JMP Academic Case Study 046

Endangered
- The Mighty Saguaro Cactus

Linear Regression and Nonparametric Tests

Produced by

Jennifer L. Verdolin, PhD

verdolin.jennifer@gmail.com

JMP STATISTICAL DISCOVERY

Endangered: The Mighty Saguaro Cactus

Linear Regression and Nonparametric Tests

Key Ideas

This case study uses linear regression, nonparametric tests and statistical inference to examine the effects of climate change on saguaro cactus in the Sonoran Desert.

Background

The saguaro is the icon of the Sonoran Desert and an ecologically important plant. It provides habitat for many species (including the highly endangered pygmy owl), nesting sites for the Gila woodpecker, and is a source of food for moths, bats, and bees. While saguaros are characterized by columnar growth, there is an unusual growth form, the crested saguaro, that is legendary. Check out this US National Park Service video detailing this mysterious growth form. As concerns over climate change grow, monitoring the timing of important lifecycle events, called phenology, becomes more critical. Increasing temperatures and reduced precipitation during the spring in the desert Southwest of the US might cause the timing of key events, such as flowering time, to shift and occur later in the year. Such a shift might negatively impact other species that rely on the saguaro for pollen or fruit.

The Task

Use open access data from the DataONE Data Catalog to investigate whether there is evidence of a change in flowering time and duration from 2004-2013 that is correlated with saguaro size, temperature, and precipitation. [Source: Renzi, J. J., Peachey, W. D., & Gerst, K. L. (2019). A decade of flowering phenology of the keystone saguaro cactus (Carnegiea gigantea). *American Journal of Botany*, *106*(2), 199-210.] (The data are open access and available online along with phenology observations at: doi.org/10.5063/F1DZ06JG. Data are licensed under CC BY 4.0)

The Data **saguaro.jmp**

The data table (which is a summarized set of the open data source mentioned above) includes the size of a saguaro and the average number of flowers produced.

Saguaro_ID	ID of the Saguaro
Average_Totalbloom(#blooms)	The average number of flowers produced
Number_of_arms	The number of arms a saguaro has

Analysis

Phenology, or the timing of important life cycle events, can be influenced by several factors, including the properties of an individual saguaro. Older saguaros are larger and have more arms, making the number of arms a good measurement of age and size. If saguaro size influences flowering, we might expect saguaros with more arms to produce more flowers. Alternatively, if saguaro size does not influence the number of flowers produced, then we might expect a similar number of flowers produced regardless of saguaro size.

To determine if there is a relationship between the number of arms an individual saguaro has and the total number of flowers it produces, we can use a linear regression to test the following hypotheses:

Ho: The total number of flowers produced is on average the same for different sized saguaros.
Ha: Larger saguaros produce, on average, more flowers

