Paper 140-2017

The Effects of Socio-Economic, Demographic Variables on US Mortality using SAS Visual Analytics with NLMS PUMS

Catherine Loveless-Schmitt

Arlington, VA

catherinelovesdata@gmail.com

ABSTRACT

The NLMS is a database developed for the purpose of studying the effects of demographic and socio-economic characteristics on differentials in US mortality rates. The NLMS is based on a multistage stratified sample of the non-institutionalized population of the United States. The NLMS consists of US Census Bureau data from Current Population Survey (CPS) and a subset of the 1980 Census, combined with state-based death certificate information to identify mortality status and cause of death. The file contains a subset of the 39 NLMS cohorts included in the full NLMS that is followed prospectively for 11 years. The file contains approximately 1,222,000 records with over 112,000 identified mortality cases. This presentation demonstrates the differential effects of mortality rates in visual displays.

INTRODUCTION

Every visualization tells a story. This is a story about mortality using the National Longitudinal Mortality Survey data (NLMS) since it contains so many different social, economic and demographic variables. A more complete introduction to the NLMS will follow in that section.

For example: Are women over 75 who live alone more prone to mortality relative to men over 75 who live alone?

Are Black men (over 75) less likely to die of cardiovascular disease compared to White men (over 75) who died of cardiovascular disease? Is this difference the same among other age groups?

NATIONAL LONGITUDINAL MORTALITY SURVEY (NLMS) DESCRIBED

The NLMS is a database developed for the purpose of studying the effects of demographic and socio-economic characteristics on differentials in US mortality rates. The NLMS is a unique research database because it is based on a multi-stage stratified sample of the non-institutionalized population of the United States. It consists of US Census Bureau data from Current Population Survey (CPS) ad a subset of the 1980 Census, combine d with death certificate information to identify mortality status and cause of death.

The 11 year follow-up file contains a subset of the 39 NLMS cohorts included in the full NLMS that can be followed prospectively for 11 years. The file contains approximately 1,222,000 records with over 112,000 identified mortality cases.

SELECTION OF THE NLMS DATA SET

The National Longitudinal Mortality Survey (NLMS) data was chosen because it is well-researched and a rich source of research and contains so many different demographic types of variables. For the purposes of this demonstration, since the proposed audience is not well-versed in statistics, nor mortality, the presentation was centered on demonstrating how versatile and easy to use the SAS VA application is. Since within SAS VA it is easy to view the data available, and load it. Remember the data must be available to access within the SAS LASR..

VISUALIZATION

SAS VA can help to demonstrate differences in mortality by different age classifications and different demographic socio-economic characteristics

For example: One can perform time series analysis with 11 years of follow-up here categorized by the different months of follow-up. Also, there is a decision tree analysis of all 112,000 cases of death records. A decision tree was created using all possible categories related to mortality. Those were drawn first into the decision tree have the strongest effect, that is the best predictor of mortality, and each subsequent branch has the next strongest effect; then the third, the fourth and so on. Also, amply demonstrated in the decision tree analysis is the ability to scope and parse the resulting decision tree to make it the branches larger, leaves smaller.

STATISTICAL ANALYSIS

Data analysis using the National Longitudinal Mortality Survey (NLMS) uses the usual customary acceptable methods of visual displays. All of those can be used to find out if data is as you expect it to be. When you start to analyze l the data a central tenet of statistical analysis is that one should have an 'a priori' hypothesis before beginning to analyze the data. In fact all we are looking for at a minimal level is .05% level of significance, which means 20% of the time it's likely to get a test that shows significance **by chance** on this data. Therefore, if a user were to run these tests over and over and over again and look for any variable A by any variable B by any variable C we could get a significant result somewhere. For the NLMS, in this example the personal data set contains a comparison between: major occupations by state by Urban or rural by foreign-born/native by living-alone by age-category by race-category by sex (gender) category. Certainly with 1.2 million people you're going to get some significant result somewhere in one of those cells! That's not how statistics was designed. Instead it's an unbiased test applied to the data to determine if there is demographic characteristics affect mortality. Indeed, this is the central reason for using the National Longitudinal Mortality

Survey (NLMS) was to determine did SAS Visual Analytics show the effects that appear in all major reports about mortality that have been produced from this same data set from the National Longitudinal Mortality Survey (NLMS).

SELECTION OF THE NLMS DATA SET

The National Longitudinal Mortality Survey (NLMS) data was chosen because it is well-researched and a rich source of research and contains so many different demographic types of variables. For the purposes of this demonstration, since the proposed audience is not well-versed in statistics, nor mortality, the presentation was centered on demonstrating how versatile and easy to use the SAS VA application is. Since within SAS VA it is easy to view the data available, and load it. Remember the data must be available to access within the SAS LASR.

VISUALIZATION

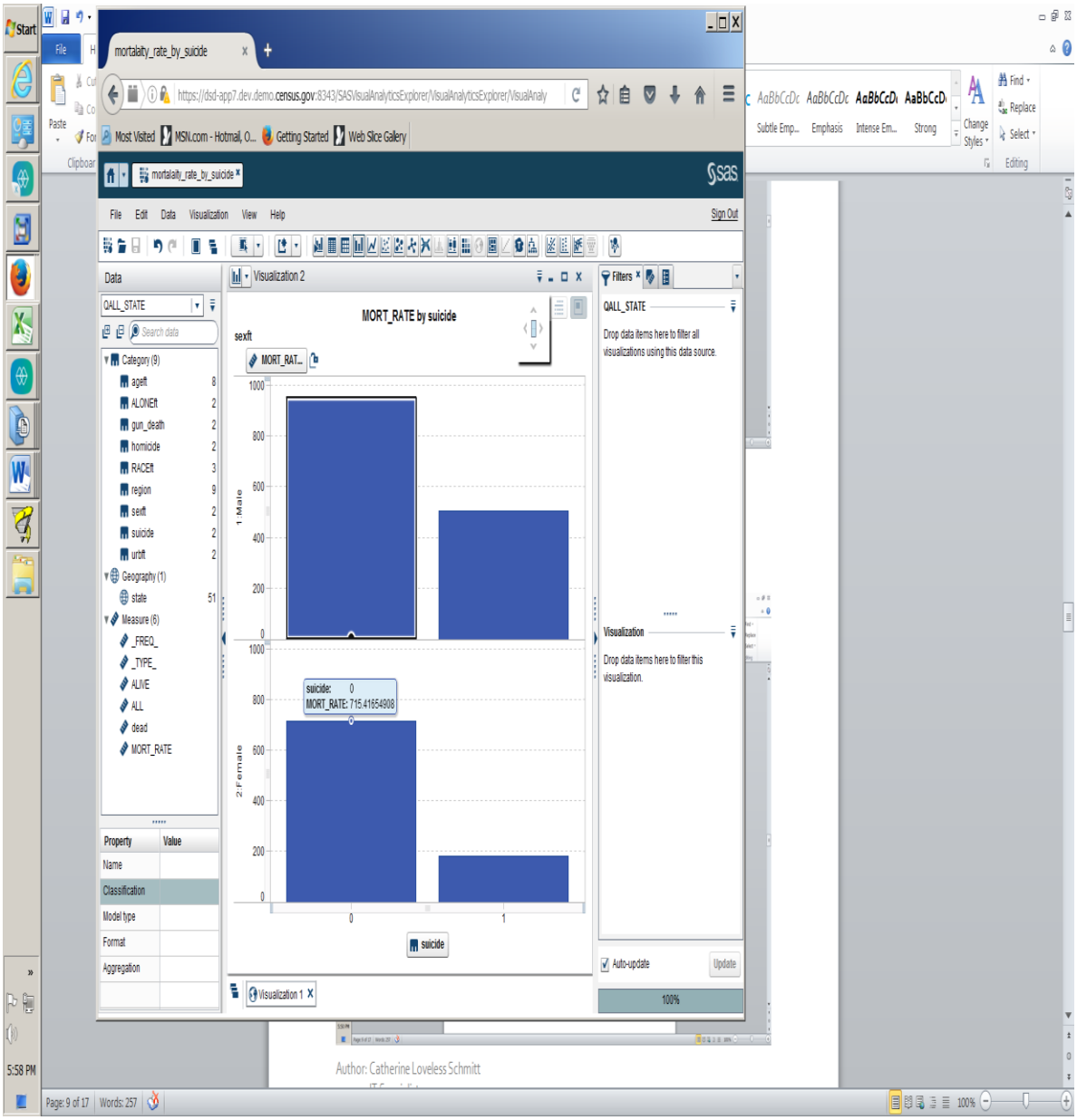


FIGURE 1. Filter: Gun Deaths (suicide or non-suicide by Gender)

