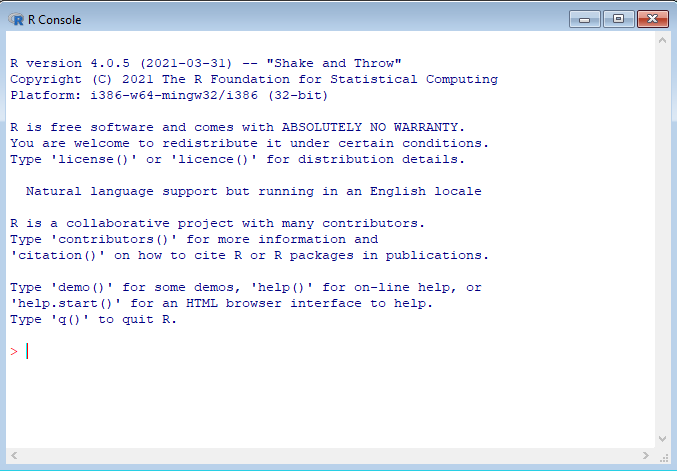
**Choice Models and Neural Networks**

**Starting R**

1. Click on the Start button in the lower left corner of Windows.
2. Click on All Programs, then click on the R folder, then R.



1. This is the command line screen. You can enter commands but need to know the syntax.
2. There is a simpler approach to running R, called Rcmdr (R Commander). If you are running a Whitman computer, Rcmdr is already installed. If not, you need to install it.

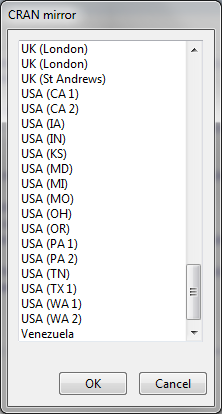
**Installing R Commander**

Follow these steps only if you do not already have Rcmdr installed.

1. In R, type the command:

install.packages("Rcmdr", dependencies = TRUE)

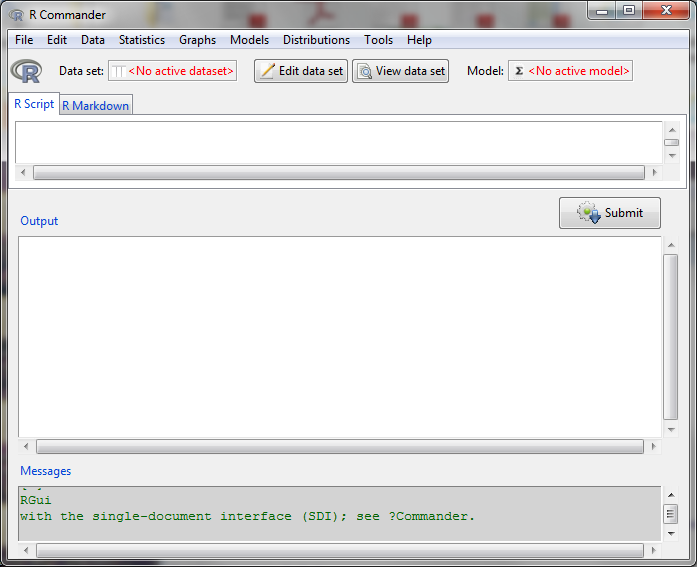
1. In the CRAN mirror, select the location closest to you; use a USA location near you, then click OK.
2. If prompted to create a personal library, click Yes.
3. If prompted to add missing packages, click Yes.



**Launch Rcmdr (R Commander)**

Rcmdr is a graphical user interface (GUI) that is easier to use than the command line. To launch Rcmdr:

1. Type library(Rcmdr).
2. If you receive a warning message that some packages are missing, it will ask if you want them installed. Click Yes.
3. On the Install Missing Packages screen, click OK.
4. R will install the necessary software.
5. The R Commander screen will appear.



**8.3.1 R: Logit Analysis Demo**

**Download Data Sets**

Download the Week8-titanic-data.xlsx file, which contains the Titanic data set.

**Loading Data**

To load data into R:

1. Click on Data at the top of the screen
2. Click on Import Data > from Excel file …
3. Enter the name that you would like to use for this data set; type in titanic
4. Click OK
5. Find the titanic file on your computer, then Open

**Viewing data fields**

In the following example, we will use passenger data from the Titanic to explore which factors relate to survival after the sinking of the Titanic. There is complete numeric data for 1,046 passengers. The variables in the data are:

Survived Survival Indicator (0 = No, 1 = Yes)

Name Passenger name

Gender Passenger’s gender

GenderNum Passenger’s numeric gender (0 = Female, 1 = Male)

Age Age in years

SiblingSpouse Number of passengers on ship who are this person’s brother, sister, or spouse

ParentChild Number of passengers on ship who are this person’s parent or child

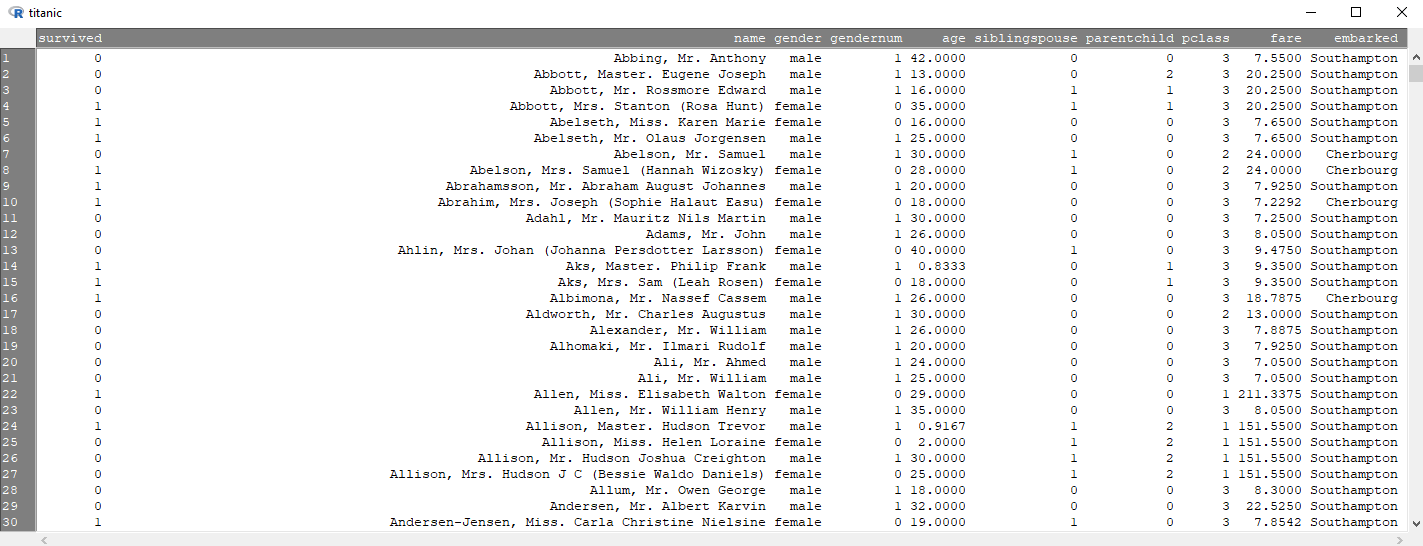
PClass Passenger class (1 = 1st, 2 = 2nd, 3 = 3rd)

Fare Passenger fare

Embarked Port of embarkation

Now return to R. To view the variables in R,

1. Click on the button View data set:



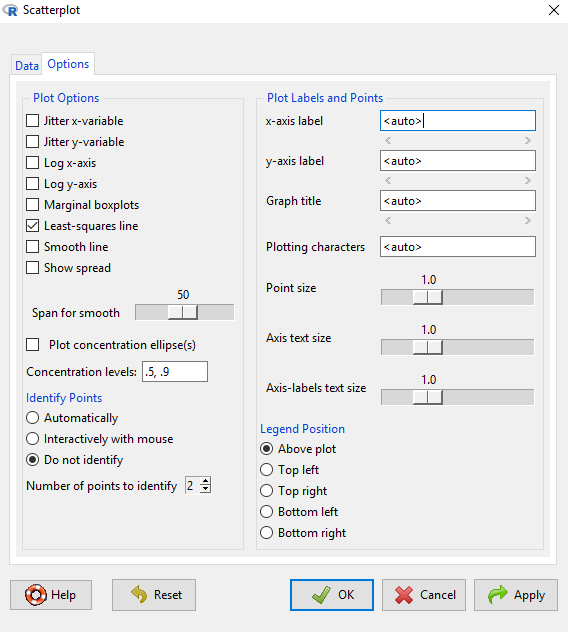
1. Which variables should affect whether someone survived? Why?
2. Click on the X in the upper right corner of the data display to close the data view.

