# TABLE OF CONTENTS

[**TABLE OF CONTENTS**](#_dju65fe7nfl1) **1**

[**Looking into Model Assumptions**](#_lvpw763iytyt) **2**

[Multicollinearity](#_o7ux9z5u57f4) 2

[Normality](#_2ugpnzpsi2c8) 3

[Linearity](#_n2sbfpn2s4ab) 7

[**Predictor By Response**](#_og4ne3x5wefb) **9**

[**Spatial Model Results**](#_nivw8ddcxf4p) **11**

[Spatial Lag Understanding](#_k2s4nhoyzlsw) 16

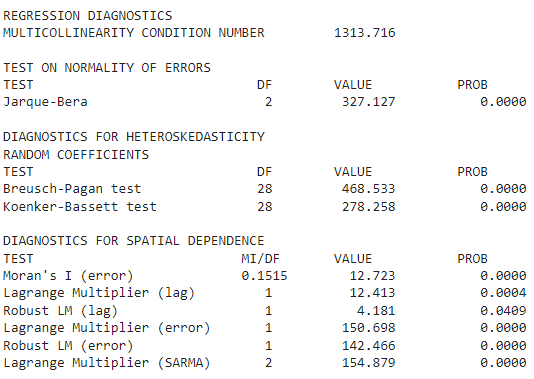
[Nathan Understanding Spatial Models Link](#_2yd0og7rhvbk) 17

[**GWR**](#_krgw0uu6rds) **18**

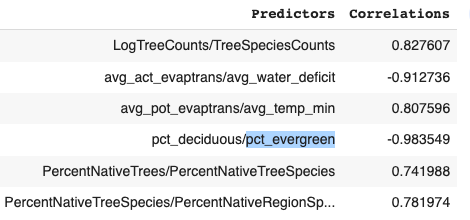
**Potential Interaction Terms:** Variables whose effect on the response changes based on a change in an interacting variable

* between overall tree counts and potentially all of the percent variables
* between evenness and richness, but that kind of represents Shannon. but still a possibility
* between tree height and tree trunk diameter, no because only 1 is sig by itself
* between tree height and canopy size, no because neither is sig by itself
* between canopy size and denseness of foliage, no because neither is sig by itself

# Looking into Model Assumptions

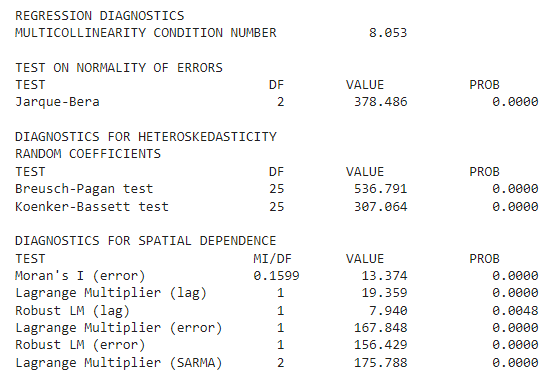


## Multicollinearity



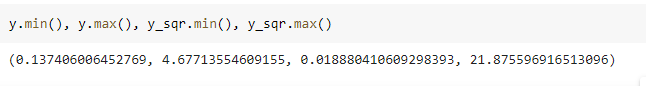
Multicollinearity reduces the precision of the estimated coefficients impacting the statistical power of the model and their significance. We want to fix the multicollinearity of the predictors that we care about interpreting: ones that are important to us.

* Deciduous and evergreen strongly, negatively correlated, so we are replacing deciduous with a less correlated partly\_deciduous
* Climatic water deficit is defined as potential evapotrans - actual evapotrans, so we don’t need all three. Additionally, potential evapotrans is also highly correlated with temp min, so we decide to remove potential and actual evapotrans
* PercentNativeTreeSpecies is highly correlated to both PercentNativeTrees and PercentNativeRegionSpecies, so we remove
* Just those three changes change the multicollinearity condition number from 1300 to 8
* **Only LogTreeCounts vs TreeSpeciesCounts has a correlation higher than 0.7 after**



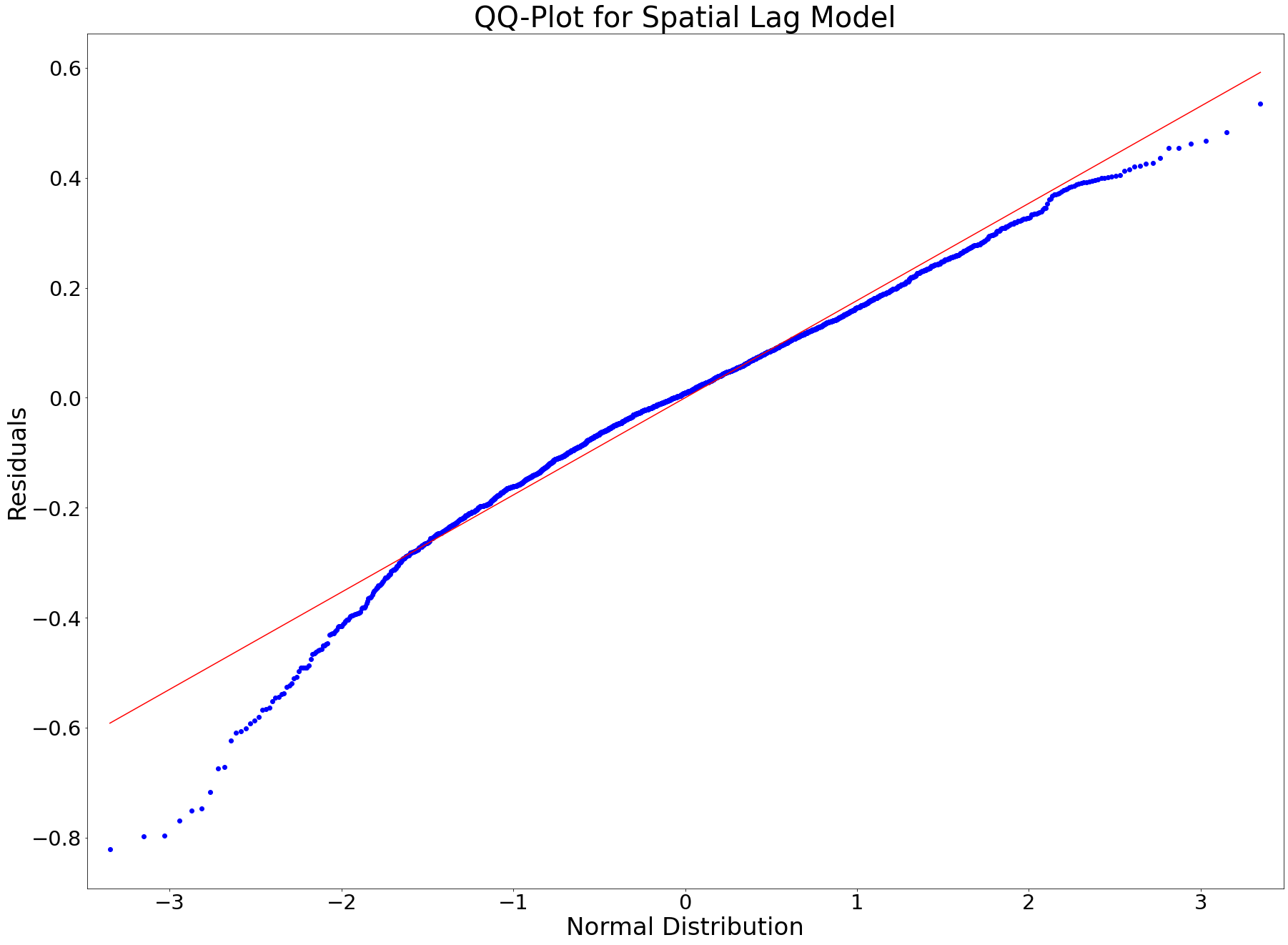
Our plan is to use spatial models which account for spatial dependence and heteroscedasticity. Therefore, our concerns are with normality and linearity. Normality is less of a concern, especially since we have a large sample size.