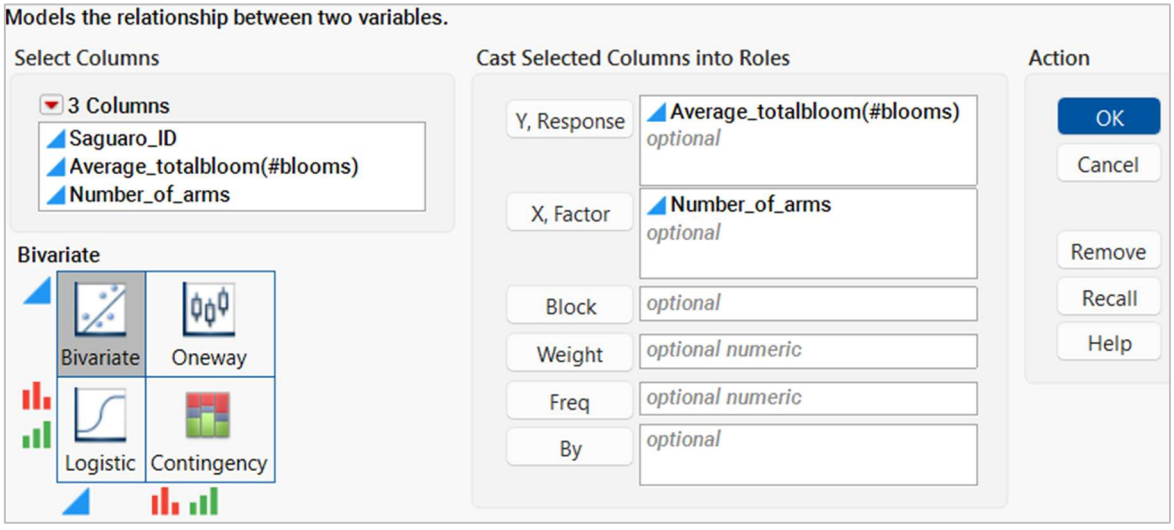
Assumptions of linear regression

Before we undertake a linear regression, we need to consider whether we meet the necessary assumptions. There are four basic assumptions:

- There is a linear relationship between the number of arms on a cactus (independent variable) and the average number of flowers produced (dependent variable).
- There is no autocorrelation present in the data.
- The variance of the residuals is constant for any value of X.
- The residuals are normally distributed.

Lets build the model and then validate the assumptions.

Exhibit 1 Setting up the Linear Regression



(To create, go to Analyze>Fit Y by X and select Average_totalbloom(#bloom) for the Y, response, and Number_of_arms for the X, factor. Click OK.)

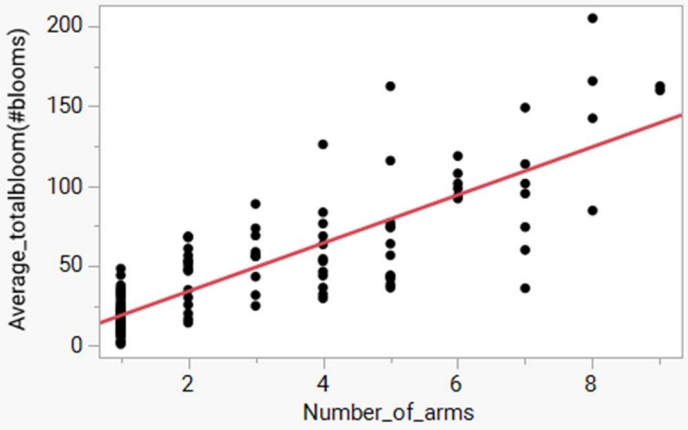
Linear regression

We can see that the data appear to have a linear relationship meeting the first assumption and the Fit Line function will be used to see if there is evidence of a significant relationship.

Exhibit 2 Regression with Average Total Bloom (Y) and Number of Arms (X)

Bivariate Fit of Average_totalbloom(#blooms) By Number_of_arms

Student Edition



Linear Fit

Linear Fit

Average_totalbloom(#blooms) = 3.8179665 + 15.090336*Number_of_arms

Summary of Fit

RSquare	0.701021
RSquare Adj	0.699014
Root Mean Square Error	21.36608
Mean of Response	44.39198
Observations (or Sum Wgts)	151

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	159487.22	159487	349.3625
Error	149	68019.87	457	Prob > F
C. Total	150	227507.09		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.8179665	2.781258	1.37	0.1719
Number_of_arms	15.090336	0.807348	18.69	<.0001*