**Scatterplots**

To generate a scatter plot,

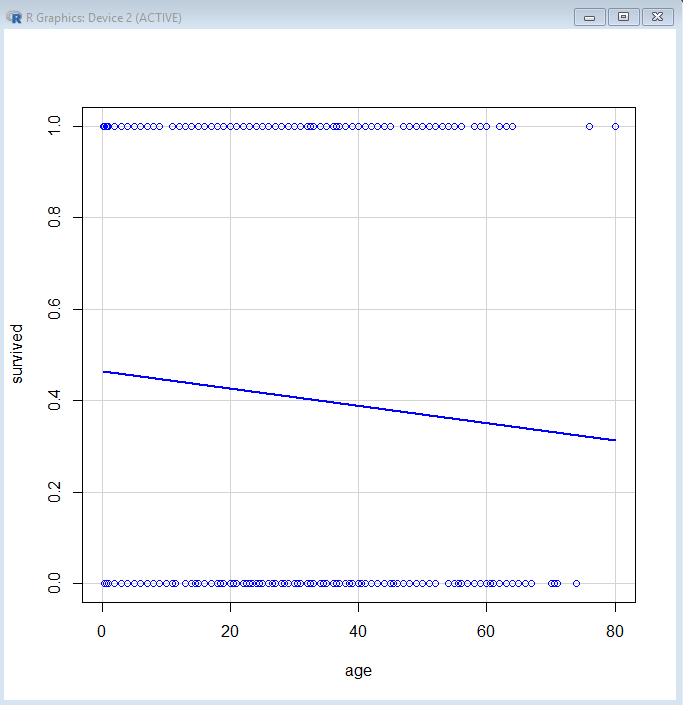
1. Click on Graphs, Scatterplot
2. Select age as the x-variable
3. Select survived as the y-variable
4. Click on the Options tab, then select check the box for Least-squares line
5. Click on OK





To interpret the chart:

1. The blue dots are the age versus survived (1) or not survived (0).
2. The blue line is the linear regression line through the data.
3. Does the linear regression line make sense?



Linear regression assumes that there is a linear relationship between the X and Y variables. In this case, that doesn’t make sense. A better solution looks like this:

Logit curve with increasing probability

Logit curve with decreasing probability

The logit regression uses the logistic function; it is either an S-shaped curve monotonically increasing or an S-shaped curve monotonically decreasing. Probit is similar but uses the normal distribution function.

**Key Differences between Logit and Probit**

Logit and probit are techniques that assume the dependent variable (Y) is zero or one, and finds the relationship between the explanatory variables (X) and the dependent variable (Y). Logistic regression and logit are based on the logistic distribution. Probit is based on the normal distribution. Logit is more sensitive to extreme values of the X variable. Probit is more sensitive to values near the mean.

The logit regression uses the logistic function to calculate the probability:

P(Y=1) = *exp(ΣβiXi)/[1+ exp(ΣβiXi)]*

The probit regression uses the normal distribution to calculate the probability:

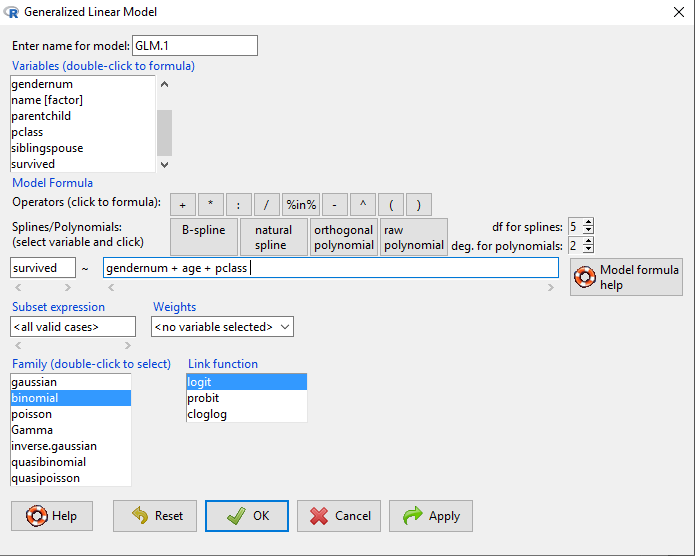
P(Y =1) = *Φ(ΣβiXi)*  where *Φ*  is the normal distribution

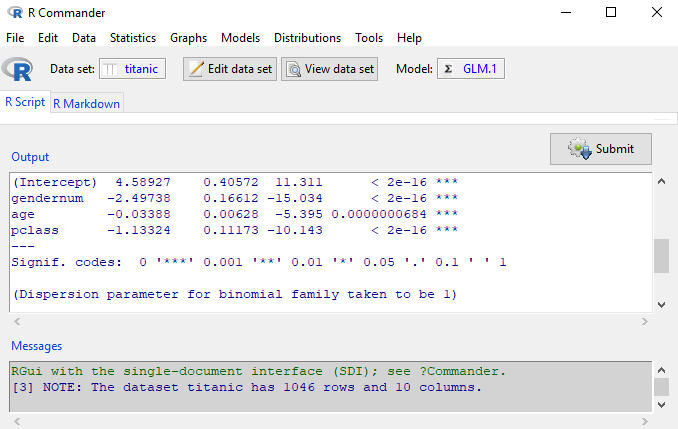
Difference between logit and probit curves

**Logit Analysis**

To perform a logit analysis on our data, where the Y variable is survived and the explanatory variables are gendernum, age, pclass:

1. Click on Statistics, Fit models, Generalized linear model.
2. Double click on survived for the dependent variable.
3. Double click on gendernum, age, pclass for the explanatory variables.
4. Select Family as binomial.
5. Select Link function as logit.
6. Click OK.