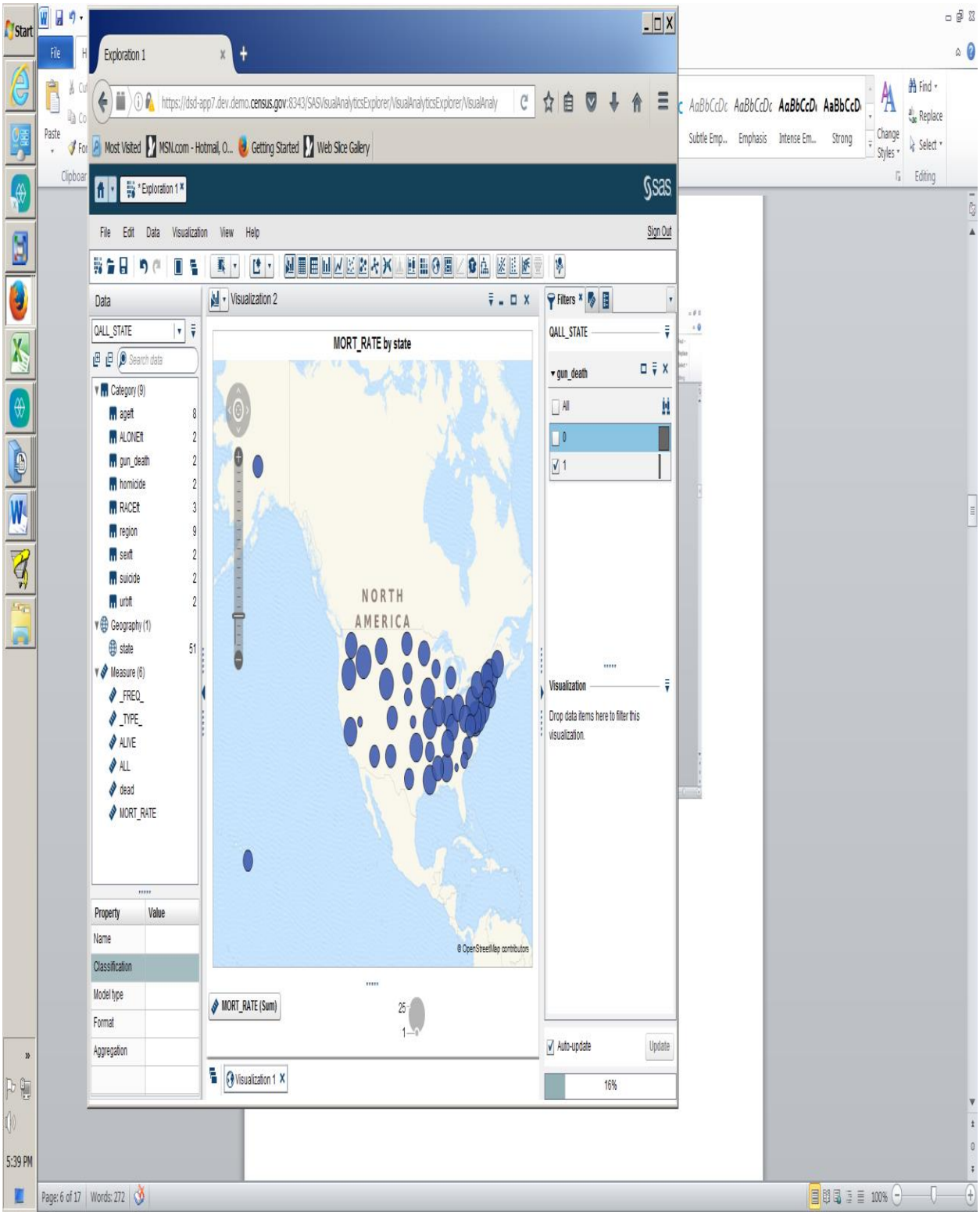


FIGURE 2. Gun deaths by State.

Note that the circles are much smaller throughout the Midwest.

PART II : STATISTICAL ANALYSIS

The National Longitudinal Mortality Survey (NLMS) National longitudinal mortality survey does allow for viewing and analysis using many different statistical tests, charts, and graphs. There are many different statistical features available:

All of these completed by simply dragging the appropriate icon into the central pane.

Crosstabs	time series analysis
Bar chart	linear regression
Box and whisker plots	logistical regression
Decisions trees	network analysis
Heat map……and much more	

All of these completed by simply dragging the appropriate icon into the central pane.

REGRESSION MODELING EXPLAINED

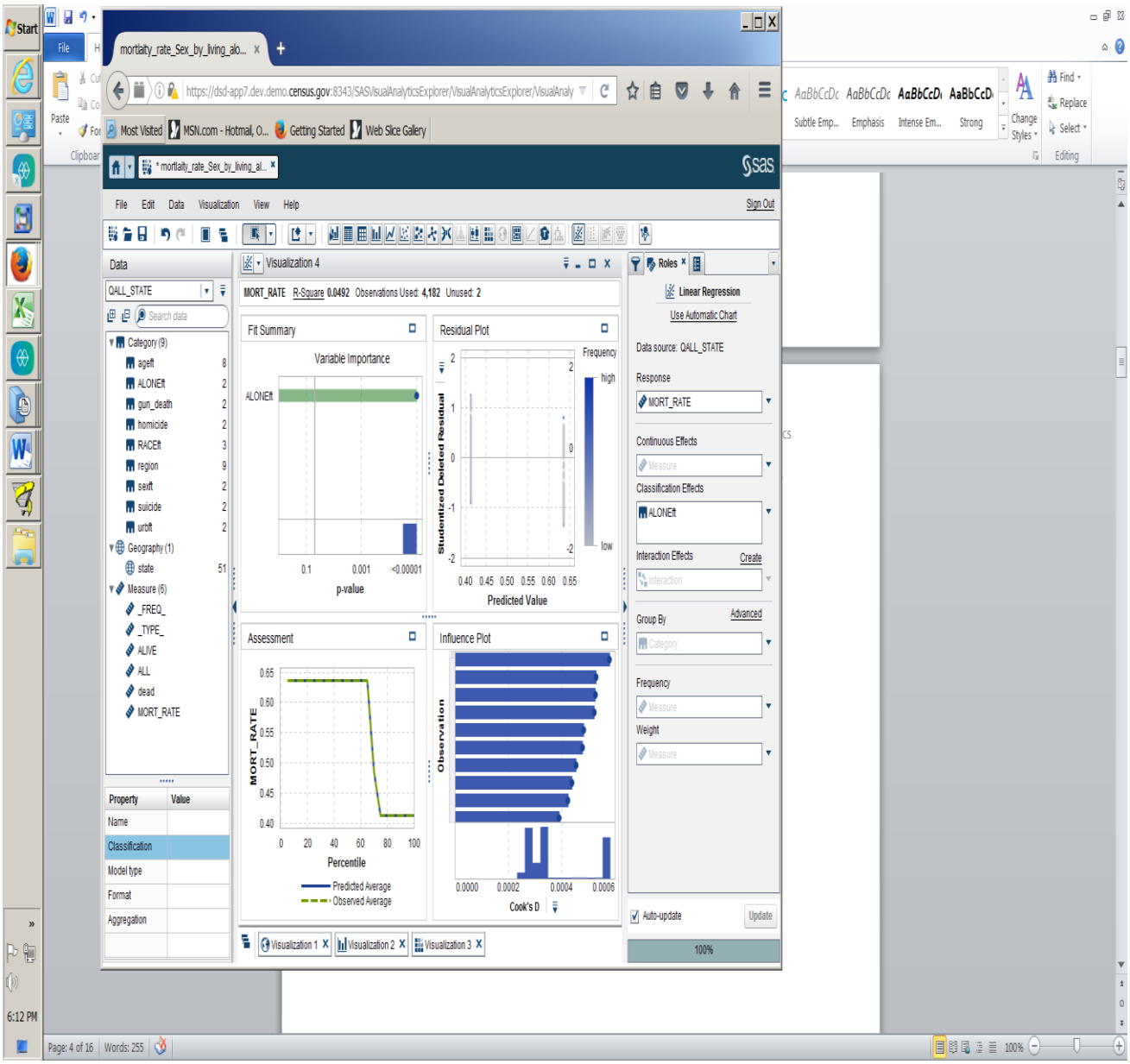


FIGURE 3. Regression demonstrates the effect of Living Alone

The four graphs shown include: Fit Summary; Quality measures ; Residual Plots; Influence Statistics

