



A User Guide to TwoRavens: An overview of features and capabilities

Vito D'Orazio*
James Honaker†

February 20, 2016‡

1 This Guide

The TwoRavens interface has been built to make data exploration quick to get going and intuitive to use. We hope that you can start exploring data just by launching the interface and browsing through the functions you see (or by watching our 3 minute introduction videos). However, this guide provides additional details about the features available in the TwoRavens interface. It also intends to explain some underlying design choices that might help to shape how you approach and interact with the interface.

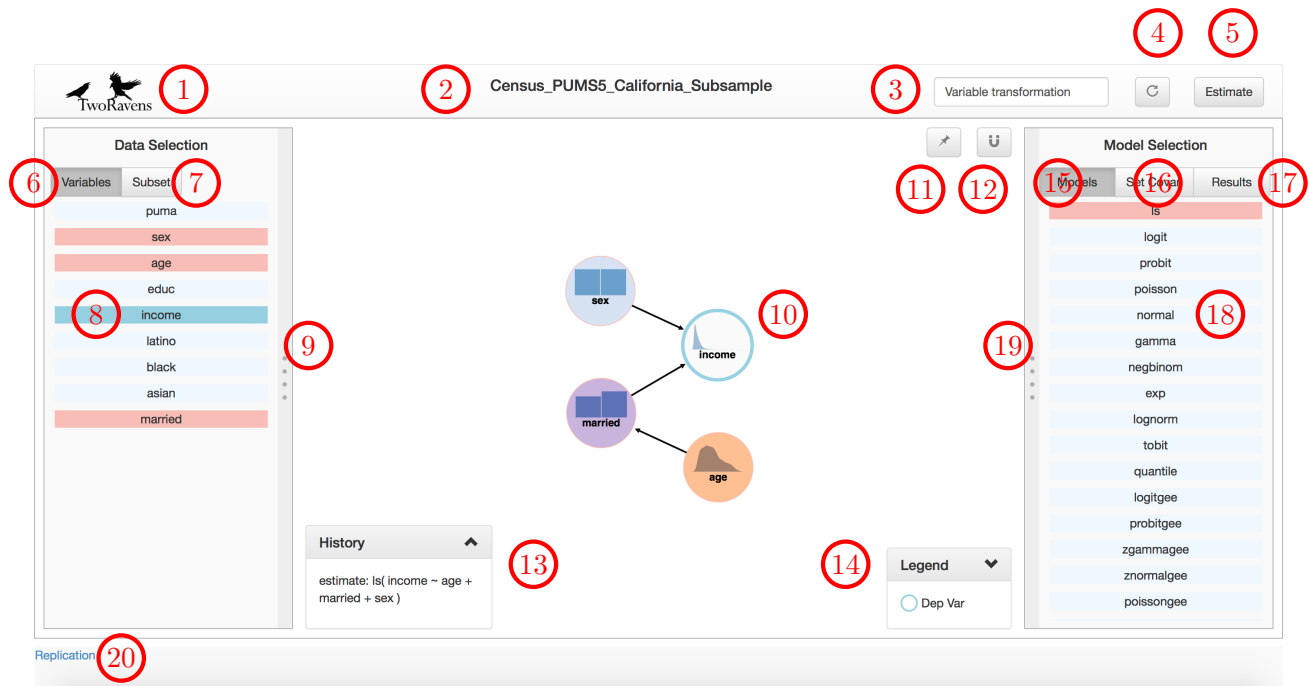
The figure on the next page labels twenty features that can be immediately seen when TwoRavens is launched with a dataset, while the accompanying table describes each with a one sentence explanation. This overview sheet may be all you need to get going. However, each of these features are explained in increased detail in the rest of the document; use the numbered references to find the appropriate section in the document, or click on the links in the table below the figure. Before detailing the features, we first give an overview of the design of the layout and controls and data handling.