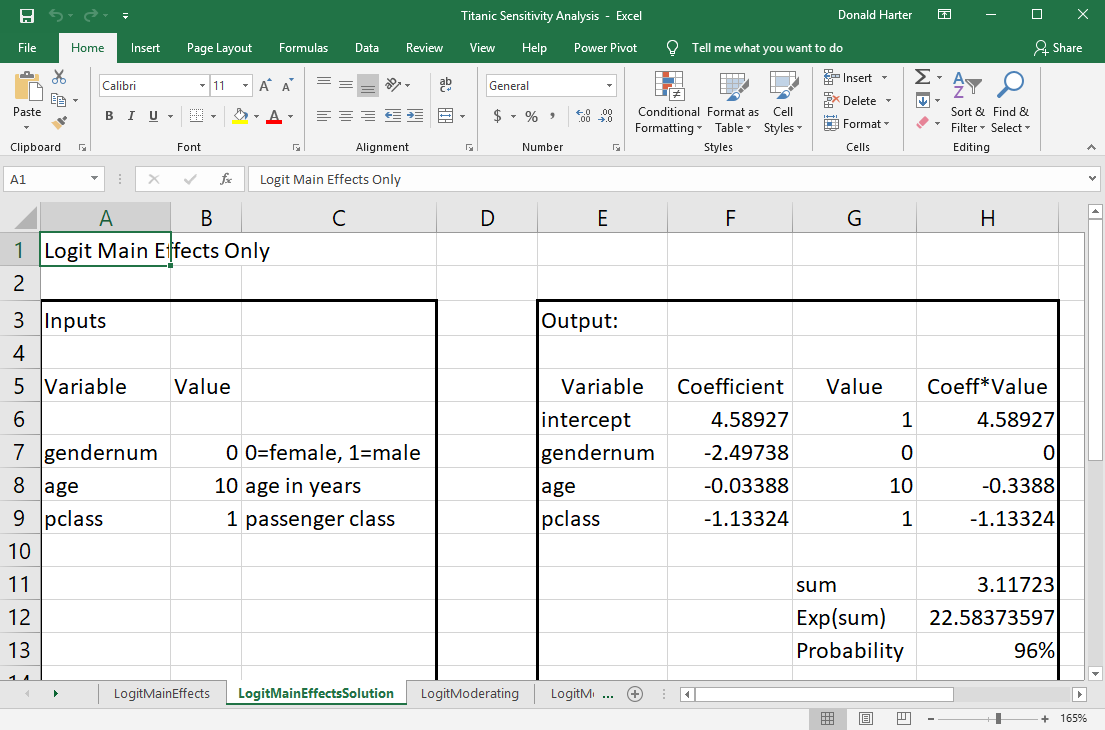


1. Are the coefficients positive or negative?
2. As gendernum goes from 0 to 1, what happens to survivability? What does this mean?
3. As age increases, what happens to survivability? What does this mean?
4. As pclass increases, what happens to survivability? What does this mean?
5. Are the coefficients statistically significant?

**8.3.2 R: Logit Predictions Demo**

You can use a spreadsheet to calculate logit probabilities for different combinations of the explanatory variable values. Let’s calculate the probability of a 10-year-old girl in 1st class surviving the Titanic. Use the data from the previous page.

1. For this exercise, use the template called Week8-titanic-sensitivity-analysis.xlsx.
2. Click on the tab LogitMainEffects.
3. The left section allows you to enter the values for gendernum, age, and pclass.
   1. In cell B7, enter 0 for a girl.
   2. In cell B8, enter 10 for 10 years old.
   3. In cell B9, enter 1 for first class.
4. Next, enter the coefficients for each variable in cells F6 through F9.
5. Enter a formula for the values in G7 through G9 pointing to the input values.
   1. In G7, enter =B7
   2. In G8, enter =B8
   3. In G9, enter =B9
   4. For G6, enter 1
6. In cells H6 through H9, enter formulas multiplying the coefficients and values.
7. The logistic function is the exponential of the sum divided by (1+ exponential of the sum).
8. First, calculate the sum. In cell H11, enter =sum(H6:H9).
9. Next, calculate the exponential of the sum. In cell H12, enter =exp(H11).
10. Finally, in cell H13, enter =H12/(1+H12). This is your probability of surviving.
11. Change the input values from male to female; change the age; change the passenger class. What happens to the probability?



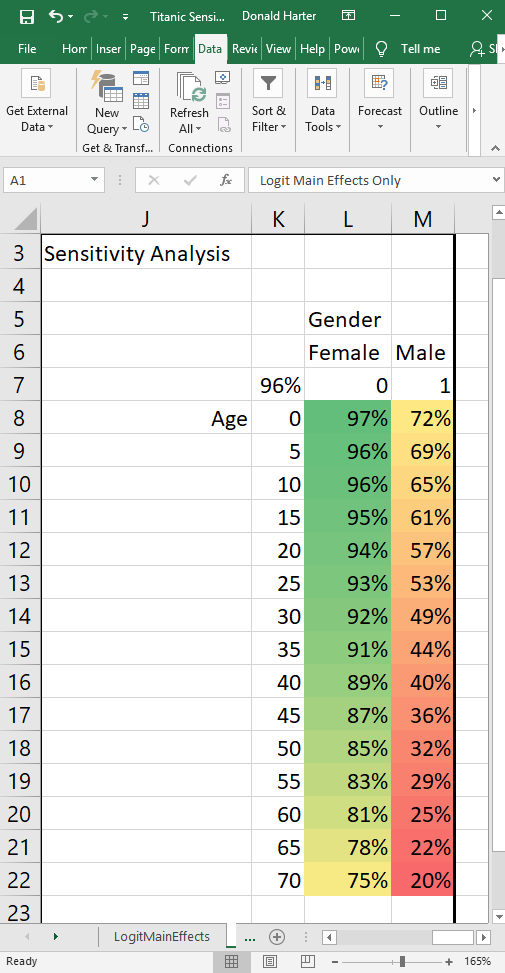
The logit regression uses the logistic function to calculate the probability:

P(Y=1) = *exp(ΣβiXi)/[1+ exp(ΣβiXi)]*

**Logit Sensitivity Analysis**

To perform a sensitivity analysis of gender versus age, use the Sensitivity section of the same spreadsheet.

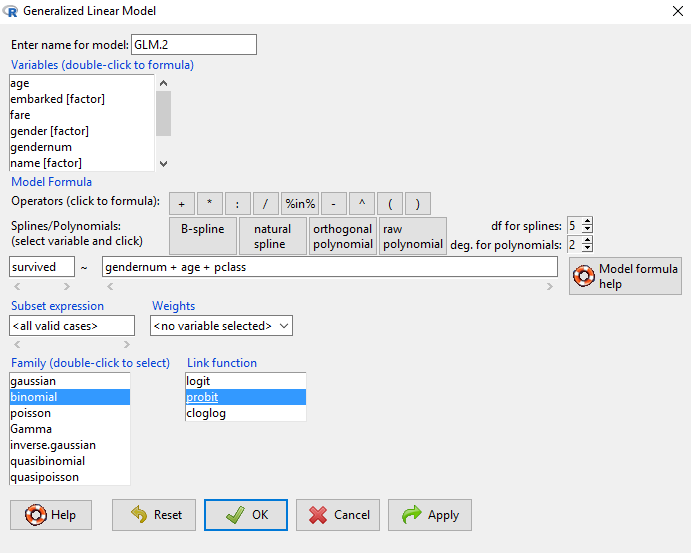
1. In general, for a sensitivity analysis, put one dimension across the top (in this case, gender), and one dimension down the side (in this case age).
2. In the cell in the corner, you must enter the formula for the probability calculation. This is already stored in cell H13. In cell K7, enter =H13.
3. Highlight K7 through M22.
4. Click on the Data tab, then What If analysis, Data Table.
5. For Row input cell, point to the cell with the value for gendernum, in this case B7.
6. For Column input cell, point to the cell with the value for age, in this case B8.
7. Click OK.
8. What pattern do you see?
9. Notice that for each gender, the curve is monotonic (always increasing or always decreasing).

