Statistical Modeling by Gesture: A graphical, browser-based statistical interface for data repositories[*]

James Honaker
Institute for Quantitative Social Science
Harvard University
1737 Cambridge St
Cambridge, MA 02138
jhonaker@iq.harvard.edu

Vito D'Orazio
Institute for Quantitative Social Science
Harvard University
1737 Cambridge St
Cambridge, MA 02138
dorazio@iq.harvard.edu

ABSTRACT

We detail our construction of *TwoRavens*, a graphical user interface for quantitative analysis that allows users at all levels of statistical expertise to explore their data, describe their substantive understanding of the data, and appropriately construct and interpret statistical models. The interface is a browser-based, thin client, with the data remaining in an online repository, and the statistical modeling occurring on a remote server. In our implementation, we integrate with tens of thousands of datasets from the *Dataverse* repository, and the large library of statistical models available in the *Zelig* package for the *R* statistical language. Our interface is entirely gesture-driven, and so easily used on tablets and phones. This, in combination with being browser-based, makes data exploration and quantitative reasoning easily portable to the classroom with minimal infrastructure or technology overhead.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Statistical Software; H.4 [**Information Interfaces and Presentation**]: User Interfaces—*graphical user interfaces*; G.4 [**Mathematical Software**]: User Interfaces

Keywords

statistical user interfaces, open data, directed graphs

1. INTRODUCTION

With the proliferation of open data sources and replicable data repositories comes the promise of increased access to scientific information and the democratization of quantitative knowledge. However, meaningful analysis of this multitude of data remains outside the grasp of analysts who lack training in statistical modeling or knowledge of and access to statistical software platforms. For those with the necessary expertise, access is still limited by data transfer bottlenecks and installation and hardware overhead. To reduce barriers to statistical and quantitative reasoning and to promote the proliferation of empirical knowledge, we introduce the TwoRavens interface.

TwoRavens is a gesture-driven, Web-based device for the analysis of statistical models and the exploration of data. It is open-source and integrates with Dataverse repositories for open archiving of data[12, 6, 7], providing users access to the more than 50,000 data files currently housed in Dataverse repositories, as well as new data that users may upload free of charge. The statistical analysis component is powered by Zelig, a library for statistical inference in *R* that provides access to dozens of statistical models through a common call structure [14, 11]. TwoRavens links the power of Zelig to one of the largest database collections anywhere, requires no installation, and is accessible to devices ranging from mobile phones to tablets to smartboards.

For ease of use, the UI is designed to mirror the common quantitative workflow. Moving left to right on the interface, users make intuitive column and row selections or subsets of the data.[1] Users may perform transformations on these selections using the input box in the top right. In the center, users specify the intended statistical model in the framework of a directed graph, and have the option to perform transformations on selected columns or tag selections with specific properties, such as "nominal." Finally, on the right, users select their statistical model and have the option to set covariate values on right-hand side variables to obtain additional bootstrapped simulations of quantities of interest. Upon estimation, statistical results are displayed using a combination of graphs and tables.

In this article we describe the UI and its integration with Dataverse and Zelig. In future research we intend to build upon this foundation in two ways. First, as users supply information about hypothesized relationships in the data, TwoRavens will synthesize their input and automatically suggest appropriate modeling decisions, such as functional form and choice of model. Second, as models are estimated, the results will be cumulatively recorded across users to form a coherent map and meta analysis of the statistical results that exist in Dataverse.

2. DESIGN GOALS AND PRINCIPLES

In designing this software, some key goals were set that we believe the next generation of statistical tools will need to achieve. The challenges and solutions to reaching these goals informed some general design principles, which we feel are important to discuss and convey for future researchers. A paramount goal was to keep the interface as a **thin client**, keeping both the data itself, and the work of statistical

[*] All code, documentation, and numerous examples for our open source *TwoRavens* software project are available at:
`http://datascience.iq.harvard.edu/tworavens`

[1] Depending on the discipline, columns are known as variables, fields, features, etc., and rows are known as observations, cases, units of analysis, etc.

processing, remote from the interface. We also required a **broadly accessible** platform that could be productively used by individuals at all levels of statistical expertise, from novice users with no statistical training, to experienced quantitative researchers. Furthermore, to make data accessible to users with different levels of available computational infrastructure, we worked to make this software **device independent** and feasible not only on traditional computers, which have previously been the sole domain of statistical software, but tablet, mobile, and other smart devices.