*School of Economic, Political, and Policy Sciences at the University of Texas at Dallas; vjdorazio@gmail.com

†Institute for Quantitative Social Science, Harvard University; jhonaker@iq.harvard.edu; <http://hona.kr>

‡Current version of this document available at <http://2ra.vn/guide>

TwoRavens Feature Overview



HEADER

- | | | |
|---|-------------------------|--|
| 1 | About Image | Mouseover to see details about this software version. |
| 2 | Dataset Name | Mouseover to see dataset citation and doi information. |
| 3 | Transformation Prompt | Construct new variables by entering transformations. |
| 4 | Reset Button | Quickly revert the screen to the initial settings. |
| 5 | Model Estimation Button | Run selected model estimation and report back results. |

DATA SELECTION PANEL

- | | | |
|---|---------------|--|
| 6 | Variable Tab | Click to list available variables. |
| 7 | Subset Tab | Click to construct subsets of the dataset. |
| 8 | Variable List | Mouseover variable names to see table of summary statistics. |
| 9 | Toggle Bar | Click on bar to open or close the Data Selection Panel. |

EXPLORATION PANEL

- | | | |
|----|---------------|--|
| 10 | Model Builder | Build a graphical representation of relationships between variables. |
| 11 | Pin Button | Press this button to hold the variable pebbles in place. |
| 12 | Wipe Button | Press this button to remove everything from the exploration panel. |
| 13 | History Log | Summary listing the procedures run in this session. |
| 14 | Model Legend | Legend explaining the graphical symbols being used in the model graph. |

MODEL SELECTION PANEL

- | | | |
|----|------------------------|---|
| 15 | Models Tab | Click to list available statistical models for this dataset. |
| 16 | Set Covariates Tab | Click to set points at which to interpret model results. |
| 17 | Results Tab | Click to see list of results from session models estimated. |
| 18 | Statistical Model List | Mouseover to see additional details about each available model. |
| 19 | Toggle Bar | Click on bar to open or close the Model Selection Panel. |

FOOTER

- | | | |
|----|----------------------|---|
| 20 | Replication Dialogue | Click to download file to replicate all analysis in this session. |
|----|----------------------|---|

2 The Layout

The features are grouped together by the five regions of the interface: the header and footer, and the three panels. We typically anticipate that TwoRavens works in a left-to-right workflow, with exploration of the variables in the left panel, followed by construction of relationships between variables in the center panel, and estimation and interpretation of statistical models in the right panel. While real workflows will inevitably cycle between these phases as the data is explored in increasing depth, that underlying model might help you remember where different features are located, or find additional functions you are looking for.

The header and footer act like parentheses to the whole analysis, with the header containing meta-information about the dataset that will never be adjusted by the user, such as the citation, and the footer containing information useful after the analysis is complete, such as links to a replication file. More details on the design of the interface can be found in [Honaker and D’Orazio \(2014\)](#).

3 Gesture Controls Across Devices

A goal of TwoRavens is to remove the infrastructure barriers to immersion in data. TwoRavens does not require any installed software that has to be maintained or kept up to date. It does not require any local storage for datasets, or any local processor power for running even the most complicated statistical models. Its intuitive gesture and visual approach allows users to quickly engage with data, on any available device, including not just computers, but smart boards, tablets and phones. All that is needed is a web browser and an internet connection. If used in a lab, no additional software needs to be installed or maintained. If used in a classroom students can use their own variety of devices. In this guide, we occasionally mention “mouseover” as a possible action, while on touch devices, this generally means click.

4 Data Storage and Manipulation

In addition to being visual and gesture based, a key part of the capabilities of TwoRavens is the fundamental change in architecture that it uses for data storage and statistical computation. Although this architecture is sophisticated, it should be seamless for the user, indeed easier than current statistical software platforms. All data is remotely stored in repositories. All statistical processing occurs remotely on servers. The TwoRavens interface allows statistical exploration of datasets, but never actually touches the data, or performs statistical computation itself. TwoRavens only understands meta-data, that is, high-level information about the data, and while the user should feel they are immersed in an exploration of the dataset, they are actually communicating with the data remotely through this meta-data, which is more interpretable, meaningful and informative. What is key to understand is that the data never actually comes to the user, only information about the data. This makes TwoRavens very powerful for exploring big data, where the data itself is too large to transfer to the user, and too computationally demanding for arbitrary user devices to process. It also allows TwoRavens to interact with data that requires privacy preservation, for example where summary statistics are allowed to be made public, but the raw data contains private information that can not be viewed.

