

UCLA

UCLA Electronic Theses and Dissertations

Title

Technology-Enhanced Statistics Education with SOCR

Permalink

<https://escholarship.org/uc/item/4mb48660>

Author

Zhou, Chaojie

Publication Date

2012

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Technology-Enhanced Statistics Education with SOCR

A thesis submitted in partial satisfaction of the
requirements for the degree Master of Science
in Statistics

by

Chaojie Zhou

2012

ABSTRACT OF THE THESIS

Statistics Education with SOCR

by

Chaojie Zhou

Master of Science in Statistics

University of California, Los Angeles, 2012

Professor Jan De Leeuw, Co-Chair

Professor Ivaylo Dinov, Co-Chair

There is an ongoing need for clear and accessible statistics teaching tools for both learners and instructors. Applications, step by step tutorials, and visualizations are extremely useful tools for teaching students to think scientifically, properly analyze the data, use proper techniques, and identify common errors. In this paper we will demonstrate technology-enhanced approaches for introductory statistics courses. Specifically we develop two different activities, using SOCR (Statistics Online Computational Resource) data, tools and resources. The first activity introduces multiple linear regression using appropriate SOCR tools. In general, linear regression is used to describe a relationship between one variable to one or several other variables. Linear regression is used extensively in practical applications such as prediction and measuring the strength of relationships between variables. Proper linear regression techniques will be demonstrated, and appropriate methods for the analysis of regression results will be discussed. The second activity demonstrates the interactive power of the SOCR Motion Charts tool. SOCR Motion Charts allow the visualization of multivariate and high-dimensional data that has time and location dimensions. Used correctly, data visualization and statistical

graphics are useful in presenting data in clear, intuitive, and engaging ways. Proper data visualization can reveal patterns and relationships that would have been hidden in other data structures, such as tables. The SOCR Motion Charts tool allows us to represent variables based on their size, time, and location attributes. With this technology we can detect patterns across time, as well as analyze the relationships of variables in terms of their magnitudes and locations. These activities and tutorials are implemented as interactive hands-on learning materials and are openly accessible on the web through the SOCR site www.socr.ucla.edu/.

The dissertation of Chaojie Zhou is approved.

Nicolas Christou

Juana Sanchez

Jan De Leeuw, Committee Co-Chair

Ivaylo Dinov, Committee Co-Chair

University of California, Los Angeles

2012

Contents

Introduction	1
Linear Regression	5
Linear Regression Tutorial	8
Exploratory Data Analyses	14
SOCR Motion Charts Tutorial (L.A. Neighborhoods Data)	16
SOCR Motion Charts Tutorial (Earthquakes Data)	19
Conclusions and Discussions	22
References	24

List of Figures

1	SOCR ANOVA Activity	2
2	SOCR Survival Graph	3
3	LA Neighborhoods data table	8
4	The Multiple Regression Analysis Activity	9
5	The "LA Neighborhoods" data table is now pasted in	9
6	Multiple Regression Analysis Mapping Tab	10
7	Multiple Regression Activity Result Tab	10
8	Income vs White Scatter plot	11
9	Residual Plot	11
10	Normal QQ Plot	12
11	Results / Calculations	13
12	Napoleon's Army in Russia	15
13	Income / White Motion Chart	17
14	Population / Diversity Motion Chart	18
15	California Earthquakes Motion Chart	20
16	All of the earthquakes on a single motion chart	21
17	Confidence Intervals in SOCR	23

ACKNOWLEDGEMENTS

The author would like to thank the following individuals for their essential support on this project: Ivaylo Dinov, Nicolas Christou.

Introduction

There is an ongoing need for clear and accessible teaching tools for both learners and instructors of statistics. Applications, step by step tutorials, and data graphics are extremely useful for teaching students to think scientifically, properly analyze the data, and use proper techniques. Technology is a great way to present difficult or abstract statistical concepts, and it is easier for students to augment their classroom learning on their own time, and at a location convenient to them (home, coffee shop, the library). Nowadays, as many fields become more data-intensive, the use of software for statistical analysis is essential. With new technologies, new data sources are coming online (internet search data, sensors in cars, GPS data in our phones), bringing with them more opportunities for data-driven decision making and exploration. Retailers look at sales data to tailor the shopping experience, police departments look at crimes data to identify hot spots, and auto manufacturers run crash simulations to develop better safety systems for their vehicles. As such, the integration of technology into the classroom has become essential.

There have been several studies that have demonstrated a positive effect when using blended instruction. In particular, one study compared using SOCR (Statistics Online Computation Resources) and traditional approaches¹. SOCR (<http://www.socr.ucla.edu>) is a suite of online educational tools, including applets, distributions, experiments, activities, data graphics tools, and various educational materials, including an online statistics textbook. The study included two statistics courses, "Statistical Methods for the Life and Health Sciences" offered in the fall of 2005 and 2006. The course offered in 2005 (control) did not have the SOCR component, while the course offered in 2006 (SOCR treatment) did. The same grading schema was used in both classes. The study

found a consistent effect of increased student satisfaction of the courses for the SOCR treatment group. In addition, the students in the SOCR treatment group displayed more homogenous and improved quantitative performance compared with the control group, which attended the same course but using traditional teaching methods. It was found that the important components of the effective use of blended instruction with technology are instructor training and the development of appropriate tools for students. Such tools include the activities included with SOCR. These activities introduce the user to statistical concepts and demonstrate their use in interactive ways. For example, the SOCR One-Way Analysis of Variance applet² is an activity which demonstrates using SOCR to perform ANOVA calculations and interpreting the results. This ANOVA applet is a part of SOCR Analyses, which also includes applets for common statistical topics such as Chi-Square goodness of fit tests and Simple Regression analysis. With the ANOVA activity, users are familiarized with the procedures for analysis of variance calculations using the SOCR applet. Users are shown how to import data (or use included example data), and generate the necessary calculations and graphs (Figure 1).

Another example is the SOCR Survival Analysis Activity³. This applet shows how to perform survival analysis using a dataset freely available online and in the free statistics tool R. Users are shown how to import data and map the variables before making the survival analysis calculations, as well as how to graph the survival curve (Figure 2).

