MIS771 Descriptive Analytics and Visualisation

Topic 6 Tutorial – Logistic Regression

Introduction

In this tutorial, you will cover logistic regression analysis.

Specifically, the aims of this tutorial are to:

- Understand where and when to use logistic regression for prediction.
- Generate logistic regression models to illustrate the key features of logistic regression modelling.
- Assess practical and statistical significance of logistic regression models.
- Interpret key outputs of logistic regression modelling.
- Use logistic regression models for prediction.
- Visualise "predicted probabilities (PP)" of a dependent variable for a given range of independent variable values.

Scenario

We continue with the analysis of 2012 US election survey conducted by ANES (American National Election Study Group). The main purpose of this survey is to ask respondents if they intend to vote for Barack Obama or Mitt Romney for U.S. President in 2012.

As an analyst working in Obama campaign, you are instructed to use ANES dataset and develop a model to predict people voting behaviour using a number of factors included in the survey such as respondent's race/ethnicity and how they feel about the Democratic and Republican Parties.

Once you developed a predictive model, you then need to advice the campaign manager based on the results of your analysis. Your recommendations should help design and deploy effective election campaigns that are customised to specific target audience (voters' subgroups) by answering the following questions:

- Are respondents with more positive feelings about the Democratic Party more likely to vote for Obama as the Democratic candidate?
- Are respondents with more positive feelings about the Republican Party less likely to vote for Obama as the Democratic candidate?
- Are those with higher incomes less likely to vote for Obama?
- Are members of any racial or ethnic minority groups (i.e., African-Americans, Hispanics, and others) more likely to vote for Obama than are White-American voters?

Open the data file and install the <u>Real Statistics Analysis Tool Pak</u>

a) Download the file ***ANES_Dataset_Tut6.xls*** from

"Content → Learning Resources → Topic 6 Folder" in Cloud Deakin. **Save it** to your hard drive.

b) Open the file in Excel.

c) Install the data <u>Real Statistics Analysis Tool Pak</u>.

Instruction:

Detailed instruction on how to download, install, and activate Real Statistics Analysis Tool Pack is provided on Cloud Deakin. You can read the instruction from:

"Discussions → General Unit Discussion Forum → Real Statistics Analysis Tool Pack Installation Guide"

1. Building a Predictive Model

a) Before developing a logistic regression model, explain why this technique should be used instead of Ordinary Least Squares Multiple Regression.

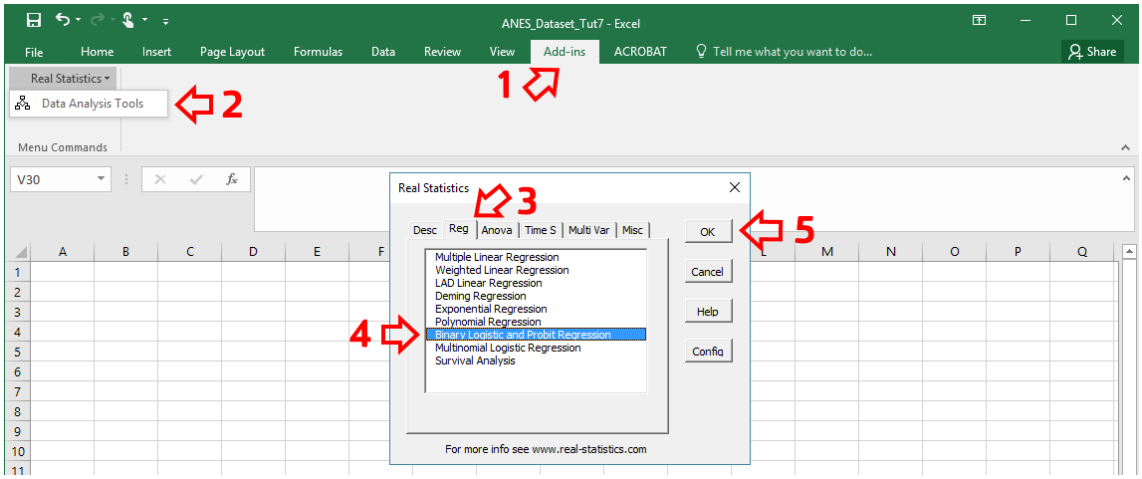
b) Decide on the Independent Variables that you need to include in your predictive model in order to answer scenario questions:

Dependent Variable = vote_obama
Independent Variables = ???

c) Run you logistic regression model by following instructions below:

Instruction:

From the *Add-ins* tab, select *Real Statistics* and then choose ***Data Analysis Tools***. The pop-up window in figure below will appear. Go to *Reg* tab and select *Binary Logistic and Probit Regression* and click "OK".