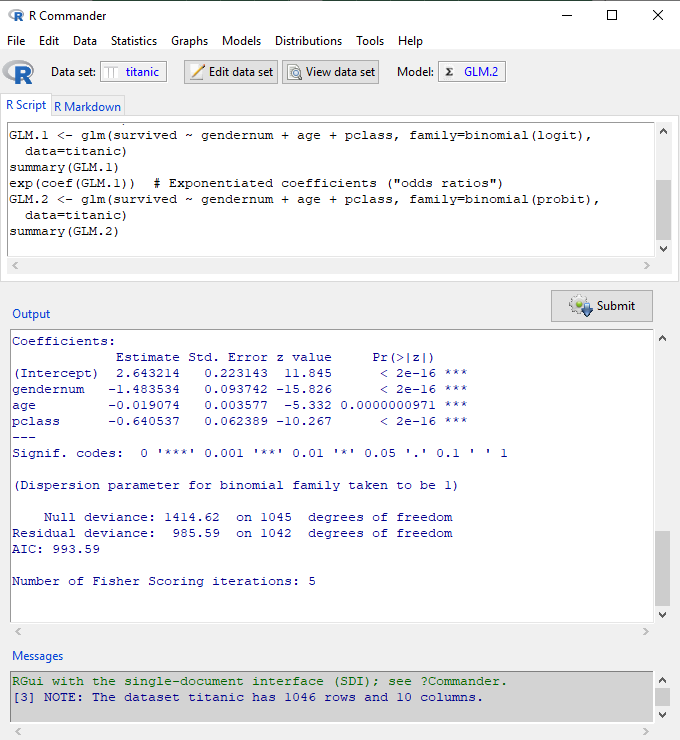


**8.4.1 R: Probit Analysis Demo**

To perform a probit analysis on our data, where the Y variable is survived and the explanatory variables are gendernum, age, pclass:

1. Click on Statistics, Fit models, Generalized linear model.
2. Double click on survived for the dependent variable.
3. Double click on gendernum, age, pclass for the explanatory variables.
4. Select Family as binomial.
5. Select Link function as probit.
6. Click OK.



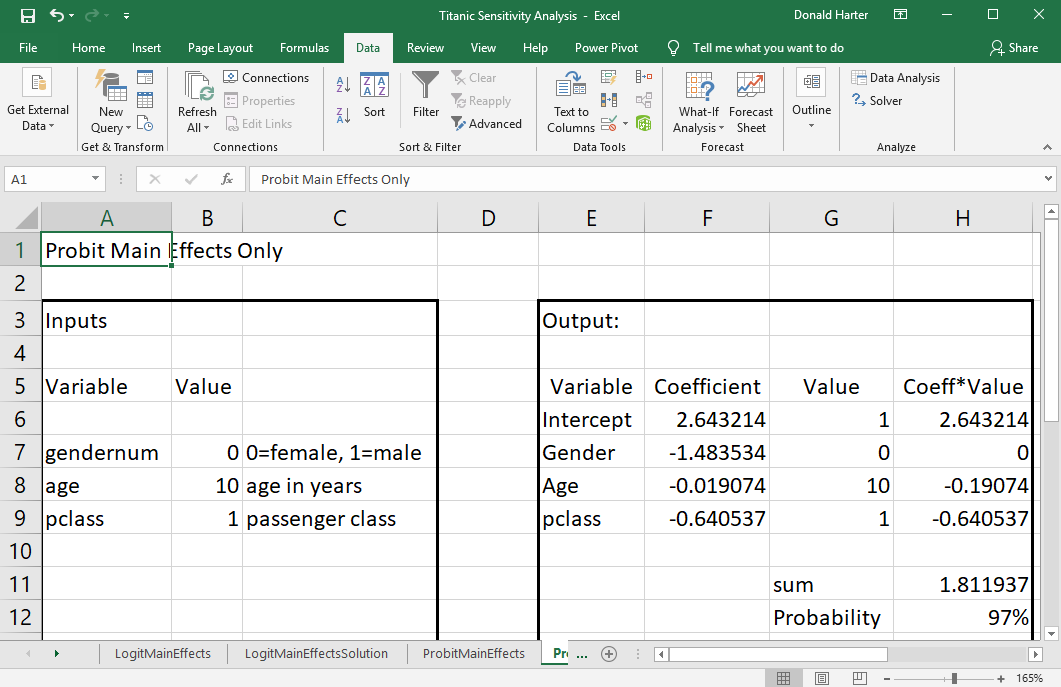


1. Are the coefficients positive or negative?
2. Are the coefficients statistically significant?
3. Is there a difference between logit and probit?

**8.4.2 R: Probit Predictions Demo**

You can use a spreadsheet to calculate logit probabilities for different combinations of the explanatory variable values. Let’s calculate the probability of a 10-year-old girl in 1st class surviving the Titanic. Use the data from the previous page.

1. For this exercise, use the template called Titanic Sensitivity Analysis.
2. Click on the tab ProbitMainEffects.
3. The left section allows you to enter the values for gendernum, age, and pclass.
   1. In cell B7, enter 0 for a girl
   2. In cell B8, enter 10 for 10 years old
   3. In cell B9, enter 1 for first class
4. Next, enter the coefficients for each variable in cells F6 through F9.
5. Enter a formula for the values in G7 through G9 pointing to the input values.
   1. In G7, enter =B7
   2. In G8, enter =B8
   3. In G9, enter =B9
   4. For G6, enter 1
6. In cells H6 through H9, enter formulas multiplying the coefficients and values.
7. The probit uses the normal function.
8. First, calculate the sum. In cell H11, enter =sum(H6:H9).
9. Next, calculate the standard normal distribution of the sum. In cell H12, enter =norm.s.dist(H11,TRUE).
10. Change the input values from male to female; change the age; change the passenger class. What happens to the probability?



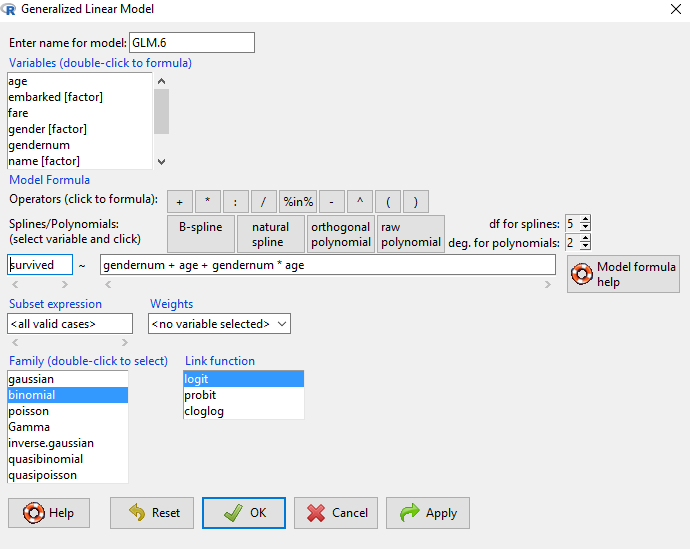
The Probit regression uses the normal distribution to calculate the probability:

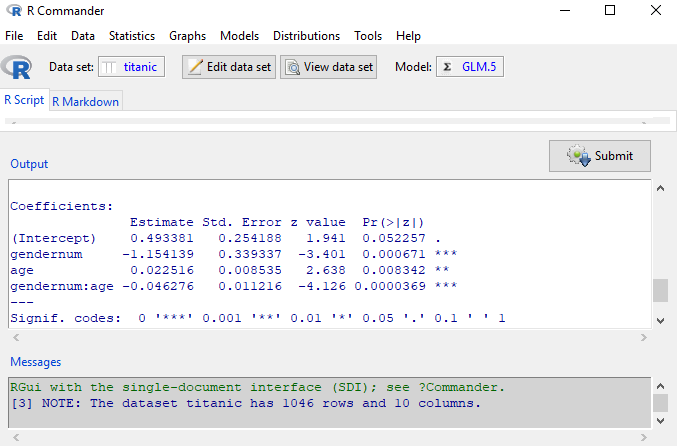
P(Y =1) = *Φ(ΣβiXi)*  where *Φ*  is the normal distribution

**8.5.1 R: Logit with Moderating Effect Demo**

To perform a logit analysis on our data, where the Y variable is survived and the explanatory variables are gendernum, age, pclass, with a moderating effect of gendernum\*age:

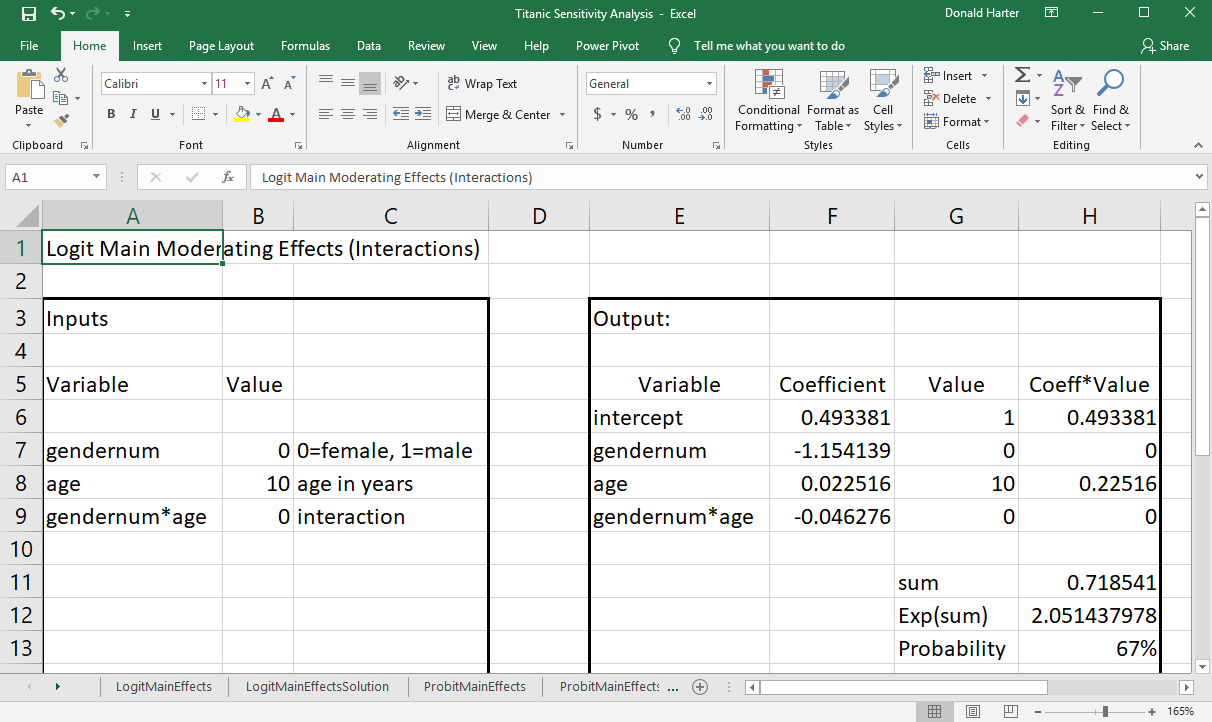
1. In R, click on Statistics, Fit models, Generalized linear model.
2. Double click on survived for the dependent variable.
3. Double click on gendernum and age for the explanatory variables; don’t include pclass.
4. To add a moderating effect, double click on gendernum again, click on \* to multiply, then double click on age.
5. Select Family as binomial.
6. Select Link function as logit.
7. Click OK.





**8.5.2 R: Logit with Moderating Effect Predictions Demo**

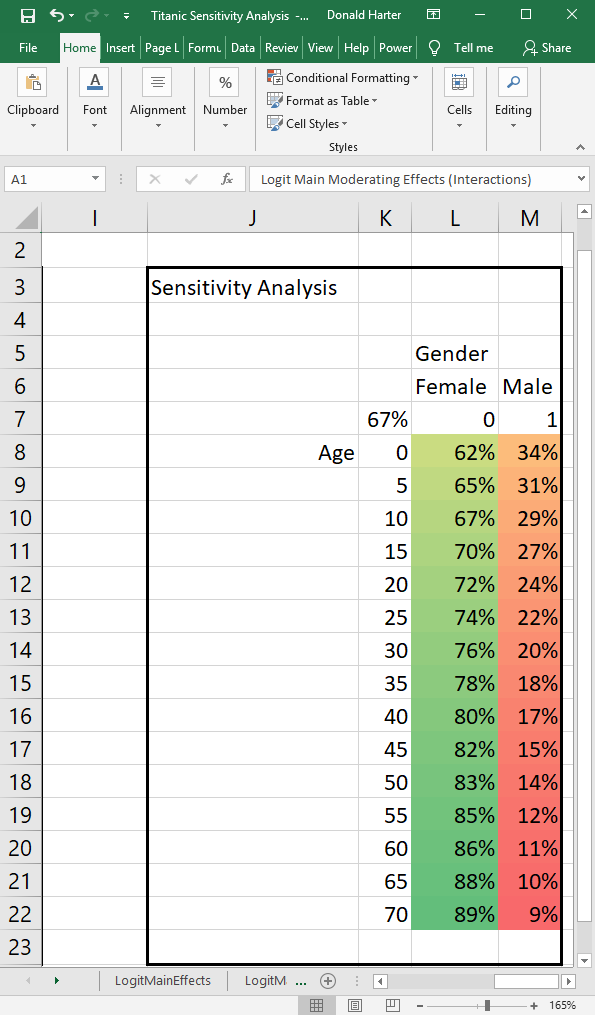
1. Use the spreadsheet tab LogitModeratingEffects to build to prediction model.
2. To predict survivability of a 10-year-old girl, enter 0 for gendernum and 10 for age in cells B7 and B8.
3. In cell B9, to calculate gendernum\*age, enter =B7\*B8.
4. Enter the coefficients in cells F6 through F9.
5. Enter the formulas for the values in cells G6 through G9; recall that the value for the intercept is 1.
6. Enter the formulas for Coefficient\*Values in cell H6 through H9.
7. In cell H11, enter the formula for the sum of H6:H9.
8. In cell H12, enter the formula for the exponential of H11.
9. In cell H13, enter the formula for the logistic function, =H12/(1+H12).



**Logit Sensitivity Analysis with Moderating Effect**

To perform a sensitivity analysis of gender versus age with a moderating effect, use the Sensitivity section of the same spreadsheet.

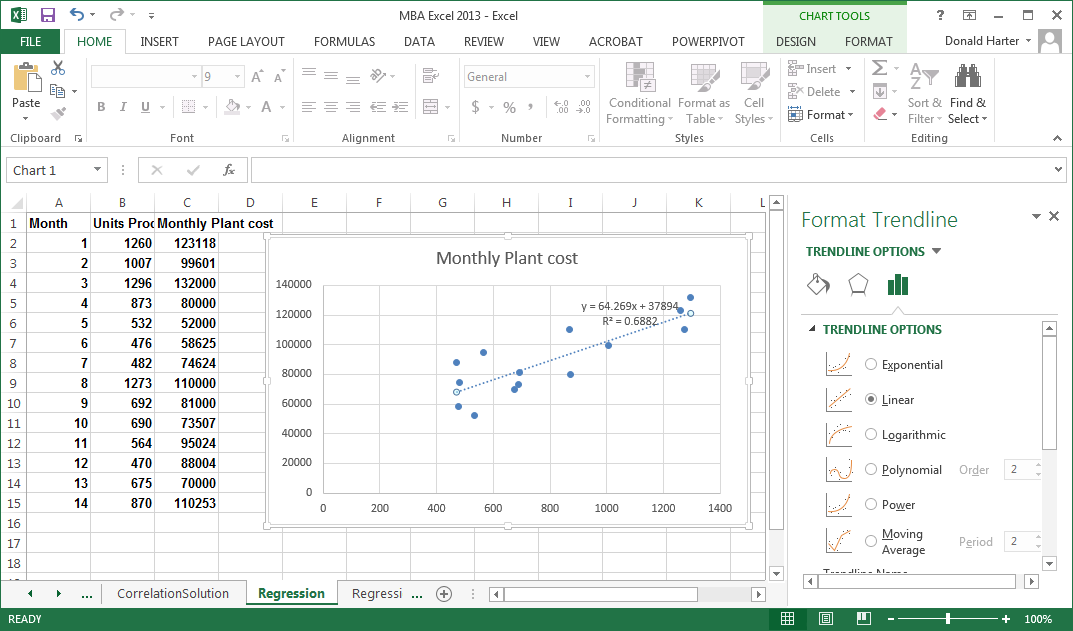
1. In general, for a sensitivity analysis, put one dimension across the top (in this case, gender), and one dimension down the side (in this case age).
2. In the cell in the corner, you must enter the formula for the probability calculation. This is already stored in cell H13. In cell K7, enter =H13.
3. Highlight K7 through M22.
4. Click on the Data tab, then What If analysis, Data Table.
5. For Row input cell, point to the cell with the value for gendernum, in this case B7.
6. For Column input cell, point to the cell with the value for age, in this case B8.
7. Click OK.
8. What pattern do you see?
9. Notice that for each gender, the curve is monotonic (always increasing or always decreasing).
10. Are the male and female curves monotonic in the same direction?

