Zelig Documentation

Release 5.0.1

The Zelig Team

August 21, 2014

CONTENTS

Zelig is a framework for interfacing with a wide range of statistical models and analytic methods in a common and simple way. Above and beyond estimation, Zelig adds considerable infrastructure to existing heterogenous R implementations by translating coefficient estimates into interpretable quantities of interest and automating statistical procedures (e.g., bootstrapping) though an intelligible call structure.

For more information about the software including goals and direction of the project, please see the *About Zelig* page.

To get started, we recommend following the *installation* guide. Additional information about supported models, including tutorials, can be found in the *userguide*.

To view the codebase, visit the source repository at https://github.com/IQSS/Zelig5 and for regular updates on releases be sure to follow us on twitter at @IQSS.

Implemented Models in Zelig 5.0.1

Inheritance Tree

CHAPTER

ONE

QUICKSTART GUIDE

Zelig 5.0 is the latest version of the Zelig framework for interfacing with a wide range of statistical models and analytic methods in the R statistical programming enviornment. This release expands the set of models available, while simplifying the model wrapping process, and solving architectural problems by completely rewriting into R's Reference Classes for a fully object-oriented architecture.

This quickstart guide is designed to get you up and running with Zelig 5.0. For more detailed tutorials see individual model vignettes in the *userguide*.

1.1 Loading Zelig

After installing R and Zelig (see *installation* page). Once installed, Zelig can be loaded like any other R package:

```
> library(Zelig5)
```

Additionally, some Zelig models require add-on packages which can be installed using `install.packages()`:

```
> install.packages("ZeligChoice") #install ZeligChoice add-on package
> libarry(ZeligChoice)
```

1.2 Running Models

Each Zelig process consists of three component methods:

1. Specify statistical model and estimate parameters: `$zelig`
2. Set explanatory variables to chosen (actual or counterfactual) values for calculating quantities of interest: `$setx`
3. Draw simulations of quantity of interest from statistical model: `$sim`

For example, to implement a least squares regression:

```
> data(cars)                              #load toy dataset
> z5 <- zls$new()                         #initialize Zelig5 least squares object
> z5$zelig(dist ~ speed, data = cars)     #estimate ls model
> z5$setx(speed = 30)                     #set speed to 30 (all other covariates set to means)
> z5$sim(num = 1000)                      #run 1000 simulations and estiamte quantities of interest
```

The same model can also be implemented using the `zelig()`, `setx()`, and `sim()` functions:

```
> z.out <- zelig(dist ~ speed, model = "ls", data = cars)
> x.out <- setx(z.out, speed = 30)
> s.out <- sim(z.out, x = x.out, num = 1000)
```

Quantities of Interest

```
> summary(sim.out) #or
> summary(s.out)
```

Plots

Coming Soon!

1.3 Zelig5 Model Reference

At present, the following models have been tested and implemented in Zelig5:

- Least Squares Regression: `zls$new()` or `model = "ls"`
- Logistic Regression: `zlogit$new()` or `model = "logit"`

The following models have been implemented **but have not been unit-tested**:

- Tobit Regression:

CHAPTER

TWO

INSTALLATION & QUICKSTART

This guide is designed to get you up and running with the current *alpha* release of Zelig 5.

Note: In code snippets,‘‘>‘‘ refers to an R terminal prompt and anything after # is a comment meant to describe what the code is doing.

2.1 Installation

Before using Zelig, you will need to download and install both the R statistical program and the Zelig package:

Installing R

To install R, go to www.r-project.org/ Select the `CRAN` option from the left-hand menu (CRAN is the Comprehensive R Archive Network where all files related to R can be found). Select a CRAN mirror closest to your current geographic location (there are multiple mirrors of this database in various locations, selecting the one closest to you will be sure to maximize your downloading speeds). Follow the instructions for downloading R for Linux, Mac OS X, or Windows.

Installing Zelig

Becuase Zelig 5.0.1 is still an *alpha* release and is not yet available on `CRAN` (with other R software packages), it must be downloaded from Github using the `devtools` package.