At present we connect to data archived on an instance of Dataverse ([King, 2007](#)) ([Crosas, 2011](#)), but are working to use many other storage modes. Statistical computation is all done in *R* ([R Core Team, 2015](#)), primarily using the Zelig statistical library ([Choirat *et al.*, 2015](#)).

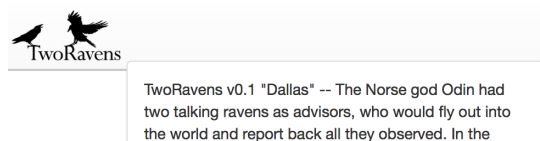
5 The Feature Descriptions

Header

The *Header* bar across the top of TwoRavens contains some preliminary information about TwoRavens and the dataset that has been opened. The header contains abstract information that is true regardless of the exploration and analysis performed, such as the name and citation to the data used, and the state of the software (for example, the version of TwoRavens being used, and the readiness of the server to run estimation).

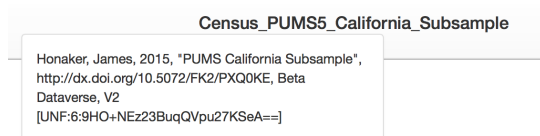
1. About Image

A TwoRavens icon can be found on the top left of the interface. On mouseover, a message will describe the current version number and release name of this instance of TwoRavens.



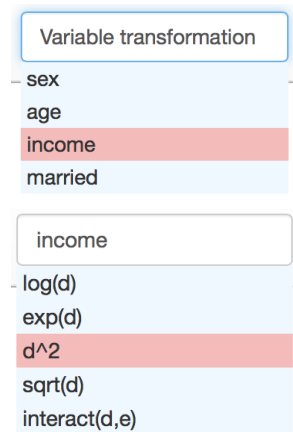
2. Dataset Name

The name given to the dataset is shown in the center of the header. This name is read from the metadata file of the dataset in the data repository. If the dataset has a defined citation or a digital object identifier (DOI) these will appear on mouseover of the name (or when the name is clicked on a touch device). An example is shown here where both are provided.



3. Transformation Prompt

The *Transformation Prompt* allows users to generate new variables from the original variables in the dataset. Common transformations, such as logarithms, squared terms and interactions, can be generated by the drop down menu. Clicking on the variable transformation prompt brings up a list of all the variables currently selected into the center panel. When a variable is selected from this list, a new list of common transformations appears. When a transformation is selected, a new variable will be generated by that transformation and added to the variable list and center panel. If more flexibility is needed, new variables can also be generated by writing expressions in this prompt in *R* syntax.



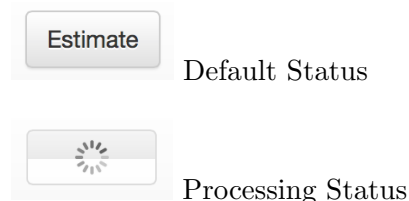
4. Reset Button

The reset button will refresh the TwoRavens browser page and return all settings, options and selections to the original state when the dataset was loaded.



5. Model Estimation Button

When clicked, the estimate button sends the constructed model to the server to estimate the outputs. During estimation, the button will show a processing icon as shown on the right, although this may occur so fast as to not be noticeable. When estimation is complete, the button will return to the default status and is ready estimate the next model. Results will then be available in the result tab. If anything necessary is missing from the model construction (for example, if a dependent variable was not selected) a reminder message will appear.

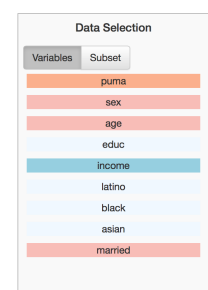


Data Selection Panel

The *Data Selection Panel* allows selection of observations and variables from the dataset for the analysis. Commonly the first step in data exploration or analysis is to understand the variables present, and select those that are interesting, as well as select the observations or population that are interest, and these tasks can be accomplished in this panel.