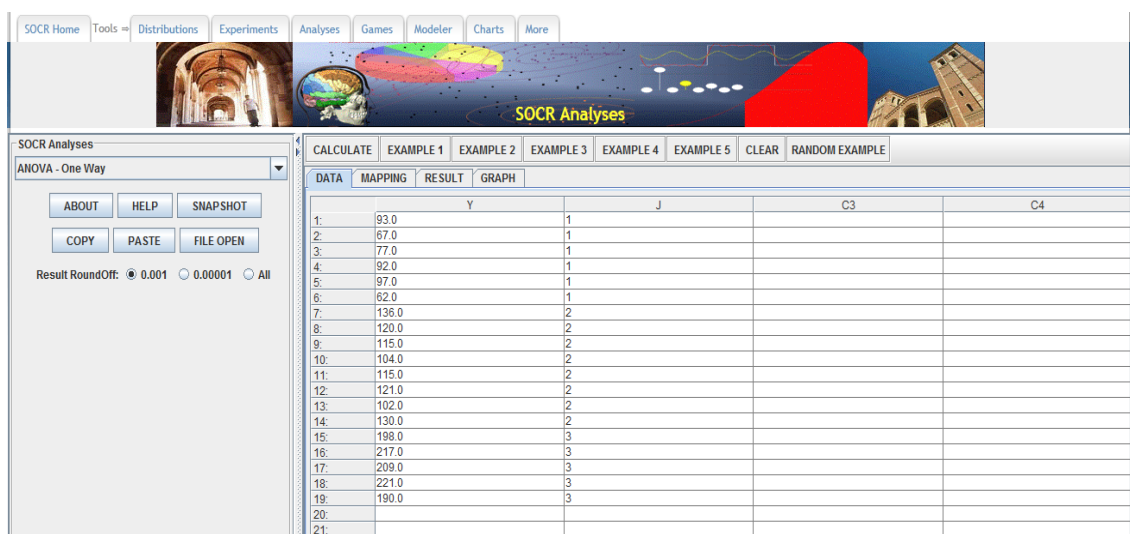


Figure 1: SOCR ANOVA Activity

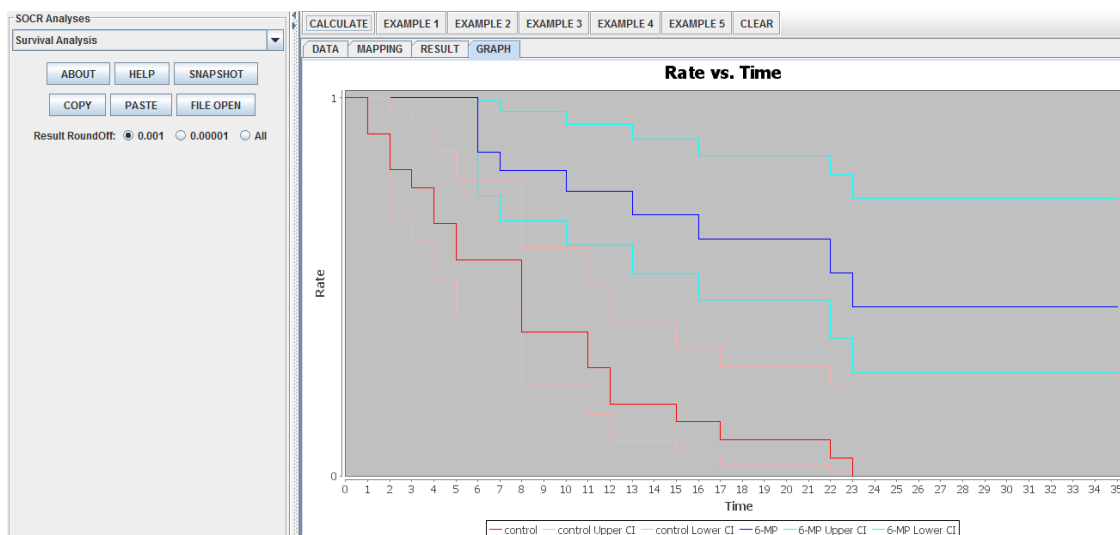


Figure 2: SOCR Survival Graph

SOCR also includes many data graphics tools which are very effective at presenting large amounts of data in a clear and intuitive way. Instead of relying on tables of data, statistical graphics and data visualization present data in pictorial form. Graphics are great for big datasets because vast amounts of complex data can be organized and presented by a single graphic, allowing previously hidden patterns and structures in the data to be revealed. In addition, certain graphics can show geography, the passage of time, or any number of attributes from the data. SOCR includes many datasets that are appropriate for statistical graphics. For example, the Los Angeles County Neighborhoods data⁴, which has demographics information as well as variables for geography, would be useful in a mapping application. Another example is the California Earthquakes data⁵, which has information on magnitudes, locations, and times of earthquakes. This dataset would be perfect for a time series graphic.

In this project we will demonstrate technology-enhanced approaches for introductory statistics courses. Specifically we develop two different activities, using SOCR (Statistics Online Computational Resource) data, tools and resources. These activities are implemented as interactive hands-on learning materials and are openly accessible on the web (URL). The first activity is a tutorial on multiple linear regression using SOCR. Proper linear regression techniques will be demonstrated, and appropriate methods for the anal-

ysis of regression results will be discussed. The data used will be the Los Angeles County Neighborhoods dataset. The relationships between income and other demographics variables will be investigated. The second activity will demonstrate the SOCR Motion Charts tool. SOCR Motion Charts allow the visualization of multivariate and high-dimensional data that has time and location dimensions. Both the Los Angeles County Neighborhoods data and the California Earthquakes data will be used.

Linear Regression

Linear regression is used to describe the relationship between one variable, and another or several other variables. For example, we may wish to investigate the relationship between household income and health, or perhaps examine a relationship between crime rate and various demographics variables. The data will be modeled with linear functions, and from this we can estimate the variable of interest. Linear regression is widely used in many practical applications such as prediction and measuring the strength of a relationship between variables. The extensive use of linear regression comes from the fact that they are easier to fit and interpret than models which are non-linearly related to its parameters, such as exponential, log, or power functions.

One example of linear regression in a real world application can be found in a study involving brain mapping. Using linear regression techniques, the study explores the relationship between a fat mass and obesity-associated (FTO) gene and brain volume⁶. Obesity is a major health concern, especially for the elderly: 40% of men and 45% of women over the age of 70 suffer from either obesity or type 2 diabetes. Obesity has been found to be a risk factor for Alzheimers disease⁶ as well. It is known that obesity and body mass index (BMI) are highly influenced by genetics: 50% to 90% of the variation in BMI can be attributed to genetics. Recently, the genetic bases of obesity and BMI have been found to be in part due to variances within the FTO gene. From the study, it was found that in 94 healthy elderly patients, BMI, fasting plasma insulin, and type 2 diabetes were associated with frontal, temporal, and subcortical atrophy⁶. Also, obese subjects showed the greatest brain tissue deficits.

Because FTO is associated with differences in BMI and is highly expressed in the

brain, there was an interest in the effect of carrying the FTO risk allele on human brain structure. The study compared these brain differences with those seen in people with high BMI. 206 elderly patients underwent MRI scans with the hypothesis that in patients carrying the FTO risk allele, structural deficits would be found in the brain in the same areas where lower brain tissue volumes were correlated with higher BMI⁶. In order to investigate these topics, linear regression was utilized.

After controlling for age and sex, there was an association found between carrying at least one copy of the FTO risk allele and higher BMI. Also, patients with higher BMI had lower regional brain volumes. From a linear regression model, it was shown that for every one unit increase in BMI, there was a 1.5% average brain tissue reduction in several brain regions. It was shown that subjects who carried the FTO risk allele showed brain tissue deficits in areas that are also associated with volume reductions in subjects with higher BMI.

Another example of using linear regression comes from finance, specifically the single index model. The single index model is commonly used in finance to measure the risk and return of a stock. The single index model states that⁷:

$$R_{it} = \alpha_i + \beta_i R_{mt} + \epsilon_{it}$$

Where R_{it} is the return of stock i at time t and R_{mt} is the return of the market at time t . The assumptions of this model are:

$$E(\epsilon_i) = 0; \text{var}(\epsilon_i) = \sigma_{\epsilon_i}^2; E(\epsilon_i \epsilon_j) = 0; \text{cov}(R_m, \epsilon_i) = 0; \text{var}(R_m) = \sigma_m^2; E(R_m) = \bar{R}_m$$

We see that the return of a stock is influenced by the expected excess return of the stock, α , the rate of return on a market index, β , and the return of the market, R_{mt} , as the independent variables. This model assumes that most stocks have a positive covariance because they all respond similarly to macroeconomic factor, such as labor costs.