Once you've successfully installed R, open it, and at the terminal prompt, type in the following commands verbatim:

```
# This installs devtools package, if not already installed
> install.packages("devtools")
# This loads devtools
> library(devtools)
# This downloads Zelig 5.0.1 from the IQSS Github repo
> install_github('IQSS/Zelig5')
```

Once you have successfully typed these commands, you will see a the following message: *"DONE (Zelig5)"*.

2.2 Quickstart Guide

Loading Zelig

After installing both R and Zelig, Zelig can be loaded by using the `library()` function:

```
> library(Zelig5)
```

2.3 Running Models

Imagine a scenario in which you want to predict how the distance a car can travel given it's speed. If we were to model this relationship using a least squares regression within Zelig we need to follow three steps:

1. First, we are going to want to specify our model, given a dataset on cars including distance and speed, and estimate the effect of speed on distance. This is done using the `$zelig()` method in the code snippet below.
2. Second, we want to translate our estimates into intepretable quanities of interest, so we can answer intuitive questions about the effect of speed on distance. For example, we may be interested into understanding how a change of speed from 10 to 20 mph affects distance versus a change from 50 to 60 mph. To do this we have to set explanatory variables in our model (i.e., speed) to simulate quantities of interest. This is done using the `$setx()` or `$setrange()` method.
3. Finally, we want to draw simualtions of quantieis of interest from our statistical model using the `$sim()` method.

The following code snippet loads a data set of cars data including speed and distance (When you install R, example datasets are also installed), and regressing distance on speed. We then go on to set speed (our main explanatory variable) and simulate quantities of interest.

```
#load toy dataset
> data(cars)
#initialize Zelig5 least squares object
> z5 <- zls$new()
#estimate ls model
> z5$zelig(dist ~ speed, data = cars)

#set speed to 30 (all other covariates set to means)
> z5$setx(speed = 30)
#or, simulate over a range of speed between 55 and 80
> z5$setrange(speed = 55:80)

#run 1000 simulations and estiamte quantities of interest
> z5$sim(num = 10)

#print model estimates
> z5

Call:
stats::lm(formula = dist ~ speed, data = cars)

Residuals:
Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
>
```

Quantities of Interest

A major feature of Zelig is the translation of model estimates into easy to interpret quantities of interest (QIs). These QIs (e.g. expected and predicted values) can be accessed via the `$sim.out` method:

```
> z5$sim.out
$range
$range[[1]]
$range[[1]][[1]]
$range[[1]][[1]]$ev
             1
 [1,] 201.2214
 [2,] 186.4815
 [3,] 187.1505
 [4,] 175.2489
 [5,] 187.9555
 [6,] 197.8107
 [7,] 199.0940
 [8,] 168.2691
 [9,] 195.7094
[10,] 222.0503

$range[[1]][[1]]$pv
             1
 [1,] 201.2214
 [2,] 186.4815
 [3,] 187.1505
 [4,] 175.2489
 [5,] 187.9555
 [6,] 197.8107
 [7,] 199.0940
 [8,] 168.2691
 [9,] 195.7094
[10,] 222.0503
>
```

Plots

A second major Zelig feature is how easy it is to plot QIs for presentation in slides or an artcle. Using the `plot()` function on the `z5$s.out` will produce ready-to-use plots with labels and confidence intervals.

Plots of QI's from binary choice model:

Plot of expected values across range of simulations:

CHAPTER

THREE

FREQUENTLY ASKED QUESTIONS

If you find a bug, or cannot figure something out after reading the frequently asked questions below, please send your question to the Zelig listserv at `https://groups.google.com/forum/#!forum/zelig-statistical-software`. Please explain exactly what you did and include the full error message, including the traceback(). You should get an answer from the developers or another user in short order.

3.1 How do I cite Zelig?

We would appreciate if you would cite Zelig as:

> Imai, Kosuke, Gary King and Olivia Lau. 2006. "Zelig: Everyone's Statistical Software," http://GKing.Harvard.Edu/zelig.

Please also cite the contributors for the models or methods you are using. These citations can be found in each individual model's vignette which can be found in the the *userguide*.