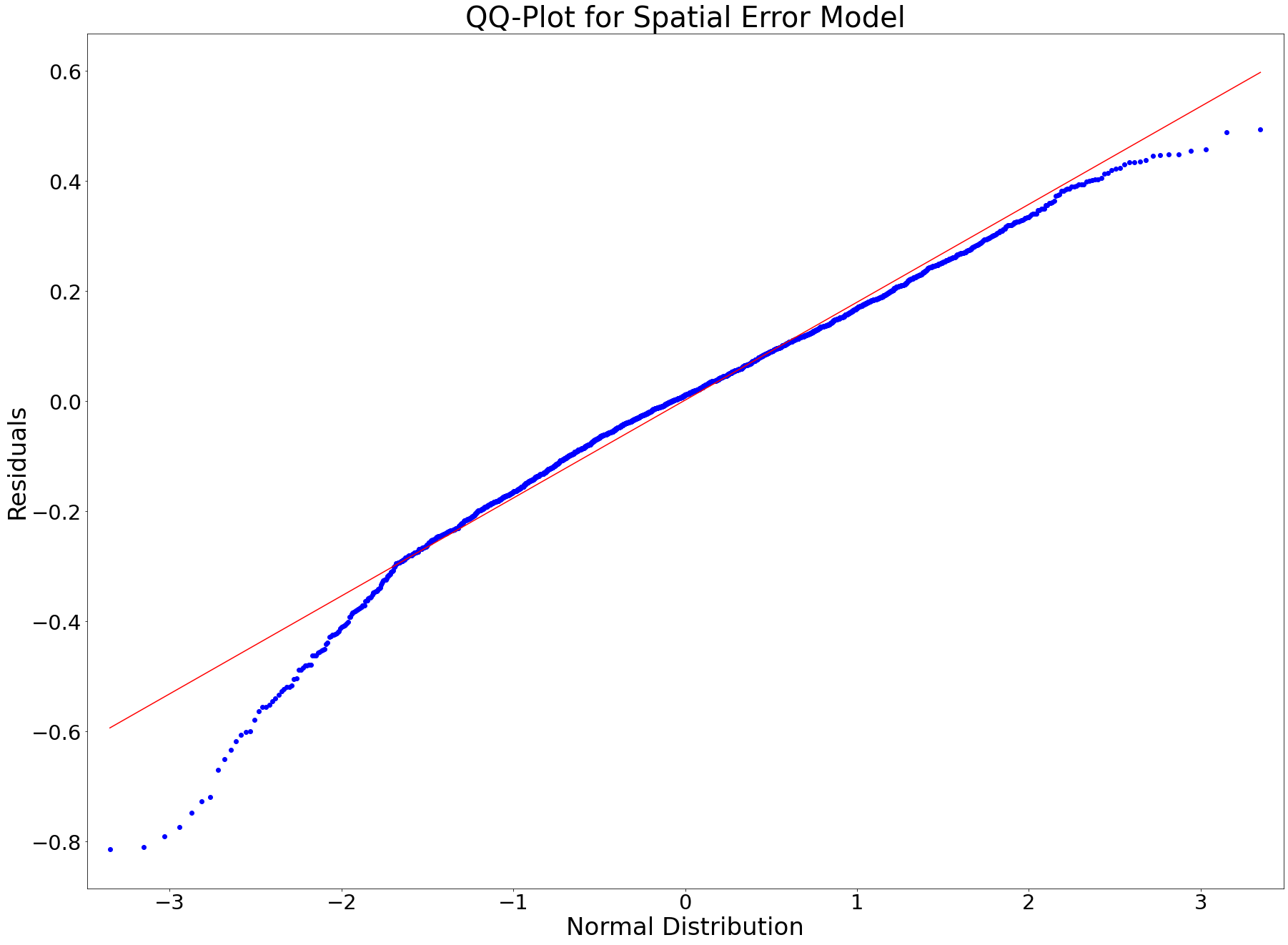
**To fix both linearity and normality, we will try squaring the response variable**