To demonstrate the usage of multiple linear regression, we develop and demonstrate a tutorial using the SOCR Multiple Regression Applet. In multiple regression, the goal is to predict a value for a dependent variable y using two or more predictors $(x_1, x_2, \dots, x_n)$. First, we will need to establish some assumptions for the linear model:

Assumptions

1. There is a linear relationship between the independent and the dependent variables. Fitting a straight line to a non-linear relationship (such as a quadratic) would not yield a good model.
2. Normality of residuals.
3. No Patterns in Residuals, and residuals should be within 2 standard deviations of the mean residual.
4. Observations should come from the same population (in other words, no outliers, as they can significantly change the regression model).

Multiple Linear Regression Models

With multiple linear regression, we predict a response variable y with multiple predictors. The general form of the model is:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \epsilon$$

Interpretation:

$\hat{y}$ is the response variable that we are predicting. β_0 is the intercept, which is the value of $\hat{y}$ when $X_1 = X_2 = \dots = X_n = 0$. The β 's are the linear rate of change in $\hat{y}$ for each unit change in the corresponding X , if all other X 's are held constant. The ϵ 's are error terms, which are independent, and with the assumptions that $E(\epsilon) = 0$ and $var(\epsilon) = \sigma^2$ (constant variance).

Linear Regression Tutorial

For this tutorial, we will be using the Multiple Regression Analysis Activity in SOCR and the LA Neighborhoods Data. This dataset comes from the 2000 U.S. Census and contains information on 110 Los Angeles neighborhoods. Various demographics variables are included, such as median income, population makeup, and homeowner information. Our goal is to predict the median income using various demographics variables. In this particular tutorial, we will predict the median income of a neighborhood by using the median age, proportion of home owners in the neighborhood, and the proportion of whites in the neighborhood.

Step 1: Importing the Data

- First, we need to import the LA Neighborhoods data into SOCR. Head to http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_LA_Neighborhoods_Data#Data_Source, and highlight the data table. Copy the data (including the headers) by pressing CTRL-C (Command-C for Macs). Alternatively, you can right-click and select "Copy":

LA Neighborhoods Data

LA Nbrhd	Income	Schools	Diversity	Age	Homes	Vets	Asian	Black	Latino	White	Population	Area	Longitude	Latitude
Adams Normandie	29606	591	0.6	26	0.26	0.05	0.05	0.25	0.62	0.06	31068	0.8	-118.30027	34.03097
Arleta	55645	712	0.4	28	0.22	0.07	0.11	0.02	0.72	0.13	31068	3.1	-118.30027	34.03097
Arlington Heights	31423	687	0.8	31	0.31	0.09	0.13	0.25	0.57	0.05	22106	1	-118.30027	34.03097
Atwater Village	53872	762	0.5	34	0.34	0.06	0.2	0.01	0.51	0.22	14888	1.6	-118.30027	34.03097
Baldwin Hills/Crenshaw	37948	556	0.4	38	0.36	0.1	0.05	0.71	0.17	0.03	30122	3	-118.30027	34.03097
Bel Air	208861	924	0.2	46	0.46	0.13	0.08	0.01	0.05	0.83	7928	6.6	-118.46398	34.09614
Beverly Crest	168104	0	0.1	45	0.45	0.1	0.04	0.02	0.03	0.88	10610	7.5	-118.42471	34.11210
Beverly Grove	63035	791	0.2	38	0.38	0.05	0.05	0.02	0.06	0.82	21417	1.7	-118.37211	34.07609
Beverlywood	105253	372	0.2	39	0.39	0.08	0.07	0.04	0.06	0.8	6080	0.8	-118.39501	34.04161
Boyle Heights	33235	688	0.1	25	0.25	0.03	0.02	0.01	0.94	0.02	92785	6.5	-118.20537	34.0337
Brentwood	112927	882	0.2	39	0.39	0.08	0.06	0.01	0.05	0.84	31344	15.2	-118.473	34.052
Broadway/Manchester	29897	556	0.5	23	0.23	0.04	0	0.39	0.59	0.01	2347	1.6	-118.27388	33.96153
Canoga Park	51601	706	0.8	30	0.3	0.06	0.11	0.04	0.51	0.31	53227	4.3	-118.598	34.201
Carthage	71398	762	0.8	37	0.37	0.04	0.05	0.13	0.16	0.58	4866	0.5	-118.369	34.059
Central Alameda	31559	568	0.1	22	0.22	0.02	0.01	0.13	0.85	0.01	40947	2.2	-118.245	33.998
Century City	95135	0	0.2	46	0.46	0.12	0.05	0.01	0.04	0.83	5513	0.7	-118.416	34.054

Figure 3: LA Neighborhoods data table

- Next, head to http://socr.ucla.edu/htmls/SOCR_Analyses.html, and find the Multiple Regression Analysis Activity in the drop down menu:

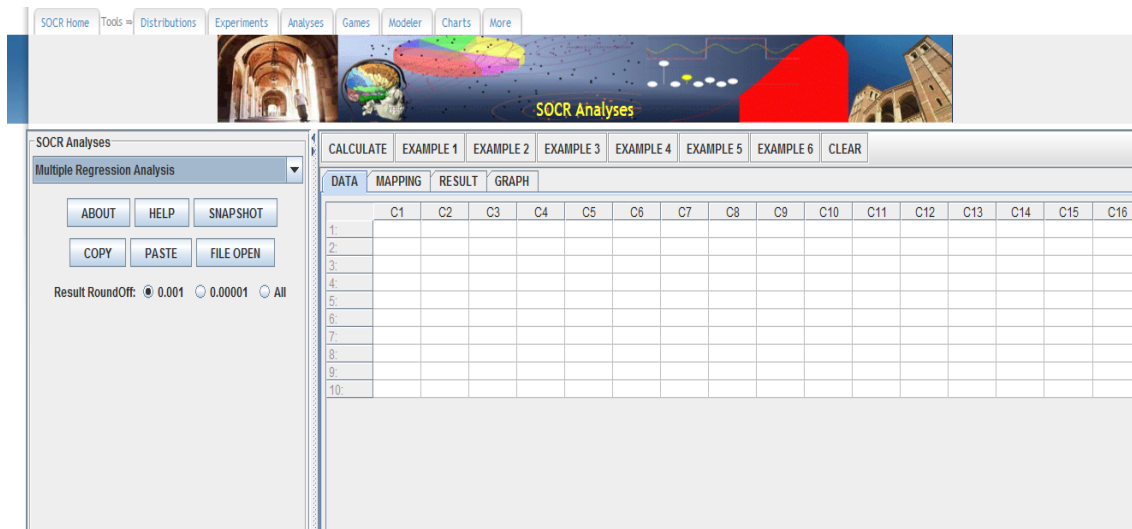


Figure 4: The Multiple Regression Analysis Activity

- Finally, click the "PASTE" button:

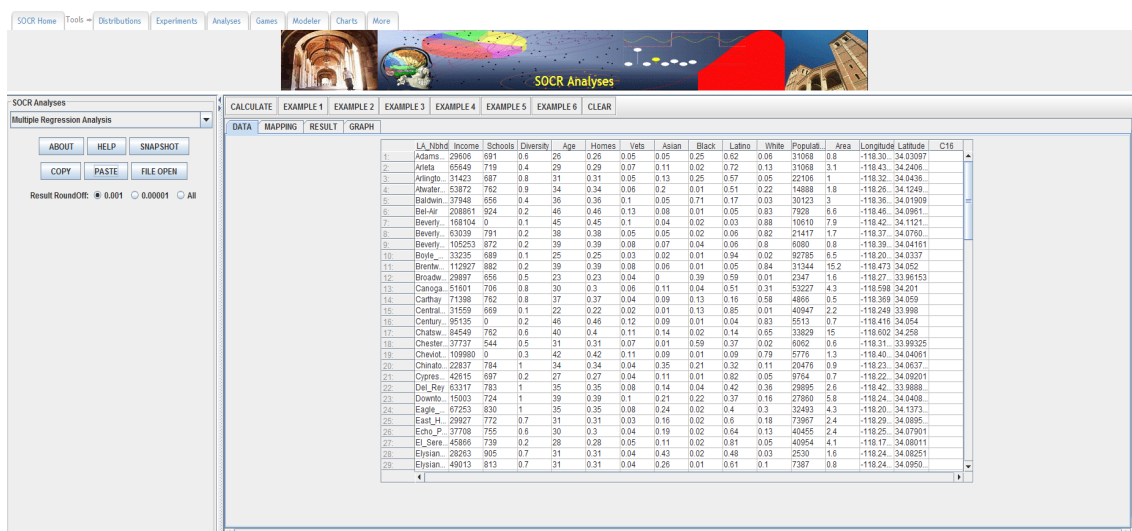


Figure 5: The "LA Neighborhoods" data table is now pasted in

Step 2: Performing the Regression

- You should now see the data in the window. Click on the "MAPPING" tab, and add "Income" to the dependent variables list and "Age", "Homes" and "White" to the independent variables list:

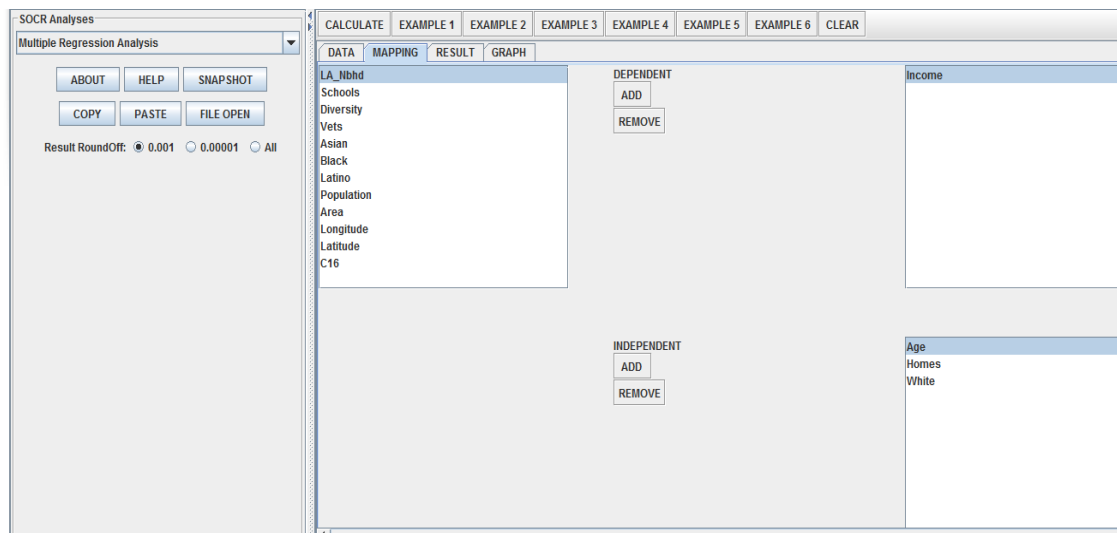


Figure 6: Multiple Regression Analysis Mapping Tab

- Click "CALCULATE". You will now be taken to the "RESULTS" tab. Here you can see the regression equation and R^2 , plus other statistics:

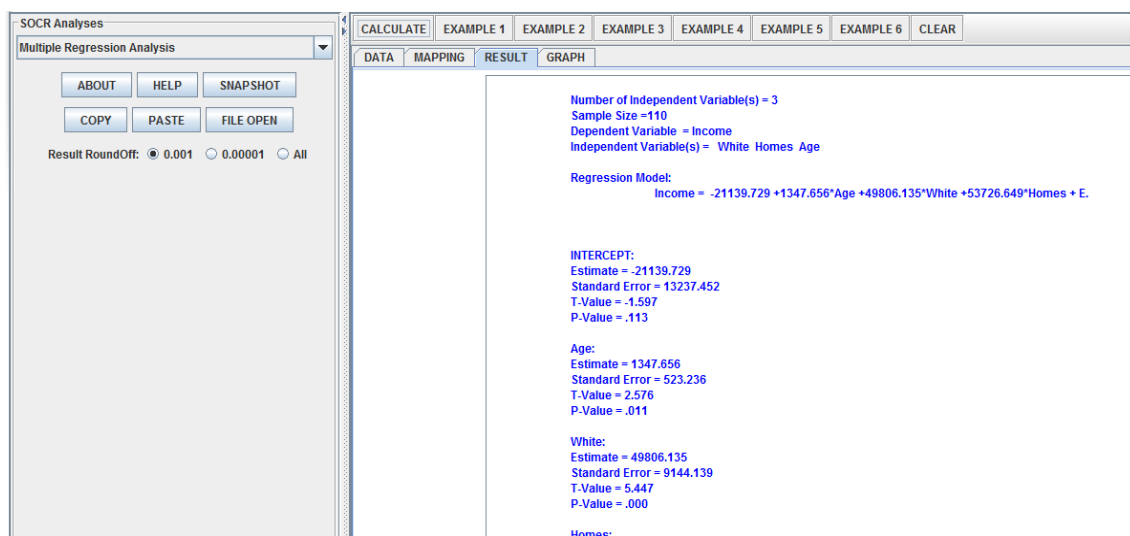


Figure 7: Multiple Regression Activity Result Tab

- Click on the "GRAPH" tab. Here you will see scatterplots of the Income variable against each of the three chosen explanatory variables, as well as the residual plots and the Normal QQ Plot as you scroll down:

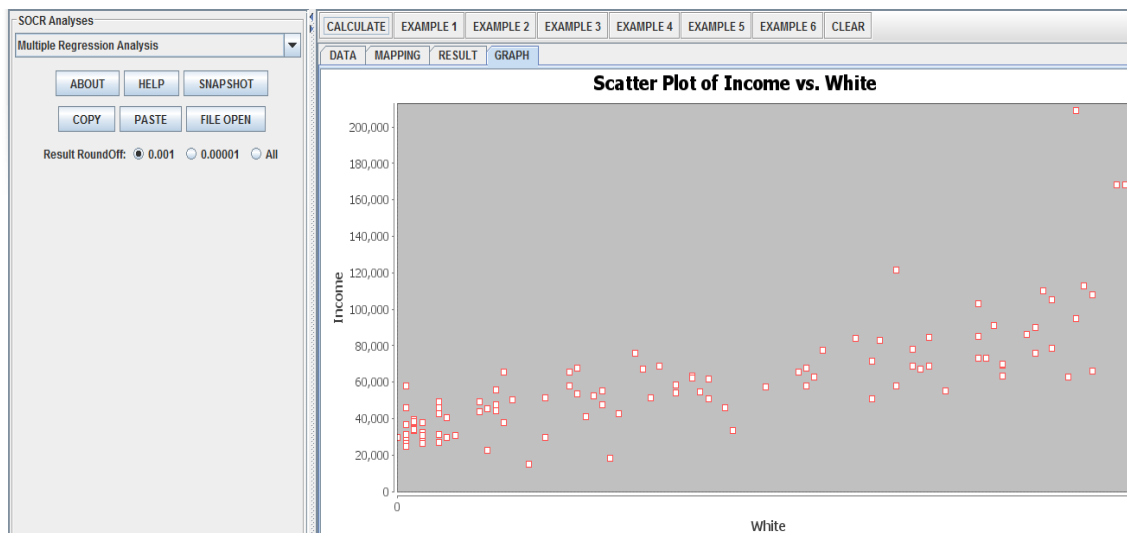


Figure 8: Income vs White Scatter plot

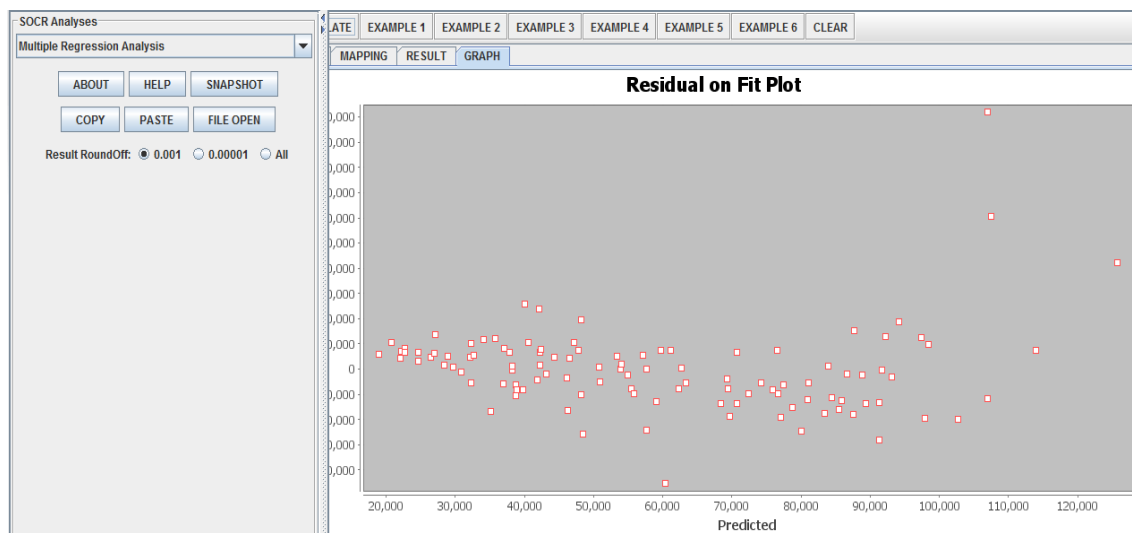


Figure 9: Residual Plot

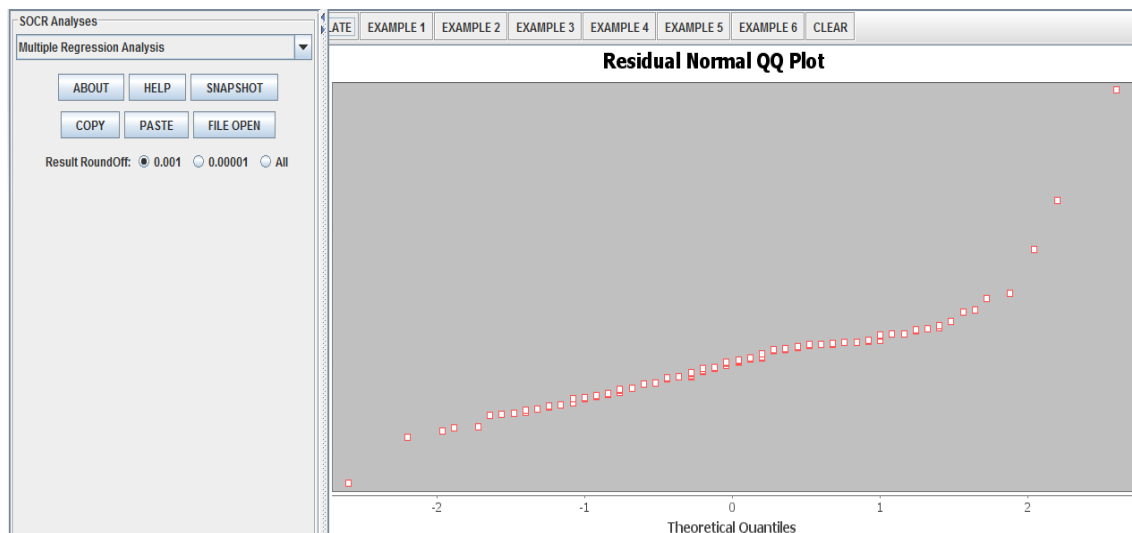


Figure 10: Normal QQ Plot

Step 3: Check Assumptions

- In order to use our linear regression model, we must check that the assumptions of linear regression are met.

Assumption 1: There is a linear relationship between the independent variables (age, homes, white) and dependent variable (income).

How to check: Make a scatter plot of income and age.

How to fix if assumption not met: Transformations (log, square roots, powers).

Linear model fits the data well.

Assumption 2: Variance is constant.

How to check: Look at plot of residuals vs. predicted values ($\hat{y}$). Make sure there is not a pattern, such as the residuals getting larger as the predicted values increase.

How to fix if assumption not met: Logging of variables, fixing underlying independence or linearity issues.

Slight increase in residuals at the top range of exploratory variables.

Assumption 3: Errors are normally distributed.

How to check: Normal QQ Plot (Should lie close to straight line).

How to fix if assumption not met: Take out outliers, if applicable. Non-linear transformations may be needed.

Assumptions met.

Step 4: Conclusions

- The assumptions are adequately met, and we can proceed to draw some conclusions from our analysis:

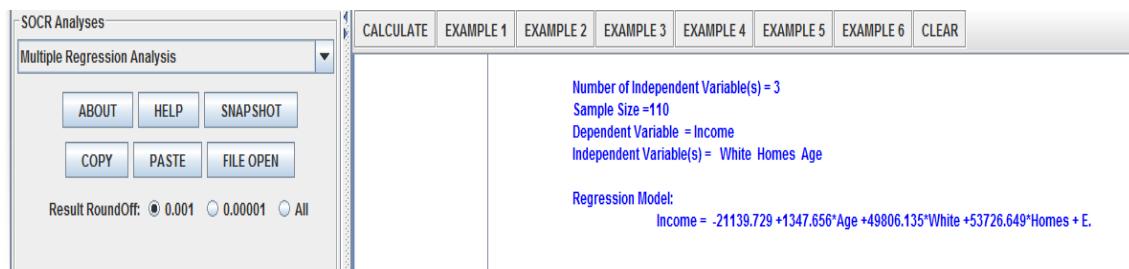


Figure 11: Results / Calculations

We can see from the results tab that the regression equation is:

$$\text{Income} = -21139.7294 + 1347.656 * \text{Age} + 49806.135 * \text{White} + 53726.649 * \text{Homes} + E.$$