2.1 Goal: Thin Client

Although we have written an interface for statistical data analysis, neither the data itself, nor any statistical analysis, occurs or resides at the client side.[2] It is a challenge to explore data, and to implement statistical models, without the the immediacy of local interaction. However, having a thin client with remote data and remote statistical processing allows some enormous, novel advantages that are increasingly useful in modern data-intensive science: 1) When datasets are large, transferring the data to the user for exploration can be the primary bottleneck for speed, which this overcomes. 2) Similarly, when datasets are large, the final analysis might require distributed processing, which again requires another transfer of the data. In our architecture, again, the code for statistical processing is instead placed close to the original data. 3) When the data has privacy implications, there are settings where it is allowable to grant access to summaries and statistical representations of the data, (or privacy preserving versions of these statistics, such as those that are provably differentially private [9, 19, 8]), but not grant access to individual observations of the data itself. This architecture, where both the data and the analysis on the data remains remote, can be an enforceable way to grant access to private data in these settings.

2.2 Goal: Broadly Accessible

Statistical consulting is increasingly a necessary service provided by research universities to aid their research faculty. Most substantive researchers are heavily invested in the collection of their data, understand their variables well, and possess expert substantive knowledge about the plausibility of various relationships. What they may not know is the set of plausible statistical models appropriate to their objectives, the statistical language to describe their goals and aims, or the software knowledge to implement these choices. Consultants often provide that link, allowing experts without statistical training to translate their knowledge and goals into appropriate statistical software routines, and interpreting the results back in a substantively meaningful fashion. Obviously, however, not every user of data has access to a trained statistician. One overarching goal of the TwoRavens project is to provide as much of this service as can be automated. At this stage of the interface, this means providing a way for users to intuitively communicate all their knowledge about the data, so an appropriate model can be constructed for the quantitative task at hand.

2.3 Goal: Device Independent

Part of expanding the set of users who have access to quantitative reasoning, means lowering the infrastructure requirements to statistical analysis. While professional researchers may have extensive computational resources, many other settings, especially community colleges and high schools, have limited resources for instructors and none for students. Even free and open source statistical packages like *R* generally require the expenses of some level of IT support to install in a networked lab. Moreover, even in well-funded universities, substantive classes that use quantitative data may have no classroom lab access. Our goal is to enable a computationally capable access to statistical exploration with very minimal infrastructure.

2.4 Resulting Design Principles

In meeting these goals, we developed the following design principles for this interface: 1) **Browser Based:** The ability to run the software entirely through a browser, with no additional installed software, is possible because of the thin nature of the interface, and helps facilitate device independence as it can run on any device that includes a browser 2) **Gesture Driven:** Common statistical packages are normally command line, script, or menu driven, but to facilitate ease of use across devices and without user expertise, we have tailored a process that allows full exploration and setup of statistical models by interacting directly with the data though gesture 3) **Graphical Representation:** Directed graphs, also called probabilistic networks, are increasing used by statisticians as a representation of complex data generation processes [10], particularly in structural equation modeling as well as casual inference [16]. Using directed graphs to convey possible relationships between variables allows novices to convey all of their substantive understanding of a dataset, substantive experts and teachers to communicate in a manner conducive to qualitative discussion, and still allow statisticians to explicitly define their hypotheses of the data generating process. 4) **Maximal computational leverage:** Most statistical packages are interactive in a command driven relationship, waiting for instructions before creating any analysis product. To facilitate the architecture of the interface, we instead have an impatient design that preprocesses any graphs and summary statistics that we can envision might be needed, so they are already loaded on the thin client, without needing data interactions or queries. Much of this preprocessing can be done when the data is ingested to the repository, before the user has even discovered the data.