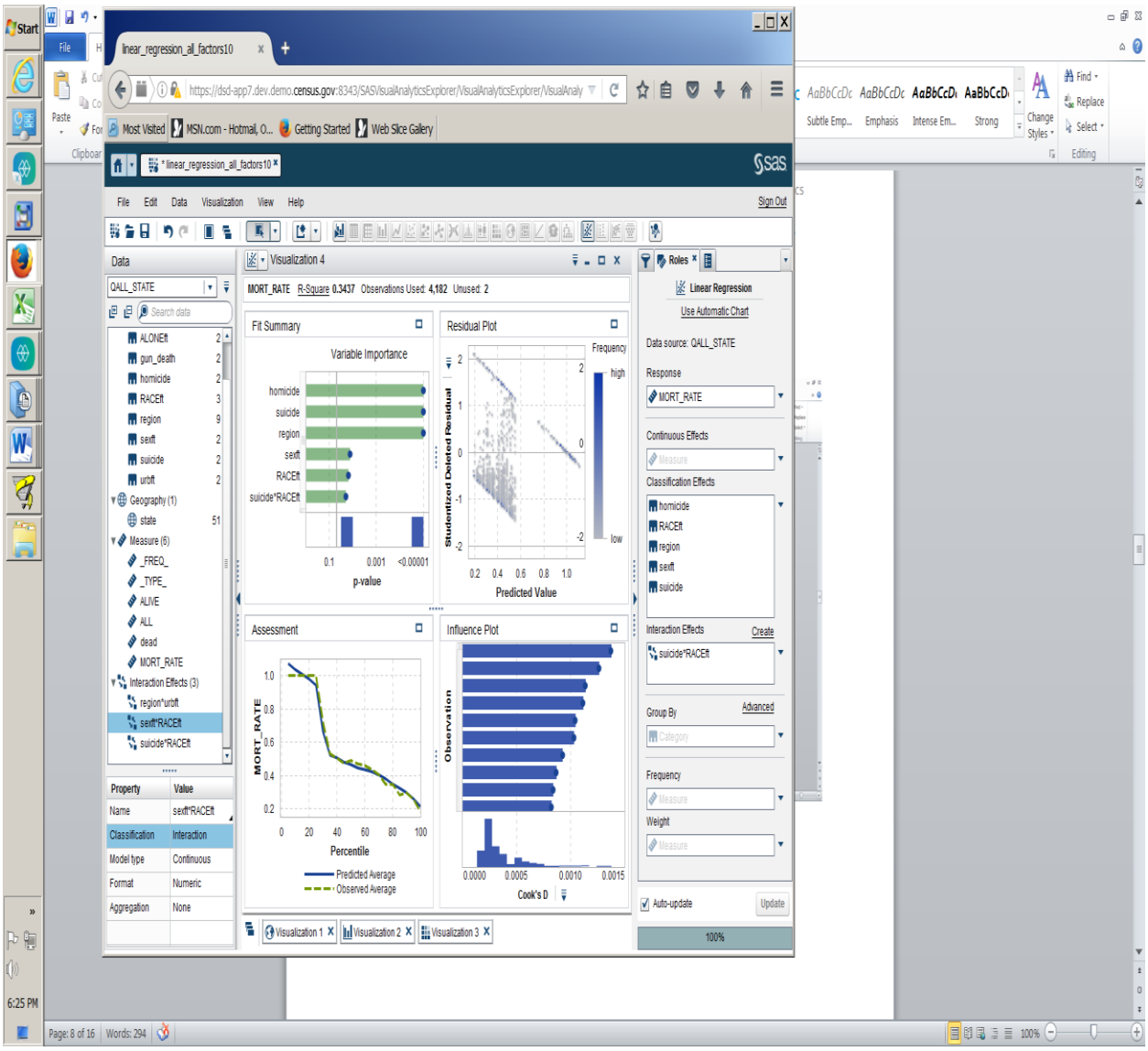


Figure 4: Homicide, race, region, sex, and suicide all provide explanatory value on mortality.

SAS performs maximum likelihood functions upon the model automatically.

SAS Exploration automatically updates the model when any changes are made.

So the user can see changes as classification effects are added

Therefore, the model is automatically updated and refitted with the classification effects

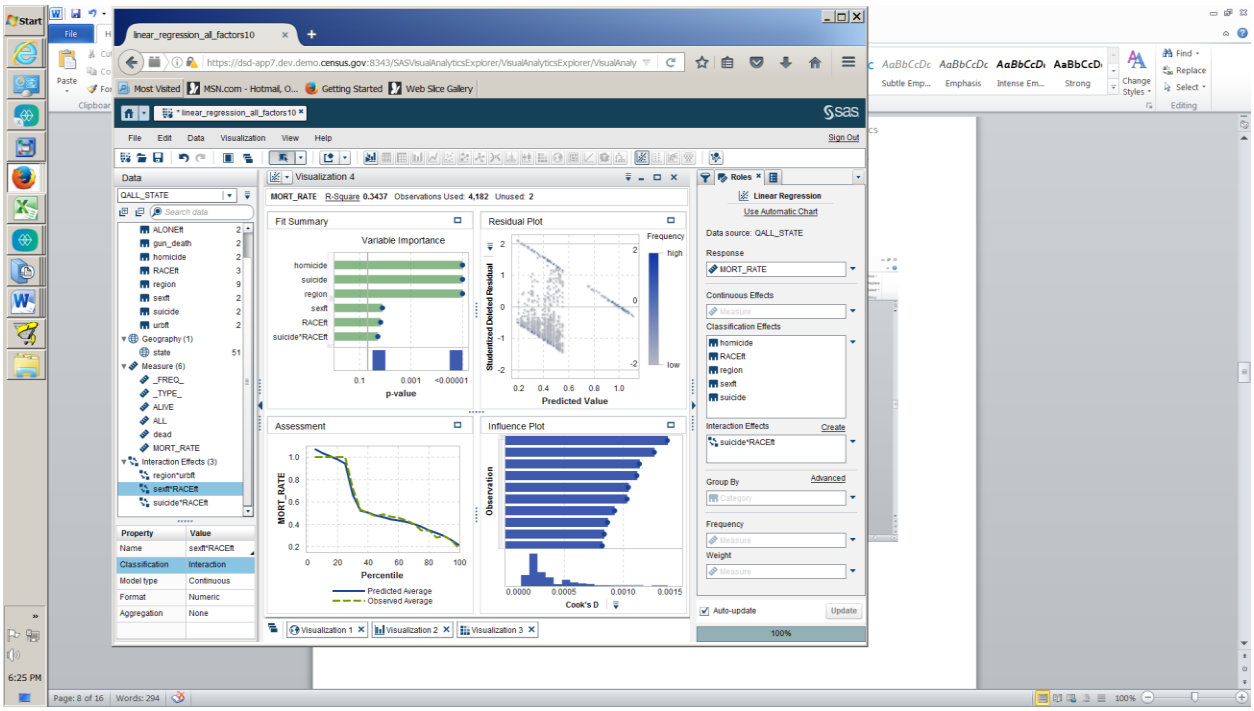


Figure 5: Suicide by Race interaction is added to the Model

Suicide and Race have an interaction.