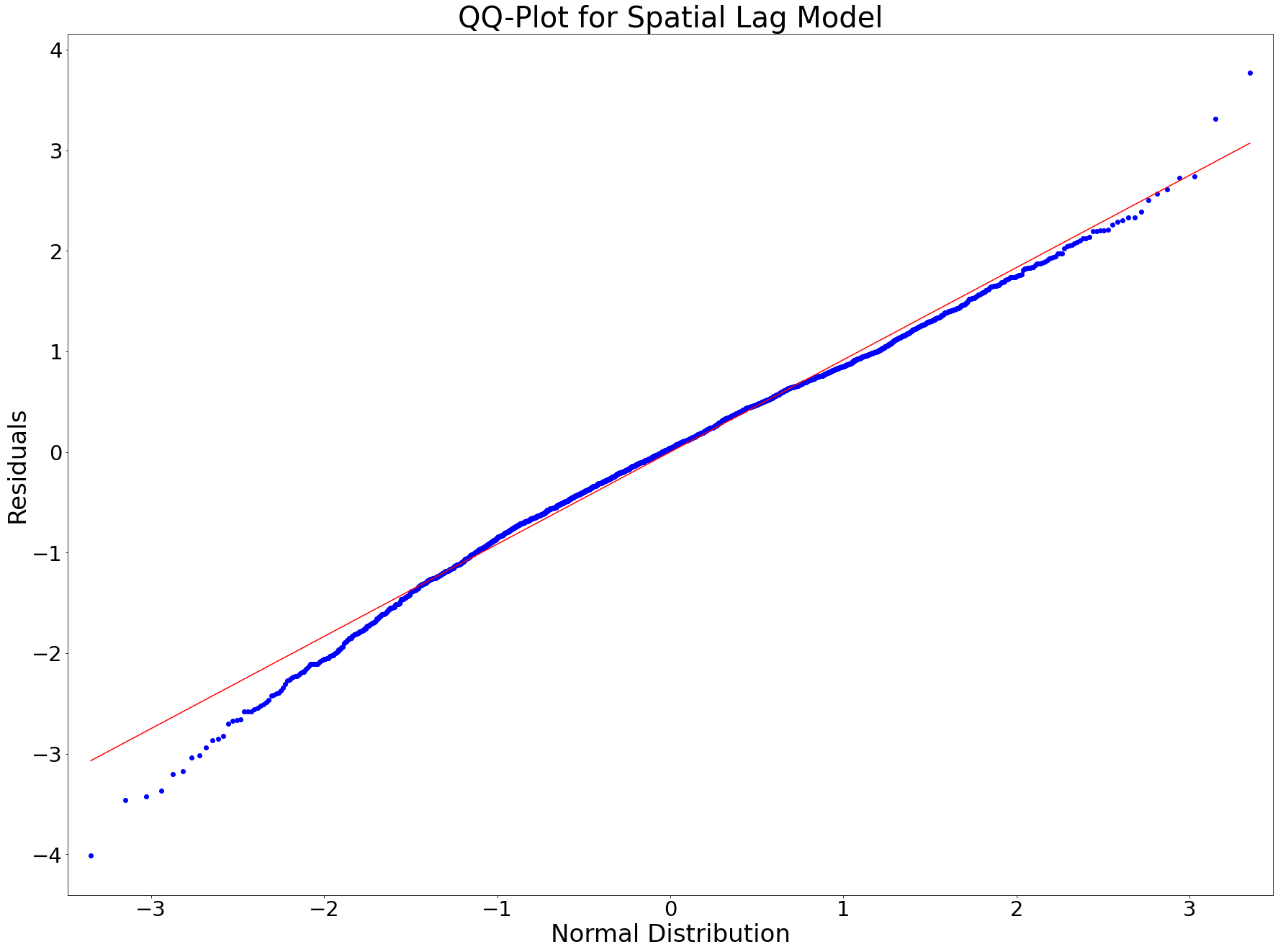
## Normality

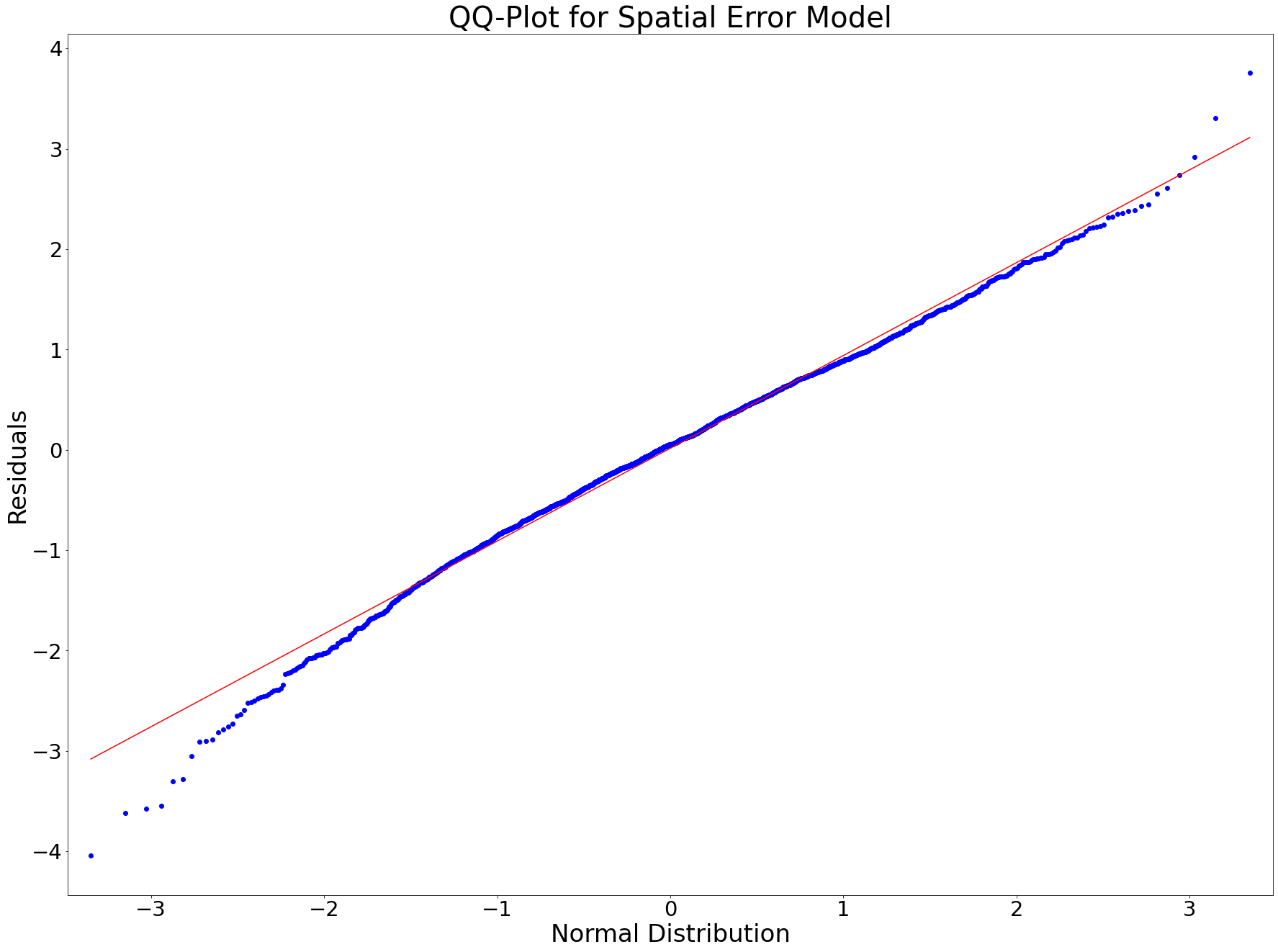
**ORIGINAL**



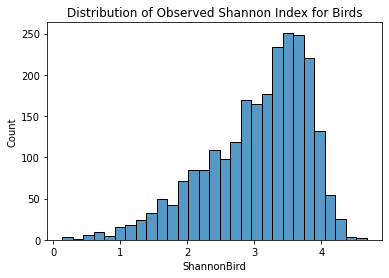


**TRANSFORMED y2**

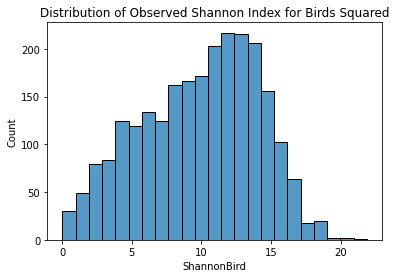
****

****

**ORIGINAL**



**TRANSFORMED y2**

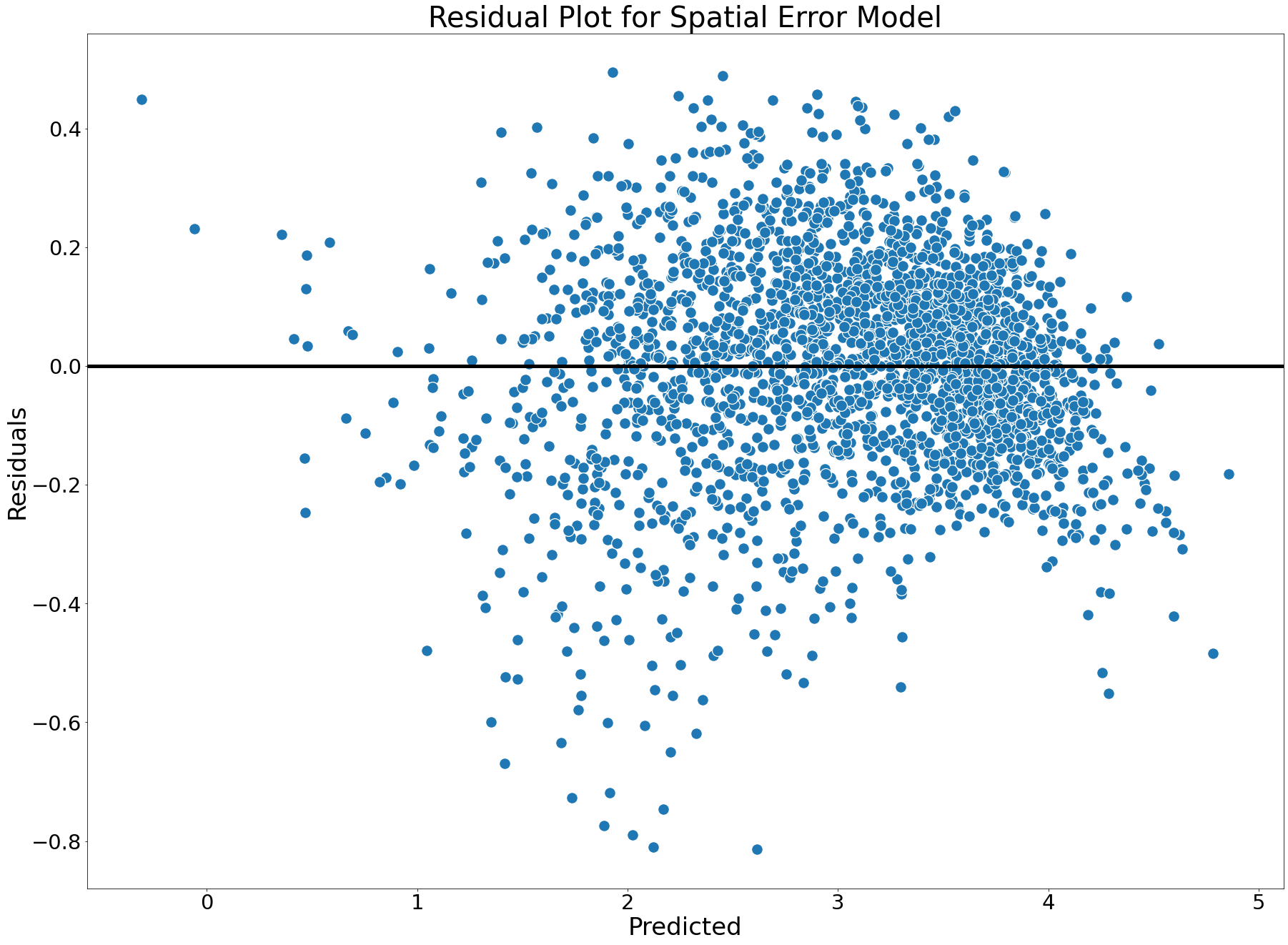
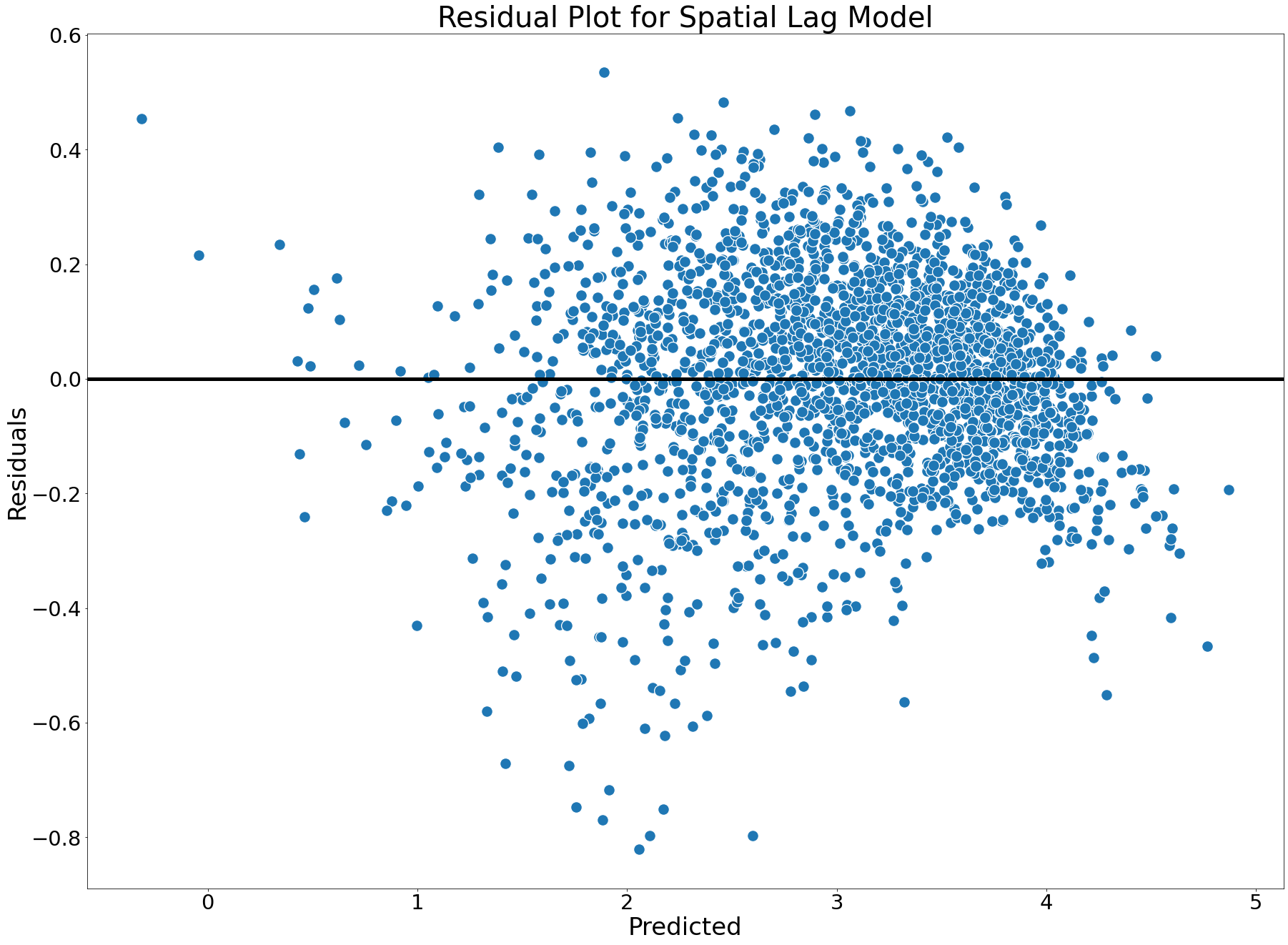