3.2 Why can't I install Zelig?

We recommend that you first check your internet connection, as you must be connected to install packages. In addition, there are a few platform-specific reasons why you may be having installation problems:

- **On Windows**: If you are using the very latest version of R, you may not be able to install Zelig until we update Zelig to work with this latest release. Currently Zelig 5.0.1 is compatible with R(3.0.2). If you wish to install Zelig in the interim, install the appropriate version of R and try to reinstall Zelig.
- **On Mac or Linux systems**: If you get the following warning message at the end of your installation:

  ```
  > Installation of package VGAM had non-zero exit status in ...
  ```

  this means that you were not able to install VGAM properly. Make sure that you have the g77 Fortran compiler. For Intel Macs, download the Apple developer tools. After installation, try to install Zelig again.

If neither solution works, feel free email the Zelig mailing list directly at: https://groups.google.com/forum/#!forum/zelig-statistical-software.

3.3 Why can't I install R?

If you have problems installing R, you should search the internet for the R help mailing list, check out technical Q & A forums (e.g., StackOverflow), or email the Zelig mailing list directly at: https://groups.google.com/forum/#!forum/zelig-statistical-software.

3.4 Why can't I load data?

It is likely that the reason you are unable to load data because you have not specfied the correct working directory (e.g., the location of the data you are trying to load). You should specify you working directory use the `setwd()` function in which you will include the the file path to your working director. For example, if I wanted to load a file that is my *Documents* folder, I must first:

```
> setwd("path/to/Documents")
```

File paths can be found by right clicking the workign directory folder in any file browser and clicking "Get Info" (on Mac) or "Properties" (on Windows). Black-slashes (\) in file paths copied from the "Properties" link on Windows machines must be replace with forward-slashes (/). For example, the Windows path: `C:\Program Files\R`, would be typed as `C:/Program Files/R`.

3.5 How do I increase the memory for R?

Windows users may get the error that R has run out of memory. If you've installed more memory on your machine, you may have to reinstall R in order to take advantage of the additional capacity.

You may also set the amount of available memory manually. Close R, then right-click on your R program icon (the icon on your desktop or in your programs directory). Select "Properties", and then select the "Shortcut" tab. Look for the "Target" field and after the closing quotes around the location of the R executible, add

```
--max-mem-size=500M
```

You may increase this value up to 2GB or the maximum amount of physical RAM you have installed. If you get the error that R cannot allocate a vector of length x, close out of R and add the following line to the "Target" field:

```
--max-vsize=500M
```

or as appropriate.

You can always check to see how much memory R has available by typing at the R prompt

```
> round(memory.limit()/2^20, 2)
```

which gives you the amount of available memory in MB.

3.6 Why doesn't the pdf print properly?

Zelig uses several special LaTeX environments. If the pdf looks right on the screen, there are two possible reasons why it's not printing properly:

- Adobe Acrobat isn't cleaning up the document. Updating to Acrobat Reader 6.0.1 or higher should solve this problem.
- Your printer doesn't support PostScript Type 3 fonts. Updating your print driver should take care of this problem.

3.7 R is neat. How can I find out more?

R is a collective project with contributors from all over the world. Their website (r-project.org.) has more information on the R project, R packages, conferences, and other learning material.

CHAPTER

FOUR

ABOUT ZELIG

Zelig is an open source project development and maintained by the Data Science group at the Data Science group at Harvard's Institute for Quantative Social Science (IQSS). It was originally concieved and created by Kosuke Imai, Gary King, and Olivia Lau in 2007. The name is borrowed from Woody Allen's movie with the same name, Zelig. Leonard Zelig is a fictional character who takes on the characteristics of any strong personality around. Likewise, the Zelig statistical software easily adapts to any statistical model written in R, and in essence, takes the characteristics of any model.

It leverages (R) code from many researchers and is designed to allow anyone to contribute their methods to it. Hence, we often refer to Zelig as "everyone's statistical software" and our aim is to make it, as well as the models it wraps, as accessible as possible. As such, Zelig comes with self-contained documentation that minimizes startup costs, automates model summaries and graphics, and bridges existing R implementations through an intelligble call structure.

