R Notebook: Descriptive Analysis

Hari Subhash

2016-12-13

Datasets used for the analysis

This analysis uses two datasets.

1. The geographic dataset on the coordinate of district centres and the distances of these from the actual GQ and NSEW highways and the straight lines between nodal districts on these highways.
2. The district characteristics dataset that contains data on the several topics for each district in the dataset.

The geographic dataset

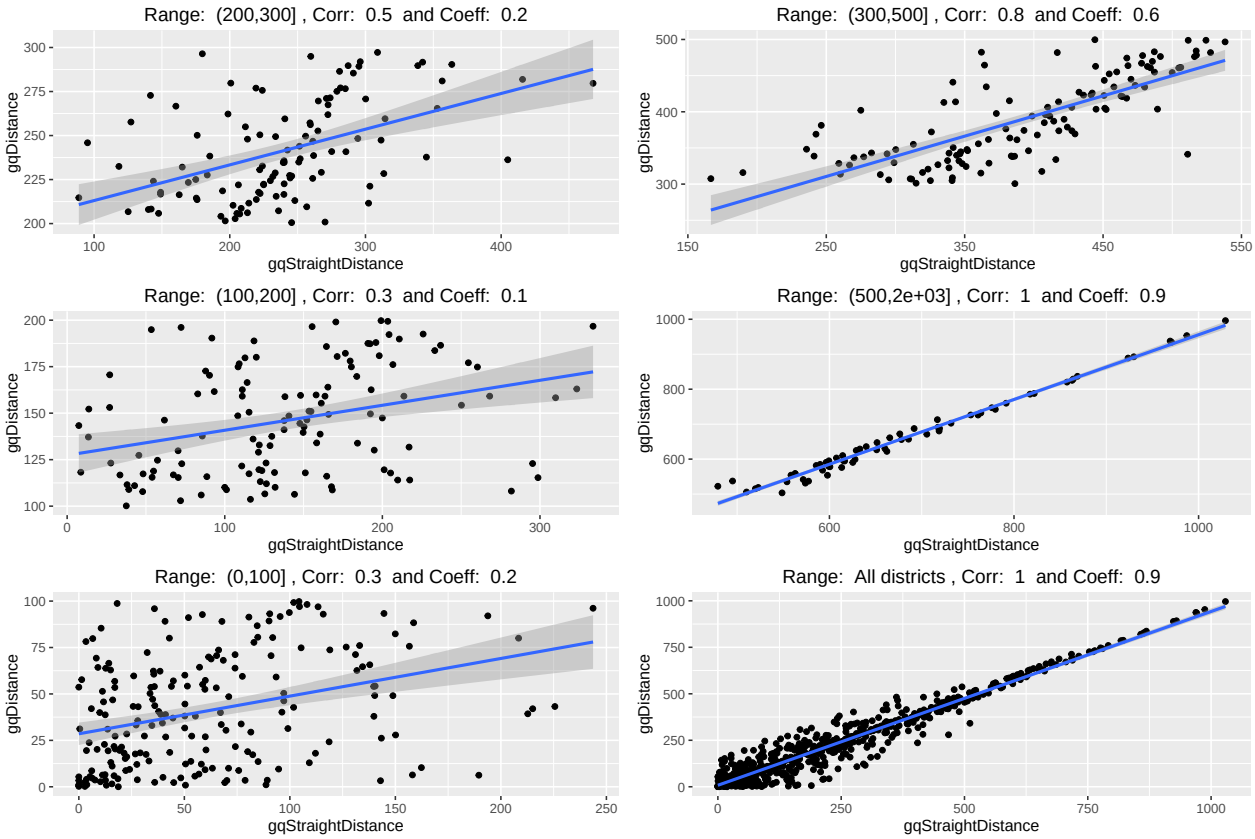
The geographic dataset contains the following variables.

1. id: This is the same id variable that is used in the south asia spatial dataset and ties the districts to all the other variables (i.e economic, demographic etc).
2. state: Name of the state
3. district: Name of the district
4. long: The longitude coordinate of the district centre
5. lat: The latitude coordinate of the district centre
6. gqDistance: The distance from the district centre to the GQ highway
7. nsewDistance: The distance from the district centre to the NSEW highway
8. gqStraightDistance: The distance from the district centres to the straight lines between GQ nodes
9. nsewStraightDistance: The distance from the district centres to the straight

The first part of this analysis is interested in the relationship between the distances from the actual highways to those between the nodal districts. The plots below show the straight line distances between the districts and the actual highway and those from the straight lines between highway nodes.