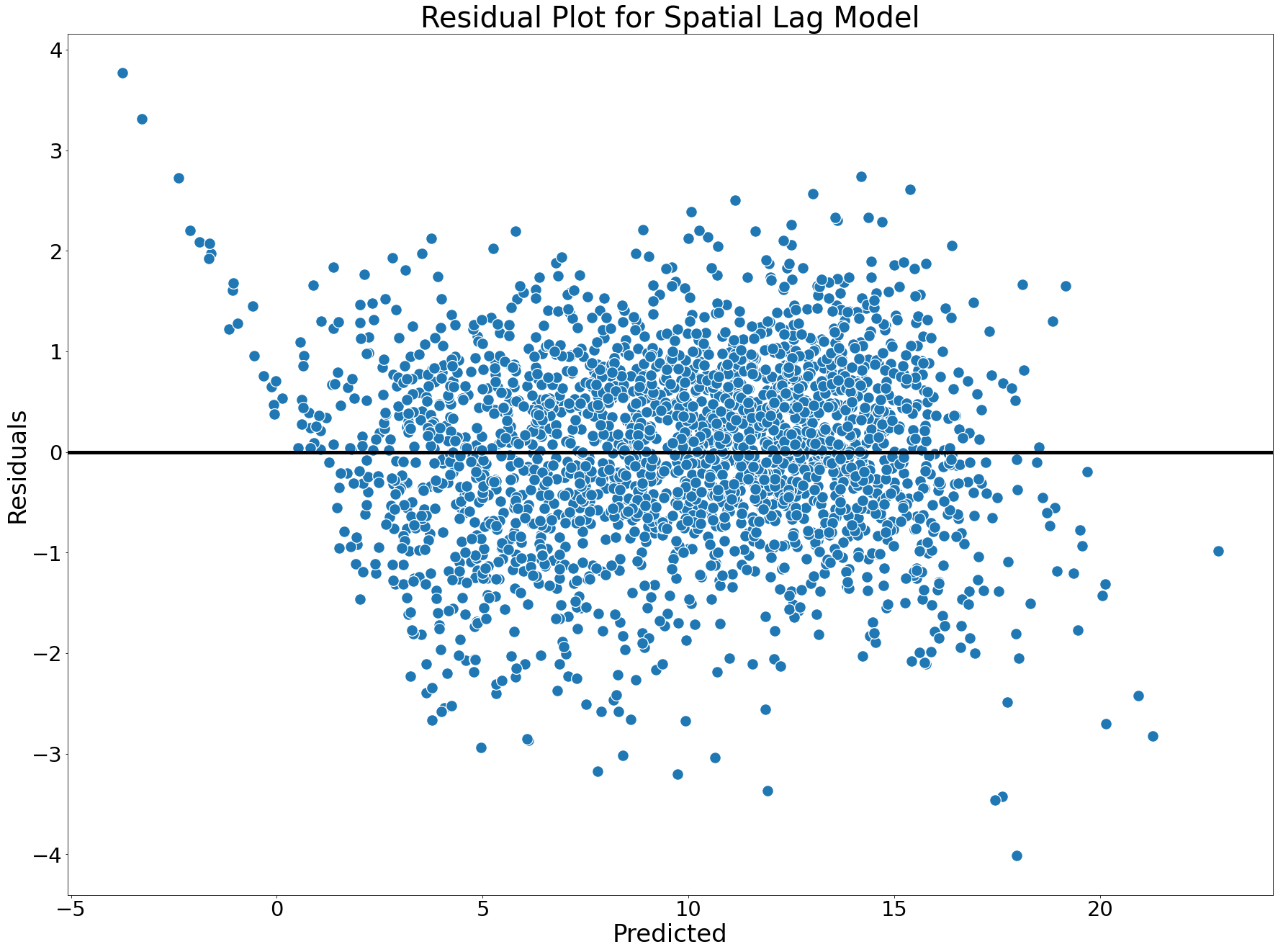
The test-statistics's value went from 374 to 82 , so although it is still a rejection region p-value, it has gotten much better. Additionally, normality is not a huge concern