6. Variable Tab

When the tab marked *Variable* is selected, a list of the names of all the variables in dataset will appear. The list is in the order of the variables in the file. Clicking on any variable adds or removes that variable from the center panel. Variables that are present in the center panel appear in colored blocks. Any variable that has been tagged with a special property that appears in the legend, will have a block with a color to match the legend. For example, on the right the variables *puma*, *sex*, *age* *income*, *married* have been added to the center panel. The *puma* variable was tagged as a nominal variable and the *income* variable was tagged as the dependent variable as so they have colors that match the legend.



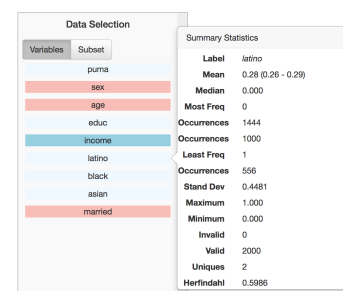
7. Subset Tab

When the subset tab is selected, the panel will show histograms or density plots of all the variables currently selected in the variable list. These graphs are interactive, and by clicking bars in histograms, or by brushing over regions of density plots, a subset of observations can be selected. In the figure on the right, the user has selected persons aged between 40 and 65 by brushing that region of the age graph, and married individuals by clicking that bar of the married histogram. When the button marked “select” in the top right is clicked, then the data will be subsetted to only those observations that meet all the selected requirements (variables that have not be selected on, for example income and sex in the figure at the right, impose no restrictions on the subset). All the summary statistics will now reflect this subset of the data, and all statistical models will only be conducted on this subset of the data.



8. Variable List

As explained in the variable tab section (6), variables names can be clicked to add or remove variables from the center panel. In addition, if a user mouseovers a variable name, a table of the summary statistics of that variable will appear. The top of this table will include any variable codebook description, if the description is available in the metadata for this dataset at the data repository. This table is useful for better understanding the variables in the dataset, and for deciding which to introduce into the center panel, and quickly moving down the list allows an inspection of all these values.

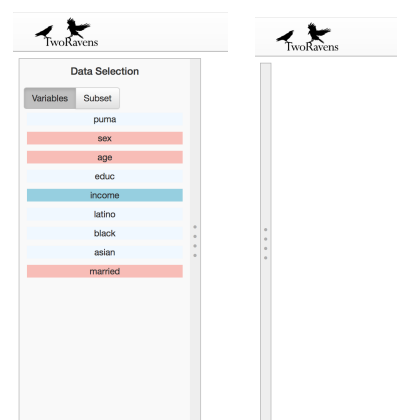


Data Selection	
Variables	Subset
puma	
sex	
age	
educ	
income	
latino	
black	
asian	
married	

Summary Statistics	
Label	income
Mean	0.28 (0.26 - 0.29)
Median	0.000
Most Freq	0
Occurrences	1444
Least Freq	1
Occurrences	1000
Stand Dev	0.4481
Maximum	1.000
Minimum	0.000
Invalid	0
Valid	2000
Uniques	2
Herfindahl	0.5986

9. Toggle Bar

The toggle bar along the right side of the Data Selection Panel can be used to open and close the panel. If the panel is open, then clicking anywhere on the toggle bar will close the panel leaving only the bar remaining at the left edge of the TwoRavens interface. This is shown on the rightmost figure to the right. Clicking the toggle bar again will return the panel to its previous state. Whatever tab was previously selected (Variables or Subset) will remain open, as well as the state of any selections in that tab. The bar is most usefully employed when a user is finished with the features in the Data Selection Panel and wishes to make more room available in the center panel.



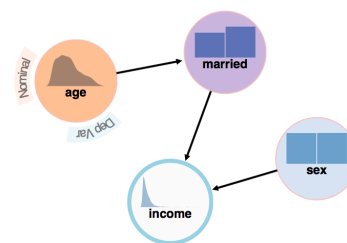
Exploration Panel

The central *Exploration Panel* is where variables, represented as *pebbles*, can be arranged into a directed graph to represent possible relationships to explore among the variables. The panel is made up *pebbles* that represent all the information in a variable, and can be arranged and connected to form a possible network of relationships between variables, called a *directed graph*.

