REGRESSION ANALYSIS OF THE DIFFERENTIAL PRIVACY ALGORITHM RUNTIME

U.S. Census Bureau – Research and Methodology Directorate

Simson Garfinkel – Senior Computer Scientist for Confidentiality and Data Access

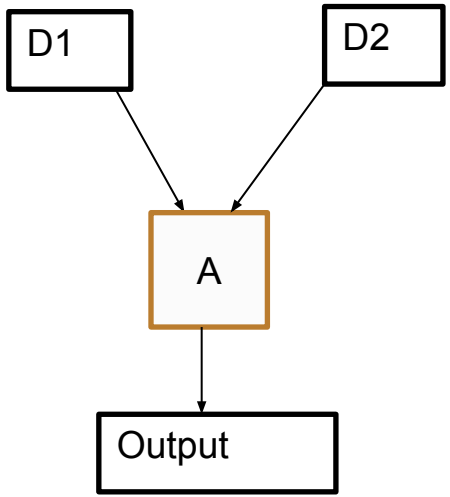
CIVIC DIGITAL FELLOWSHIP

United States™ Census Bureau

VIKRAM RAO
The George Washington University
Systems Engineering

INTRODUCTION

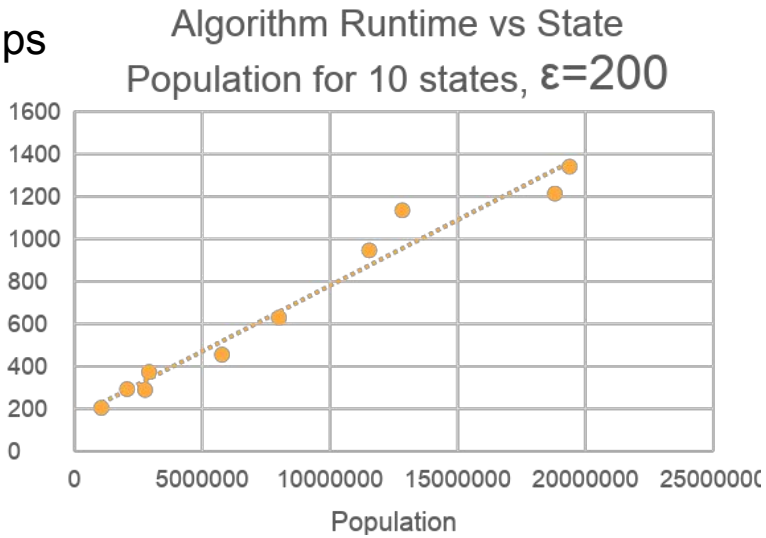
- I am working with the Disclosure Avoidance Team (R&M Directorate), to implement the Differential Privacy Algorithm
- Differential Privacy works by adding random noise to data sets so that individual records can't be identified
- The DP Algorithm is currently undergoing development
- It's important to understand the algorithm runtime
- I use regression to model runtime as a function of 5 variables



CIVIC DIGITAL FELLOWSHIP

REGRESSION MODEL

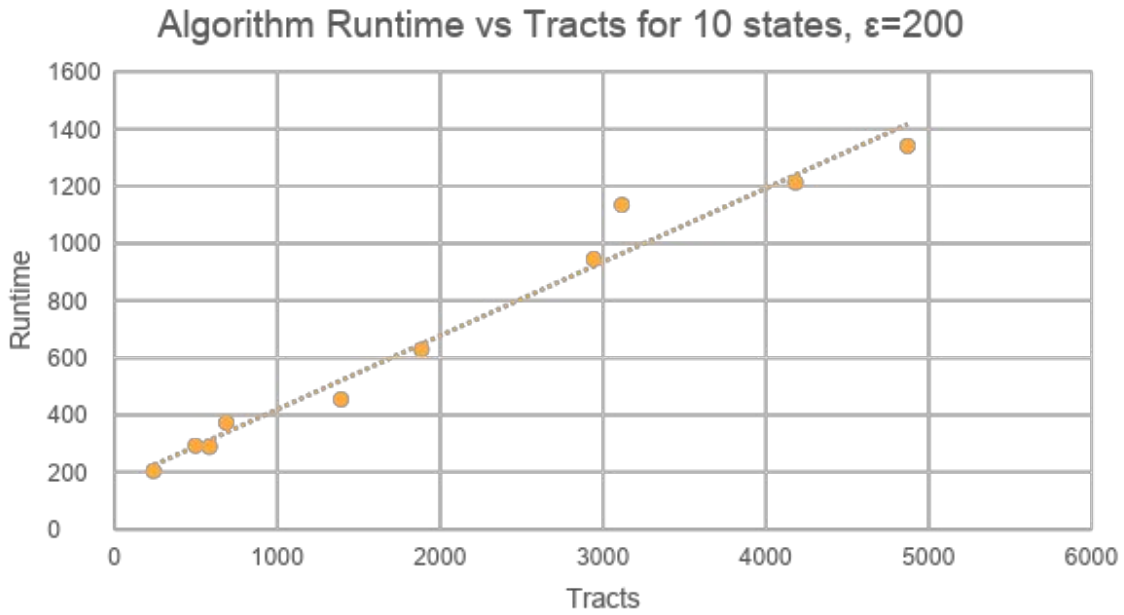
- I have been running the DAS algorithm for various states, and collecting runtimes.
- 3 public variables: population, tracts, and block groups
- linear regression: Y = mX+b
- I also make predictions for other states and compare vs actual runtimes.