License: GPL-2 | GPL-3 [expanded from: GPL (2)]

Contact: For questions, please join the Zelig mailing list: https://groups.google.com/forum/#!forum/zelig-statistical-software

The Zelig Team:

- Gary King *(Principle Investigator)*
- James Honaker *(Project Lead)*
- Christine Choirat *(Lead Author)*
- Kosku Imai
- Olivia Lau

4.1 Technical Vision

Zelig is a framework for interfacing a wide range of statistical models and analytic methods in a common and simple way. Above and beyond estimation, Zelig adds considerable infrastructure to existing heterogenous R implementations by translating hard-to-interpret coefficients into quanities of interest (e.g., expected and predicted values) through a simple call structure. This includes many specific methods, based on likelihood, frequentist, Bayesian, robust Bayesian and nonparametric theories of inference. Developers are encouraged to add their R packages to the Zelig toolkit by writing a few simple bridge functions.

Additional features include:

- Dealing with missing data by combining multiply imputed datasets
- Automating statistical bootstrapping

- Improving parametric procedures by leveraging nonparametric matching methods
- Evaluating counterfactuals
- Allowing condtional population and super population inferences
- Automating the creation of replication data files

CHAPTER

FIVE

MODEL VIGNETTES

5.1 Zelig-ls

Least Squares Regression for Continuous Dependent Variables

Use least squares regression analysis to estimate the best linear predictor for the specified dependent variables.

5.1.1 Syntax

```
z.out <- zelig(Y ~ X1 + X2, model = "ls", data = mydata)
x.out <- setx(z.out)
s.out <- sim(z.out, x = x.out)
```

5.1.2 Examples

Basic Example with First Differences

Attach sample data:

```
library(Zelig5)
```

```
## Loading required package: methods
## Loading required package: MASS
## Loading required package: survival
## Loading required package: splines
## Loading required package: VGAM
## Loading required package: stats4
## Loading required package: jsonlite
##
## Attaching package: 'jsonlite'
##
## The following object is masked from 'package:utils':
##
##     View
##
## Loading required package: AER
## Loading required package: car
##
## Attaching package: 'car'
##
```

```
## The following object is masked from 'package:VGAM':
##
##     logit
##
## Loading required package: lmtest
## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
##
##
## Attaching package: 'lmtest'
##
## The following object is masked from 'package:VGAM':
##
##     lrtest
##
## Loading required package: sandwich
##
## Attaching package: 'AER'
##
## The following object is masked from 'package:VGAM':
##
##     tobit
##
## Loading required package: plyr
## Loading required package: dplyr
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:plyr':
##
##     arrange, desc, failwith, id, mutate, summarise, summarize
##
## The following object is masked from 'package:MASS':
##
##     select
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## Warning: replacing previous import by 'VGAM::show' when loading 'Zelig5'
## Warning: replacing previous import by 'AER::tobit' when loading 'Zelig5'
## Warning: replacing previous import by 'quantreg::untangle.specials' when loading 'Zelig5'
## Warning: replacing previous import by 'dplyr::arrange' when loading 'Zelig5'
## Warning: replacing previous import by 'dplyr::desc' when loading 'Zelig5'
## Warning: replacing previous import by 'dplyr::failwith' when loading 'Zelig5'
## Warning: replacing previous import by 'dplyr::id' when loading 'Zelig5'
## Warning: replacing previous import by 'dplyr::mutate' when loading 'Zelig5'
## Warning: replacing previous import by 'dplyr::select' when loading 'Zelig5'
```

```
## Warning: replacing previous import by 'dplyr::summarise' when loading 'Zelig5'
## Warning: replacing previous import by 'dplyr::summarize' when loading 'Zelig5'
## Warning: replacing previous import by 'methods::setRefClass' when loading 'Zelig5'
```

```
data(macro)
```

Estimate model:

```
z.out1 <- zelig(unem ~ gdp + capmob + trade, model = "ls", data = macro)
```

```
## How to cite this model in Zelig:
##   Kosuke Imai, Gary King, and Olivia Lau. 2007.
##   ls: Least Squares Regression for Continuous Dependent Variables
##   in Kosuke Imai, Gary King, and Olivia Lau, "Zelig: Everyone's Statistical Software,"
##   http://datascience.iq.harvard.edu/zelig
```