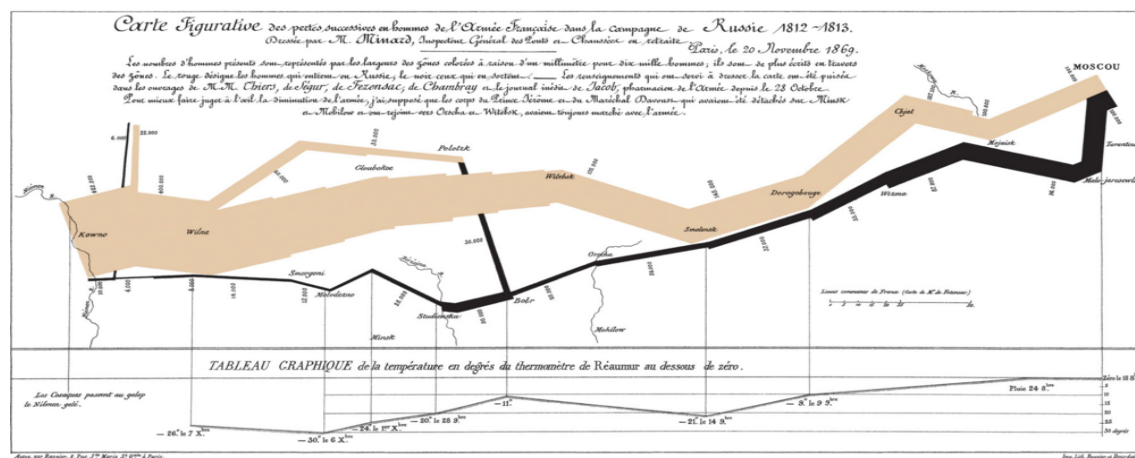
The "E" is the error term. "Income" is the predicted value, and "Homes", "Age", and "White" are the explanatory variables.

This model states that for every 1% increase in homeowner proportion, with everything else held constant, the median household income will increase by \$537.26. For every 1 year increase in median age, with everything else held constant, the median household income will increase by \$1,347.66. For every 1% increase in the proportion of whites in the population, with everything else held constant, the median household income will increase by \$498.06.

Exploratory Data Analyses

SOCR includes many data graphics tools which are very effective at presenting large amounts of data in a clear and intuitive way. Instead of relying on tables of data, statistical graphics and data visualization methods present data in pictorial forms. Graphics are great for big datasets because vast amounts of complex data can be organized and presented by a single picture, allowing previously hidden patterns and structures in the data to be revealed. For example, a single time series graphic can show an easy to interpret summary of several complicated and dense tables. In addition, any seasonal patterns in the data would be much easier to detect in a graphic, leading to a better understanding of the data. Statistical graphics are a great way to present information which otherwise would be unwieldy. In addition, certain graphics can show geography, the passage of time, magnitude, and many other attributes from the data. Good statistical graphics of multivariate data can be especially useful because they promote comparisons of the data. SOCR includes many datasets that are appropriate for statistical graphics. For example, the Los Angeles County Neighborhoods data, which has demographics information as well as variables for geography, would be useful in a mapping application. Another example is the California Earthquakes data, which has information on magnitudes, locations, and times of earthquakes. This dataset would be perfect for a time series graphic.

To illustrate an example of good use of statistical graphics, we present a classic graphic created by Charles Joseph Minard⁸:



SOCR Motion Charts Tutorial (L.A. Neighborhoods Data)

To demonstrate the use of graphical tools in SOCR, we will use the Motion Charts Activity. SOCR Motion Charts is useful in analyzing multivariate and longitudinal data, allowing the visualization of high dimensional datasets. Variables can be mapped to size, time, colors and many other characteristics.

Step 1: Importing the Data

- We will use the L.A. Neighborhoods data, which can be found at:
`http://wiki.stat.ucla.edu/socr/index.php/SOCR\_Data\_LA\_Neighborhoods\_Data`.
- Find the table with the data and copy it, along with the headers, as we did in the linear regression tutorial.
- To find the SOCR Motion Charts Activity, go to: `http://socr.ucla.edu/SOCR\_MotionCharts`.
- Click on the "Data" tab, and paste all of the data (including the headers) from the L.A. Neighborhoods page into the data field in the Motion Charts application. To do this, select the first cell and press CTRL+V (Command+V on Macs).

Step 2: Setting Up the Motion Chart

- Click on the "Chart" tab. On the right side, there is a panel called "Mappings".
- *Key* is the time variable. We want to see all of our data at the same time, so select the blank.
- The *X-Axis* and *Y-Axis* attributes should be the Longitude and Latitude variables, respectively. Neighborhoods now will show up as bubbles on the motion chart, at the correct coordinates.
- *Size* is the attribute which will govern how big each bubble is on the motion chart. We will choose "income" as the size variable. This way, the higher the neighborhoods median income, the bigger its bubble will be.
- *Color* is the attribute which will govern the color of each bubble on the motion chart. We will choose "white" as the color variable. Neighborhoods with high percentages of Caucasians will be red.

- *Category* is the attribute which differentiates each bubble. We want each neighborhood to be its own bubble, so we choose "LA_Nbhd" as the category attribute. Also, when we hover our cursor over a bubble, the neighborhood name will be displayed. If a dialog box pops up, click "OK".

Step 3: Analyze the Motion Chart

- Now look at our motion chart:

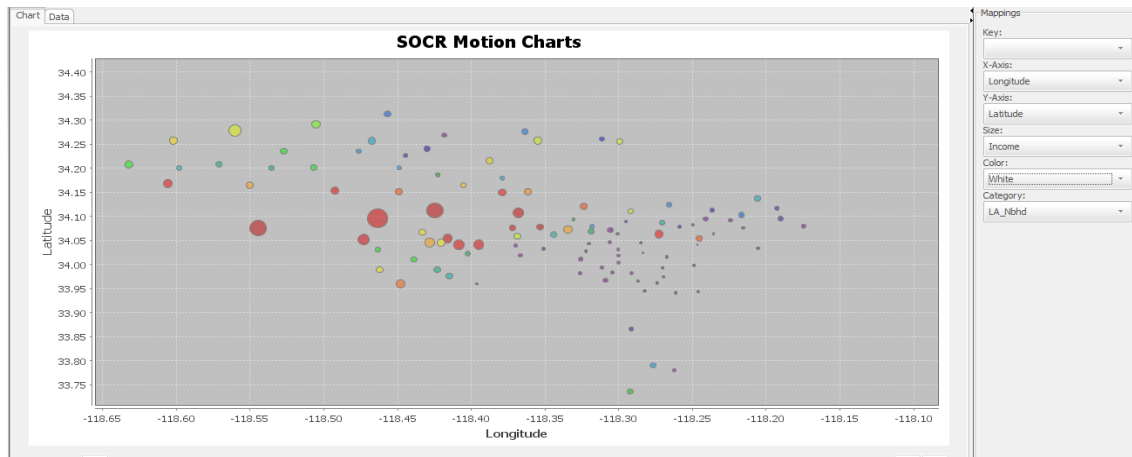


Figure 13: Income / White Motion Chart

- Each neighborhood in our dataset is represented by one bubble. Putting your cursor over a bubble will reveal the name of the neighborhood.
- Because we have longitude and latitude data, these bubbles are positioned in their correct coordinates.
- The size of each bubble is proportional to the median income of the neighborhood: the larger the bubble, the higher the median income.
- Bubbles representing neighborhoods with high percentage of whites will be red.
- Using this information, we can see some interesting things in this graph. For example, the Westwood bubble is very small, meaning it has a low median income. This is likely due to the fact that Westwood is a college town right outside of UCLA, and students generally have low incomes. Bel-Air and Beverly Crest are both represented by large red bubbles, reflecting the high median income and high percentage of whites in those neighborhoods. We can see a correlation between median income and the percentage of whites in the population. The bigger bubbles (neighborhoods with higher median income)

tend to be red (high percentage of whites in the population).