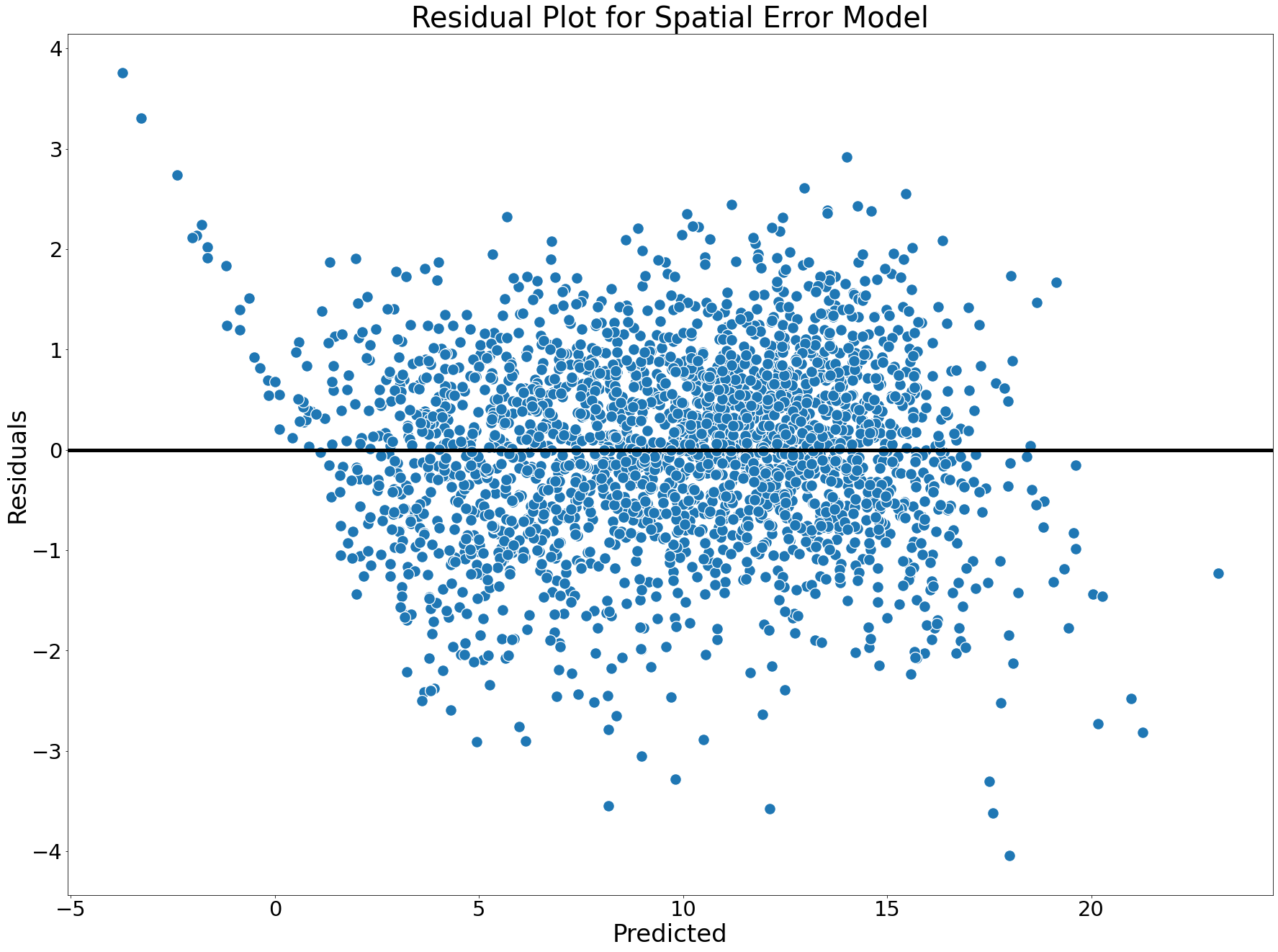
## Linearity

**ORIGINAL**

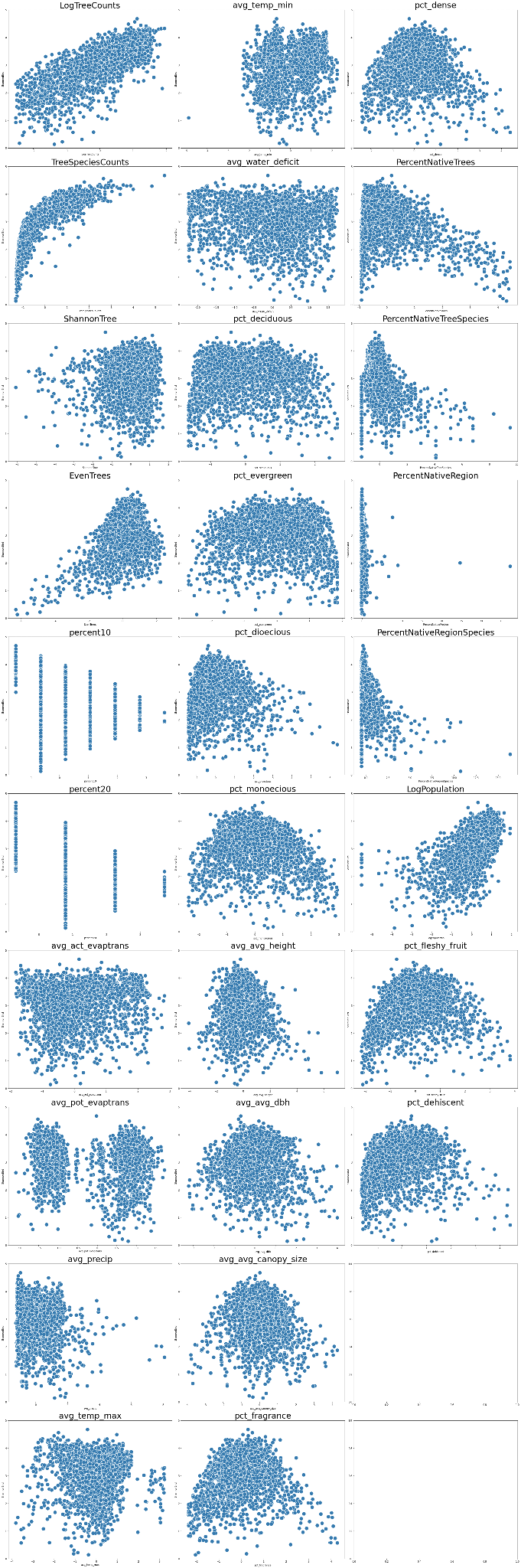


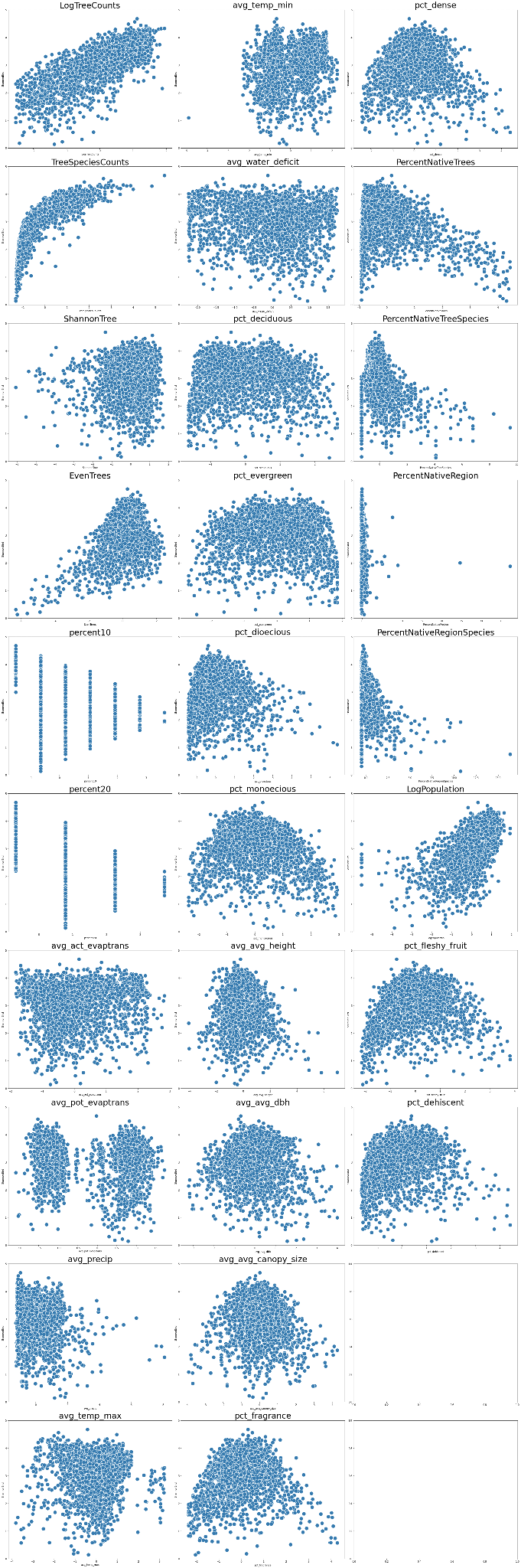
**TRANSFORMED : y2**



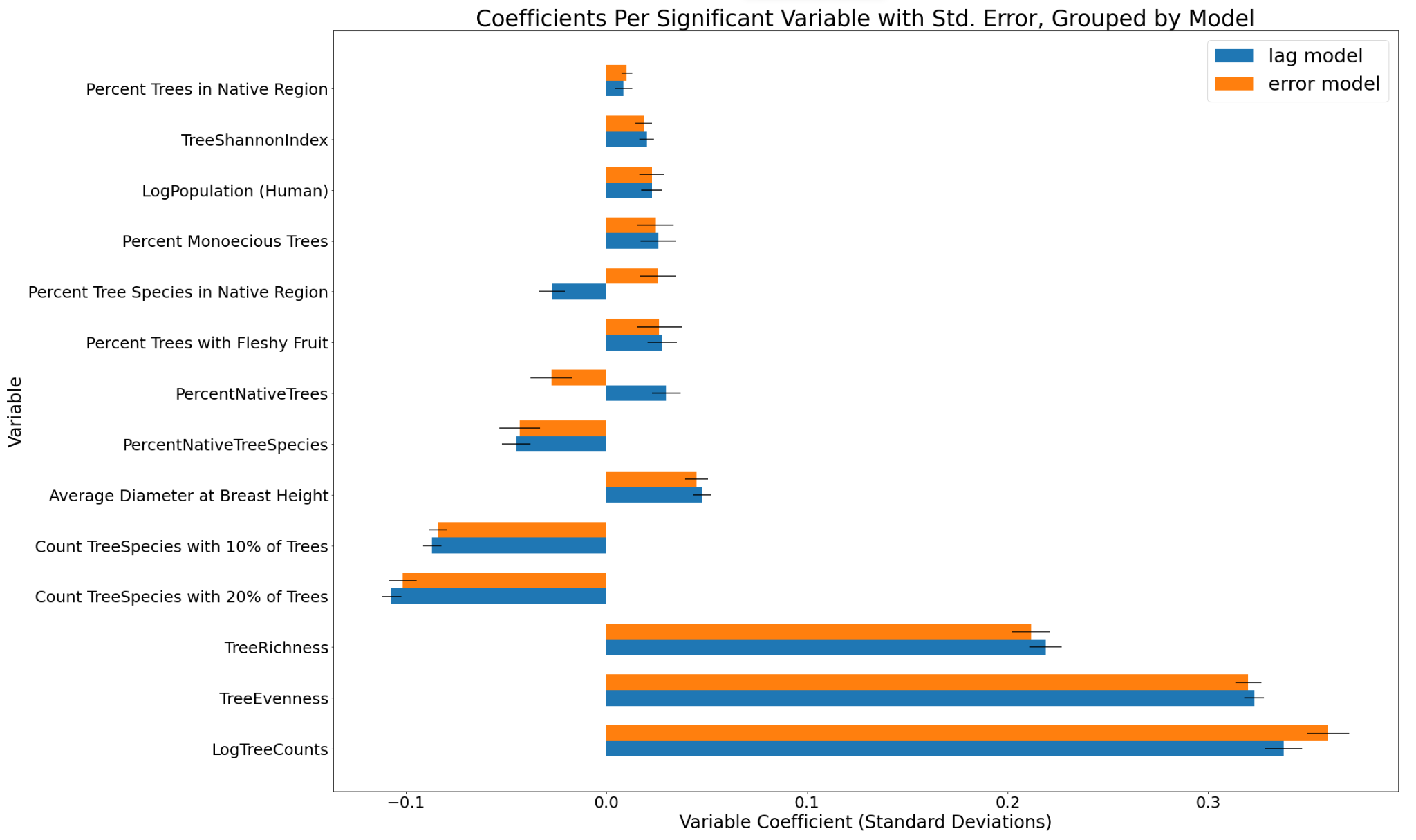


# Predictor By Response

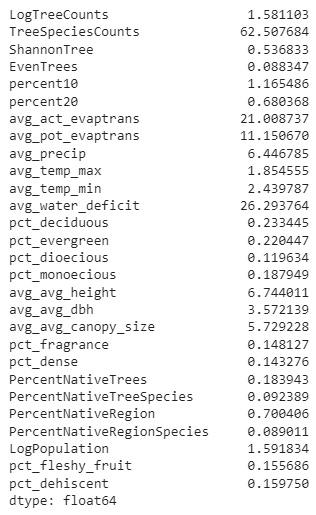
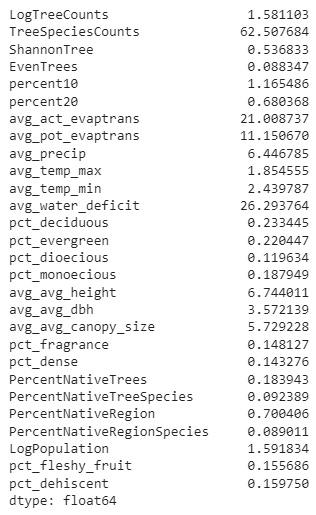