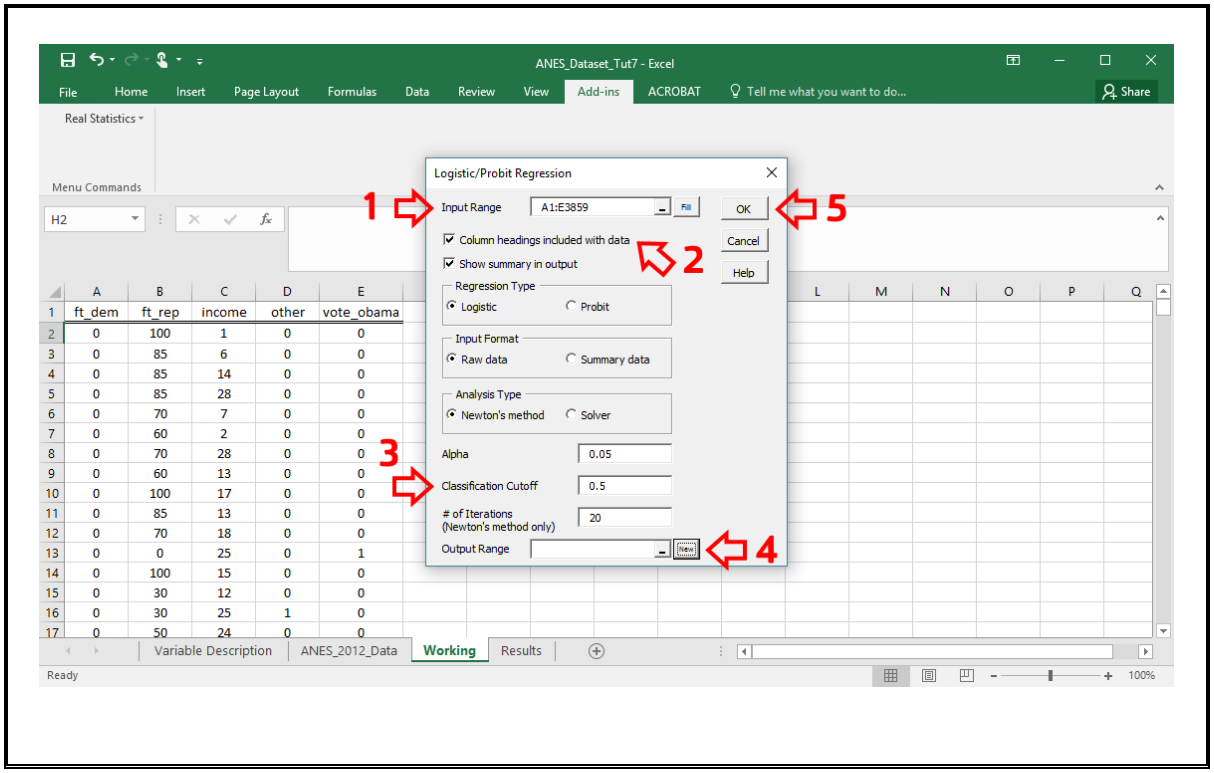
In "Logistic/Probit Regression" window, specify *Input Range* (i.e., the range of values for all variables – both independent and dependent variable).

Unlike Analysis Tool Pack, in Real Statistic Tool Pack we do ***NOT*** *specify independent and independent variables values separately.* ***Real Statistics assumes that the last column to the right holds Dependent Variable values****. Thus, make sure you place your dependent variable on the rightmost column before specifying 'Input Range'.*

If variable labels are included in the *Input Range*, make sure you select '*Column headings included with the data*' option.

