**Outbreak of *Salmonella* Typhimurium in the UK**

Case-control study analysis: Part Two

# Logistic regression (and a bit of binomial regression)

# Stata guide

## Developed by:

Ranya Mulchandani (C2019) and Alicia Barrasa Blanco (Scientific Coordinator)

## April 2021

# Background

The scenario presented in this case study is based on an investigation of a cluster of *Salmonella* Typhimurium cases identified in the UK through the analysis of whole genome sequencing data in 2020.

The information presented in this case study and the associated data files have been deliberately changed to facilitate the acquisition of the learning objectives.

The aim of this case study is to analyse the data collected as part of a foodborne outbreak investigation, using either Stata or R software.

The case study is formed of two parts. This exercise is the second of the two parts (it is expected that the first part, the pre-module exercise, has already been completed)

# Learning outcomes

By the end of this exercise, participants should be able to:

1. Analyse data from a case-control study using logistic regression;
2. Understand the respective roles played by several exposures.

# Pre-requisites

Participants are expected to be familiar with data management as well as descriptive and stratified analysis in STATA or R and have completed the pre-course exercise (which forms part one of this case study).

# Reminder of part one of the case study…

Remember the scenario by checking the pre-module homework.

Univariate analysis results suggested that several food items were associated with illness, with the consumption of Nut bar Brand A showing the strongest association. These associations could be real (e.g. multiple items being contaminated) or due to another variable which acts as a confounding factor.

Stratified analysis showed that Nut bar Brand A was confounding the effect of several other food items, including chicken, cabbage, peanuts, hazelnuts, Nut bar Brand B, pecans and pistachios. In addition, there appears to be potential effect modification between consumption of brazil nuts and Nut bar Brand A.

Interpretation of results should be done with caution due to the small number of cases involved in this stratified analysis.

# Part 1: Logistic regression: adjusting for confounding

## **Question 1: How would you explore the effect of several risk factors?**

Proposed steps for the analysis:

* Using **salm\_v1.4.csv**, start by performing logistic regression analysis with only one dichotomous exposure variable (i.e. exposure = nut\_bar\_a, outcome = case), interpret the results and calculate the odds
* Compare the outcome of the logistic regression when using age as a categorical variable or as a continuous one and interpret the outcomes.
* Add more variables to the model, discuss the meaning of the constant term for each one of them
* Write down the model for each one of the above steps
* Start again with a simple model (one independent variable) and add more variables in a step-by-step fashion
* Compare each new nested model with the previous one (assessing the contribution of the new variable you add each time) by using the likelihood ratio test (commands estimates store and lrtest)
* Assess the fit of each model and try to identify the most parsimonious model

**Help question 1.**

Let’s remember some of the findings from the stratified analysis:

ccinter case nut\_bar\_b, by( nut\_bar\_a)

Number of obs = 139 , Missing = 0

nut\_bar\_a = Exposed

------------------------------+

nut\_bar\_b Cases Controls|

------------------------------| Odds Ratio **1.13 [0.06-78.49]**

Exposed 3 1 | Attrib.risk.exp 0.11 [-17.15-0.99]

UnExposed 8 3 | Attrib.risk.pop 0.03

------------------------------+

Total 11 4

Exp % 27% 25%

nut\_bar\_a = Unexposed

------------------------------+

nut\_bar\_b Cases Controls|

------------------------------| Odds Ratio **16.62 [0.78-988.61]**

Exposed 2 1 | Attrib.risk.exp 0.94 [-0.29-1.00]

UnExposed 13 108 | Attrib.risk.pop 0.13

------------------------------+

Total 15 109

Exp % 13% 1%

Test of Homogeneity (M-H) : pvalue : **0.1364003**

Crude OR for nut\_bar\_b : 13.21 [1.95-143.21]

MH OR for nut\_bar\_b adjusted for nut\_bar\_a : **3.67 [0.60-22.57**]

Adjusted/crude relative change : -**72.23** %

The stratum specific ORs are the same (note the M-H test for homogeneity p-value), and there is more than a 70% difference between the Mantel-Henzel OR for Nut bar Brand B adjusted by Nut bar Brand A and the crude OR (univariable). This means that the effect of the Nut bar Brand B is confounded by the effect of Nut bar Brand A (or vice versa) and the OR to report is the adjusted one.

There are more variables confounded by Nut bar Brand A, and there may be other confounders (since we didn’t do all the potential stratifications).

The objective of your multivariable analysis is to identify variables independently associated with the outcome and to control for confounding.

To prepare your dataset for multivariable analysis, you need to decide on the variables of interest based on your previous descriptive and stratified analysis and you might need to create or recode variables (age groups, dummy variables, etc...)

Logistic regression with only one exposure dichotomous variable.