10. Model Builder

The model builder in the exploration panel is the heart of the TwoRavens interface. Here, every variable that has been selected in the variable selection panel is represented as a circular icon called a *Pebble*. A pebble is more than just the name of a variable, it should be thought of as a container for all the information in that variable. Pebbles typically have graphs of the distribution of their variable, to emphasize that one is manipulating all the observations of a variable, and not just a name. These graphs also help make the display in this panel more informative and intuitive.

On the mouseover of a pebble, options for that pebble will appear as *arcs* or tabs around the border. These will contain possible attributes about that the user can assign to that variable. For example, the user can make a variable the outcome, or dependent variable, of the analysis, or state that a variable is nominal (categorical). Variables that have been assigned attributes will be given colored *halos* to represent this information, and a legend will build that explains the meaning of these colors. (These colors will also map back to the variables names in the variable list in the Data Selection Panel.)



Pebbles can be connected by arrows. Arrows are initiated by a two-finger or right-click. On some touch devices, the arrow is initiated by selecting an arc labelled *connect*. If this is dragged to another variable an arrow will be constructed between those two variables. Arrows represent possible causal relationships, that is, an arrow from A to B may mean A causes B or the event of A leads to B . Arrows may simultaneously point in both directions, for example an arrow from A to B and also an arrow from B to A . In such situations these are created as two separate arrows. Together, the set of all arrows is called a *directed graph*. Clicking on any arrow, deletes it from the graph.

11. Pin Button

By default, the graph in the model moves like a force diagram, that is, it acts as though the pebbles have some repulsive force keeping them apart, and the arrows act like springs. This generally moves the pebbles into a useful array. However, if more precise control of the pebble location is desired, pressing the pin button will lock all pebbles in place. Afterwards, any pebble can be dragged to any location in the panel, and it will remain in that new location. Reclicking the pin button will revert to the force effect where the pebbles adjust themselves automatically.



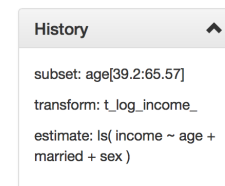
12. Wipe Button

The wipe button, whose icon is a magnet, when clicked will remove all pebbles from the exploration panel, leaving a blank panel.



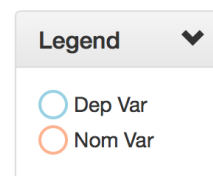
13. History Log

The history log lists operations that have occurred on the dataset, including subsetting of the dataset, transformations of variables, and estimation of statistical models. If no such operations have occurred the history log will not be visible. The history log can be minimized by clicking the down arrow in the history title bar.



14. Model Legend

When tags have been applied to the pebbles of variables, to describe their attributes, they are given a colored halo as a visual identifier of the attributes of that variable. These colored halos will then appear in the legend box, together with a description of their meaning, to aid in interpretation of the relationship that has been described in the exploration panel. In the example to the right, a *dependent variable* has been selected, and the legend shows this has a blue halo, while a categorical variable has also been tagged appropriately as *nominal* and the legend explains such variables are identified by an orange halo. The legend can be minimized by clicking the down arrow in the legend title bar.

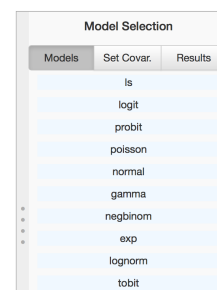


Model Selection Panel

The *Model Selection Panel* contains the features for selecting and interpreting statistical models, and comparing results across alternate models.

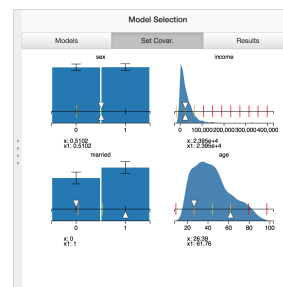
15. Models Tab

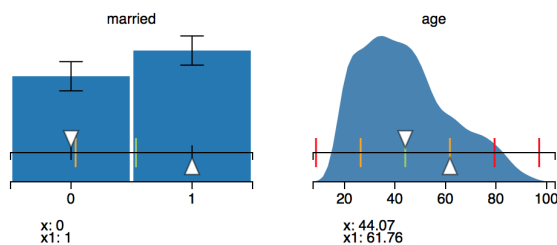
When the models tab is selected in the Model Selection Panel, a list of the available statistical models that TwoRavens can run on this dataset is presented. The desired model can be selected by clicking that model name. Mouseover of any model name will give a short description of that statistical model.