****

**8.6 Perceptron and Neural Network Overview**

**8.6.1 Review of Linear Regression**

The purpose of linear regression was to identify the linear relationship between a dependent (response) variable and one or more independent (explanatory) variables. For the data below, we created the regression line that best fits the data.



For this data, the line that best fits the data is:

Monthly Plant Cost = 37894 + 64.269 \* Units Produced

Another way to view this equation is:

1

Units

Produced

Monthly

Plant Cost

64.269

37894

∑

**8.6.2 Review of Logit and the Logistic Function**

The purpose of logit regression was to identify the relationship between a binary dependent variable using a logistic function and explanatory variables. The form of the logistic function is:

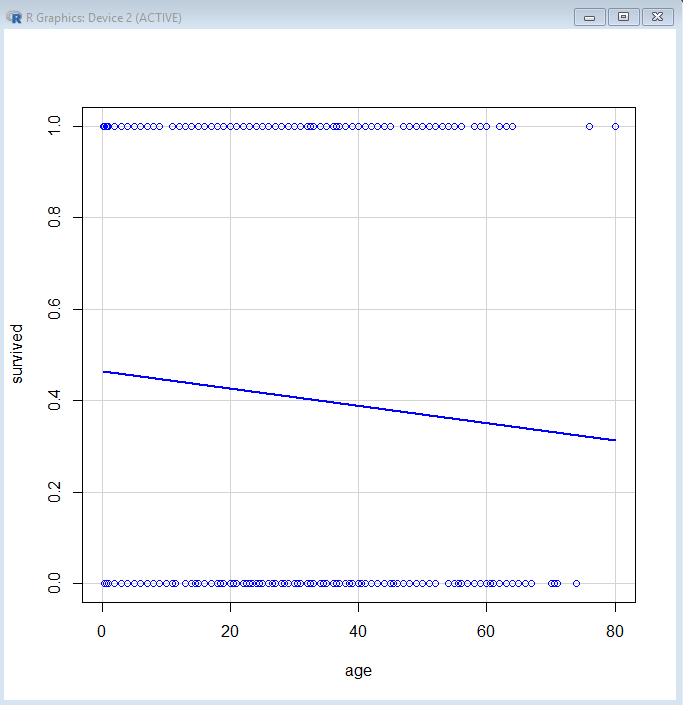
f(X) = exp(∑βiXi) / (1 + exp(∑βiXi))

Where:

βi = the coefficients in a logit regression

Xi = the variables in a logit regression

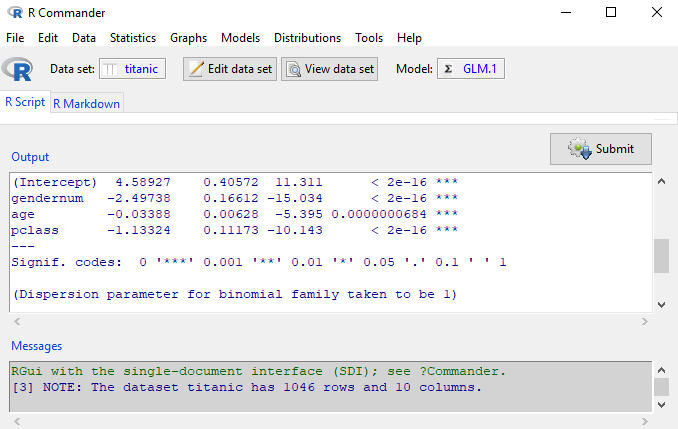
We are trying to find a function that fits the data below. A linear regression is not appropriate when the dependent variable is zeroes and ones.



A better solution looks like this:

Logit graph with decreasing probability

For the logit regression below, we created the logistic function that best fits the data.



The logistic function that predicts the probability of surviving is:

P(survived) = exp(4.589 - 2.497\*gender - 0.034\*age – 1.133\*pclass)

1 + exp(4.589 - 2.497\*gender - 0.034\*age – 1.133\*pclass)

Pictorially, this equation looks like:

1

4.589

gender

-2.497

-0.034

age

pclass

exp(∑)\_

1+exp(∑)

-1.133

Prob(Surv)

**8.6.3 Perceptrons**

Reference:

Rosenblatt, Frank. “The Perceptron—A Perceiving and Recognizing Automaton.” Report 85-460-1, Cornell Aeronautical Laboratory, 1957.

Minsky, M. L., and S. A. Papert. *Perceptrons*. Cambridge, MA: MIT Press, 1969.

A perceptron is a technique developed in the 1950s and 1960s to linearly classify data points into two groups. In the example below, red triangles identify customers who purchased a Lexus, blue dots are customers who did not purchase. The line is the linear perceptron classification line that separates the two groups.

Perceptron on age versus income indicating purchases of a Lexus

But the perceptron could not classify all types of problems. Computer scientists recognized that it could classify data point formed by the classic “AND” condition. In the following example, the X marks the condition when a customer is both rich and young; O when this is not true.

|  |  |  |
| --- | --- | --- |
|  | Rich | Not Rich |
| Young | X | O |
| Not Young | O | O |

A line can be used to classify the Xs and Os.

|  |  |  |
| --- | --- | --- |
|  | Rich | Not Rich |
| Young | X | O |
| Not Young | O | O |

Similarly, the classic “OR” condition can be classified.

|  |  |  |
| --- | --- | --- |
|  | Rich | Not Rich |
| Young | X | X |
| Not Young | X | O |

With the result:

|  |  |  |
| --- | --- | --- |
|  | Rich | Not Rich |
| Young | X | X |
| Not Young | X | O |

But researchers used the Exclusive OR condition to demonstrate the weakness of a perceptron. An Exclusive OR means rich or young, but not both.

|  |  |  |
| --- | --- | --- |
|  | Rich | Not Rich |
| Young | O | X |
| Not Young | X | O |

In this case, there is not straight line that separates the groups.

**8.6.4 Neural Networks**

Reference:

Rumelhart, D. E., and James McClelland. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge: MIT Press, 1986.

David Rumelhart and Jay McClelland recognized the limitation of a linear perceptron and proposed two innovations in 1986.

1. Use the logistic function to represent nonlinear behavior.
2. Add another layer (called the hidden layer) to produce more complex functions and represent more complex relationships.