Depending on the research context and purpose, you can change the '*Classification Cutoff*' value if you like. By default, it is set to **0.50**.

Finally, specify the *Output Range* (i.e. where you would like to have your analysis output) and click *'OK'* (see figure below).

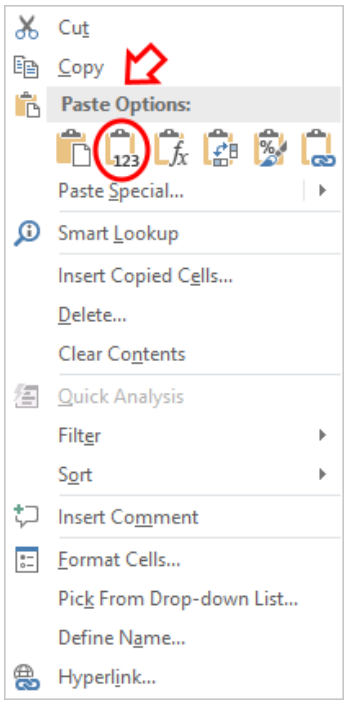


	A	B	C	D	E
1	ft_dem	ft_rep	income	other	vote_obama
2	0	100	1	0	0
3	0	85	6	0	0
4	0	85	14	0	0
5	0	85	28	0	0
6	0	70	7	0	0
7	0	60	2	0	0
8	0	70	28	0	0
9	0	60	13	0	0
10	0	100	17	0	0
11	0	85	13	0	0
12	0	70	18	0	0
13	0	0	25	0	1
14	0	100	15	0	0
15	0	30	12	0	0
16	0	30	25	1	0
17	0	50	24	0	0

2. Checking for Practical Significance of the Logistic Regression Model.

a) Based on logistic regression output (Q1c) comment on the practical significance of the model.

Once regression output was created, copy all relevant outputs and *paste* them **as "values"** to remove all background computations (formulas) from your output (see below).



3. Checking for Statistical Significance of the Logistic Regression Model.

a) Evaluate the overall fit of the logistic regression model (-2*LL* and its significance).

b) Comment on the explanatory power of the logistic regression model (Pseudo R^2 values).

c) Interpreting Logistic Regression Coefficients (B_i).

   i. Explain the directions of relationships between independent variables and the outcome variable (and then answer the case study questions).

   ii. Discuss the magnitude of coefficients by first calculating Exponentiated Regression Coefficients (EXP B_i) (i.e. Odds Ratios)

Instruction:

To calculate Exponentiated Regression Coefficients from coefficients produced by RealStatistics Tool Pack, use the following excel equation:

F40 = EXP (E40), where EXP is exponential function (e^{value})

Use this function to calculate exponentiated values of all logistic regression coefficients.

iii. Determine the significance of logistic regression coefficients by calculating:
   a. Standard Errors (*SE*)
   b. Wald Statistics (*Wald*)
   c. Significance of Logistic Coefficients (*p*-value).

Instruction:

- ***Calculating Standard Errors of Logistic Regression Coefficients:***

  Standard Errors can be calculated from the ***Covariance Matrix*** produced by RealStatistics Tool Pack.
  ***Note*:** that covariance matrix is in essence a correlation matrix wherein values are not normalised (i.e., scaled between -1 to +1).

  Standard Error (SE) of a coefficient for a given variable is equal to square root of its variance (i.e. its corresponding value on covariance matrix diagonal).

  To calculate a standard error (e.g. SE of the Intercept), type the following function:

  N54 = SQRT(D54), where N54 is SE value and D54 contains intercept variance in the covariance matrix.

  Use this function to calculate standard errors of all independent variables regression coefficients.

- ***Calculating Wald Statistics***

  Like *t*-statistic in linear regression, *Wald* statistic tells us whether the B coefficient is significantly different from zero:

  $$\textbf{Wald Statistic } (z^2) = (B/SE)^2$$

  To calculate Wald statistic (e.g. Wald of the Intercept), type the following function:

  O54 = (L54/N54)^2, where L54 contains intercept coefficient value, and N54 includes intercept standard error.