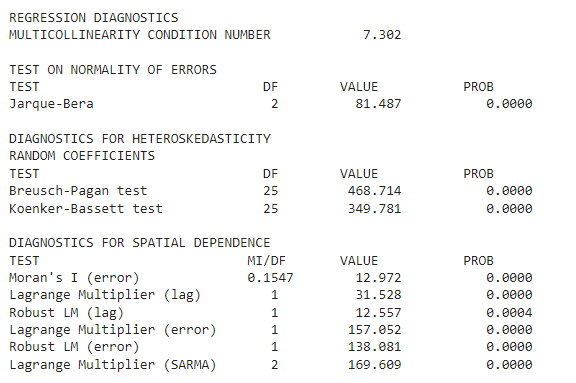


# Spatial Model Results

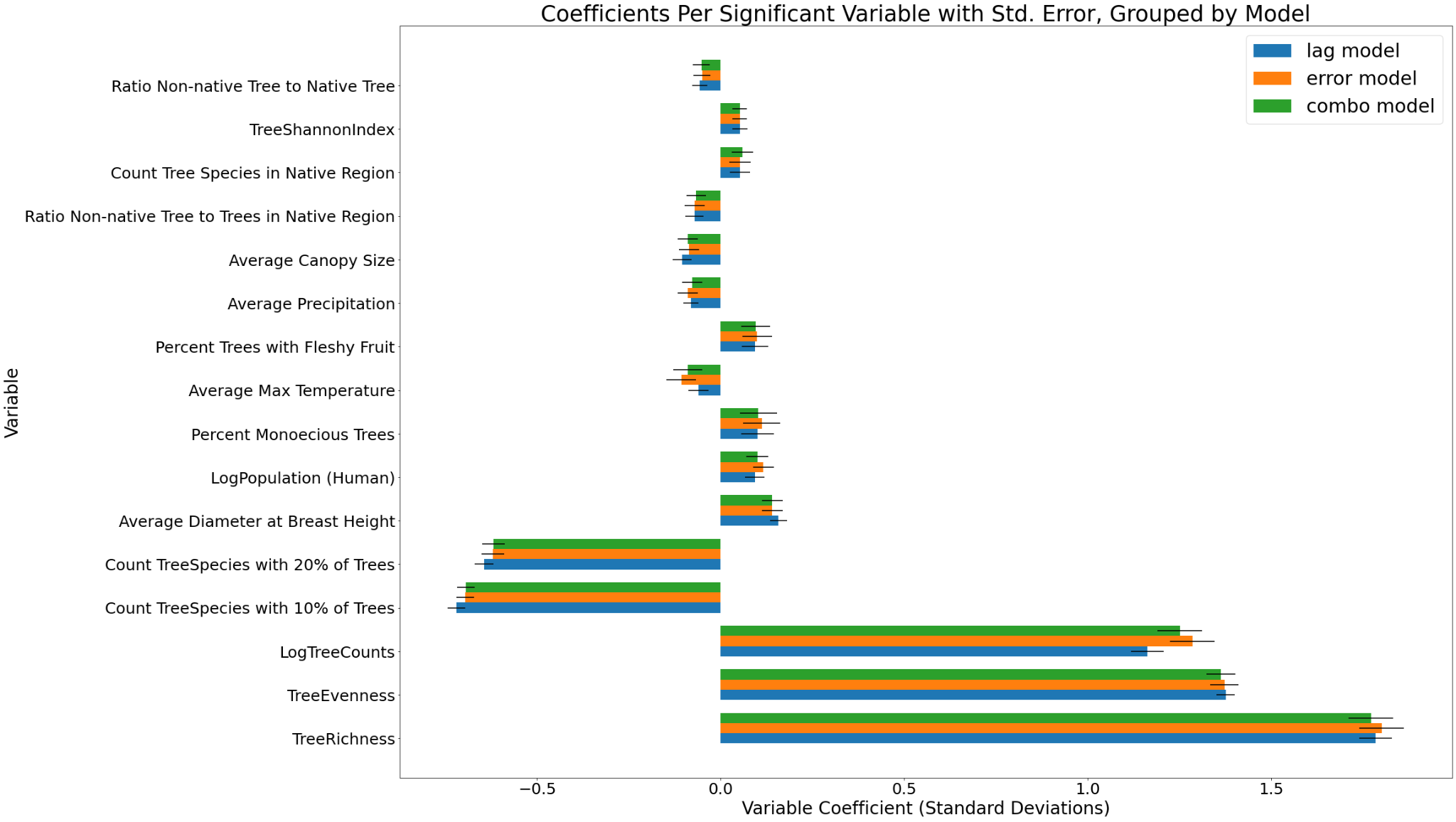


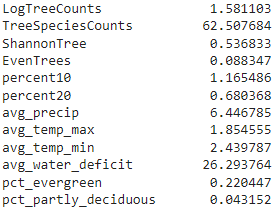
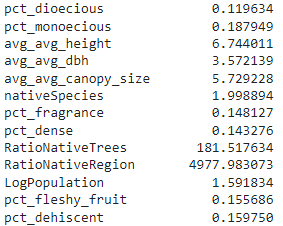
Standard deviations of the variables

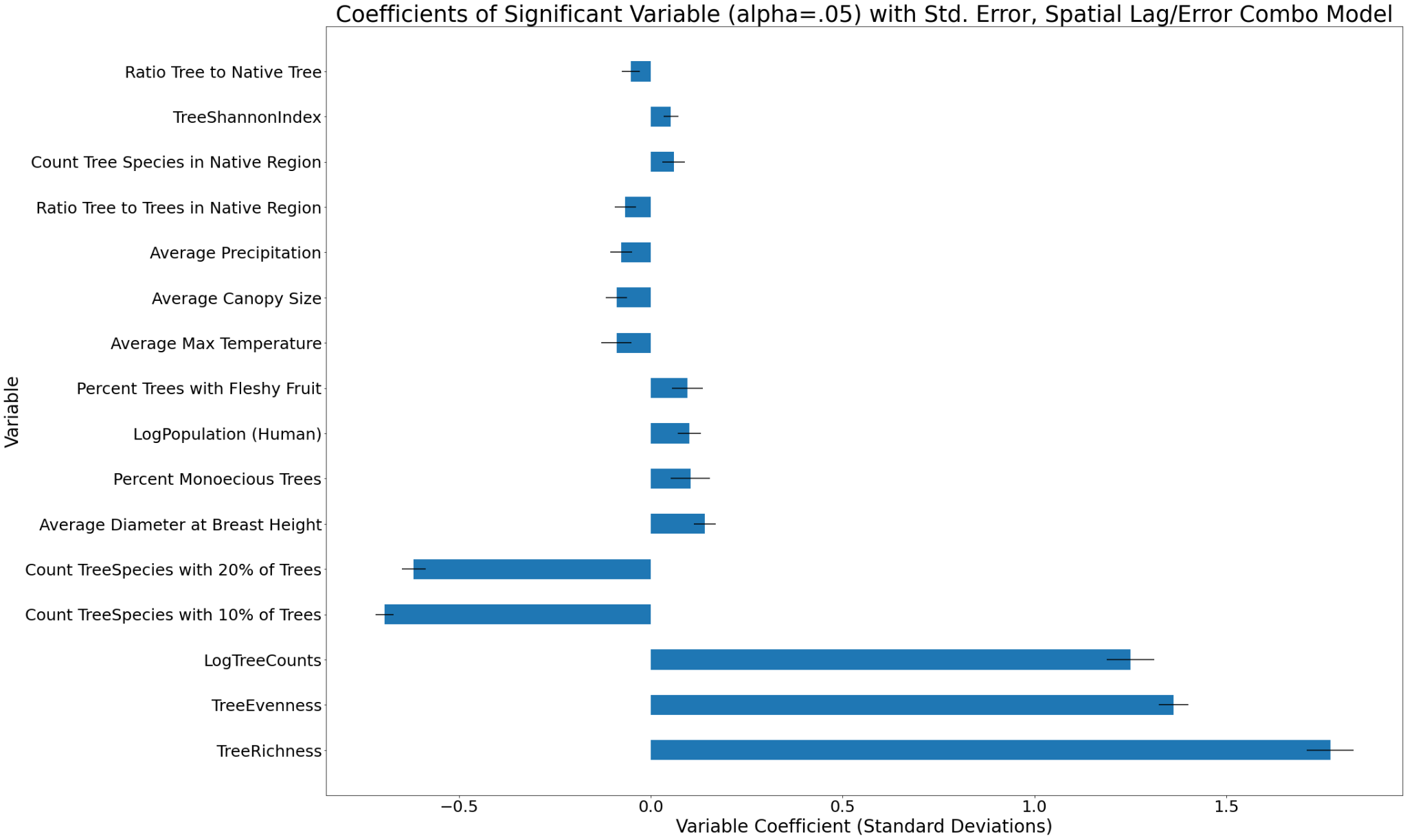




using Y^2, with removed intercorrelated variables, native ratios instead of native percents

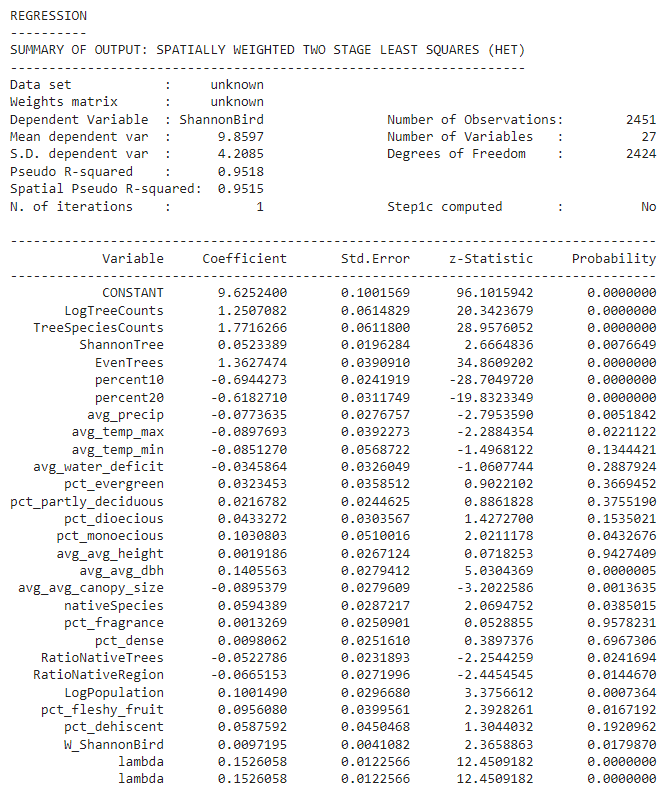
alpha=0.05, using GM\_Combo\_Het model which combines lag and error 





Combo Model Results

We get the same variables if we use alpha = 0.10



Both Spatial Lag (W\_ShannonBird) and Spatial Error (lambda) terms are significant

## Spatial Lag Understanding

In a regression context, spatial effects pertain to two categories of specifications. One deals with spatial dependence, or its weaker expression, spatial autocorrelation, and the other with spatial heterogeneity. 3 The latter is simply structural instability, either in the form of non-constant error variances in a regression model (heteroskedasticity) or in the form of variable regression coefficients

In the standard linear regression model, spatial dependence can be incorporated in two distinct ways: as an additional regressor in the form of a spatially lagged dependent variable (Wy), or in the error structure (E[εiεj ] ≠ 0). The former is referred to as a spatial lag model and is appropriate when the focus of interest is the assessment of the existence and strength of spatial interaction. Spatial dependence in the regression disturbance term, or a spatial error model is referred to as nuisance dependence. This is appropriate when the concern is with correcting for the potentially biasing influence of the spatial autocorrelation, due to the use of spatial data (irrespective of whether the model of interest is spatial or not).