That is, depending on your race it is more likely, that you will commit suicide.

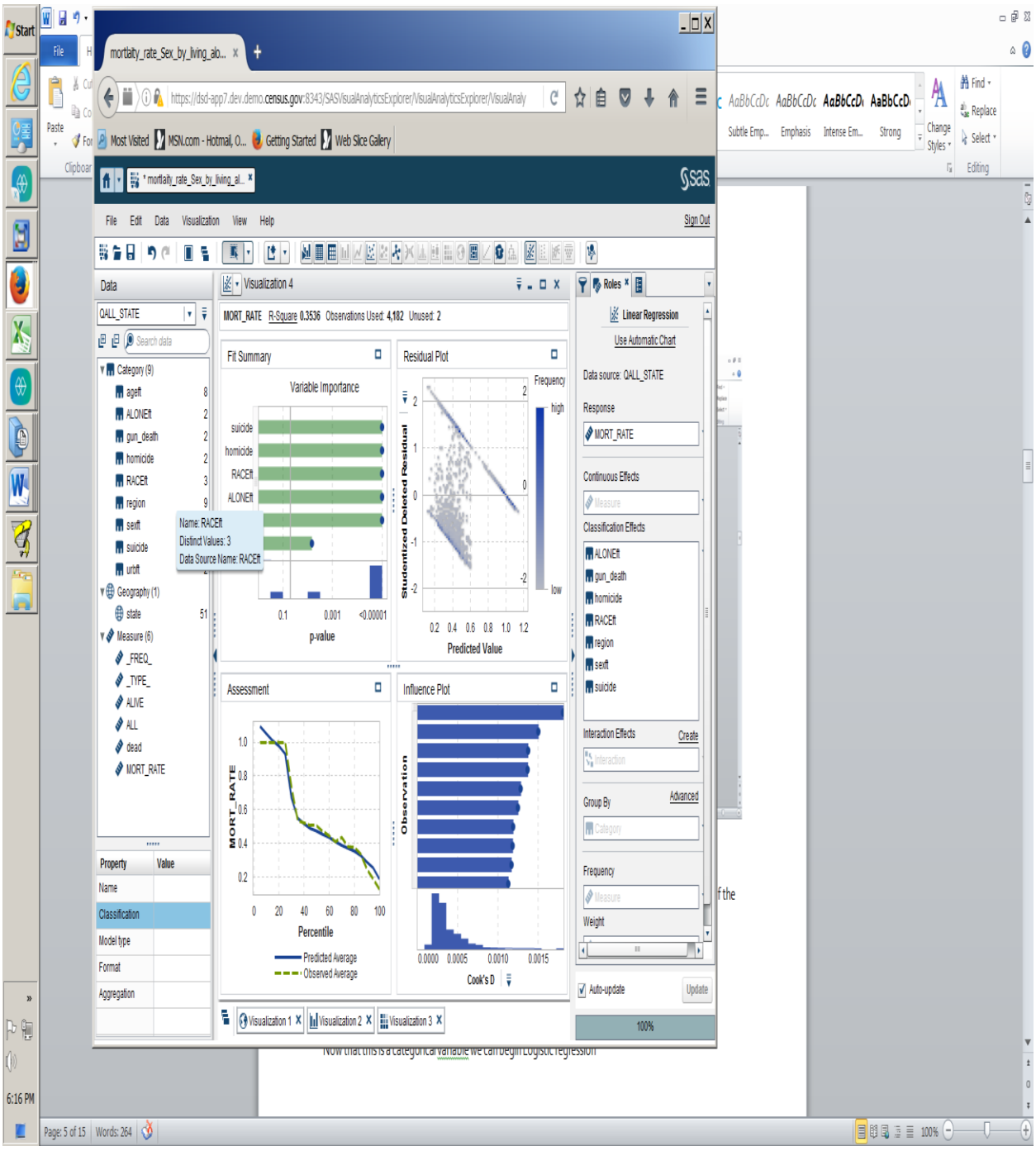


FIGURE 6. Logistic Regression (Race, Suicide/Homicide, sex, Rural/Urban)

Look at the relative strength of the variables.

Look at how they change when a stronger predictor is added (that provides more explanatory power of the model to the user)