3. TWORAVENS SOFTWARE

The interface is written in Javascript and incorporates aspects of interactive graphics, Web-based statistics, and customized R applications. The interactive visualizations use Data-Driven Documents (D3), which has been influenced by tools such as Protovis and Processing [4, 3, 17]. Although components of TwoRavens appear as a Web interface for statistical software (comparable to R-fiddle or SAS OnDemand), its interaction with R is closer to that of applications created using tools such as RStudio's Shiny [18]. In its entirety, TwoRavens is comparable to GleamViz, which contains an interactive statistical modeling tool and whose remote servers handle the necessary processing [5]. Where TwoRavens distinguishes itself from GleamViz is that it does not require a desktop client and it is a general-purpose statis-

[2] Or slightly rephrased, we have built a package for statistical analysis of data, that can not handle data, nor analyze any statistics.

tical modeling tool, whereas GleamViz is tailored for modeling infectious diseases. Thus, as a Web-based, gesture-driven tool for statistical modeling, TwoRavens is unique.

Furthering its distinction from existing statistical software, it integrates with Dataverse, providing instant access to tens of thousands of datasets by simply launching TwoRavens from any Dataverse repository page, and with Zelig, delivering analysts meaningful yet easily interpretable statistical estimates and graphics. Given the availability of open access data in repositories such as Dataverse, the open source power of R, and the trend towards Web-based, interactive visualizations, such a device ties together many threads of modern quantitative computing.

3.1 Dataverse and Zelig Integration

In the last decade, the Data Science team at Harvard's Institute for Quantitative Social Science (IQSS) has developed software infrastructure and tools to facilitate and enhance data sharing, preservation, citation, reusability and analysis [13]. Over that time, the team has continuously developed two software products now widely used by the research community: *Dataverse*, a repository infrastructure for sharing research data, and *Zelig*, a statistical package for *R*.

Dataverse is "an open source data repository, which allows one to publish, share, reference, extract, and analyze research data" [7]. The Harvard Dataverse Network contains more than 52,000 studies with more than 700,000 files, and is the world's largest collection of social science data sets (`http://thedata.org`). The primary connection between TwoRavens and Dataverse is an API accessing metadata that complies with formatting standards set by the Data Document Initiative (DDI), an organization that promotes diligent data management [1]. In addition to keeping it thin, this facilitates the deployment of TwoRavens to any data repositories that comply with DDI standards.

Zelig is a wrapper and interface that allows a large body of different statistical models in the R statistical language to be used from a unified call structure [11, 14]. It is also a modeling architecture that interprets these statistical models in a substantively meaningful fashion [15]. By integrating with Zelig, TwoRavens has minimal backend manipulations, other than to map the information that has been entered by the user into an appropriate Zelig call. Thus, new functionality in Zelig may be translated to new functionality in TwoRavens rather seamlessly. Both R and Zelig are open source and freely available.

4. THE USER INTERFACE

Javascript is lightweight, and allows us to run the interface entirely through an internet browser. For ease of use and cross-platform portability, all functionality has gesture-driven capability, so statistical models on datasets stored in online repositories could be run from a tablet or mobile device without a keyboard. This ability expands the set of ways individuals use archived data and quantitative analysis, including bringing real-time data analysis into the classroom without using a computer laboratory setting.

A screenshot of TwoRavens is shown in figure 2. The workflow of developing a statistical analysis moves from left to right, first examining and selecting variables in the dataset, then constructing a framework of possible relationships, and then choosing and interpreting an appropriate statistical model for that framework. In figure 2, the user has selected a model, tagged `ti_cpi` as the dependent variable, and graphed an appropriate statistical model.

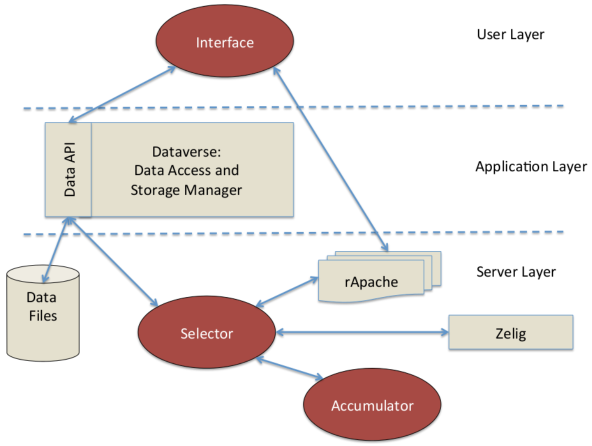


Figure 1: Architecture, including integration with *Dataverse* archival data repository, and *Zelig* library of statistical models for *R*.