Summarize regression coefficients:

```
summary(z.out1)
```

```
## Model: 1
## Call:
## stats::lm(formula = unem ~ gdp + capmob + trade, data = .)
##
## Coefficients:
## (Intercept)          gdp       capmob        trade
##      6.1813      -0.3236       1.4219       0.0199
##
## Next step: Use 'setx' method
```

Set explanatory variables to their default (mean/mode) values, with high (80th percentile) and low (20th percentile) values for the trade variable:

```
x.high <- setx(z.out1, trade = quantile(macro$trade, 0.8))
x.low <- setx(z.out1, trade = quantile(macro$trade, 0.2))
```

Generate first differences for the effect of high versus low trade on GDP:

```
s.out1 <- sim(z.out1, x = x.high, x1 = x.low)
```

Using Dummy Variables

Estimate a model with fixed effects for each country (see for help with dummy variables). Note that you do not need to create dummy variables, as the program will automatically parse the unique values in the selected variable into discrete levels.

```
z.out2 <- zelig(unem ~ gdp + trade + capmob + country,
                model = "ls", data = macro)
```

```
## How to cite this model in Zelig:
##   Kosuke Imai, Gary King, and Olivia Lau. 2007.
##   ls: Least Squares Regression for Continuous Dependent Variables
##   in Kosuke Imai, Gary King, and Olivia Lau, "Zelig: Everyone's Statistical Software,"
##   http://datascience.iq.harvard.edu/zelig
```

```
z.out3 <- zelig(unem ~ gdp + trade + capmob + as.factor(country),
                model = "ls", data = macro)
```

```
## How to cite this model in Zelig:
##   Kosuke Imai, Gary King, and Olivia Lau. 2007.
##   ls: Least Squares Regression for Continuous Dependent Variables
##   in Kosuke Imai, Gary King, and Olivia Lau, "Zelig: Everyone's Statistical Software,"
##   http://datascience.iq.harvard.edu/zelig

x.US <- setx(z.out3, country = "United States")
```

Set values for the explanatory variables, using the default mean/mode values, with country set to the United States and Japan, respectively:

```
x.US <- setx(z.out2, country = "United States")
x.Japan <- setx(z.out2, country = "Japan")
```

Simulate quantities of interest:

```
s.out2 <- sim(z.out2, x.US, x.Japan)

plot(s.out2)
```

5.1.3 Model

- The *stochastic component* is described by a density with mean mu_i and the common variance $sigma^2$

  $$Y_i ; sim ; f(y_i mid mu_i, sigma^2).$$

- The *systematic component* models the conditional mean as

  $$mu_i = x_i beta$$

  where x_i is the vector of covariates, and $beta$ is the vector of coefficients.

  The least squares estimator is the best linear predictor of a dependent variable given x_i, and minimizes the sum of squared residuals, $sum_{i=1}^n (Y_i-x_i beta)^2$.

5.1.4 Quantities of Interest

- The expected value (*qi$ev*) is the mean of simulations from the stochastic component,

  $$E(Y) = x_i beta,$$

  given a draw of $beta$ from its sampling distribution.

- In conditional prediction models, the average expected treatment effect (*att.ev*) for the treatment group is

  **$$frac{1}{sum_{i=1}^n t_i}sum_{i:t_i=1}^n left{ Y_i(t_i=1) -** E[Y_i(t_i=0)] right},$$

  where t_i is a binary explanatory variable defining the treatment ($t_i=1$) and control ($t_i=0$) groups. Variation in the simulations are due to uncertainty in simulating $E[Y_i(t_i=0)]$, the counterfactual expected value of Y_i for observations in the treatment group, under the assumption that everything stays the same except that the treatment indicator is switched to $t_i=0$.

5.1.5 Output Values

The output of each Zelig command contains useful information which you may view. For example, if you run *z.out <- zelig(y ~ x, model = ls, data)*, then you may examine the available information in *z.out* by using *names(z.out)*, see the *coefficients* by using *z.out$coefficients*, and a default summary of information through *summary(z.out)*. Other elements available through the *$* operator are listed below.