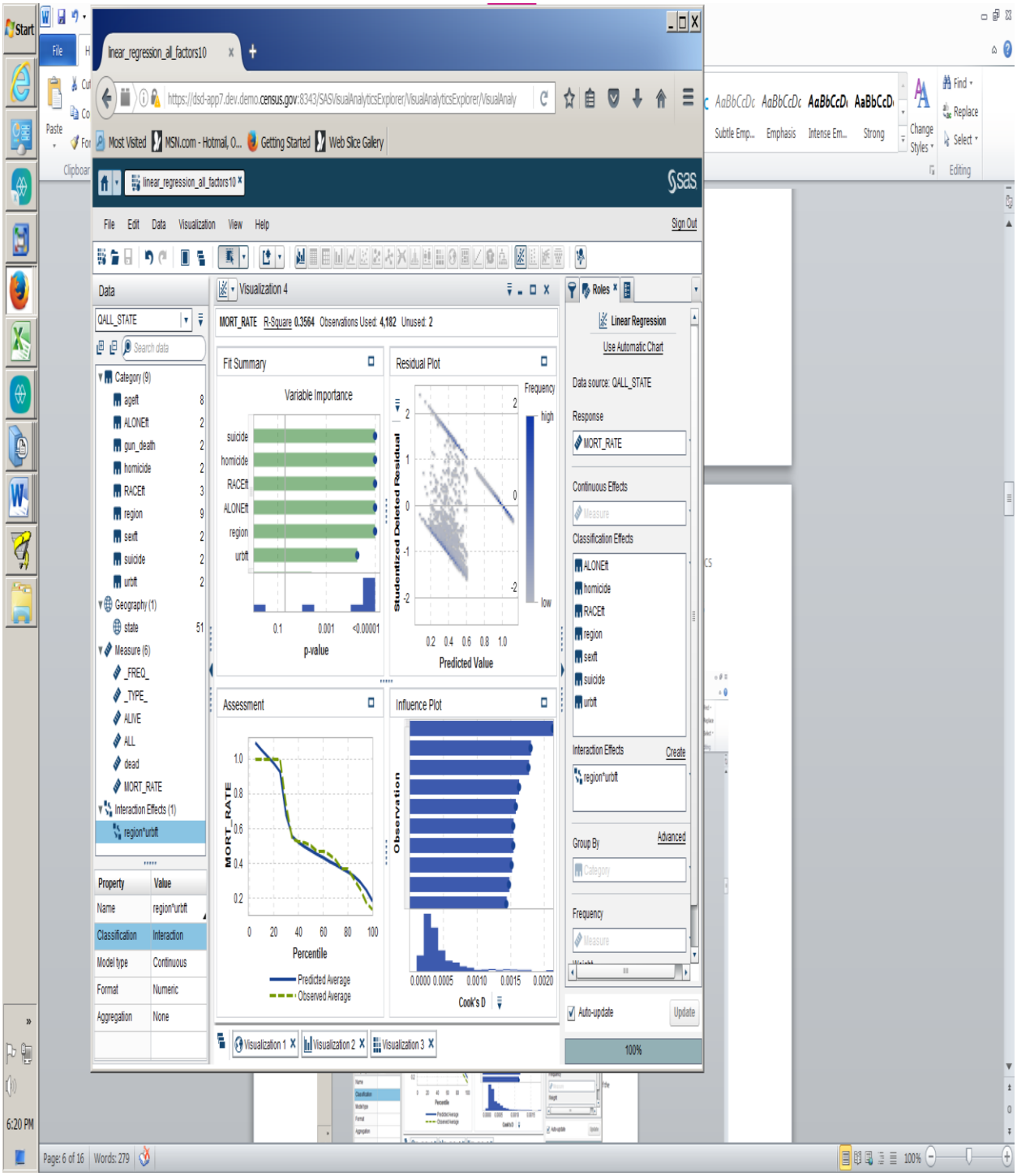
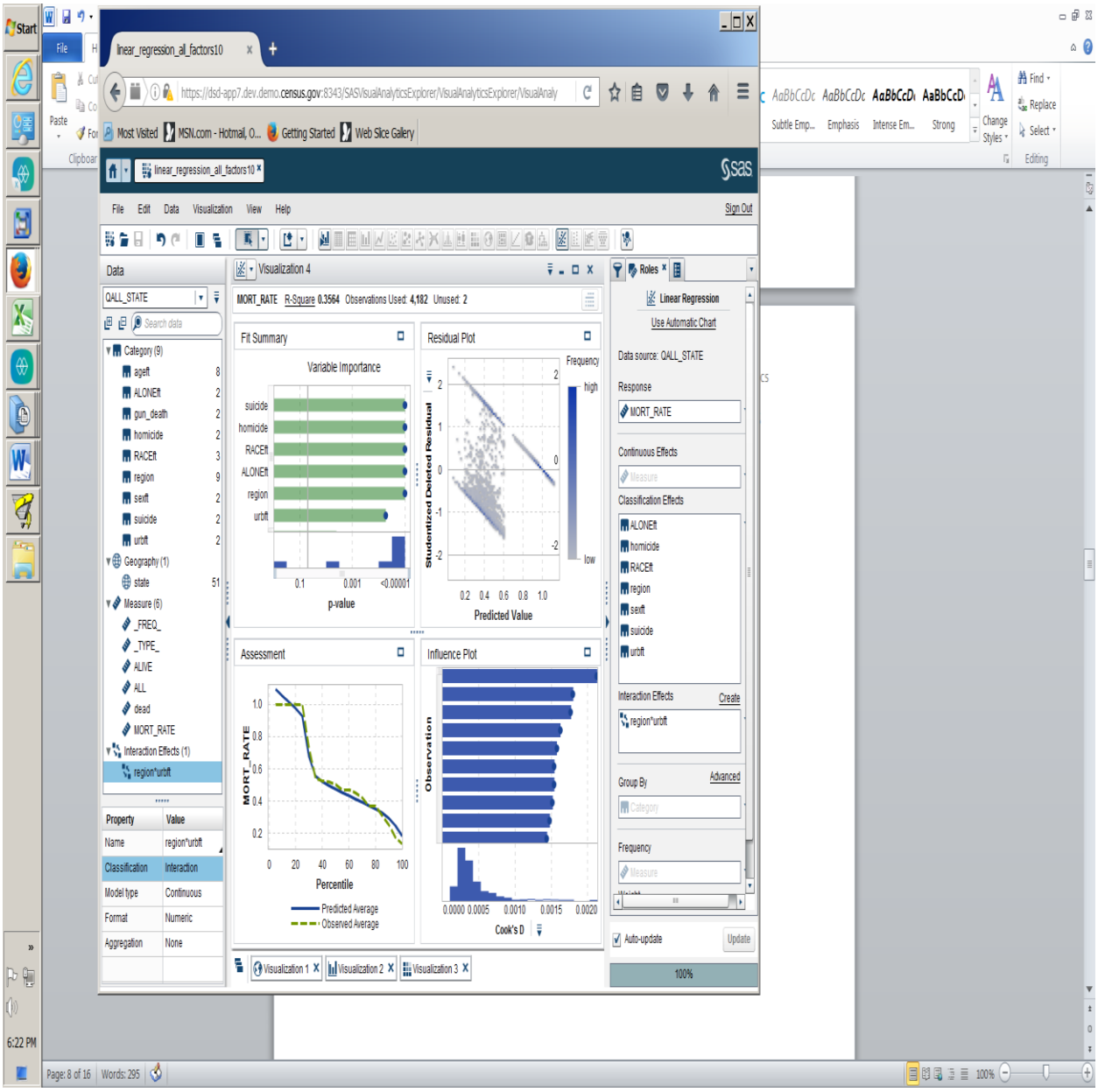


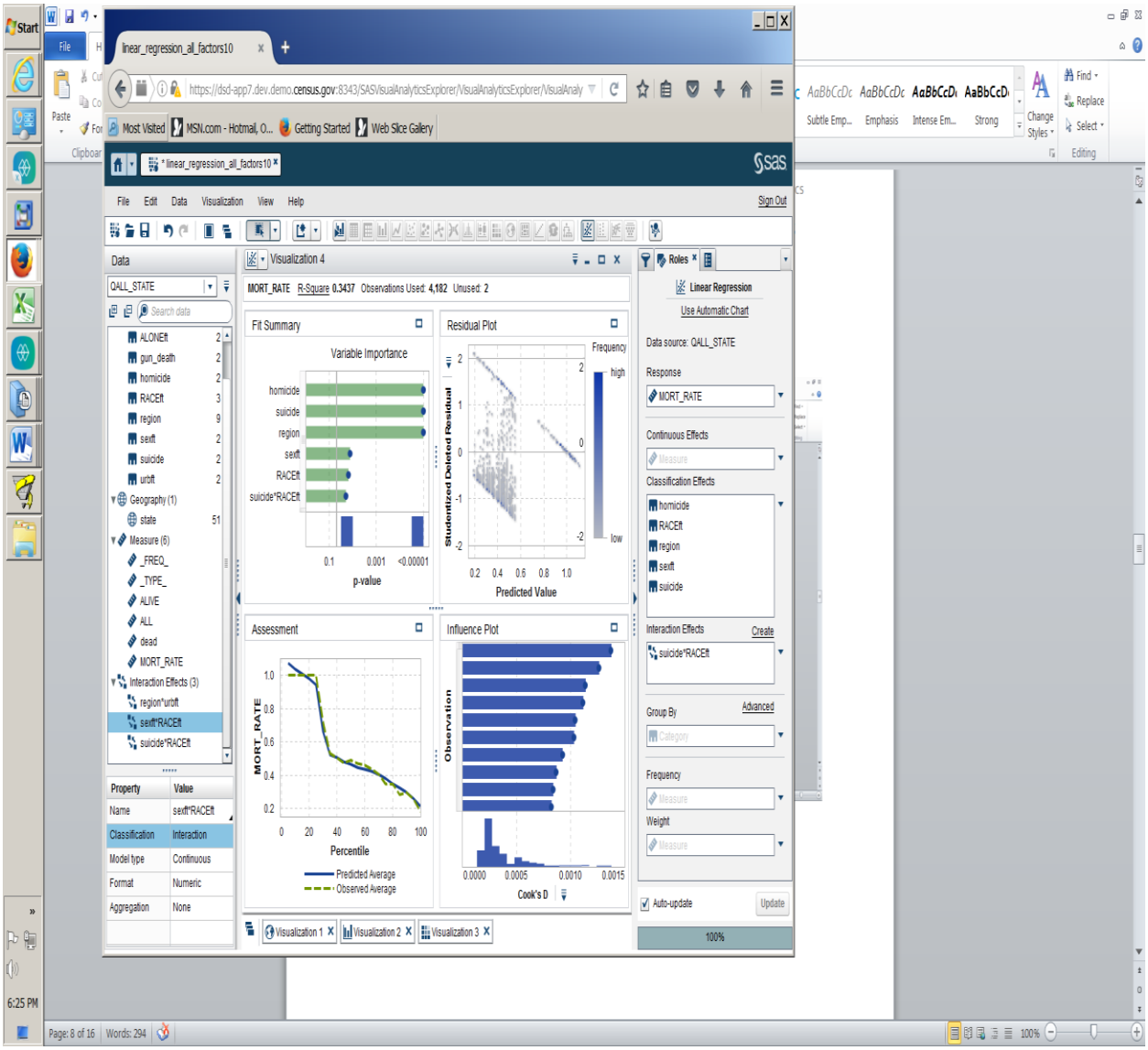
Figure 7: Linear Regression with 7 explanatory variables to Control for interaction effects:

Quickly integrates the new term (the interaction of 2 , or more variables, then updates the model. **When a classification variable is added to test for an interaction between Region and Urban/Rural, we do not find a relationship.**



Adding Rural/Urban does not add more explanatory power the shift key to select multiple

variables at once.