A logistic regression follows the formula:



Where:

ln(p/1-p)) is the log of the odds for the outcome

α is the log of the odds in the unexposed

β is the log of the OR for exposure x

Stata command logit gives you these coefficients

logit case nut\_bar\_a

Logistic regression Number of obs = 139

LR chi2(1) = 25.10

Prob > chi2 = 0.0000

Log likelihood = -54.43597 Pseudo R2 = 0.1874

------------------------------------------------------------------------------

case | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

nut\_bar\_a | 2.994899 .6455618 4.64 0.000 1.729621 4.260177

\_cons | -1.983298 .2753924 -7.20 0.000 -2.523057 -1.443538

------------------------------------------------------------------------------

You can write down the above model by substituting α and β in the formula with the coefficients above:

log odds = -1.98+(2.99 \* nut\_bar\_a)

The OR for nut\_bar\_a is:

di exp(2.994899)

19.983341

The logit command with the ,or option or the logistic command (no option needed) gives you the ORs.

logistic case nut\_bar\_a

Logistic regression Number of obs = 139

LR chi2(1) = 25.10

Prob > chi2 = 0.0000

Log likelihood = -54.43597 Pseudo R2 = 0.1874

------------------------------------------------------------------------------

case | Odds Ratio Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

nut\_bar\_a | 19.98333 12.90048 4.64 0.000 5.638515 70.82248

\_cons | .1376147 .037898 -7.20 0.000 .080214 .2360909

------------------------------------------------------------------------------

This model corresponds to the equation:

odds = exp(α + βX) = cons\*exp(βX) = cons\*exp(β)X

The \_cons in the regression output corresponds to exp(α), which in cohort studies can be interpreted as the odds of being a case among the unexposed; in case control studies the interpretation is meaningless. Note that even if this \_cons is shown in the OR column, it is not an OR. This odds needs to be multiplied with the correct odds ratios for each exposure group to produce the odds of being a case for each exposure combination.

The OR=19.98 corresponds to exp(β) in the equation above: The odds of being a case among those that consumed Nut bar A is almost 20 times higher (95% CI: 5.64-70.82) that among those that didn’t.

Logistic regression using age as a categorical variable.

For categorical variables, you can create dummy variables for each level of the variable (minus 0, usually the reference level of that exposure). You can directly include age in groups (agegr) as categorical variables when running any regression model by using the i. prefix. This will result in Stata considering variables as categorical and directly create dummy variables. The lowest value ofagegris automatically set as a reference category (tabulate agegr to make sure you know how the numbers were labelled).

Logistic regression Number of obs = 136

LR chi2(4) = 10.01

Prob > chi2 = 0.0403

Log likelihood = -56.80688 Pseudo R2 = 0.0809

------------------------------------------------------------------------------

case | Odds Ratio Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

agegr |

35-44 | .1029412 .1196441 -1.96 0.050 .0105505 1.004393

45-54 | .6176471 .5217511 -0.57 0.568 .1179496 3.234329

55-64 | .9032258 .6249165 -0.15 0.883 .2327416 3.505248

64-more | 1.441176 1.042349 0.51 0.613 .3492003 5.947845

|

\_cons | .2857143 .1619848 -2.21 0.027 .0940472 .8679963

------------------------------------------------------------------------------

We can however ask Stata to change the reference level (for example use 3 instead of 0)

Use the following line of commands and check what happens:

logistic case ib3.agegr

Logistic regression using age as a continuous variable

What would have happened if we had included instead the continuous variableage? Try it and interpret the OR:

logistic case age

Logistic regression Number of obs = 136

LR chi2(1) = 3.09

Prob > chi2 = 0.0789

Log likelihood = -60.266009 Pseudo R2 = 0.0250

------------------------------------------------------------------------------

case | Odds Ratio Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

age | 1.032846 .0196184 1.70 0.089 .9951014 1.072022

\_cons | .0767036 .0497411 -3.96 0.000 .0215191 .2734054

------------------------------------------------------------------------------

Remember that the logistic equation can be expressed as:

odds = cons + exp(βX) = cons + exp(β)X

Though age does not appear to be associated with the outcome, the coefficient 1.033 represents the increase in the OR with one unit increase in age. What would be the OR for a 10-unit increase in age?