4.1 Left Panel - Data Selection

Each variable in the dataset can be introduced as a node in the center space by clicking on the variable name in the left panel. Users may not be familiar with each variable in the dataset, so on mouseover information pertaining to that variable appears, such as primary summary statistics, graphs, metadata and short descriptions of the variable.[3] This information is precalculated by Dataverse when a new dataset is uploaded, and stored in a json file that is compatible with the DDI schema [1].

Although not explicitly in the left panel, TwoRavens includes an option to perform transformations on variables, such as taking the log of a variable or multiplying it by some factor, using the input box in the top right. Users may transform variables in two ways: (1) by manually entering text into the transformation input box; or (2) by clicking on the input box, selecting a variable from the drop down list, and then selecting a transformation from the function list. In this way, the gesture-driven functionality is preserved, while allowing additional flexibility for users familiar with *R* functions.

Users may be interested in subsetting the dataset to examine only observations that have specific values (for example, only European countries, or only respondents over the age of 60). For such cases, we include a *Subset* tab that shows the distribution of each variable. These distributions are either a density plot or a bar plot, depending on the variable's level of measurement and its number of unique values. By brushing the plots with the pointer, users select ranges of that variable upon which the data is to be subsetted. The numbers associated with the range are shown so that users may be precise in the ranges they specify. All metadata is remotely recalculated for the new subset using the same Dataverse ingest routines, and a new space that represents the subsetted data is added to the carousel in the center space of the interface.

[3] For devices not compatible with mouseover, we enable this feature via click and hold.