Formally, a spatial lag model, or a mixed regressive, spatial autoregressive model is expressed as y = ρWy + Xβ + ε, (14.9) where ρ is a spatial autoregressive coefficient, ε is a vector of error terms, and the other notation is as before.Unlike what holds for the time series counterpart of this model, the spatial lag term Wy is correlated with the disturbances, even when the latter are iid. Consequently, the spatial lag term must be treated as an endogenous variable and proper estimation methods must account for this endogeneity (OLS will be biased and inconsistent due to the simultaneity bias). The endogeneity of the spatially lagged dependent variable can be addressed by means of an instrumental variables or two-stage least squares (2SLS) approach. Under a set of reasonable assumptions that are easily satisfied when the spatial weights are based on contiguity, the spatial two-stage least squares estimator achieves the consistency and asymptotic normality properties of the standard 2SLS

A spatial error model is a special case of a regression with a non-spherical error term, in which the off-diagonal elements of the covariance matrix express the structure of spatial dependence. Consequently, OLS remains unbiased, but it is no longer efficient and the classical estimators for standard errors will be biased. The model allows for spatial lags in the dependent variable, the exogenous variables, and disturbances. The innovations in the disturbance process are assumed to be heteroskedastic with an unknown form. We formulate multi step GMM/IV-type estimation procedures for the parameters of the model. We also give the limiting distributions for our suggested estimators and consistent estimators for their asymptotic variance-covariance matrices. We conduct a Monte Carlo study to show that the derived large-sample distribution provides a good approximation to the actual small-sample distribution of our estimators.

The GMM estimation is carried out in multiple steps. The basic rationale is the following. First, an initial estimation yields a set of consistent (but not efficient) estimates for the model coefficients. The initial consistent estimates provide the basis for the computation of a vector of residuals, say u (here, we do not use separate notation to distinguish the residuals from the error terms, since we always need residuals in practice). The residuals are used in a system of moment equations to provide a consistent (but not efficient) estimate for the error autoregressive coefficient λ. The consistent estimate for λ is used to construct a weighting matrix that is necessary to obtain the optimal (consistent and efficient) GMM estimate of λ in a second iteration. A third step then consists of estimating the regression coefficients (β and ρ, if appropriate) in a spatially weighted regression, using spatially filtered variables that incorporate the optimal GMM estimate of λ. At this point, we could stop the estimation procedure and use the values of the regression coefficients, the corresponding residuals, and λ to construct a joint asymptotic variance-covariance matrix for all the coefficients

Finally, the estimation procedure as outlined in K-P-D only corrects for the presence of spatial autoregressive errors, but does not exploit the general structure of the heteroskedasticity in the estimation of the regression coefficients. The main contribution of K-P-D is to derive the moment equation such that the estimate for λ is consistent in the presence of general heteroskedasticity.

## Nathan Understanding Spatial Models [Link](https://link.springer.com/content/pdf/10.1007/978-94-015-7799-1.pdf)

Uses Spatial two stage least squares (S2SLS) with results and diagnostics [Link](https://cran.r-project.org/web/packages/ivreg/vignettes/Diagnostics-for-2SLS-Regression.html)

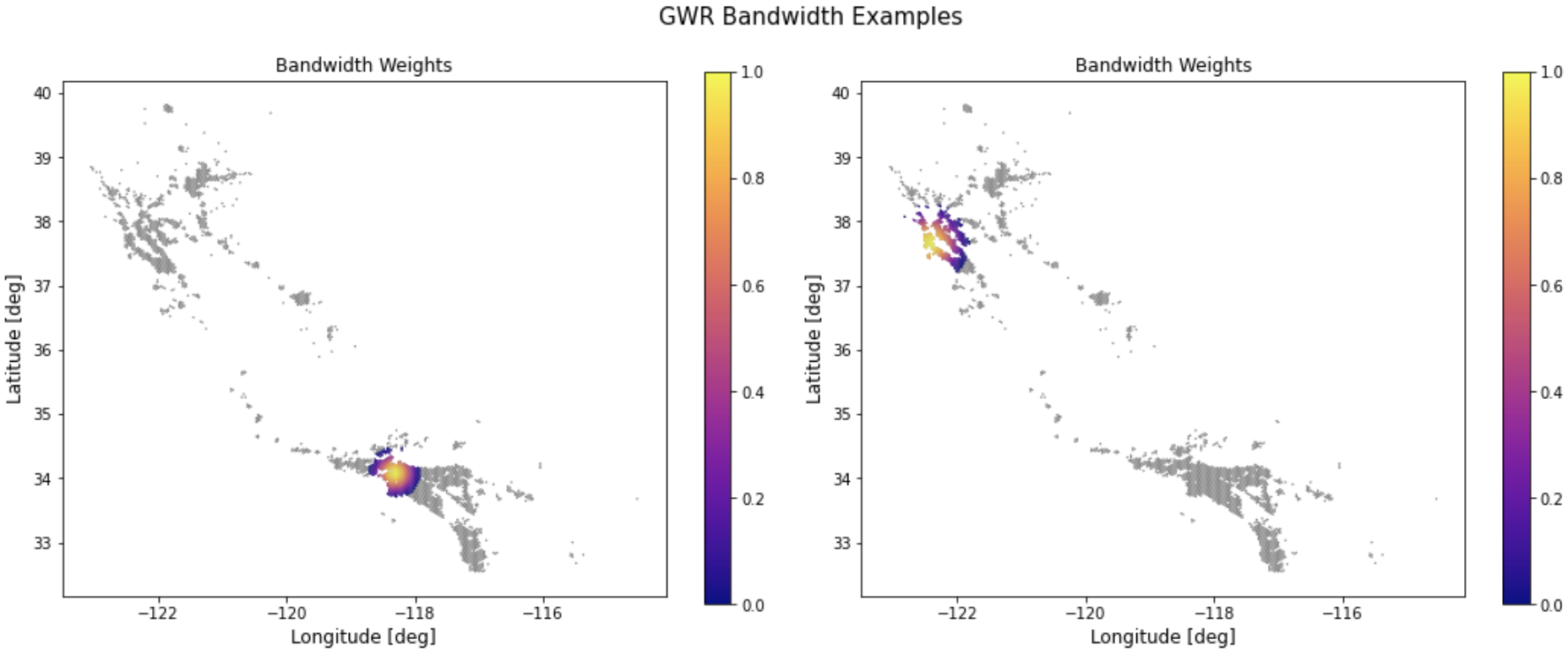
“For the spatial models considered here, the various conditions are typically satisfied when the structure of spatial interaction, which is expressed jointly by the autoregressive coefficient and the weight matrix, is nonexplosive. Formally, this can be assessed by studying the properties of the Jacobian associated with each model, e.g., det (l-pW) in the simple spatial autoregressive formulation.” (60)

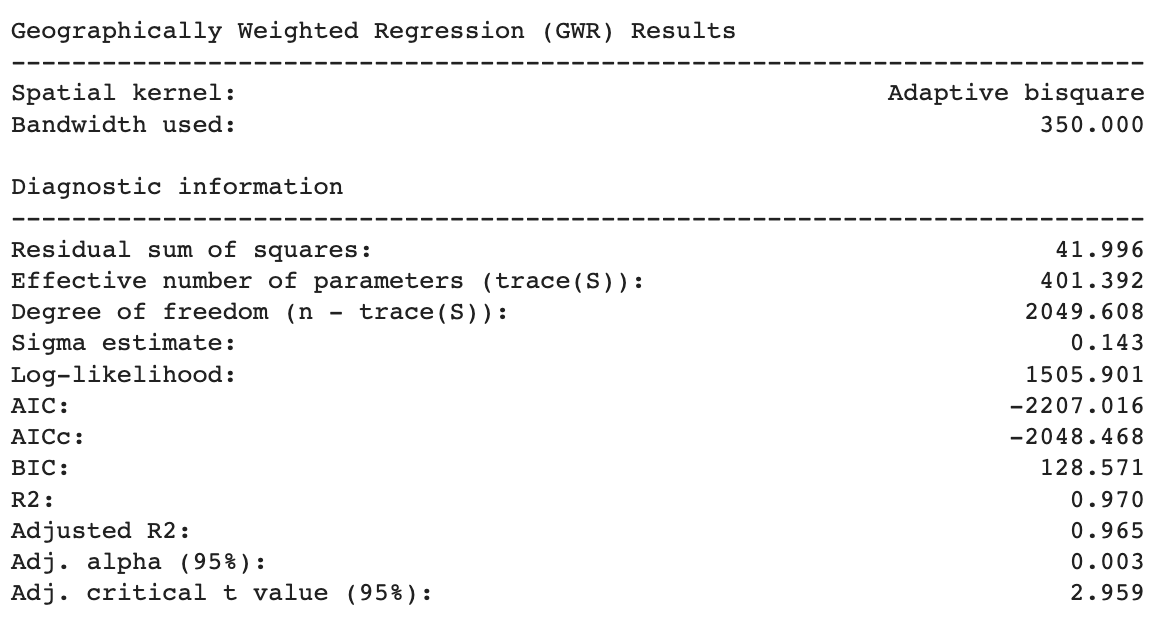
For the spatial model, these conditions are essentially satistied by a non-explosive structure of interaction in the spatial weight matrix, and by imposing non-negativity constraints on the diagonal elements of the estimated error covariance matrix, i.e., on the coefficients in the function for heteroskedasticity(61)