- You can also experiment with other variables to find other correlations and patterns. For example, the population could govern the bubble sizes and diversity could govern the bubble color:

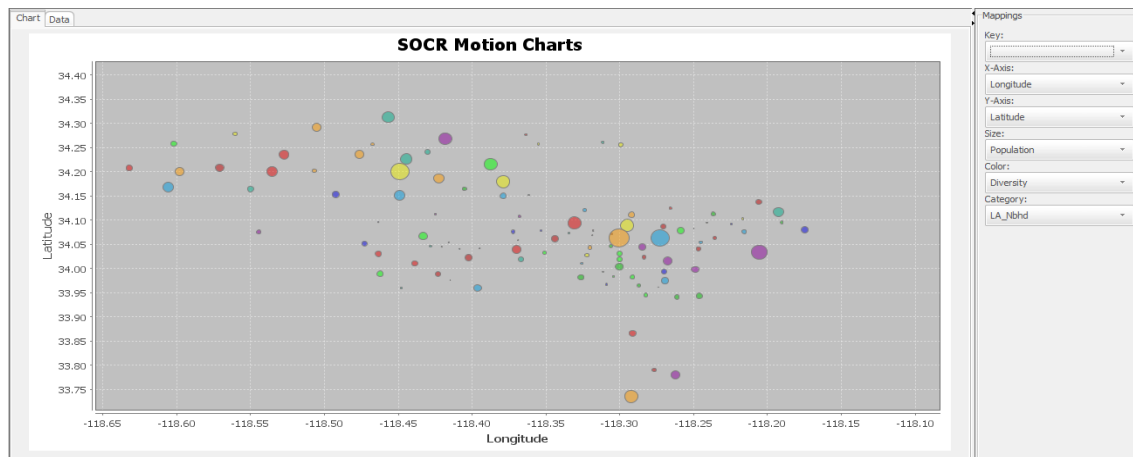


Figure 14: Population / Diversity Motion Chart

The SOCR Motion Charts activity and the interactive applets on the SOCR website makes it easy for users to explore different combinations of inputs and quickly and easily simulate different situations and models.

SOCR Motion Charts Tutorial (Earthquakes Data)

We will now look at another dataset and use it with SOCR Motion Charts. This time, we will utilize a time variable, so that we can see changes throughout time.

Step 1: Importing the Data

- We will be using the "California Earthquakes" data. This dataset contains variables on the date of the earthquake, its location, magnitude, depth, among other characteristics. The data can be found at:

http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_021708_Earthquakes

- After navigating to the link, find the table with the data and copy it (including the headers).
- Head to the SOCR Motion Charts page at:

http://socr.ucla.edu/SOCR_MotionCharts/.

- Click on the "Data" tab, and paste all of the data (including the headers) from the Earthquakes Data page into the data field in the Motion Charts application as we did in the previous tutorial.

Step 2: Setting Up the Motion Chart

- Click on the "Chart" tab. On the right side, there will be a panel called "mappings" with drop-down menus.

- *Key* is the time variable. To see all of the data on the Motion Chart at the same time, we can select the blank. For this tutorial, we want to introduce a time variable, so we will choose "Date_(YYYY/MM/DD)" as the *Key* variable.

- For *X-axis* and *Y-axis*, choose the Longitude and Latitude variables, respectively. Earthquakes will show up as bubbles on the chart, at the correct coordinates.

- *Size* is the variable which will govern how large each bubble is drawn on the Motion Chart. We will choose "Mag" as the size variable. Now the size of each bubble will represent the magnitude of each earthquake: the larger the bubble, the larger the magnitude.

- *Color* is the variable which determines the color of each bubble. Choose "Depth" as the color variable. The darker the purple, the deeper the depth of the earthquake.

- *Category* is the variable that differentiates each bubble. Choose "EventID" for this variable so that each bubble represents a different earthquake.

Step 3: Analyze the Motion Chart

- Look at the resulting Motion Chart:

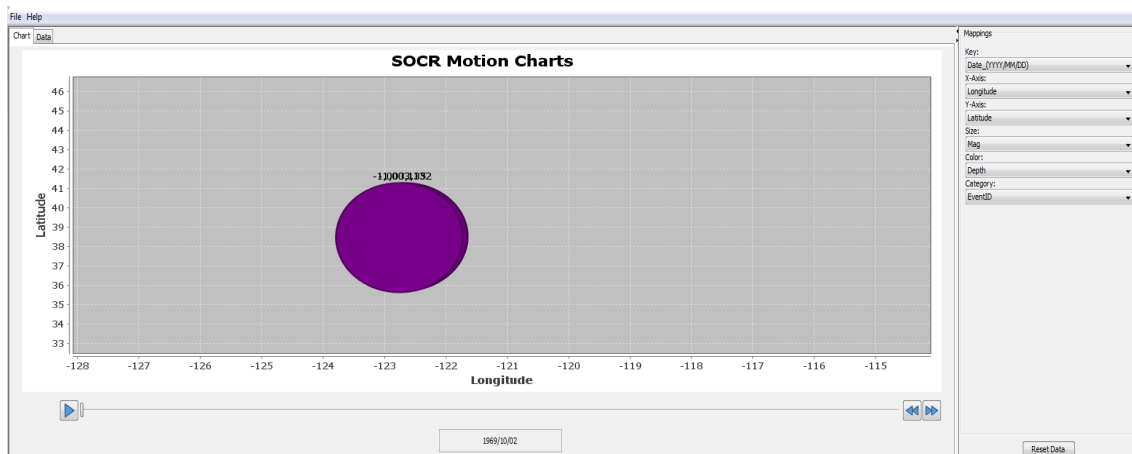


Figure 15: California Earthquakes Motion Chart

- There is a scrollbar at the bottom of the chart, which controls the time variable. Moving it to the right will advance time forward. The first date is 10/02/1969. Notice that there are actually two overlapping bubbles, representing two different earthquakes which both happened on that date.
- Moving the scrollbar to the right or clicking the "Fast Forward" button will advance through the data in time and show the earthquakes that happened on each date. The size of the bubble represents its magnitude, and the shade of the bubble represents the depth. Each earthquake bubble will also pop up in the correct longitude and latitude coordinates.
- Move the scrollbar all the way to the left to start from the beginning of the dataset. Now press the play button. As the Motion Chart scrolls through time, earthquakes will pop up at the correct times and locations.
- To see all of the earthquakes at the same time, select the blank as the *Key* variable:

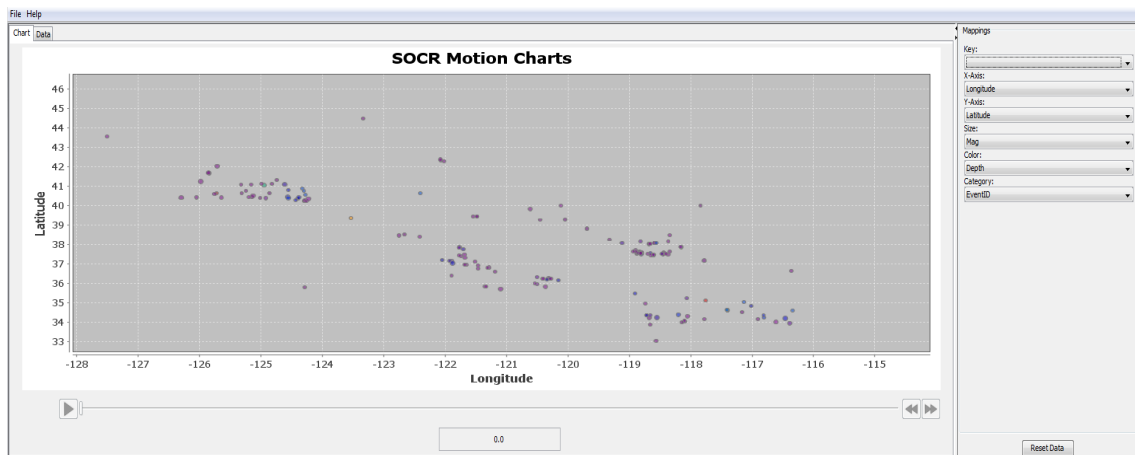


Figure 16: All of the earthquakes on a single motion chart

Conclusions and Discussions

Today, the amount of data available is huge, and the use of technology for statistical analysis work is essential. As such, the use of technology in teaching statistics is especially important. Interactive applets and data graphics are extremely useful for helping students understand more abstract or difficult concepts. Tools such as visualizations, graphs and simulations allow an additional and more intuitive way to present material. In addition, having such resources freely available on the internet allows students to augment their classroom learning on their own time and at a location convenient to them. Specifically, the Statistics Online Computational Resource (SOCR) website allows instructors to supplement traditional teaching materials with hands-on computer applets, simulations, and interactive graphical displays of statistical concepts.

SOCR includes many interactive applets, graphics tools, and instructional materials. There are applets for common statistical analyses techniques, such as Analysis of Variance (ANOVA), Chi-Square Goodness of Fit, Confidence Intervals, T-tests, and Survival Analysis. These applets allow statistical analysis to be accessible to anyone with an internet connection since there is no need for any special statistics software to be installed. For example, the SOCR Confidence interval applet allows users to run confidence intervals in an interactive way. The applet allows the user to sample from any of the 70+ distributions of SOCR, set specific parameters of the distribution, select the appropriate confidence interval parameter (μ, σ, p) , and then choose the sample size, confidence level, and number of intervals to construct. The above variables can all be adjusted and changed easily in the applet to allow for experimentation of confidence intervals. The Java applet can be found at:

http://www.socr.ucla.edu/htmls/SOCR_Experiments.html

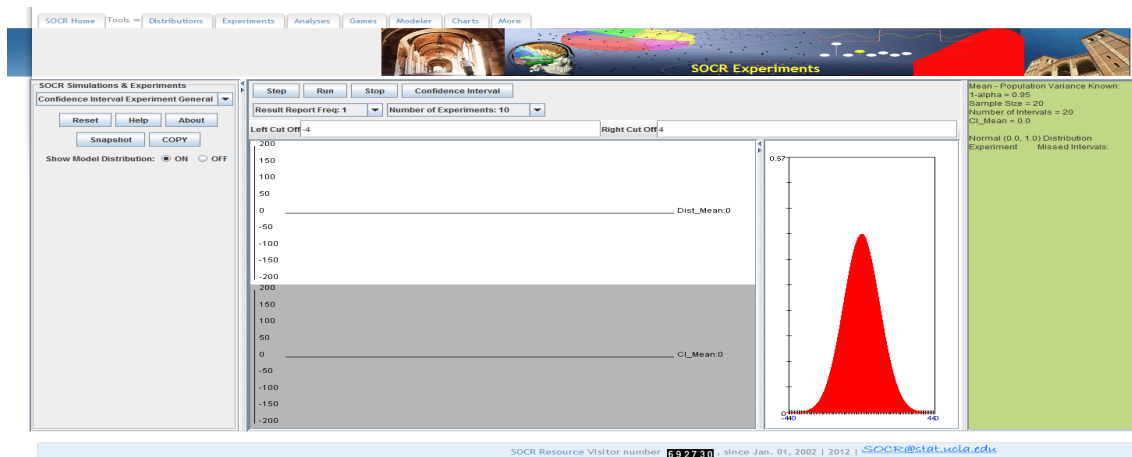


Figure 17: Confidence Intervals in SOCR

There are also many Experiments available on the SOCR website⁹. These include popular and well known statistics and probability experiments such as the Ball and Urn Experiment¹⁰, the Birthday Experiment¹¹, and the Dice Experiment¹². These SOCR experiments allow the user to observe many repeated runs of a simulation in a relatively short amount of time, helping them understand common statistical concepts.

In this paper, we have demonstrated a Multiple Regression Applet and a Motion Charts Applet. The Multiple Regression Applet Tutorial demonstrates proper procedures for using regression models. Important aspects of correctly utilizing regression tools, such as establishing assumptions for the model and interpretation of the betas are presented. The SOCR Multiple Regression Applet makes it easy to import data and to investigate different combinations of dependent and independent variables in the regression model. Also, the resulting regression model, along with relevant model attributes, is displayed clearly. The fit of the model is also easily ascertained by the provided residual and Normal QQ plots.

The Motion Charts tutorials utilize data graphics to improve understanding and introduce data visualization in support of the instruction of key statistical concepts. The tutorials provide an additional layer of interactivity with the material, with graphics that summarize large amounts of data in an easily digestible form. Patterns and structures previously clouded when the data was in a non-graphical format are now readily apparent. In addition to the questions that were investigated in the tutorials, the interactive nature of the Motion Charts applet encourages the user to explore other research questions, such as the relationships between income and home ownership, or between population and diversity.

References

1. Christou, N., Dinov, I. (2010) A Study of Students' Learning Styles, Discipline Attitudes and Knowledge Acquisition in Technology-Enhanced Probability and Statistics Education. JOLT 6:546 - 572.
2. SOCR ANOVA (2011) http://wiki.stat.ucla.edu/socr/index.php/SOCR_.\EduMaterials_AnalysisActivities_ANOVA_1.
3. SOCR Survival Analysis (2011) http://wiki.stat.ucla.edu/socr/index.php/tSOCR_EduMaterials_AnalysisActivities_Survival.
4. SOCR L.A. Neighborhoods Data (2011) http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_LA_Neighborhoods_Data.
5. SOCR Earthquakes Data (2011) http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_021708_Earthquakes.
6. Ho, A., et al. (2009) A Commonly Carried Allele of the Obesity-Related FTO Gene is Associated With Reduced Brain Volume in the Healthy Elderly. PNAS Early Edition 1-6.
7. Christou, N. (2011) Single Index Model http://www.stat.ucla.edu/~nchristo/statistics_c183_c283/statc183c283_single_index.pdf.
8. Tufte, Edward R. (2001) *The Visual Display of Quantitative Information*. Cheshire, Connecticut: Graphics Press LLC.
9. SOCR Experiments. (2012) http://www.socr.ucla.edu/htmls/SOCR_Experiments.html.
10. SOCR Ball and Urn Experiment. (2012) http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_BallAndRunExperiment.
11. SOCR Birthday Experiment. (2012) <http://wiki.stat.ucla.edu/socr/index>.

php/SOCR_EduMaterials_Activities_Birthday.

12. SOCR Dice Experiment. (2012) http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_DiceExperiment.