[](http://upload.wikimedia.org/wikipedia/commons/e/e4/Artificial_neural_network.svg)

exp(∑)\_

1+exp(∑)

exp(∑)\_

1+exp(∑)

exp(∑)\_

1+exp(∑)

exp(∑)\_

1+exp(∑)

exp(∑)\_

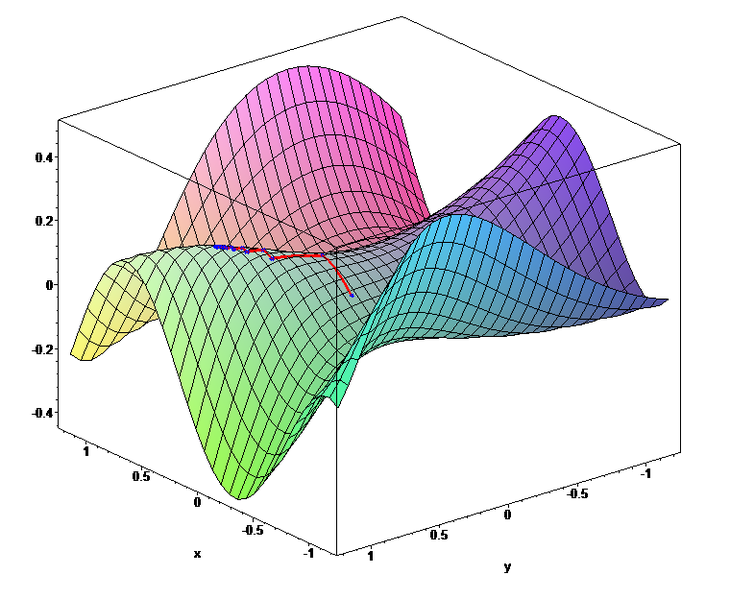
1+exp(∑)

exp(∑)\_

1+exp(∑)

Why use the logistic function (used in logit) rather than the normal distribution (used in probit)? The logistic function has a very simple derivative. Why is this important?

Neural network searches use gradient search. Imagine that you are climbing a hill. To reach the peak in the shortest amount of time, look at where you are standing and find the direction with the steepest slope. Head in that direction, then decide in a new direction.

[](http://upload.wikimedia.org/wikipedia/commons/6/68/Gradient_ascent_(surface).png)

The risk is that there might be multiple high points, where some are local optima. Gradient search uses multiple starting points to find the global optimum.

So, why is a simply derivative for the logistic function important? The derivative gives you the slope so you can determine search direction. Other functions could be used, but the logistic function is the most popular because the derivative is easy to calculate.

If f(x) is the logistic function, then the derivative is f(x) \* (1 – f(x)).

**8.6.5 R: Neural Network Demo**

**Installing neuralnet**

Follow these steps to install neuralnet.

1. In R, type the command:

install.packages("neuralnet," dependencies = TRUE)

1. If prompted for a CRAN mirror, select the location closest to you; use a USA location near you, then click OK.
2. If prompted to create a personal library, click Yes.
3. If prompted to add missing packages, click Yes.

**Launch neuralnet**

neuralnet is the R software that performs neural network calculations:

1. Type:

library(neuralnet)

1. If you receive a warning message that some packages are missing, it will ask if you want them installed. Click Yes.
2. On the Install Missing Packages screen, click OK.
3. R will install the necessary software.

**Neural Network Analysis**

To run the neural network on loan defaults with inputs of loan to income ratio (LTI) and age, copy the command into R:

titanicnet <- neuralnet(survived ~ gendernum + age, titanic, hidden=2, lifesign="minimal", linear.output=FALSE, threshold=0.01)

where:

titanicnet stores the results

neuralnet program which runs the neural network analysis

survived dependent variable (0 = did not survive, 1 =survived)

gendernum independent variable – gender (0 = female, 1 = male)

age independent variable – age in years

hidden number of hidden nodes

lifesign amount of output

linear.output whether you want linear or non-linear model

threshold error term threshold

The neural network algorithm will perform a gradient search to find a solution that minimizes the error.

**Neural Network Model**

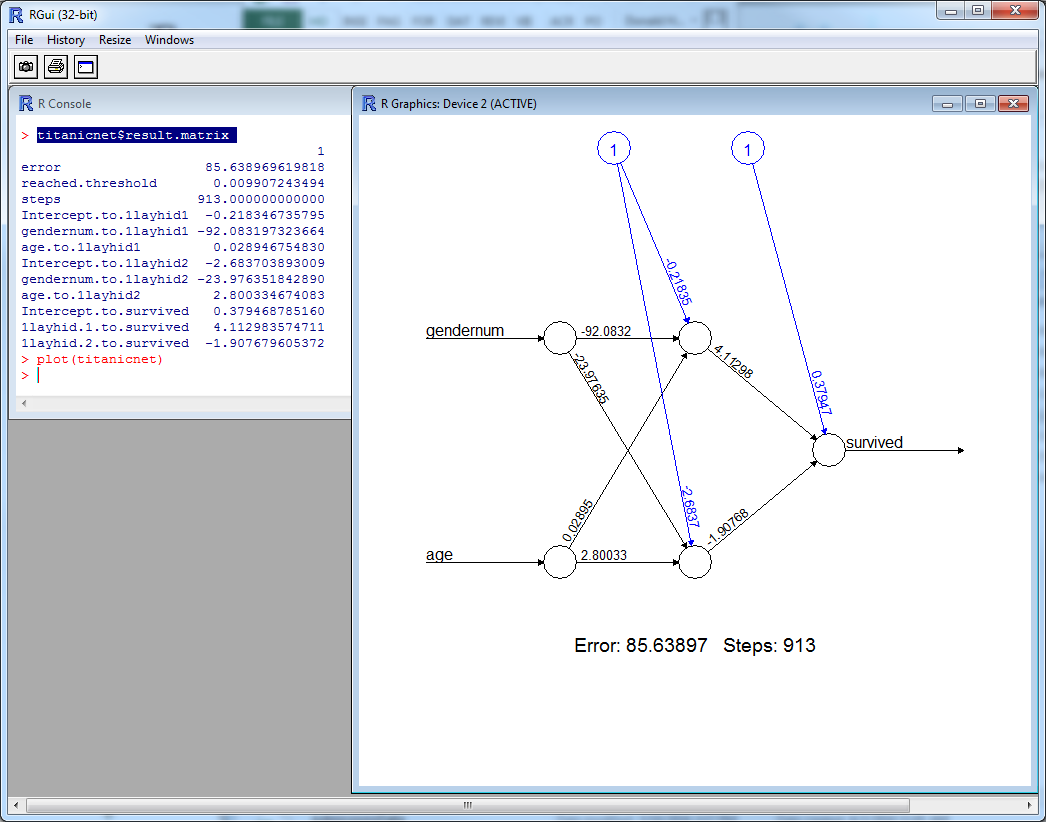
The result of the model can be displayed by plotting the model

1. To list the coefficients, type the command:

titanicnet$result.matrix

1. To generate the graph, in the R console (RGui), type the command:

plot(titanicnet)