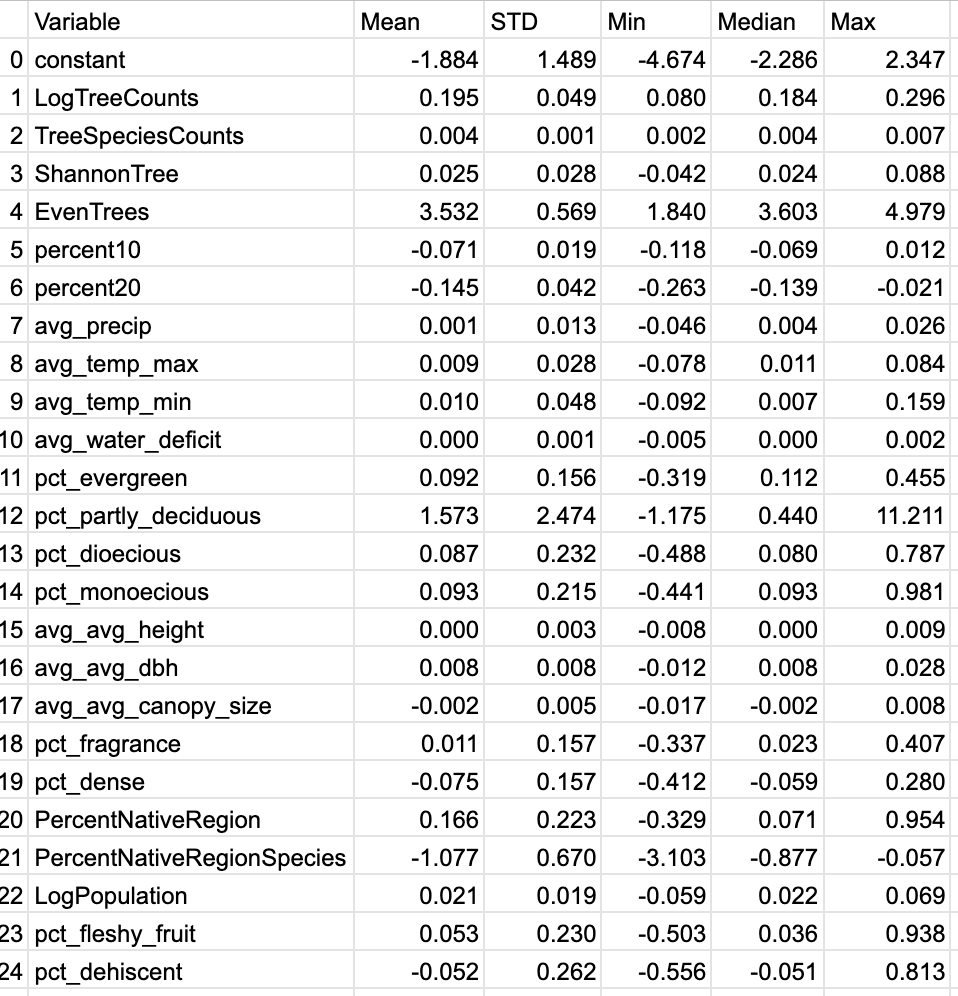
Anselin’s model uses Lagrange Multiplier test (also score test?) for their hypothesis test… and Anselin is the source for the Lag Model

* OLS should suffice for hypothesis testing(70)
* LM Statistic consists of two parts, one pertaining to the heteroskedastic component, the other to the spatially dependent component
  + Implying that equal variance isn’t needed?

# GWR

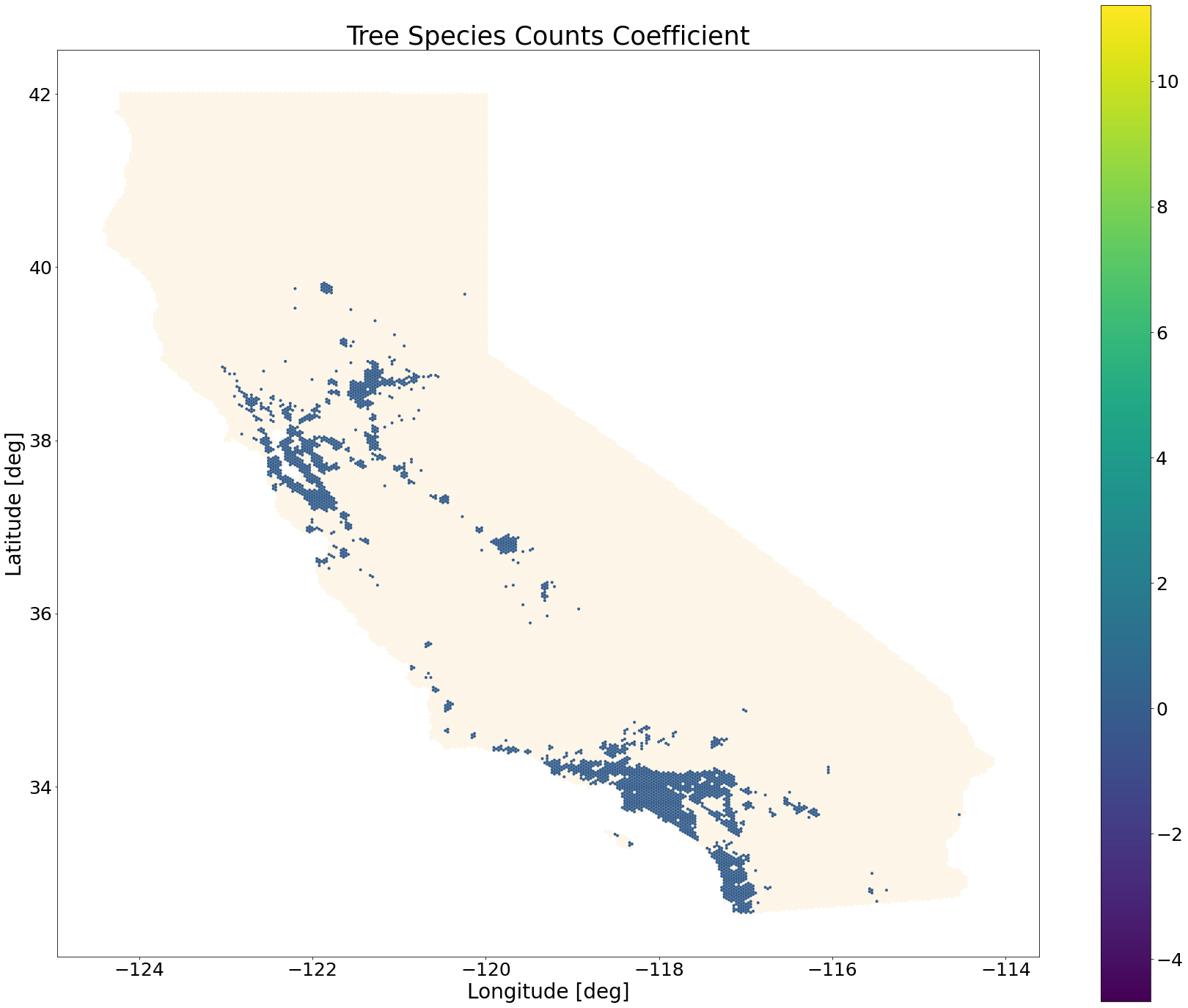






* Things to note from distributions of coefficients
  + ‘EvenTrees’ has largest average coefficient (both mean and median)
    - But not the largest overall
    - The max coefficient of ‘pct\_partly\_deciduous’ (11.211) is the largest single coefficient
      * That value is somewhat concerning as it doesn’t make a ton of sense (may make sense in the context of that single regression)
  + ‘pct\_partly\_deciduous’ has the most variation in its constants, but this may be due to the large max value skewing the data
    - Will look into potential solutions to this, or figuring out if this is incorrect
* Universal color scale
  + Would need use global min and global max of all coefficients in order for the single scale to capture all coefficients
  + May work well for some variables
  + But variables with small variation (in comparison to scale) would show very little on the maps
    - The plots below show the coefficients for the same variable (Tree Species Counts) with two different scales
  + Including 0 in each of the legends at the same point in the scale would result in similar problems due to the differences in the min values between variables
  + Will be best to have different color scales for each variable and make sure to include footnotes to indicate the differences

**With universal color scale (-4.674, 11.211)**



**Color scale found for variable:**

