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

Ashley Hall and Kimihiro Noguchi

Department of Mathematics, Western Washington University

Mathematics Continued Conference 2020

October 24, 2020

Abstract: 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, via simulation, we verify that the fourth-root transformation is considered best for identifying significant predictor variables affecting park visitation. Specifically, we achieve that by comparing the probabilities of selecting fourth-rooted predictor variables to the untransformed and log-transformed predictor variables using the least absolute shrinkage and selection operator (LASSO) regression. Finally, we apply these three transformations to various TCMA distance-type predictor variables to demonstrate that the significance of distance to the nearest bus stop improves dramatically with the fourth-root or log transformation.

Background Information:

With a desire to understand and analyze what amenities nearby city parks most impact visitation, we analyze five predictor variables using the distances between the features and the nearest park in the Twin Cities Metropolitan Area (TCMA). These predictor variables include: bike paths (**Dist2Bike**), bus stops (**Dist2Bus**), Minneapolis-St. Paul downtown areas (**Dist2MSP**), hiking trails (**Dist2Trail**), and water features (**Dist2WtrMC**). These predictor variables are approximately gamma distributed with shape parameter values close to 1, resembling exponential distribution. The response variable is the log-transformed visitation density obtained from Twitter, which is approximately normally distributed.

We applied three transformations (**identity**, **fourth-root**, and **log**) on our (nearly) exponential predictor variables to assess the accuracy of model selection using the least absolute shrinkage and selection operator (LASSO).

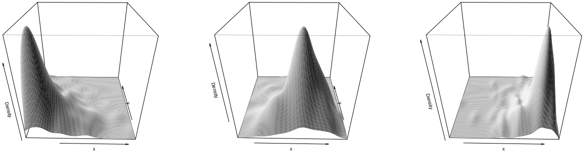

Figure: Perspective plots showing the bivariate density of a pair of response and transformed Dist2WtrMC predictor variables. The fourth-root transformation on the Dist2WtrMC variable (middle) improves the bivariate normality of the original untransformed pair (left) better than the log transformation (right).

Model and Simulation Results:

Model: The multiple linear regression models of the form

$$Y_i = \beta_0 + \beta_1 g_j(X_{1,i}) + \cdots + \beta_k g_j(X_{k,i}) + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2),$$

$i = 1, \ldots, n$, where $g_1(x) = x$, $g_2(x) = x^{1/4}$, and $g_3(x) = \log(x)$, $x > 0$, are considered. The LASSO simultaneously performs coefficient estimation and selection of the predictors using the residual sum of squares with a constraint on the ℓ_1-norm of the coefficients after standardizing all the variables.

Simulation Results: We generate 10,000 sets of $\boldsymbol{Z}_i := (Z_{0,i}, Z_{1,i}, \ldots, Z_{5,i})^T$, $i = 1, \ldots, 300$, from the multivariate normal distribution with $\mathrm{Var}(Z_{t,i}) = 1$, $\mathrm{Cov}(Z_{t_1,i}, Z_{t_2,i}) = 0.2$ if $t_1 \leq 3$, $t_2 \leq 3$, $t_1 \neq t_2$, and 0 otherwise. Then, we set $Y_i = Z_{0,i}$ and $X_{t,i} = F_{\chi^2}^{-1}(\Phi(Z_{t,i}))$, where $F_{\chi^2_2}(\cdot)$ and $\Phi(\cdot)$ denote the cumulative distribution functions of the scaled exponential (χ^2_2) and $N(0,1)$ distributions.

The LASSO regression model is then applied to each set to calculate the probability of retaining each variable for each transformation. Note that $X_{1,i}$, $X_{2,i}$, and $X_{3,i}$ are correlated with Y_i in theory while $X_{4,i}$ and $X_{5,i}$ are not. The table below shows that the fourth-root transformation is best, closely followed by the log transformation.

Transformation	$X_{1,i}$	$X_{2,i}$	$X_{3,i}$	$X_{4,i}$	$X_{5,i}$
Identity	0.3244	0.3361	0.3269	0.0039	0.0035
Fourth-Root	0.4393	0.4410	0.4356	0.0044	0.0034
Log	0.4120	0.4063	0.4009	0.0045	0.0030

Table: Probabilities of retaining simulated predictor variables. They are the highest for $X_{1,i}$, $X_{2,i}$, and $X_{3,i}$ with the fourth-root transformation while $X_{4,i}$ and $X_{5,i}$ have similar low probabilities.

Application:

Based on $10,000$ random subsets of one-third of the original TCMA park data ($n = 890$), the probability that the LASSO model retains each of the five distance-type predictor variables (Dist2Bike, Dist2Bus, Dist2MSP, Dist2Trail, and Dist2WtrMC) for each transformation is calculated.

The table below clearly suggests that the Dist2Bus predictor variable becomes much more significant with the fourth-root and log transformation (0.7934, and 0.9572, respectively) compared to the untransformed one (0.0133). Thus, the results suggest that the distance to the nearest bus stop is significantly correlated with the park visitation, which is revealed only after the transformation.

On the other hand, there is no large change in these probabilities for non-significant predictor variables (Dist2Bike, Dist2Trail, and Dist2WtrMC). Lastly, Dist2MSP seems to be mildly correlated with park visitation.

Transformation	Dist2Bike	Dist2Bus	Dist2MSP	Dist2Trail	Dist2WtrMC
Identity	0.0016	0.0133	0.3734	0.0020	0.0086
Fourth-Root	0.0040	0.7934	0.3821	0.0022	0.0063
Log	0.0031	0.9572	0.4278	0.0041	0.0058

Table: Probabilities of predictor variables being chosen via the Lasso regression model.