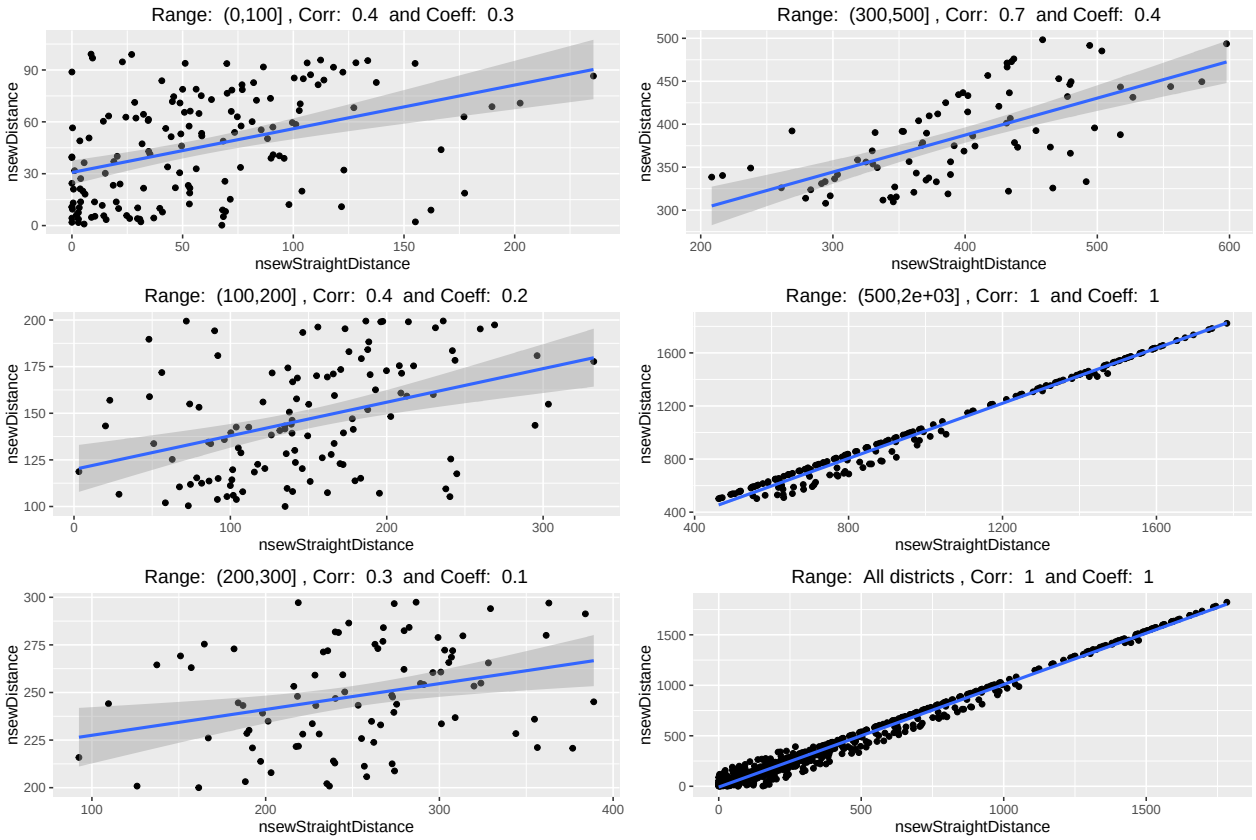
Before considering the relationship between the IV and the distances from the highways, the island and India/China districts are removed from the dataset. The data is also split into 5 ranges based on the distance of a district from the actual highway: 0 to 100, 100 to 200, 200 to 300, 300 to 500 and 500 to 2000. This would give us a better sense for how well the IV predicts the distance from the actual highway for districts that are at different distances from the highway.

GQ Actual vs Straight line distances



The plots above show the relationship between the IV and the actual distance from the highway. It seems that the IV is a better predictor of the distances from the highway when the districts are further away from the actual highway. Overall, the relationship is positive across all ranges of distance from the highway.

NSEW Actual vs Straight line distances



The plots for NSEW show a similar trend as GQ. However, the slopes are smaller.