16. Set Covariates Tab

The *Set Covar(iates)* tab provides a powerful way of interpreting statistical models, particularly non-linear models where the parameters themselves have no direct interpretation. For every variable that has been included in the center exploration panel, there is graph. For categorical variables, or variables with few unique values, this is a histogram, while for others this is a density plot over the range of the variable. These graphs are interactive and can be adjusted by the user to describe points of interest for which the user would like interpretation of the statistical model.





The graphs on the left show two of these interactive figures in detail. The black horizontal line is a slider on which the white triangles can be dragged. The triangle above the line (downward pointing) represents a baseline value of the variable, and the triangle below the line (upward pointing) represents the new value this variable changes to. Marked along the slider are the mean (in green) and colored marks each one standard deviation from the mean (yellow marks are one standard deviation, red marks are two or more). Variables with only a few unique values additionally show those unique values as black marks.

In the left graph, the baseline of the married variable is set at 0 (unmarried), and changed to 1 (married). In the right graph the baseline of age is set to the mean, the green mark (which is noted below as 44 years), and the change is to one standard deviation above the mean, the yellow mark (which is given as 61 years). Thus the results will report what happens to the outcome variable, when we move from an individual who is unmarried and 44 years old, to a similar individual who is 61 and married. This is a very interpretable story for any audience, regardless of their statistical sophistication, in a way that the statistical parameters of the model may not be.

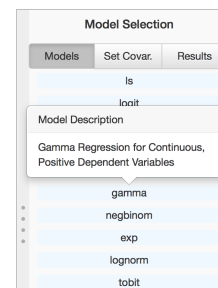
17. Results Tab

Every time a statistical model is estimated in a TwoRavens session, the results are added to a list that appears when the Results tab is selected in this panel. Models are enumerated in the order in which they were estimated. Clicking on any model name in the list reveals the parameter estimates from that model, and the graphs that have been constructed to interpret that model. Clicking between model names allows for quick comparisons across results in the same dataset.



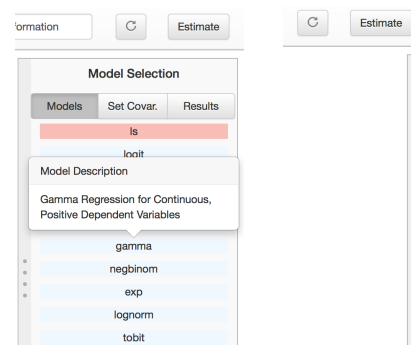
18. Statistical Model List

When the models tab is selected in the Model Selection Panel, a list of the available statistical models that TwoRavens can run on this dataset is presented. The desired model can be selected by clicking that model name. Mouseover of any model name will give a short description of that statistical model.



19. Toggle Bar

The toggle bar along the left side of the Model Selection Panel can be used to open and close the panel. If the panel is open, then clicking anywhere on the toggle bar will close the panel leaving only the bar remaining at the right edge of the TwoRavens interface. This is shown on the rightmost figure to the right. Clicking the toggle bar again will return the panel to its previous state. Whatever tab was previously selected (Models, Set Covar., or Results) will remain open, as well as the state of any selections in that tab. The bar is most usefully employed when a user wants to close this panel to free up space in the window.



Footer

The footer contains information about the completed analysis. This includes a link to a replication file that can recompute all the results and analysis. The empty space in the footer is also sometimes used as a ticker to communicate information about the session and possible improvements to model.