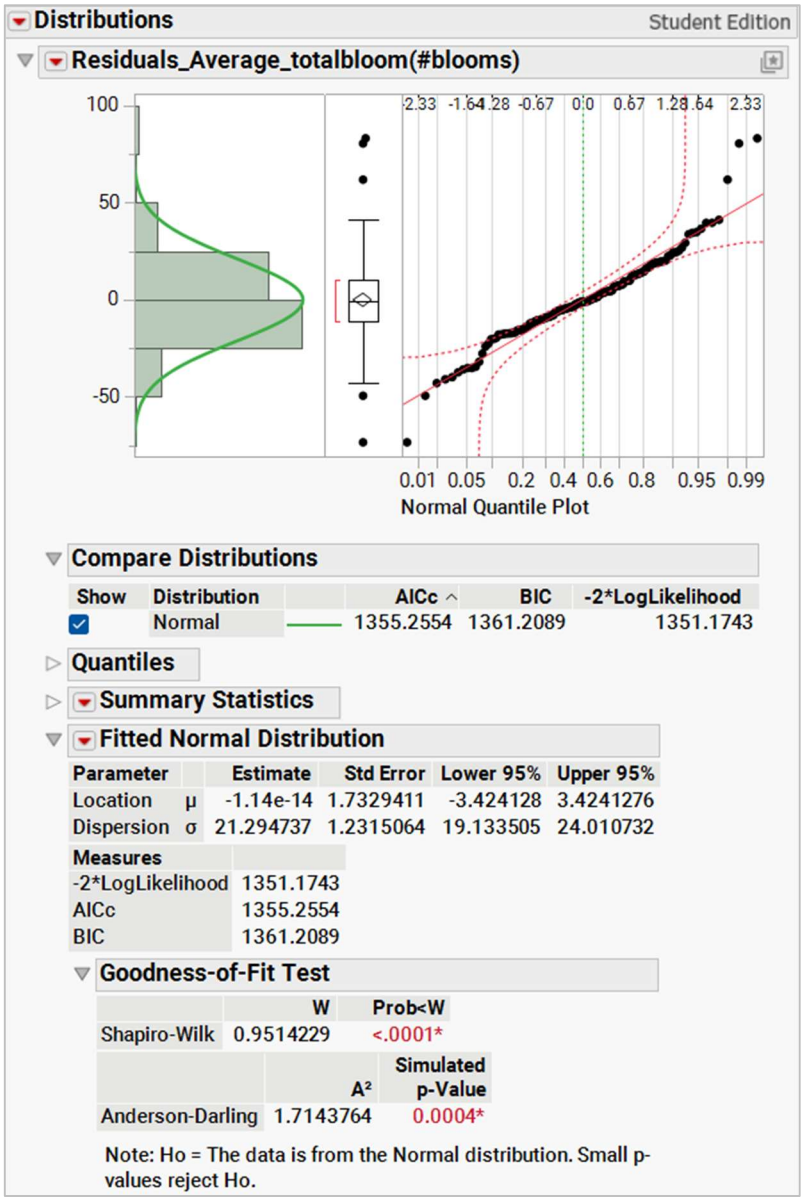
(To create, click on the red triangle and select Fit Line.)

Exhibit 2 demonstrates that 70% of the variability in the number of blooms is attributed to variation in the number of arms of a saguaro (RSquare = 0.701021), suggesting a statistically significant relationship between saguaro size and flowering.

Let's check and see if we meet another assumption by looking at the residuals. In a linear regression, the residuals should follow a normal distribution. To do this, click on the red triangle next to Linear Fit and select Save Residuals. You will then have a new column in the By_Saguaro data set called Residuals_Average_totalbloom(#blooms).

By looking at the distribution of the residuals and using a goodness of fit test, we can check whether we meet the normality assumption. In a goodness of fit test, our null hypothesis, or Ho, is that the data is from a normal distribution. Thus, if our results support this, the p-value will not be significant.