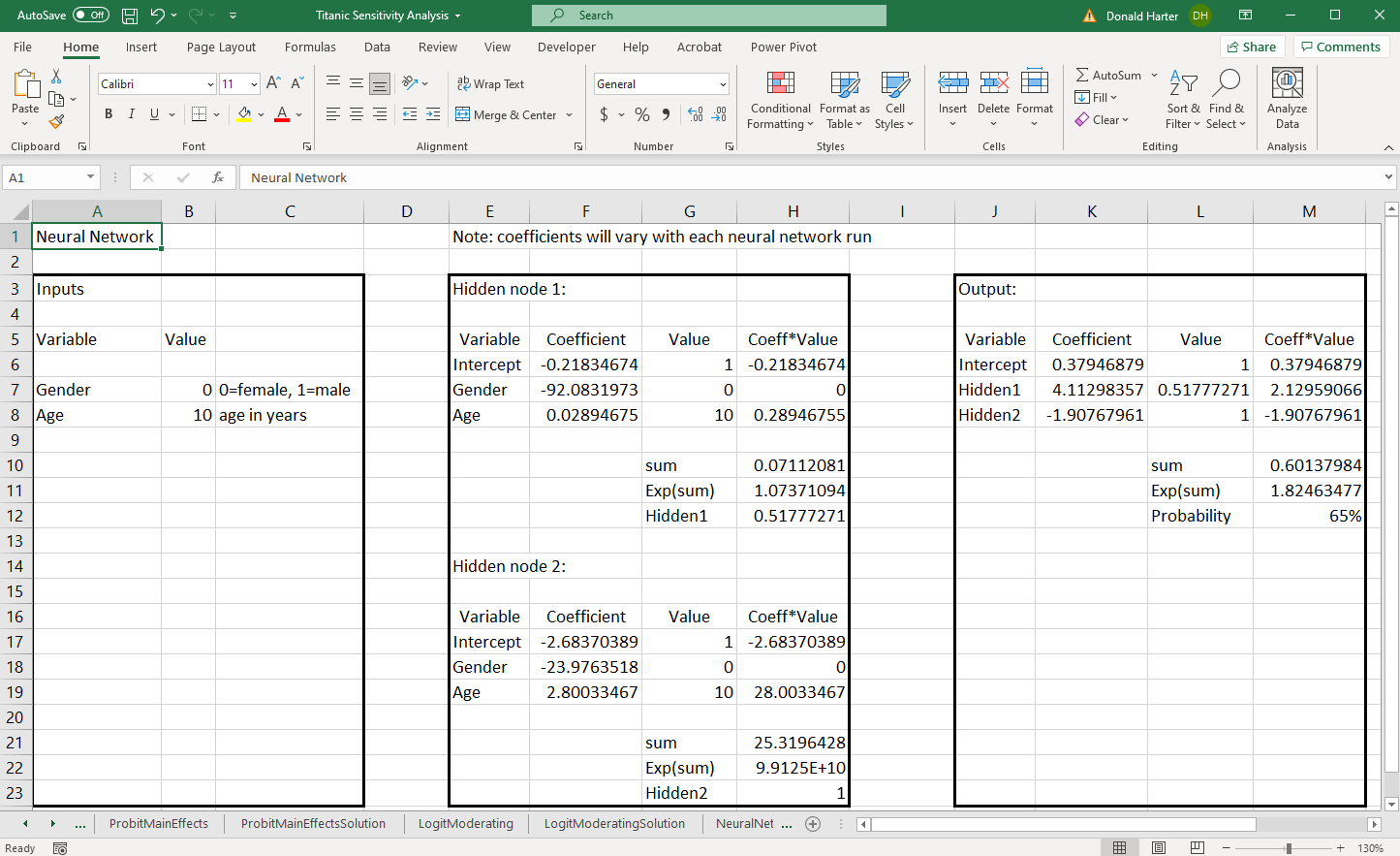


**8.6.6 Neural Network Prediction**

Interpreting the results of a neural network is often easier to do through a spreadsheet prediction and sensitivity analysis. We will take the coefficients from the neural network and build an Excel spreadsheet to calculate the predictions of survivability of Titanic passengers, using gender and age.

Let’s build the neural network calculations:

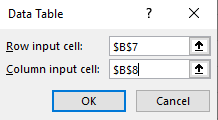
1. Click on the NeuralNetwork tab in the Titanic Sensitivity Analysis Excel workbook.
2. The two input variables which we used in the neural network were Gender and Age.
   1. Type in a sample gender number into B7 (0 for female).
   2. Type a sample age into B8 (in this case, 10 for 10 years old).
3. For Hidden node 1, in cells F6:F8, enter the coefficients.
4. In cells G6:G8, enter the value formulas (intercept value is 1).
5. In cells H6:H8, enter coefficient\*value formula.
6. In cell H10, enter sum formula.
7. In cell H11, enter exponential of the sum.
8. In cell H12, enter =H11/(1+H11) for the logistic function.
9. Repeat the process for Hidden node 2.
10. In the Output section, enter the coefficients in cells K6:K8.
11. For the Values:
    1. Enter 1 for the intercept.
    2. For Hidden 1, enter =H12 to point to the calculation for hidden node 1.
    3. For Hidden 2, enter =H23 to point to the calculation for hidden node 2.
12. In cells M6:M8, enter coefficient\*value formula.
13. In cell M10, enter sum formula.
14. In cell M11, enter exponential of the sum.
15. In cell M12, enter =M11/(1+M11).

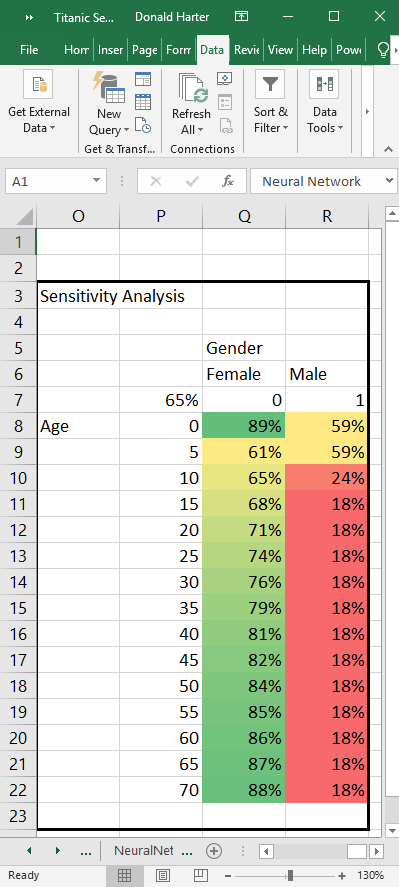


**Sensitivity Analysis**

Now that the calculations have been created for the neural network, let’s develop a two-way sensitivity analysis.

1. Across the top of the sensitivity analysis is gender.
2. Down the side of the sensitivity analysis is age.
3. The formula is in the corner cell, P7. Point to the final formula output =M12.
4. Click on the Data tab, What-If-Analysis, then Data Table.
5. Since Gender varies across the row, enter B7 for Row input cell.
6. Since Age varies down the column, enter B8 for column input cell.
7. Click OK.





1. How is this pattern different from the logit?

**8.7.1 Deep Learning**

Neural networks are not limited to one hidden layer. Neural networks have more than one hidden layer are called deep neural networks and can learn much more subtle patterns and strategies. An example of a deep neural network might look like:

Inputs

Hidden Layer 1

Hidden Layer 2

Outputs

exp(∑)\_

1+exp(∑)

exp(∑)\_

1+exp(∑)

exp(∑)\_

1+exp(∑)

exp(∑)\_

1+exp(∑)

exp(∑)\_

1+exp(∑)

exp(∑)\_

1+exp(∑)

exp(∑)\_

1+exp(∑)

**8.7.2 R: Deep Learning Demo**

Using the Titanic data once again, let’s use two inputs (gender, age), two hidden layers (with two nodes each), and one output. The command becomes:

titanicnet <- neuralnet(survived ~ gendernum + age, titanic, hidden=c(4,3), lifesign="minimal", linear.output=FALSE, threshold=0.01)

titanicnet stores the results

neuralnet program which runs the neural network analysis

survived dependent variable (0 = did not survive, 1 =survived)

gendernum independent variable – gender (0 = female, 1 = male)

age independent variable – age in years

hidden number of hidden nodes at each level: c(2,4) would mean 2 in hidden layer 1, 4 in hidden layer 2

lifesign amount of output

linear.output whether you want linear or non-linear model

threshold error term threshold

The result of the model can be displayed by plotting the model.

1. To generate the graph, in the R console (RGui), type the command:

plot(titanicnet)

