AET (Energy) when the minimum recorded temperature (x^2) is zero. However, it is unknown to this author if the Ponderosa Pine can grow at such values (or if such values exist in the United States). As a result, the minimum recorded temperature should be the minimum temperature seen in the USA where this pine species is located.

Slope: The value of 0.06 represents the slope with respect to minimum recorded temperature. For each unit increase in the minimum recorded temperature, the AET (Energy) will increase by 0.06 units on average. Since the slope is positive, this suggests that as the minimum recorded temperature increases, the AET (Energy) level of the pine trees will also increase.

x^2 : This variable represents the minimum recorded temperature. Each value of x^2 can be used to get a predicted value of the AET (Energy). (Please note, for clarity the author calls this the x^2 variable, but since there is only one x in this equation, it could be notated as x^1).

$\hat{y}$: This value is the predicted AET (Energy) for each tree at a specific minimum recorded temperature.

Summary Output:
Regression - AET (Energy, Y Var) with Latitude (X1 Var) and Min Temp (X2 Var)

Regression Statistics	
Multiple R	0.39
R Square	0.15
Adjusted R Square	0.12
Standard Error	8.05
Observations	51.00

ANOVA

	df	SS	MS	F	Significance F
Regression	2.00	558.55	279.28	4.31	0.02
Residual	48.00	3,111.56	64.82		
Total	50.00	3,670.11			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	32.82	10.00	3.28	0.00	12.71	52.93
Latitude (X1 Var)	0.13	0.25	0.50	0.62	(0.38)	0.63
Min Temp (X2 Var)	0.06	0.02	2.92	0.01	0.02	0.11

Regression Equation Prediction: $\hat{y} = 32.82 + 0.13(x^1) + 0.06(x^2)$

Explanation:
This equation predicts the AET (Energy) of Ponderosa Pines in the USA based on their Latitude (x^1) and their Minimum Recorded Temperature (x^2).

Y-int: The value of 32.82 is the y-intercept and represents the predicted AET (Energy) when both the latitude (x^1) and the minimum recorded temperature (x^2) is zero. (**Note:** For the practical restrictions on the values, please refer to the explanations provided in the prior equations for each separate variable.)

Slope:
- The value of 0.13 represents the slope with respect to the latitude. For each unit increase in the latitude, the AET (Energy) will increase by 0.13 units on average (assuming minimum recorded temperature is constant). Since the value is positive, as latitude increases, the AET (Energy) levels of the pine trees will also increase.

- The value of 0.06 represents the slope with respect to the minimum recorded temperature. For each unit increase in the minimum recorded temperature, the AET (Energy) will increase by 0.06 units on average (assuming latitude is constant). Since the value is positive, as minimum temperature increases, the AET (Energy) levels of the pine trees will also increase.

x^1 and x^2: The x^1 variable represents the latitude, while the x^2 variable represents the minimum recorded temperature. Each value of x^1 and x^2 can be used to get a predicted value of the AET (Energy).

$\hat{y}$: This value is the predicted AET (Energy) for each tree given a specific combination of latitude and minimum recorded temperature values.

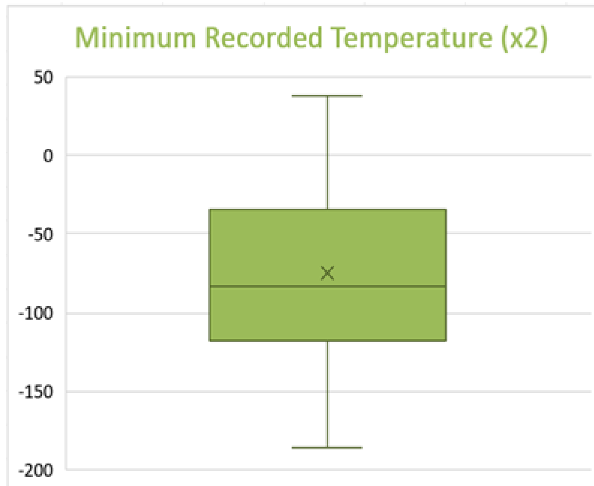
Z Score calculations for best model (7a)

Data Continued:

Min Temp (X2 Var)	Z-Scores ((x - μ) / σ)	Min Temp (X2 Var)	Z-Scores ((x - μ) / σ)
(186.00)	(2.06)	(3.00)	1.33
(144.00)	(1.28)	(2.00)	1.35
(143.00)	(1.26)	8.00	1.53
(141.00)	(1.22)	10.00	1.57
(141.00)	(1.22)	11.00	1.59
(140.00)	(1.21)	16.00	1.68
(137.00)	(1.15)	21.00	1.77
(137.00)	(1.15)	38.00	2.09
(134.00)	(1.09)		
(128.00)	(0.98)		
(123.00)	(0.89)		
(121.00)	(0.85)		
(118.00)	(0.80)		
(118.00)	(0.80)		
(113.00)	(0.71)		
(112.00)	(0.69)		
(107.00)	(0.60)		
(103.00)	(0.52)		
(103.00)	(0.52)		
(97.00)	(0.41)		
(94.00)	(0.36)		
(91.00)	(0.30)		
(91.00)	(0.30)		
(86.00)	(0.21)		
(85.00)	(0.19)		
(83.00)	(0.15)		
(80.00)	(0.10)		
(80.00)	(0.10)		
(79.00)	(0.08)		
(79.00)	(0.08)		
(66.00)	0.16		
(64.00)	0.20		
(62.00)	0.24		
(53.00)	0.40		
(47.00)	0.51		
(47.00)	0.51		
(36.00)	0.72		
(35.00)	0.74		
(35.00)	0.74		
(28.00)	0.86		
(23.00)	0.96		
(14.00)	1.12		
(9.00)	1.22		

5-Number Summary	
Smallest(Min)	-186.00
Q1	-118.00
Q2(Median)	-83.00
Q3	-35.00
Largest(Max)	38.00
Range	224.00
IQR	83.00

Descriptive Statistics	
Min Temp (X2 Var)	
Mean	(74.78)
Standard Error	7.57
Median	(83.00)
Mode	(47.00)
Standard Deviation	54.09
Sample Variance	2,925.77
Kurtosis	(0.81)
Skewness	0.30
Range	224.00
Minimum	(186.00)
Maximum	38.00
Sum	(3,814.00)
Count	51.00



Standard Deviations from the Mean (7b and 7c)					
Confidence Interval	Equation	Negative	Positive	Actual %	Outliers
~ 68 % (+/- 1 SD)	μ - 1σ	(128.87)	(20.69)	62.75%	n/a
~ 95 % (+/- 2 SD)	μ - 2σ	(182.97)	33.40	96.08%	(186.00), 38.00
~ 99.7 % (+/- 3 SD)	μ - 3σ	(237.06)	87.49	100%	none

Comparison of Regression Models

Model Option	Model Description	Adjusted R Square (coefficient of determination)	Significance F (overall significance)	P-Value
1	**AET (Energy, Y Var) with Latitude (x^1 Var)** $\bar{y} = 35.87 - 0.07(x^1)$	0.02	0.80	0.80
2	**AET (Energy, Y Var) with Min Temp (x^2 Var)** $\bar{y} = 37.71 + 0.06(x^2)$	0.13	0.01	0.01
3	**AET (Energy, Y Var) with Latitude (x^1 Var) and Min Temp (x^2 Var)** $\bar{y} = 32.82 + 0.13(x^1) + 0.06(x^2)$	0.12	0.02	0.62(x1) and 0.01(x2)

Table 1: Comparison of Regression Models for Best Fit

Best Model Selection (6a)

Due to the below rationale, Model 2 is selected as the best model. In other words, the variation of minimum temperature of Ponderosa Pines in the United States can best explain the variation in the tree's stored energy (AET). The equation $\bar{y} = 37.71 + 0.06(x^2)$ is the best equation to use from Table 1 for making predictions.

Use of Best Model to Predict Y Values (AET, Energy) (6b)

	Given X = -94	**Given X = -35**
Predicted Data	$\hat{y} = 37.71 + 0.06(x^2)$ $\hat{y} = 37.71 + 0.06(-94)$ $\hat{y} =$ **32.07**	$\hat{y} = 37.71 + 0.06(x^2)$ $\hat{y} = 37.71 + 0.06(-35)$ $\hat{y} =$ **35.61**
Actual Data	y = **30.94**	y = **33.18**

Table 2: Use of best model (model 2) to predict Y values

Discussion of Selection of Best Model (6c)

Adjusted R Squared Values

Adjusted R squared values indicate the proportion of the variance in the dependent variable that is predictable from the independent variable (AET, Energy). In other words, it is the measure of how well the independent variables (x values) explain the variability of the dependent variable (y value).

Since we want to be able to make predictions of y based on x, a higher adjusted R squared value is desirable. The first model did not have a high value, so it was not selected. The other two models had comparable adjusted R squared values, so both were still potential best models.

Significance F

Significance F (or the F-statistic) indicates the overall significance of the regression model. It tells us if the model is actually useful. For statistical significance in this study, we chose an alpha of 0.05. As such, we want the significance F-value to be lower than alpha, since this corresponds with statistical significance.

Model 1 was ruled out, but it is noted that it had a poor significance F-value (a high value at 0.80), further indicating that the equation was a poor fit as a regression model. Both models 2 (at 0.01) and 3 (at 0.02) had low F-values, so they both remain as potential models.