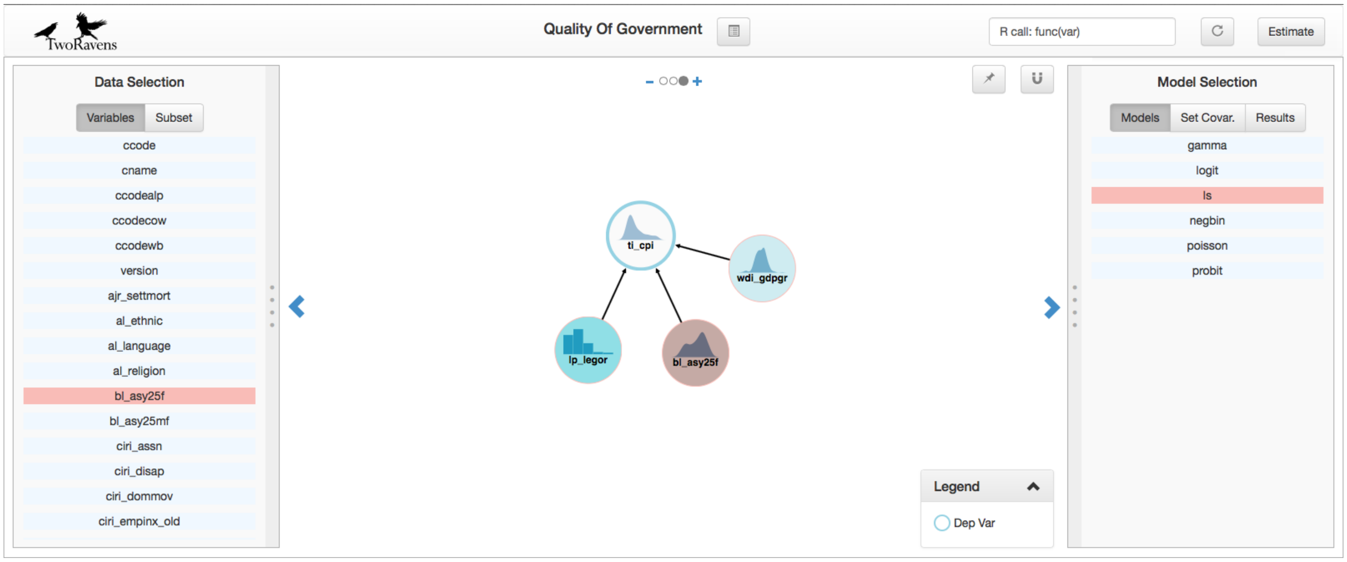


Figure 2: Basic UI with Dependent Variable and Model Selected

4.2 Center Space - Relationship Mapping

In the focal, central space, the substantive knowledge of the researcher is easily communicated by drawing a directed graph representation. With a two-finger click, arrows are drawn connecting nodes, depicting the possible relationships between variables. The nodes and arrows are an application of D3's force layout. In the simulated forces that propel the visualization, each node has a gravity, which draws all the nodes together, as well as a charge that repels them from getting too close. The arrows act as simple springs, and are removed when clicked. As the graph between the variables is built up by the researcher, it dynamically rearranges itself by these simulated forces. The researcher can also physically drag the pieces around, which acts as an additional force in the dynamic visualization, or completely turn off the simulated forces, and place every node manually for complete control of the representation (using the *force* toggle icon, represented by a pin).

Dependent variables, time, cross-sectional, and nominal identifiers are properties that may be tagged, by the user, to a node. Each property is denoted with their own colored halo. To tag a property, a user mouses over a node in the directed graph, at which point buttons appears as arcs around the perimeter of the node. Each arc is colored and labeled, and clicking on the arc associates that property to that node. At this point, the halo is colored appropriately and a legend appears, reminding users what the color of the halo represents. Additionally, the background color of the tagged variable in the left panel is changed to reflect the color of the tagged property.

The portion of the center space that is visible is actually just one element of a carousel, and users have the option to add and remove new workspaces (hereafter, elements). For example, if a user subsets the data, then an additional element is added to the carousel that corresponds to the subsetted data. Users toggle between elements by clicking a chevron or by clicking and swiping in the direction they want the carousel to move. Their current element is shown by highlighting its corresponding dot in the top-center of the center space. By clicking on the plus sign to the right of the dots, users are duplicating the current carousel, and placing its representation at the right of the element array. By clicking the minus sign, users are dropping this element from the carousel. To clear a modeling space but not drop it from the carousel, users may select the Erase icon, represented with a magnet.