  Use this function to calculate Wald Statistics of all independent variables regression coefficients.

- ***Calculating statistical significance (p-values) of regression coefficients***

  Wald statistic, which is analogous to t-statistic, follows a special distribution known as "chi-square" distribution. To calculate *p*-values (significance) of Wald statistics (and thus, logistic regression coefficients), use the following formula in excel:

  P54 = CHIDIST(O54,1), where P54 contains calculated *p*-value; O54 includes *Wald* ratio of the constant. *df* (degrees of freedom) in this test is ALWAYS equal to 1.00.

  Use this function to calculate significance (*p*-values) for all independent variables regression coefficients.

d) Review and explain Receiver Operating Characteristic (ROC) curve output.

The ROC Curve is a plot of values of the False Positive Rate (FPR) versus the True Positive Rate (TPR) for all possible cut-off values from 0 to 1.

NOTE: The higher the ROC curve the better the fit. In fact, the area under the curve (AUC) can be used for this purpose. The closer AUC is to 1.00 (the maximum value) the better the fit. Values close to 0.50 show that the model's ability to discriminate between success and failure is due to ***chance***.

4. Using Logistic Regression Model for Prediction.

a) Predict the following probabilities:

   i. The predicted probability of a White-American with positive feeling towards Republicans (i.e. score of 80 on ft_repablican thermometer) and neutral feeling towards Democrats (i.e. score of 50 on ft_democrat thermometer) with an income within $27,500–29,999 range (category 10).
   ii. The predicted probability of a minority voter with positive feeling towards Democrats (i.e. score of 80 on ft_democrat thermometer) and neutral feeling towards Republicans (i.e. score of 50 on ft_repablican thermometer) with an income within $27,500–29,999 range (category 10).
   iii. Comment on predicted probabilities of voting for Obama for these two groups of voter.

5. Visualising Predicted Probabilities (PP) and Recommendations to Obama Campaign

Obama campaign manager would like to gain a deeper knowledge of White-American voters' behavior in the upcoming presidential election.

He is specifically interested in understanding probability of voting for Barack Obama for individuals who meet these two criteria:

a. Don't feel favorable towards Democrats (those with scores of 10, 20, 30, and 40 on ft-Democratic thermometer),
b. Feel neutral towards Republicans (with score of 50 on ft-Republican thermometer)

He believes that this group of voters will play a key role in Obama's success (or failure) in the upcoming presidential campaign and thus, future campaigns should specifically focus on delivering a right message to these voters.

Prior research also shows that individuals from higher income brackets more likely to swing to Republicans campaign if a Democratic candidate does not address concerns of these voters. Accordingly, the campaign manager would like to take into account voters' income level in predicting the voting behavior of above-mentioned group of voters:

> "White-American voters with **negative feeling towards Democrats** and **neutral feeling towards Republicans** across **different income levels** (Level 1 to Level 28)"

Your job is now to predict the probabilities of voting for Obama using the logistic regression model developed earlier by considering the above-mentioned criteria. You then need to plot these predicted probability values using appropriate visualization techniques:

a) Calculate probability of a White-American voter who scored 10 on ft-Democrat thermometer, 50 on ft-Republican thermometer, with category 1 income (i.e., under $5,000):

   i. Calculating Logit = $b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + b_4X_{4i}$
   ii. Calculating Odds = e^{logit}
   iii. Calculating Probability = $\frac{\text{Odds}}{(1+\text{Odds})}$

b) Repeat the above calculations for individuals at different levels of income (category 2 to 28).

c) Repeat this exercise (steps a and b) for those scored 20, 30, and 40 on ft-Democratic thermometer.

d) Plot the predicted probabilities of voting for Obama and interpret the resulting visualisation.

e) Advise Obama campaign manager using Predicted Probability Plot. Specifically, explain which group of voters are more likely to vote for Obama and thus should be targeted in upcoming election campaign.

HOMEWORK

Now it is your turn. Try producing your own logistic model by adding the following independent variables: sex, white, black, Hispanic.