20. Replication Dialogue

The replication link found on the left of the footer leads to a file which can be downloaded and saved. This file contains all the information one would need to make an exact replication of all the work in the session as executable *R* code. This is a very complete replication archive that includes the exact version numbers of all packages used in *R* and the random seeds used for analysis. Users might take this file to attach to replication archives they want to distribute with any published results. Users might also use this code as a way to re-run all the results they have generated so far in an *R* session, if they want to use some additional *R* functions not available through the interface, or collaborate with colleagues who are *R* users.

```

data file citation from DataFave:
Makar, James, 2015, "PUNS California Subsample", http://dx.doi.org/10.5072/FX2/PXQ0XE, Beta Dataaverse,
V2 [UNF4:9W0 H82t3Bggquq27K5e-J=]

Dataset Information:
R version 3.2.10 (2015-12-10)
Platform x86_64-redhat-linux-gnu (64-bit)
Running under CentOS release 6.7 (Final)

locale:
 [1] C

attached base packages:
 [1] splines  stats95  stats  graphics  grDevices  utils      datasets
 [8] methods  base

other attached packages:
 [1] feiig_3.0-5      netxik_1.2-4      nlswtools_0.6-16  anelvia_1.7-3
 [4] RMCNCA_1.1-6    Hmisc_2.11-1      code_0.17-1      gepack_1.0-0
 [7] qregTest_5.1-1  SparseM_1.0-1     dplyr_5.4.2      plot_1.8-3
 [10] Intert_5.2-4     flow_17-12        car_2.2-4         car_2.2-4
 [13] VGAM_0.9-8       survival_2.38-1  MASS_7.3-5       sjovdmlib_2.3-3
 [16] DescTools_0.99-11 manipulate_1.0-1 devtools_1.8-0
 [19] DescTools_2.2-15 Rook_1.1-1

loaded via a namespace (and not attached):
 [1] nlplotr_1.0-4    gfitr_0.0.1       tools_3.2.3       boot_1.3-17
 [4] digest_0.6-8    jpeg_1.1-6        nlsw_3.1-122      nlsw_3.1-122
 [7] lattice_0.20-33 mgw_1.8-9         Matrix_1.2-11     DT_0.3-1
 [10] curl_3.2-1      parallel_1.2.3    nlsw_1.1-1        nlsw_1.1-1
 [13] rversions_1.2-3 fontTools_2.3.2   netx_7.3-11       R6_2.0-1
 [16] foreign_0.8-64 Formula_1.2-4     nlsw_1.1-4         mgw_1.8-9
 [19] codetools_0.2-14 assertthat_0.1   pkthktest_0-2     brew_1.0-0

Replication code for Twokerns session 602448-477: 4455-1lib3350c3a. Note that unless your session
information is identical to that described above, it is not guaranteed the results will be identical.
Please download Rooktools.R from https://github.com/QJBB/Twokerns/tree/master/rook and ensure that you
have Rooktools.R in your working directory.

library(Rook)
library(rjron)
library(sjovdmlib)
library(devtools)
install_github("QJBB/feig")
library(feiig)
source(rooktools.R)

download.file("https://beta.dataaverse.org/api/access/datafile/2070key?
  &file=+data%204448-477%2445-1lib3350c3a",
  method = "curl", extra = c("--insecure"))
mydata <- read.delim(file = "data%204448-477%2445-1lib3350c3a.tab")
library(RookTools)
mydata <- as.data.frame(mydata)
mydata <- as.data.frame(mydata, history = history)
mydata <- as.data.frame(mydata, sub=myvars, varnames=myvars, plot=plot)
usdata <- refactor(usdata)
plot <- selgit(formula=myformula, model=mymodel, data=usdata)
eval(parse(text = "x<=0 <- netx(sx01)"))
mydata <- "x<=0 <- netx(sx01)"
plot <- sim(x=0, xrow=0, xrow=alt)

```

References

- Choirat C, Honaker J, Imai K, King G, Lau O (2015). “Zelig: Everyone’s Statistical Software, Version 5.0-7.” URL <http://ZeligProject.org>.
- Crosas M (2011). “The Dataverse network: An open-source application for sharing, discovering and preserving data.” *D-Lib Magazine*, **17**, 1–2. Doi:1045/january2011-crosas.
- Honaker J, D’Orazio V (2014). “Statistical Modeling by Gesture: A graphical, browser-based statistical interface for data repositories.” In *Extended Proceedings of ACM Hypertext 2014*. ACM.

- King G (2007). “An introduction to the Dataverse network as an infrastructure for data sharing.” *Sociological Methods and Research*, **36**, 173–199.
- R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.