Adding a second variable to the model

Coming back to the model with Nut bar Brand A as the only exposure, let’s add now Nut bar Brand B. From the stratified analysis you know that the effect of Nut bar Brand B was confounded by Nut bar Brand A. Can you see that in the model?

logistic case nut\_bar\_a nut\_bar\_b

Logistic regression Number of obs = 139

LR chi2(2) = 28.21

Prob > chi2 = 0.0000

Log likelihood = -52.880315 Pseudo R2 = 0.2106

------------------------------------------------------------------------------

case | Odds Ratio Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

nut\_bar\_a | 15.51258 10.35136 4.11 0.000 4.194568 57.36944

nut\_bar\_b | 6.057862 6.214743 1.76 0.079 .8110997 45.24437

\_cons | .1275962 .0363346 -7.23 0.000 .073021 .2229605

------------------------------------------------------------------------------

Odds[illness] = exp(α + β1Χ1 + β2Χ2 + β3Χ3) = exp(a)\*exp(β1Χ1)\*exp(β2Χ2) = \_cons\* exp(β1\* nut\_bar\_a)\*exp(β2\* nut\_bar\_b)

Note that, in the above expression, nut\_bar\_a and nut\_bar\_b can get the values 0 or 1, according to whether they consumed Nut bar Brand A or Nut bar Brand B respectively.

\_cons = 0.12 is the odds of illness among the unexposed, i.e. among those who consumed neither Nut bar Brand A nor Nut bar Brand B.

exp(β1) = 15.51 is the OR for Nut bar Brand A adjusted by Nut bar Brand B. The odds of illness among those who consumed Nut bar Brand A but did not consume Nut bar Brand B is 15.5 (95% CI: 4.19-57.37) times higher compared to those who consumed neither Nut bar Brand A nor Nut bar Brand B.

exp(β2) = 6.058 is the OR for Nut bar Brand B adjusted by Nut bar Brand A. The odds of illness among those who consumed Nut bar Brand B but did not consume Nut bar Brand A is 6 time higher the odds of those who consumed neither neither Nut bar Brand A nor Nut bar Brand B, however, this finding is not statistically significant (p = 0.079).

Adding a third variable to the model.

logistic case nut\_bar\_a nut\_bar\_b brazil

Logistic regression Number of obs = 139

LR chi2(3) = 42.43

Prob > chi2 = 0.0000

Log likelihood = -45.773048 Pseudo R2 = 0.3167

------------------------------------------------------------------------------

case | Odds Ratio Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

nut\_bar\_a | 14.6955 10.79984 3.66 0.000 3.480403 62.04961

nut\_bar\_b | 4.506198 5.298396 1.28 0.200 .4497447 45.14967

brazil | 8.935365 5.181566 3.78 0.000 2.867514 27.84319

\_cons | .0728186 .0273474 -6.98 0.000 .0348794 .1520251

------------------------------------------------------------------------------

Try to write down the model and interpret all its coefficients.

Adding variables in a step-by-step fashion using the likelihood ratio test to compare different models.

Variables to be included in a multivariable regression model are selected on the basis of the results of the crude and stratified analyses. Variables showing an association with the outcome and having a p-value less than 0.2 are often considered eligible. The cut-off should be chosen depending on the specific situation (it should be pre-defined in your analysis protocol *before* you start the analysis). Often it is between 0.25 and 0.1 but higher p-values can sometimes be justified. However, if you have any reason to believe a specific variable (exposure) should be in the model (i.e. because it might be a confounder or is a primary hypothesis), you should include it in the model anyway. There is no golden rule in the final inclusion of variables in a multivariable analysis model, especially in outbreak investigations.

To be able to check statistically if the inclusion of a variable improves the model significantly, the models need to have the same number of observations. If any variable had missing values this means that these will have different number of observations, so you need to drop all the missings.

drop if agegr == .

(in case you have many variables with missing values there is and ado file that allows to do this in on single like of command, dropmissingthat you may need to download)

To include variable in the model here are two possible strategies:

* to start off with a model that includes only one independent variable and add others one by one (sometimes referred to as a “forwards stepwise approach”)
* to start with a full model (including all eligible variables) and, one at a time, remove variables that do not seem relevant (sometimes referred to as a “backwards stepwise approach”)

We will begin with only one independent variable.

logistic case nut\_bar\_a

Logistic regression Number of obs = 139

LR chi2(1) = 25.10

Prob > chi2 = 0.0000

Log likelihood = -54.43597 Pseudo R2 = 0.1874

------------------------------------------------------------------------------

case | Odds Ratio Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

nut\_bar\_a | 19.98333 12.90048 4.64 0.000 5.638515 70.82248

\_cons | .1376147 .037898 -7.20 0.000 .080214 .2360909

------------------------------------------------------------------------------

The likelihood ratio test (lrtest) helps identify whether the addition of variables contribute significantly to the model, but first you need to save the model statistics:

estimates store m1 // (will store estimates in the model m1, nothing will be shown on the display)

Now do a second model with one additional variable.