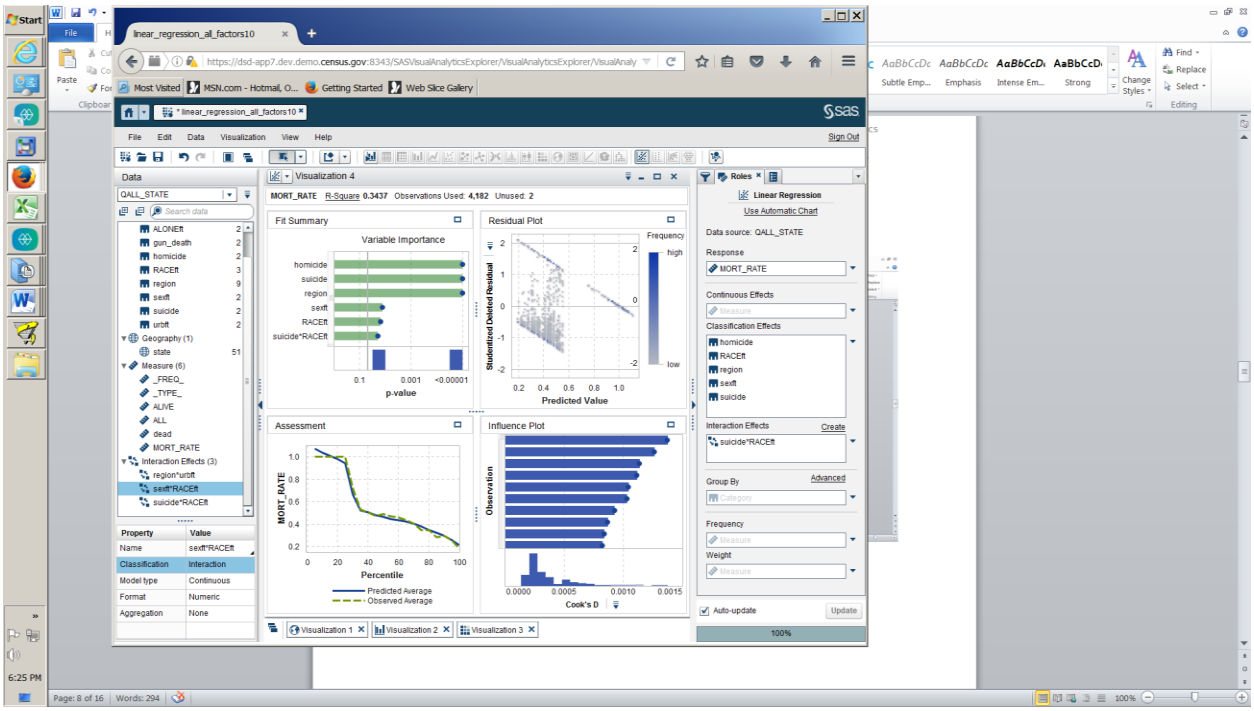


SAS performs maximum likelihood functions upon the model automatically.

Homicide, race, region, sex, and suicide all provide explanatory value.

Suicide and Race have an interaction.

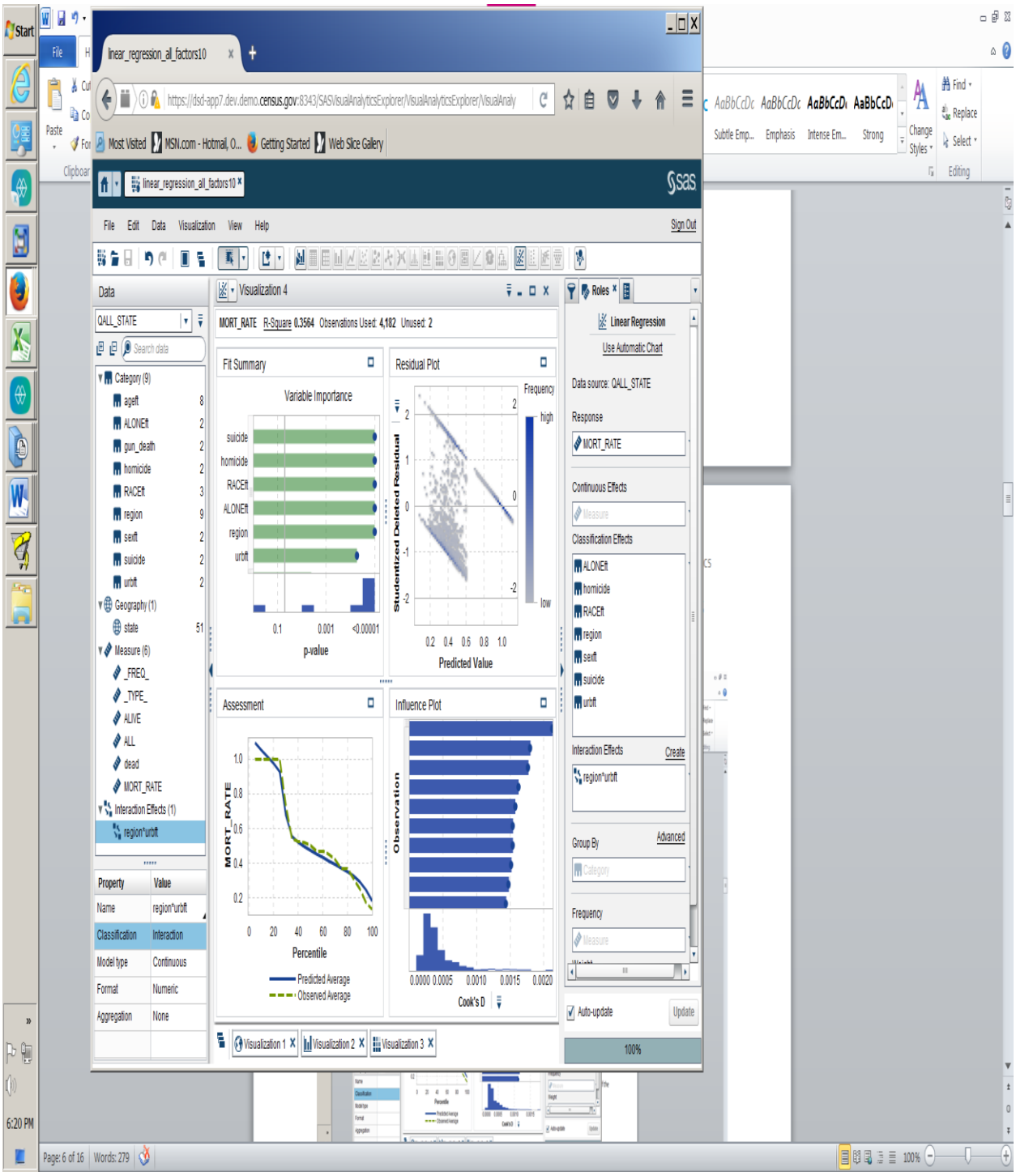
That is, depending on your race it is more likely, that you will commit suicide.



SAS Exploration automatically updates the model when any changes are made.

So the user can see changes as classification effects are added

Therefore, the model is automatically updated and refitted with the classification effects


linear_regression_all_factors10
File Edit Data Visualization View Help
Sign Out
Data
QALL_STATE
Visualization 4
MORT_RATE R-Square 0.3564 Observations Used: 4,182 Unused: 2
Fit Summary
Variable Importance
suicide
homicide
RACEflt
ALONEflt
region
urbflt
p-value
Residual Plot
Studentized Deleted Residual
Predicted Value
Frequency
high
low
Assessment
MORT_RATE
Percentile
Predicted Average
Observed Average
Influence Plot
Observation
Cook's D
Roles
Linear Regression
Use Automatic Chart
Data source: QALL_STATE
Response
MORT_RATE
Continuous Effects
Classification Effects
ALONEflt
homicide
RACEflt
region
sexflt
suicide
urbflt
Interaction Effects
Create
region*urbflt
Group By
Advanced
Frequency
Auto-update
Update
100%
Visualization 1
Visualization 2
Visualization 3
Property
Value
Name
region*urbflt
Classification
Interaction
Model type
Continuous
Format
Numeric
Aggregation
None

;

The four graphs shown include: Fit Summary displaying the most significant effects of the variables.