Exhibit 3 Normal Quantile Plot and Test of Normality

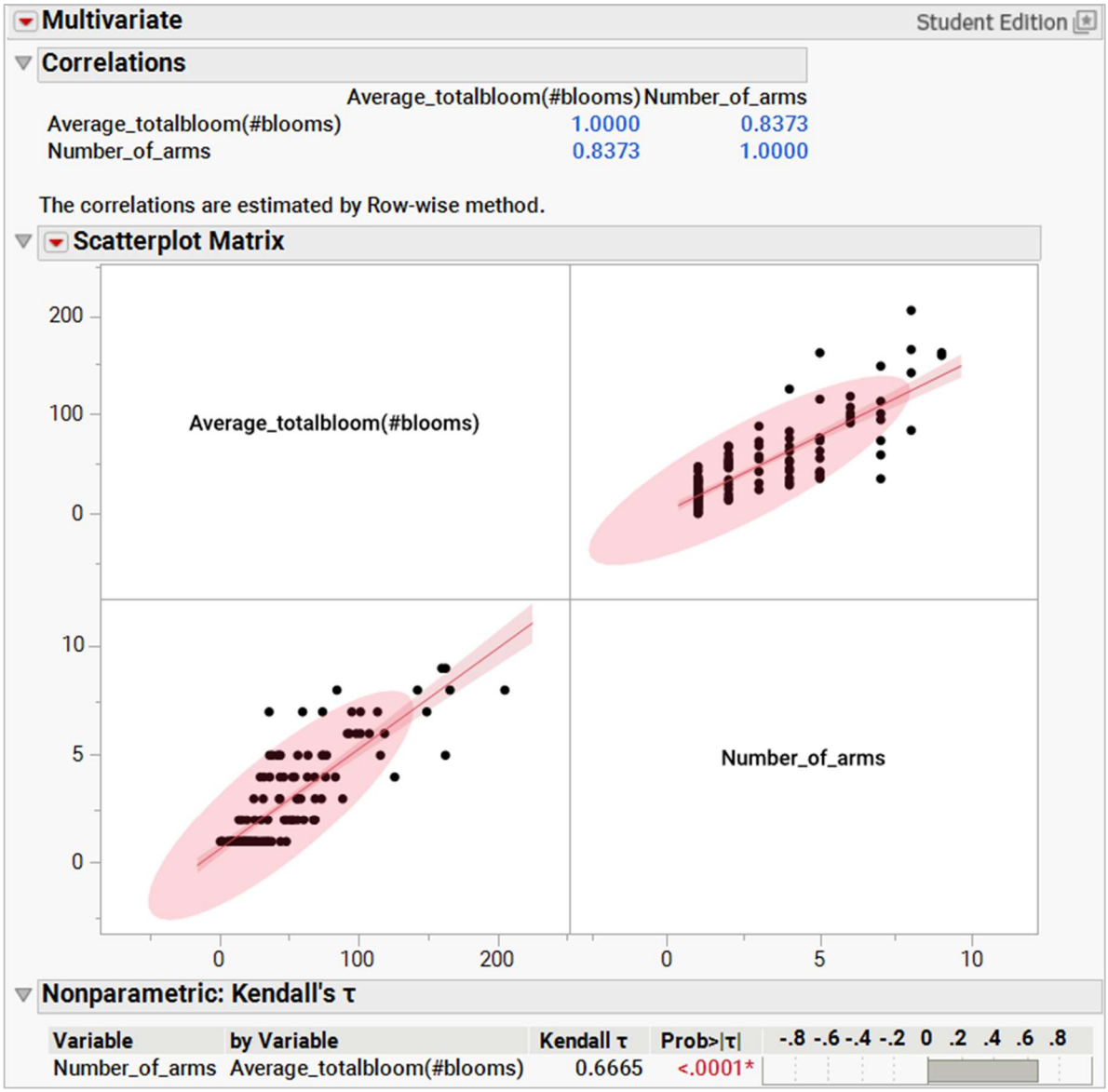


(To create, select Analyze>Distribution then select Residuals_Average_totalbloom(#blooms), add it to the Y, Columns, and click OK. Click the red triangle next to Residuals and select Normal Quantile Plot. Click on the red triangle next to Residuals and select Continuous Fit>Normal. Click on the red triangle next to Fitted Line and select Goodness of Fit Test.)

We can see that the residuals are not normally distributed. By visually examining the normal quantile plot, it is evident that there is some skew in the data; this is further confirmed by the significant p-value in the goodness of fit test. But how robust is the result in light of the violation of this assumption? One way we can explore this is to use a nonparametric correlation test, such as Kendall's Tau, to see if we maintain the integrity of the direction and strength of the relationship uncovered in the linear regression. An important difference is that the regression is predictive or causal in nature, while the correlation is limited to quantifying the direction and strength of the relationship between two numeric variables.

Kendall's Tau is the nonparametric alternative and allows for relaxation of the assumptions.