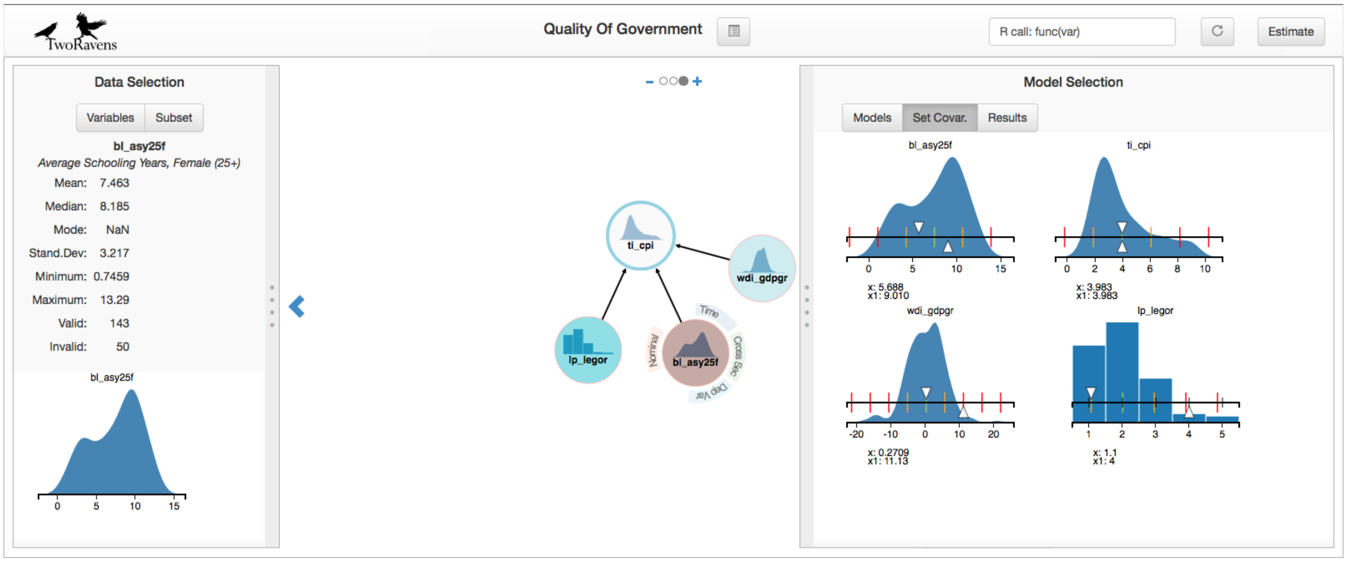
4.3 Right Panel - Model Implementation

After examining the data and constructing a diagram of relationships, in the right panel users begin to investigate statistical models. The *Models* tab provides a list of the available statistical models that can be employed by Zelig. On mouseover, users see a brief description of the model so that they have some guidance as to which to select.

The next tab, labeled *Set Covar.*, provides the ability to interpret any estimated model by means of predicted and expected values at chosen values of the covariates, as well as first differences created by the changes in the predictions across changes in the covariates [15]. As shown in figure 3, users may choose values of the covariates at which to interpret the model by means of sliders superimposed on the densities of the variables. The slider positions initially default to the mean of each variable, while the scale of the slider marks each standard deviation away from the mean within the range of the variable. For bar plots, the value of the bar is also placed on the slider's scale.

When this information is complete, the researcher can estimate the model by clicking the *Estimate* button. At this time, the information extracted from the user is passed to an instance of an *R* application hosted on a remote server.[4] This remote application first builds a formula representation of the model from the graph connections the researcher has constructed. In the present version, this is all variables that have a path to the dependent variable, but more complex graphs can include intermediate or post-treatment variables as well as consequences of the dependent variable, which are

[4] We developed our *R* application in Rook and host it in rApache.



bl_asy25f

Average Schooling Years, Female (25+)

Mean:	7.463
Median:	8.185
Mode:	NaN
Stand.Dev:	3.217
Minimum:	0.7459
Maximum:	13.29
Valid:	143
Invalid:	50

x: 5.688
x1: 9.010

x: 3.983
x1: 3.983

x: 0.2709
x1: 11.13

x: 1.1
x1: 4

Figure 3: Node Description (left panel) and Set Covariate Values (right panel)

useful for forecasting and imputation, but omitted from the formula for a causally oriented analysis. Using this formula, the *R* application calls the applicable statistical model from the Zelig library. The results from Zelig are asynchronously returned to the browser interface. As can be seen in figure 4, the plots that Zelig produces, as well as a table of estimates, are available for viewing inside the *Results* tab.

5. EXAMPLE APPLICATION

The Quality of Government (QoG) represents one database for which the primary benefits of our tool might be recognized for two target audiences, the novice user and the classroom instructor. The QoG is a collection of country-year datasets whose variables include data on population, respect for human rights, and political regime type [20]. Its primary objective "is to address the theoretical and empirical problem of how political institutions of high quality can be created and maintained" [2].

A simple empirical exploration of this question using the QoG data, however, can be fraught with difficulties for novice users. Minimally, one needs to be familiar with some statistical software package and to understand enough statistics to be able to analyze the data. The QoG data would have to be downloaded locally, as would the software being used.