	Actual	Predicted
State #1	183.34	215.86
State #2	442.26	446.71

CIVIC DIGITAL FELLOWSHIP

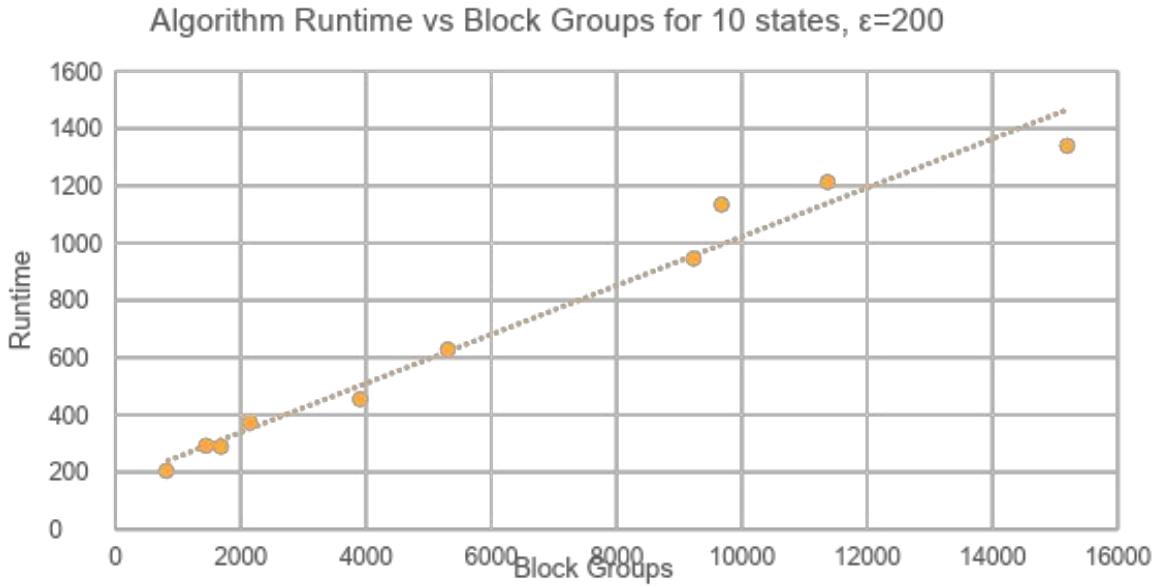
REGRESSION MODEL 2



	Actual	Predicted
State #1	183.34	217.16
State #2	442.26	443.04

CIVIC DIGITAL FELLOWSHIP

REGRESSION MODEL 3



	Actual	Predicted
State #1	183.34	217
State #2	442.26	427.88

CIVIC DIGITAL FELLOWSHIP

DISCUSSION

- Seems that Number of Tracts/Number of Block Groups are better independent variables to use – we get higher R^2 values.
- Now consider 2 more variables:
- The algorithm works by solving a series of optimization problems, at state level, tract level, and block level.
- There are thousands of problems to solve – we can examine two key characteristics, *file size* and *number of non-zero values*
- *File size* tells us the size (in megabytes of all files)

  Non-zero values tells us the number of non-zero values in the optimization problem's constraints – the more non-zeroes, the more difficult to solve

REGRESSION MODEL 4

- Using these 2 new variables:
- For file size, we get an R^2 value of 0.9454

	Actual	Predicted
State #1	183.34	167.48
State #2	442.26	490.34

- For Non-zeroes, we get an R^2 value of 0.963

	Actual	Predicted
State #1	183.34	165.58
State #2	442.26	480.74

CIVIC DIGITAL FELLOWSHIP

MULTIPLE REGRESSION

- We can perform multiple regression with all 5 variables:
  $Y = m_1 * X_1 + m_2 * X_{2.....} + b$

- We get an R^2 value > 0.99

Predictions	Actual	Predicted
State #1	183.34	173.77
State #2	442.26	434.28

CIVIC DIGITAL FELLOWSHIP

CONCLUSIONS/FUTURE WORK

- Troubleshooting the DAS algorithm is ongoing work – regression helps diagnose which variables are important to consider

- Additional work performed:
  Replaced coefficients in optimization problem to test effect on performance
  Isolated specific optimization problems to check their importance

- Future work:
Continue to use data science and statistical methods to examine algorithm performance

CIVIC DIGITAL FELLOWSHIP