Scholars Week 2020

May 18th, 12:00 AM - May 22nd, 12:00 AM

Modeling Park Visitation Using Transformations of Distance-Type Predictor Variables with LASSO

Ashley Hall
Western Washington University

Follow this and additional works at: https://cedar.wwu.edu/scholwk

Part of the Statistics and Probability Commons

Hall, Ashley, "Modeling Park Visitation Using Transformations of Distance-Type Predictor Variables with LASSO" (2020). Scholars Week. 44.
https://cedar.wwu.edu/scholwk/2020/2020/44

This Event is brought to you for free and open access by the Conferences and Events at Western CEDAR. It has been accepted for inclusion in Scholars Week by an authorized administrator of Western CEDAR. For more information, please contact westerncedar@wwu.edu.
Modeling Park Visitation Using Transformations of the Distance-Type Predictor Variables with LASSO

Ashley Hall and Kimihiro Noguchi
Department of Mathematics, Western Washington University, Bellingham, WA 98225

Introduction
We examine three common transformations (identity, fourth-root, and log) to determine the most suitable transformation for evaluating the importance of certain common features surrounding the Twin Cities Metropolitan Area (TCMA) city parks on park visitation. The distances between these features and city parks are approximately exponentially distributed by noting that their relative locations closely follow the spatial Poisson process. Because a fourth-root transformation improves the normality of exponential random variables, we verify that the fourth-root transformation is considered best by comparing correlation coefficients of the fourth-rooted data to the untransformed and log-transformed data via simulation. Using the TCMA city parks data, we also confirm that the fourth-root transformation improves the bivariate normality. Finally, we show that the fourth-root transformation of distance-type variables improves the probability of selecting the most important features affecting the park visitation using the least absolute shrinkage and selection operator (LASSO) regression.

Data
Response Variable: Mean annual Twitter-User-Days (TUD) counts [1], a proxy for park visitation counts per year, for each TCMA park between 2012 and 2014, is divided by the surrounding population and log-transformed (LogTUDDensity). The LogTUDdensity variable is approximately normal.

Predictor Variables: Distances between the following features and the nearest city park: Bike path (Dist2Bike), bus stop (Dist2Bus), Minneapolis-St. Paul downtown area (Dist2MSP), hiking trail (Dist2Trail), and water features (Dist2WtrMC).

These variables are approximately gamma distributed with shape parameter values close to 1, resembling exponential distribution.

Note: Parks with zeros in any of the above variables were removed. In total, 890 parks are included in our study.

Simulated Correlation Coefficient Comparisons
The correlation coefficient measures the strength of a linear association between the response and predictor variables.

To confirm why the fourth-root transformation is favorable, we firstly generated 10000 pairs of observations from the standard bivariate normal distribution with various correlation coefficients. Then, we transformed the second observation in each pair into a standard exponential observation and tested the three transformation on them. We confirmed that the fourth-root transformation tends to retain the original correlation coefficient better than the untransformed and log-transformed ones based on the mean squared error (MSE) and bias criteria (see Fig. 1).

Bivariate Normality Comparisons
Bivariate normality is a desirable property for a pair of response and predictor variables to reliably estimate their correlation coefficient. The contour plots and perspective plots below confirm that the fourth-rooted predictor variables improve the bivariate normality (see Fig. 2).

Application of the LASSO Regression to the TUD Data
LASSO is a popular statistical method for selecting the most important predictor(s) in linear regression. Table 2 summarizes whether or not each of the five predictors is included using LASSO for the three models (identity, fourth-root, and log transformation). Our results show that the fourth-root and log transformation tend to select more variables as significant.

Conclusion and Future Work
The fourth-root transformation of distance-type variables tends to improve the accuracy of multiple linear regression results. Specifically, the transformation is expected to improve the correlation coefficient estimation and bivariate normality. Furthermore, the fourth-root transformation seems to help LASSO perform reasonable model selection compared to the identity and log transformation. It may result in selecting too few or too many distance-type predictors.

References

Acknowledgments
We would like to thank Dr. Spencer Wood and Ms. Marie Donahue at the Natural Capital Project for kindly providing us the TCMA park data. This project was supported by the Western Washington University Washington Space Grant Academic Research Award for Future Science Teachers.

Fig. 1 Comparison of correlation coefficients using bias and MSE criterion. 1 = no transformation, 0.25 = fourth-root transformation, 0.1 = log transformation.

Fig. 2 Perspective (top) and contour (bottom) plots of predictor variable Dist2WtrMC showing the fourth-root transformation (middle) improves the bivariate normality of the normally transformed data (left) better than the log transformation (right).

Table 1 Variable inclusion/exclusion for TUD data.

<table>
<thead>
<tr>
<th>Transformation</th>
<th>Identity</th>
<th>Fourth Root</th>
<th>Log</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Included</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Omitted</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 2 The probability of each predictor variable being selected via LASSO for each transformation.

<table>
<thead>
<tr>
<th>Transformation</th>
<th>Identity</th>
<th>Fourth Root</th>
<th>Log</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>0.0885</td>
<td>0.0780</td>
<td>0.4142</td>
<td>0.0358</td>
</tr>
<tr>
<td></td>
<td>0.0320</td>
<td>0.2764</td>
<td>0.3486</td>
<td>0.0041</td>
</tr>
</tbody>
</table>

Although Table 2 suggests the lower the power, the higher the probability of selecting each distance predictor, it may imply that the log transformation could end up choosing too many predictors. The fourth-root transformation seems to provide a good balance of selection and omission of these distance-type predictors.