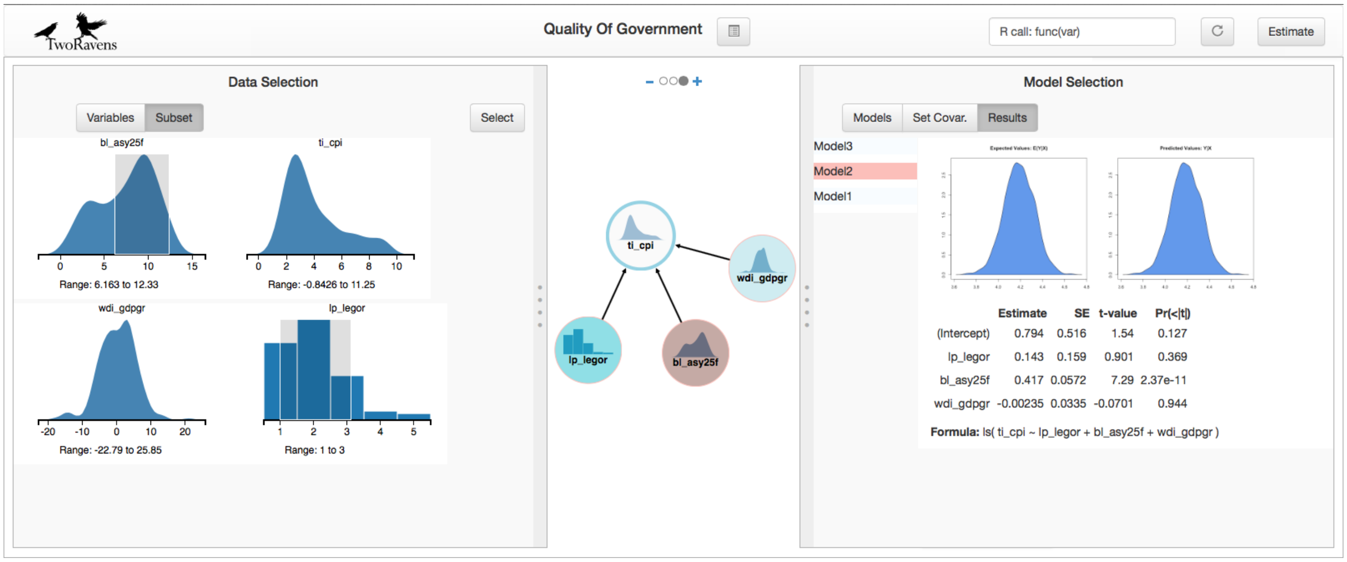
Our tool improves accessibility by reducing or removing these initial barriers. Users interact with the data in visual, gesture-based ways, and so the time spent learning how to use TwoRavens is negligible in comparison to what is necessary to analyze data with R, for example. Statistical models are represented visually as a directed graph, and users may instantaneously view each variable's distribution and summary statistics by hovering over a node in the graph. For example, Figure 3 shows a description of the QoG variable `bl_asy25f` in the left panel. The variables are listed in the left panel, and are added and removed from the modeling space by clicking on the variable name. The modeling space, the center of the figure, shows a representation of a user-defined statistical model.

For instructional purposes, the QoG represents a type of dataset that may be used for a class project. It contains 746 columns of data in a time-series, cross-sectional format. Many of these columns are substantively interesting dependent variables, while others appear as explanatory variables in many models in quantitative Political Science. For instructors to teach how to use data to study questions in politics, requires some degree of expertise in a statistical software package, a projector, and a computer with the rights to that software. For student projects, each student would have to download the data and some statistical software to their personal computer, and the instructor would have to provide materials and guidance on software usage. Most statistical software is proprietary, so students are often restricted to labs that are equipped with the necessary licenses. TwoRavens removes these barriers, and all that is necessary for the instructor to bring data to the classroom, and for students to analyze the data, is an internet connection and a Web browser.

6. CONCLUSIONS

TwoRavens is a Web-based tool for statistical analysis that is lightweight, broadly accessible, and device independent. It integrates with Dataverse, providing access to the tens of thousands of data repositories that exist there, and leverages the power of Zelig, an R library that provides a common call structure for a large number of statistical models. Entirely gesture-driven and intuitive for users of all levels, TwoRavens reduces barriers to statistical analysis and promotes the proliferation of empirical research.