Quality measures

Residual Plots

Influence Statistics

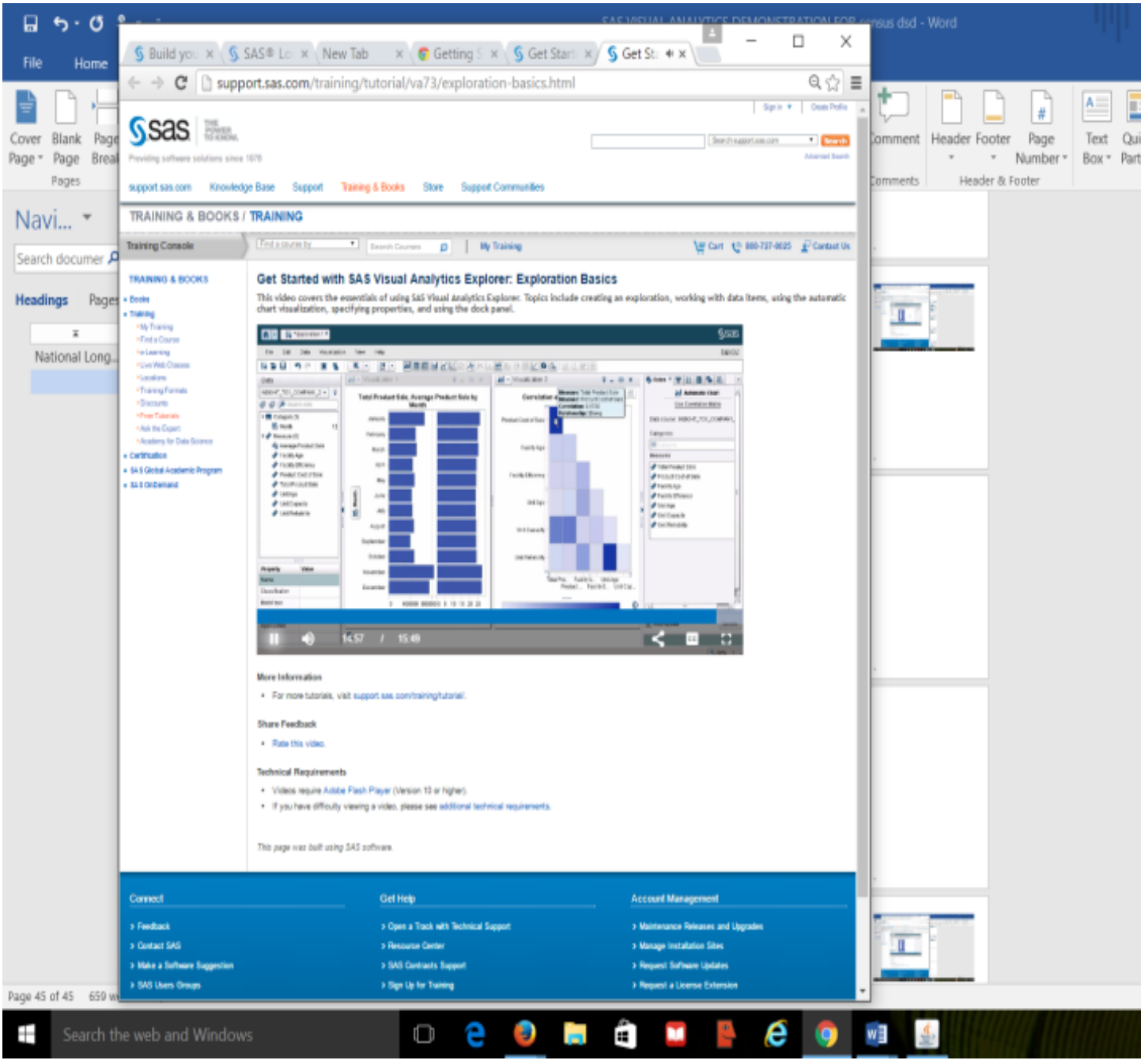
The user may choose to look at the results of statistical test by reviewing the summary data,

There are several more statistical options available include:

Filtering the data

Using a scorecard.

Performing follow-up analysis.



DECISION TREES

A decision tree was created using all possible categories related to mortality. Those were drawn first into the decision tree have the strongest effect, that is the best predictor of mortality, and each subsequent branch has the next strongest effect; then the third, the fourth

and so on. Also, amply demonstrated in the decision tree analysis is the ability to scope and parse the resulting decision tree to make it the branches larger, leaves smaller

DECISION TREES

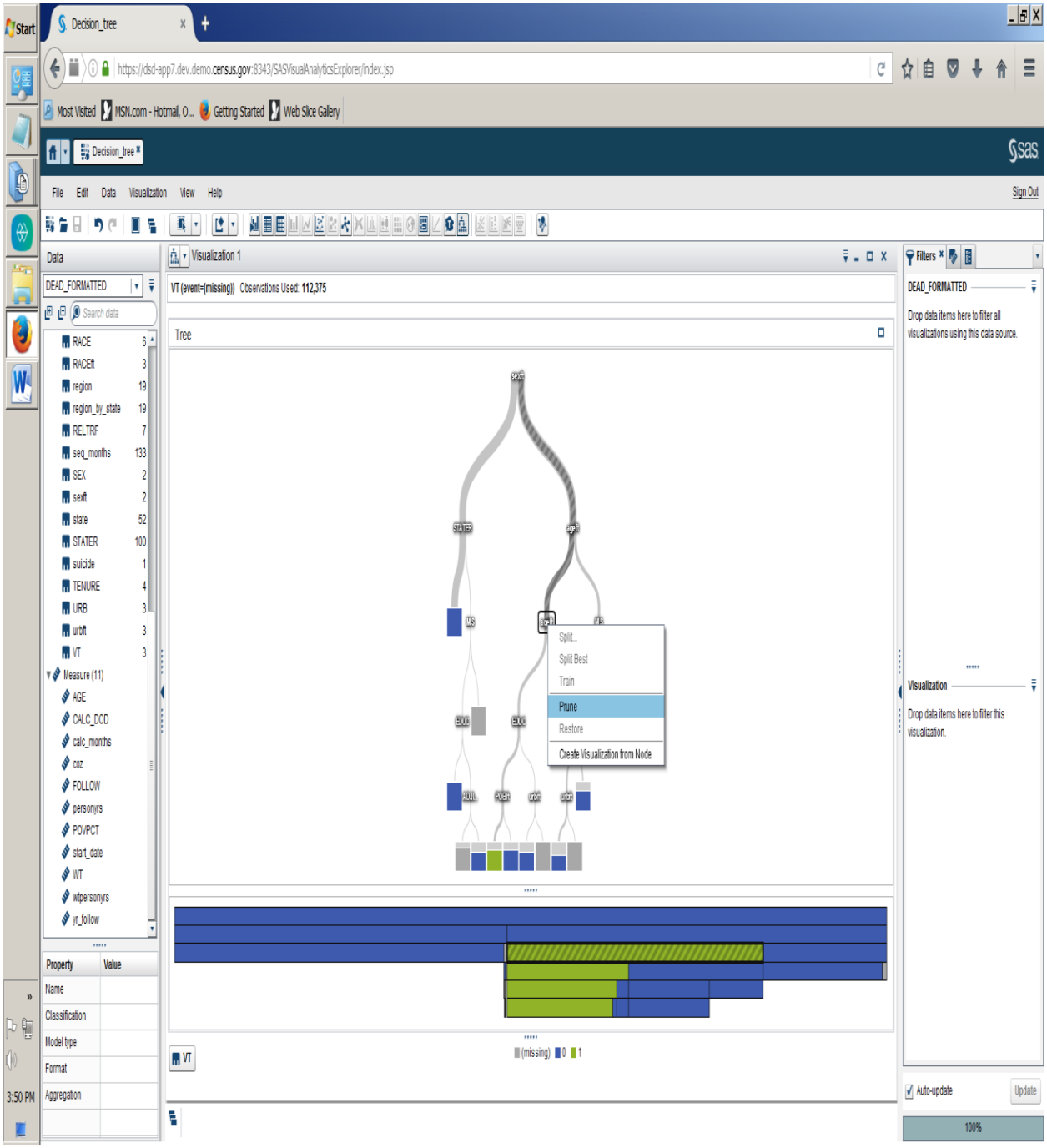


Figure 8: Decision with 22 variables added to the model. Living Alone is the strongest predictor of mortality.