Exhibit 4 Scatter Plot Matrix and Kendall's Tau



(To create, select Analyze>Multivariate Methods>Multivariate and add both variables to the Y, Columns, and click OK. Click the red triangle next to Multivariate and select Nonparametric Correlations> Kendall's τ. Under Scatterplot Matrix, use the red triangle to select Fit Line and shaded ellipses.)

Based on the correlation results, we see above that the direction of their relationships and their significance (i.e., $p < 0.0001$) was retained with nonparametric Kendall's Tau correlation, suggesting that although the residual was not normally distributed in the regression, the result is robust.

Both the simple linear regression and the Kendall Tau's correlation produce very small p-values and strong evidence that larger saguaros produce, on average, more flowers. When it comes to reporting the results, which one should be reported? In general, the more rigorous approach is to report the results that best meet the necessary assumption.

Summary

Statistical insights

Here we utilized both parametric and nonparametric approaches to assess the direction and strength of the relationship between variables.

There are several similarities between linear regression and correlation analysis, but there are some significant differences as well. A material difference lies in the interpretation. With regression, one is generating an equation that allows us to determine how much and in which direction a variable will change as a function of the other variable. It is used to make predictions. Correlation analysis is simply revealing the degree of linear association between two variables.

It is important to be aware of the assumptions that different statistical procedures make. For instance, linear regression does not require that the variables be normally distributed, but it does assume that the residuals (error) are normally distributed. How normal these need to be is subjective and the goodness of fit test provides an objective assessment of how normal the residuals really are. In this case, a nonparametric test may be the best approach.

JMP features and hints

In this case study we used the Fit X by Y platform to conduct a simple linear regression. We used the Distribution Platform and goodness of fit test to assess the distribution of the residuals. We used the Multivariate Methods>Multivariate platform to perform nonparametric correlations.

Exercises

We know that temperature and precipitation can greatly influence when a plant blooms, including saguaro cacti. Typically, saguaros bloom in the spring before the monsoon season arrives; warmer spring temperatures may lead to earlier bloom times. While precipitation is important, increased precipitation may lead to a later onset in blooming. This exercise involves exploring the timing of saguaro flowering using temperature and precipitation data from 2004-2013.

Data all-data.jmp

Source: Renzi, J. J., Peachey, W. D., & Gerst, K. L. (2019). A decade of flowering phenology of the keystone saguaro cactus (Carnegiea gigantea). *American Journal of Botany*, *106*(2), 199-210.

The data are open access and available online, along with phenology observations at doi.org/10.5063/F1DZ06JG. Data are licensed under CC BY 4.0

1. Since temperature and moisture influence the timing of flower blooming in saguaro catci in opposing ways, generate a hypothesis for each.
2. Using linear regression, test the influence of winter and spring T_{min}/T_{min} and $FallPrecip_{mm}$ and $WinterPrecip_{mm}$ on Onset_DOY.
3. How much variation in Onset_DOY is explained by the temperature and precipitation variables you explored? Which variables do you think explain more variation?
4. Check for normality of the residuals and perform a nonparametric Kendall's Tau correlation as needed.
5. In those cases where you used Kendall's Tau, were the results congruent in direction and significance with the linear regressions?
6. After running these analyses, what can you say about the role that temperature and precipitation have on the timing of flowering in the saguaro cacti measured?

Data saguaro-by-year.jmp; all-data.jmp; Metadata.jmp

Source: Renzi, J. J., Peachey, W. D., & Gerst, K. L. (2019). A decade of flowering phenology of the keystone saguaro cactus (Carnegiea gigantea). *American Journal of Botany*, *106*(2), 199-210.

The data are open access and available online along with phenology observations at website: doi.org/10.5063/F1DZ06JG. Data are licensed under CC BY 4.0

1. Using the analytical skills you've developed, test if there has been a significant change in the onset of blooming over time.
2. Using the metadata.jmp file to guide you in understanding what all the variables are in the All_Data.jmp file, develop an additional hypothesis to test.
3. Perform the analyses you've learned on this new hypothesis and discuss the results.
4. What additional abiotic variables do you think might impact flowering time?
5. How will pollinators be impacted by changes in flowering time?
6. What new questions should be investigated and what data would you need to evaluate this future hypothesis?