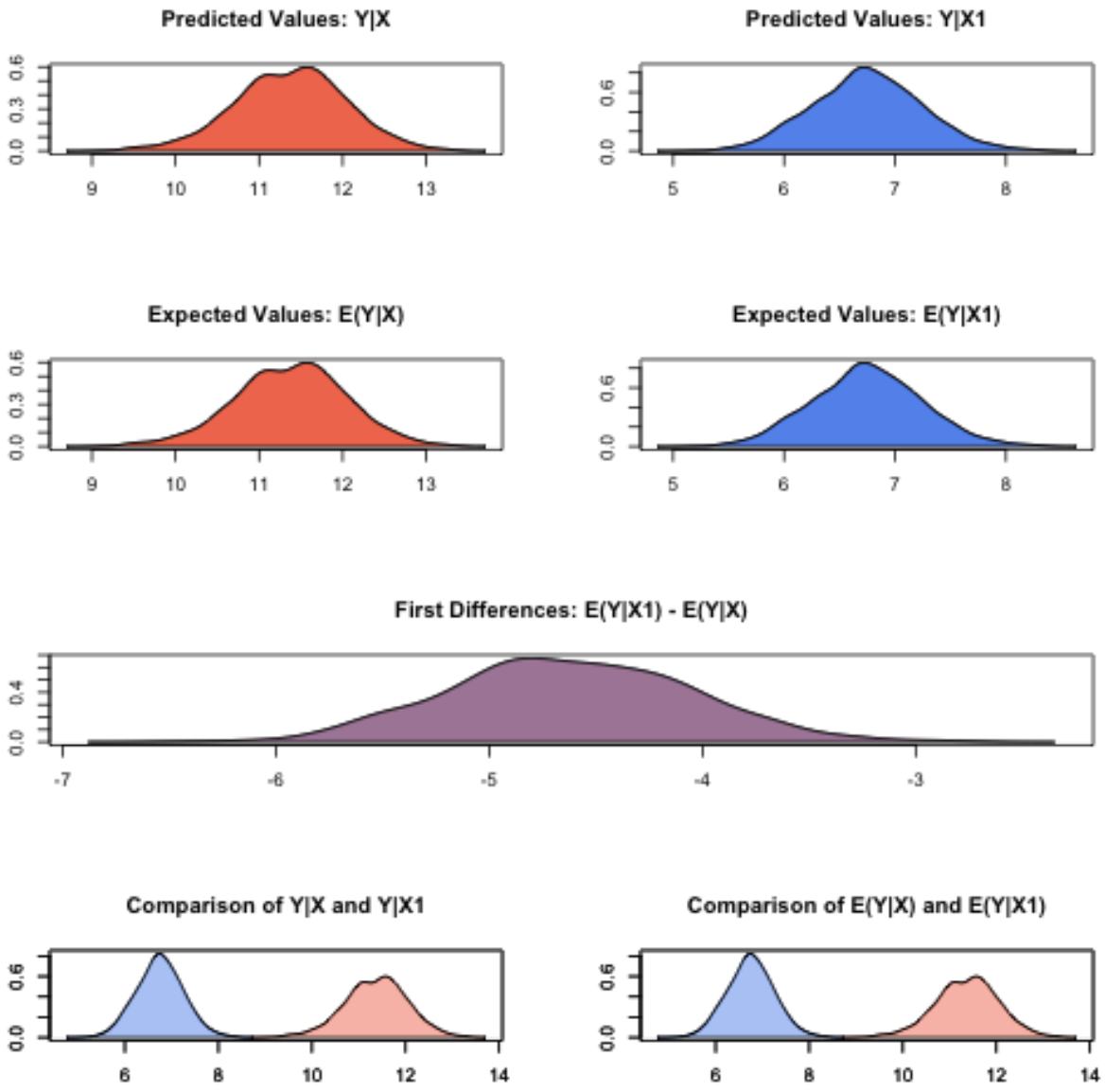

Figure 5.1: plot of chunk unnamed-chunk-12

5.1.6 See also

The least squares regression is part of the stats package by William N. Venables and Brian D. Ripley [@VenRip02]. In addition, advanced users may wish to refer to *help(lm)* and *help(lm.fit)*. Robust standard errors are implemented via the sandwich package by Achim Zeileis [@Zeileis04]. Sample data are from @KinTomWit00.