P-Value

The P-value tests the significance of individual predictors in the model. Values lower than our alpha of 0.05 indicate that the predictor is statistically significant.

Model 1 is ruled out, but it is noted that it had a high P-value (at 0.80) which further indicates that that equation was a poor fit. Model 2 has a low P-value (0.01), suggesting that it may be the best model. Model 3 had two P-values (since it was a combined equation for both x^1 and x^2). As a result, the model 3 P-values are 0.62 (for x^1, a poor fit) and 0.01 (x^2, a good fit). Taken as a whole, model 3 could not be said to be better than model 2 due to the presence of a high P-value for x^1.

Discussion of Results Compared to Predicted Theory (8a)

In the initial theory, it was predicted that Latitude (x^1) can influence the amount of sunlight, and thus the solar energy a location receives, which directly impacts the energy levels (AET) of pine trees. While locations closer to the equator (at lower latitudes) receive more direct sunlight compared to locations closer to the poles (higher latitudes), this model appears to be a poor fit.

Since the sample size of 51 is relatively small, it is hard to make a definitive conclusion whether our conclusion regarding latitude is truly the case in reality. However, the author notes that there is significant variation in altitudes (among other factors) that the pine trees can grow at, which might explain why the variation in latitude did *not* significantly contribute towards the variation in energy (AET).

In the initial theory, it was also predicted that the Minimum Recorded Temperature (x^2) would be a large factor for plant metabolism and growth. Since metabolic processes are often directly impacted by environmental temperatures, it follows that the available energy (AET) would in turn be dependent on temperature. As evidenced by the best fit regression model 2, it *does* appear that the variation in Minimum Recorded Temperature (x^2) could account for some of the variation in Energy (AET). This part of the predicted theory is supported by the best fit regression model.

Discussion of Selection of Best Regression Model (8b)

Two potential independent variables were selected and weighed as contributors towards the variation in the dependent variable. To determine the best regression model, descriptive statistics, regression statistics, and an ANOVA table were created using Excel. The best model was then selected according to the values obtained for the Coefficient of Determination (adjusted R squared value), F significance, and P-value. These metrics provide a holistic approach to evaluating whether the resulting equations are suitable for making predictions about the y variable.

Discussion of Model Positive Aspects and Drawbacks (8c)

There are several positive aspects of the chosen model. AET (Energy) represents both the combined loss of water (through evaporation from the soil surface) and transpiration (water movement through plants). These processes are directly influenced by temperature, which is reflected by the model.

Evaluation of Variable: Minimum Recorded Temperature (x^2)

If temperatures drop very low, pine trees may respond by closing leaf pores to conserve water, resulting in lower AET (Energy) values. Conversely, if the temperatures increase, the leaf pores may open to release more water, resulting in higher AET (Energy) values. These trends are reflected in the collected data and the resulting model.

Drawbacks are also present with the chosen model. This regression model assumes a linear relationship between the dependent variable (y value) and the independent variables (x values). In real life, there are likely additional factors that result in a particular AET (Energy) reading besides the temperature.

Evaluation of Variable: Latitude (x^1)

The variation in latitude did not sufficiently explain the variation in AET (Energy). There are many potential reasons for this. Variations in altitude could change the growing environment substantially. Even though a plant may be closer to the equator, it may be situated high enough where the temperature is too cool to perform proper metabolic reactions for energy. Conversely, extreme high temperatures could damage the plant and lead to poor energy production. There is likely an ideal environment for growth of these trees, which may appear on the map as a scattered and discontinuous micro-climate, rather than a continuous band at a specific latitude.

Discussion of Suggestions for Improved Regression Model (8d)

By examining more data, we can potentially create a more accurate and precise regression equation that can be used to predict pine tree AET (Energy). For instance, we could look at additional variables such as the tree age, soil composition, degree of interspecies competition, nutrient availability, rainfall, pest presence, air movement, solar radiation, humidity, and the presence of plant diseases. These factors (and likely many others) could impact how much energy a pine tree can store and utilize for growth. By accounting for as many factors as possible, we can create a model that best matches our collected data.

Discussion of Final Model Reliability (8e)

There is sufficient evidence to suggest that the final chosen model (Model 2) can be used somewhat reliably to predict the AET (Energy) of Ponderosa Pines growing in the United States. Despite the simplicity of the equation, it does pass multiple tests, including the Coefficient of Determination, Significance F, and P-Value at the 95% confidence level. However, due to numerous environmental factors that impact plant metabolism, it is suggested that the equation be improved upon to account for additional variables as outlined in the Improvement section. By accounting for multiple variables, we can create a strong regression equation to predict data, which is an important tool for climate monitoring and environmental protection efforts.