This article details the design of the user interface and its applicability for use by statistical novices and users not familiar with or who do not have access to statistical software. Although integral and foundational to the TwoRavens project, the UI is only the first of three layers. In future research, the TwoRavens project will be adding the model "selector" and results "accumulator" layers to provide more automated guidance on model selection and specification. The selector synthesizes the user input and automates suggestions, such as potential omitted variables, functional form,



	Estimate	SE	t-value	Pr(<|t|)
(Intercept)	0.794	0.516	1.54	0.127
lp_legor	0.143	0.159	0.901	0.369
bl_asy25f	0.417	0.0572	7.29	2.37e-11
wdi_gdpgr	-0.00235	0.0335	-0.0701	0.944

Formula: ls(ti_cpi ~ lp_legor + bl_asy25f + wdi_gdpgr)

Figure 4: Results Shown in Graphs and Table

and choice of statistical model. The accumulator stores all models that have been estimated on each dataset, and provides users with feedback on existing research using that dataset and datasets judged to be similar.

7. ACKNOWLEDGMENTS

The authors would like to thank Mercè Crosas and Gary King for extensive feedback and ideas, Leonid Andreev, Michael Heppler and Elizabeth Quigley for continued development assistance, and Dwayne Liburd for creating our logo.

8. REFERENCES

[1] Data document initiative. `http://www.ddialliance.org`, Jun 2014.

[2] Quality of Government. `http://www.qog.pol.gu.se/research/`, May 2014.

[3] M. Bostock and J. Heer. Protovis: A graphical toolkit for visualization. *IEEE Trans. Visualization and Comp. Graphics,*, 15(6):1121–1128, 2009.

[4] M. Bostock, V. Ogievetsky, and J. Heer. D3 data-driven documents. *IEEE Trans. Visualization and Comp. Graphics,*, 17(12):2301–2309, 2011.

[5] W. V. Broeck, C. Gioannini, B. Gonçalves, M. Quaggiotto, V. Colizza, and A. Vespignani. The gleamviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale. *BMC infectious diseases*, 11(1):37, 2011.

[6] M. Crosas. The dataverse network: An open-source application for sharing, discovering and preserving data. *D-Lib Magazine*, 17(1-2), 2011. Available at: `http://j.mp/12yqVCZ`.

[7] M. Crosas. A data sharing story. *Journal of eScience Librarianship*, 1(3):173–179, 2013.

[8] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pages 265–284. Springer Berlin Heidelberg, 2006.

[9] C. Dwork, M. Naor, O. Reingold, G. N. Rothblum, and S. Vadhan. On the complexity of differentially private data release: efficient algorithms and hardness results. In *Proceedings of the 41st annual ACM symposium on Theory of computing*, pages 381–390. ACM, 2009.

[10] D. Edwards. *Introduction to graphical modelling.* Springer, 2000.

[11] K. Imai, G. King, and O. Lau. Toward a common framework for statistical analysis and development. *Journal of Computational Graphics and Statistics*, 17(4):892–913, 2008.

[12] G. King. An introduction to the Dataverse Network as an infrastructure for data sharing. *Sociological Methods and Research*, 36:173–199, 2007.

[13] G. King. Restructuring the social sciences: Reflections from Harvard's Institute for Quantitative Social Science. *PS: Political Science and Politics*, 47(1):165–172, 2014.

[14] G. King, K. Imai, and O. Lau. Zelig: Everyone's statistical software, 2007. http://zeligproject.org.

[15] G. King, M. Tomz, and J. Wittenberg. Making the most of statistical analyses: Improving interpretation and presentation. *American journal of political science*, 44(2):347–361, 2000.

[16] J. Pearl. *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press, 2000.

[17] Processing.js. `https://www.processing.org`, Jul 2014.

[18] Shiny. `http://shiny.rstudio.com`, Jul 2014.

[19] A. Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the 43rd annual ACM symposium on Theory of computing*, pages 813–822. ACM, 2011.

[20] J. Teorell, N. Charron, S. Dahlberg, S. Holmberg, B. Rothstein, P. Sundin, and R. Svensson. The quality of government basic dataset made from the quality of government dataset, 2013. Available at: `http://www.qog.pol.gu.se` [Accessed: 15 May 2014].