Look at how the tree structure has changes with a simple click.

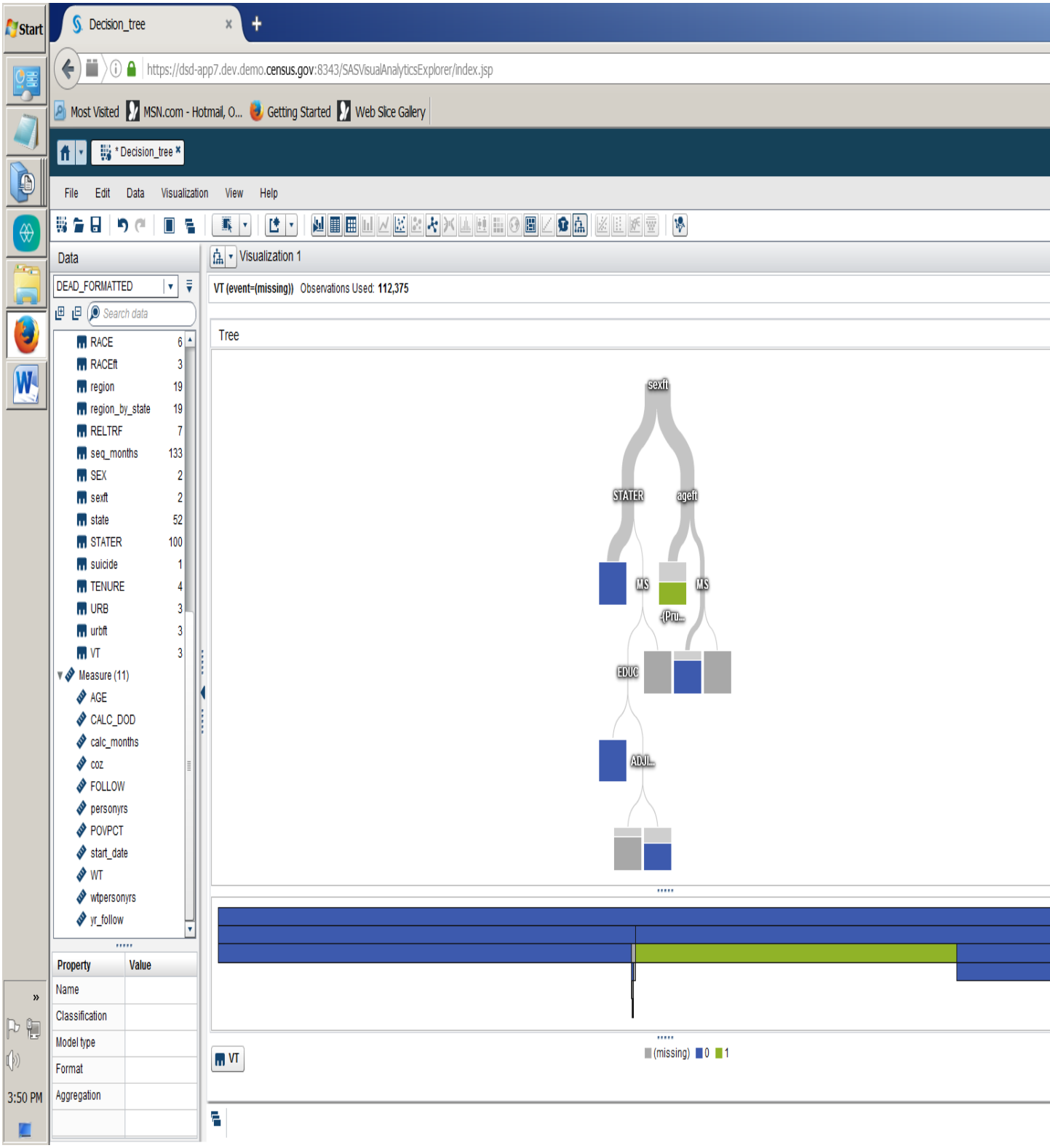


Figure 9. The decision tree analysis too is flexible enough to drill down within variables.

CONCLUSION

This is a demonstration of how to effectively review and check data to demonstrate to use visualizations combined with proper statistical techniques for better data analysis.

Proposals for future use

The advantage of a process such as SAS Visual Analytics is that SAS Visual Analytics is so effective at manipulating large datasets that outliers can be found using tools such as: data mining, machine learning, box and whiskers; correlation Matrix; linear/logistical regression(s). Edits of data that are encoded incorrectly and not found at the onset are costly. Likewise, if there is problem with a Field Representative: incorrect, invalid responses, poorly skilled or simply poorly trained.

Data visualizations are pretty pictures but it is necessary to tell a story and the inclusion of statistic makes it incumbent on the data viz expert to understand to make sure that it tells an accurate story. Here I have demonstrated how to effectively review and check data to demonstrate how to use visualizations combined with proper statistical techniques

On the other hand, visual analytics allows us to do some forensic analysis of use cases that are outliers. There is a worry that big data is pattern seeking of data on a large scale and by doing so, they are overlooking small, anomalous but significant effects.

Are we were only looking for large effects or high numbers in bar charts? Visual Analytics is a tool to re-engineer how to find outliers that have not been caught in previous edits.

Proposals for broader, future use

Data editing for data that is incorrectly and not found at the onset are costly. Likewise, if there is problem with a Field Representative: incorrect, invalid responses, poorly skilled or simply poorly trained. The advantage of a process such as SAS VISUAL ANALYTICS is that SAS VISUAL ANALYTICS is so effective at manipulating large datasets that outliers can be found using tools such as: data mining, machine learning, box and whiskers; correlation matrix; linear/logistical regression(s).

REFERENCES

Rogot E, Sorlie PD, Johnson NJ, Loveless CA. *A Mortality Study of 1.3 Million Persons by Demographic, Social and Economic Factors: 1979-1985 Follow-up. Second Data Book*. NIH Publication No 92-3297 ed. National Institutes of Health, PHS, DHHS; 1992.

Sorlie P, Rogot E, Anderson R, Johnson NJ, Backlund E. Black-white mortality differences by family income. *Lancet* 1992 August 8;340(8815):346-50.

Sorlie PD, Backlund E, Johnson NJ, Rogot E. Mortality by Hispanic status in the United States. *JAMA* 1993 November 24;270(20):2464-8.

Johnson NJ, Backlund E, Sorlie PD, Loveless CA. Marital status and mortality: the national longitudinal mortality study. *Ann Epidemiol* 2000 May;10(4):224-38.

Kposowa AJ. Marital status and suicide in the National Longitudinal Mortality Study. *J Epidemiol Community Health* 2000 April;54(4):254-61

ACKNOWLEDGMENTS

This section is not required. If you include this section, do not change the heading style or the text "ACKNOWLEDGMENTS" of the preceding heading.

This work is the culmination of 26 years of collegial collaboration and mentorship from Norman Johnson, Phd of the CENSUS Bureau. He is the caretaker of the National Longitudinal Mortality Study (NLMS) and its data integrity. Also, I would like to acknowledge the importance of Eric Backlund and Chip Alexander in my personal growth and development as a statistician and data scientist as I was balancing graduate school, raising children, caring for elderly parents and renovating houses until I emerged as a full-fledged data scientist.

RECOMMENDED READING

- *SAS® Visual Analytics Tutorial*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Catherine Loveless-Schmitt

703-200-7336

catherinelovesdata@gmail.com