logistic case nut\_bar\_a nut\_bar\_b

Logistic regression Number of obs = 139

LR chi2(2) = 28.21

Prob > chi2 = 0.0000

Log likelihood = -52.880315 Pseudo R2 = 0.2106

------------------------------------------------------------------------------

case | Odds Ratio Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

nut\_bar\_a | 15.51258 10.35136 4.11 0.000 4.194568 57.36944

nut\_bar\_b | 6.057862 6.214743 1.76 0.079 .8110997 45.24437

\_cons | .1275962 .0363346 -7.23 0.000 .073021 .2229605

------------------------------------------------------------------------------

And store its estimates too.

estimates store m2 // (will store estimates in the model m2, nothing will be shown on the display)

Finally test for the difference in log likelihood

lrtest m2 m1

Likelihood-ratio test LR chi2(1) = 3.11

(Assumption: m1 nested in m2) Prob > chi2 = 0.0778

If thelrtestis statistically significant, this suggests that the addition of the new variable in the model significantly improves the likelihood of this model.

The results of thelrtest(p = 0.0778) suggest a borderline significance (at the 0.05 level) for the addition of the variablenut\_bar\_b*.* Remember this might be a confounder, so this may be a sufficient reason for which you may want to keep it in the model regardless of its p-value in the likelihood ratio test.

Now extend to other variables and store estimates in m3. Proceed similarly to extend or drop the model according to thelrtestresults.

Keep or drop other variables as needed. Take the likelihood ratio test, p values, magnitude of OR, and proportion of cases exposed into account in order to decide.

Assessing the fit of each model, try to identify the most parsimonious model

You can acquire the same coefficients for the different logistic regression models using glmcommands. This command can be used for any generalised linear model, including logistic regression, as long as you specify the link function which is appropriate for the type of model (i.e. logistic regression) that you are trying to fit. Because it provides different post-estimations to thelogistic*,* logitorregresscommands, and because it can compare models with different number of observations it is sometimes preferred. For logistic regression the line of commands will be as follows:

glm case nut\_bar\_a, link(logit) family(binomial) eform

Generalized linear models No. of obs = 139

Optimization : ML Residual df = 137

Scale parameter = 1

Deviance = 108.8719391 (1/df) Deviance = .7946857

Pearson = 139 (1/df) Pearson = 1.014599

Variance function: V(u) = u\*(1-u) [Bernoulli]

Link function : g(u) = ln(u/(1-u)) [Logit]

AIC = .8120283

Log likelihood = -54.43596953 BIC = -567.151

------------------------------------------------------------------------------

| OIM

case | Odds Ratio Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

nut\_bar\_a | 19.98333 12.90048 4.64 0.000 5.638515 70.82248

\_cons | .1376147 .037898 -7.20 0.000 .080214 .2360909

------------------------------------------------------------------------------

Using theglmcommand and the post-estimationestat ic command the value of the AIC is shown. You can compare AIC between models to decide which model is the most parsimonious; to do this, you need to show the model information criteria:

estat ic

-----------------------------------------------------------------------------

Model | Obs ll(null) ll(model) df AIC BIC

-------------+---------------------------------------------------------------

. | 139 . -54.43597 2 112.8719 118.7409

-----------------------------------------------------------------------------

Now do a second model including nut\_bar\_b and display the AIC command:

glm case nut\_bar\_a nut\_bar\_b, link(logit) family(binomial) eform

You can now add more variables to the model and compare the different AIC; the model with the lowest AIC value will be the most parsimonious.

Note that the AIC obtain in the output of the glm command takes into account the number of observations and therefore it differs from the AIC that we will use and that is obtained with estat ic.

# Part two: Logistic regression: including interactions

Remember that your stratified analysis showed that the effect of brazil nuts was not the same depending on whether Nut bar Brand A was present or not, meaning that the ORs significantly differ between the two strata and effect modification may be present.

## **Question 2. How would you account for effect modification?**

Proposed steps for the analysis

* Using **salm\_v1.4.csv** perform a stratified analysis using logistic regression to check for interactions.
* Add an interaction term to the model.
* Check whether the interaction term improves the fit of the model.

**Help question 2.**

First let’s remember what we saw in the stratified analysis.

ccinter case brazil, by(nut\_bar\_a)

Number of obs = 139 , Missing = 0

nut\_bar\_a = Exposed

------------------------------+

brazil Cases Controls|

------------------------------| Odds Ratio 0.83 [0.04-15.78]

Exposed 5 2 | Attrib.risk.exp 0.17 [-14.78-0.96]

UnExposed 6 2 | Attrib.risk.pop 0.08

------------------------------+

Total 11 4

Exp % 45% 50%

nut\_bar\_a = Unexposed

------------------------------+

brazil Cases Controls|

------------------------------| Odds Ratio 16.67 [4.07-69.06]

Exposed 9 9 | Attrib.risk.exp 0.94 [0.75-0.99]

UnExposed 6 100 | Attrib.risk.pop 0.56

------------------------------+

Total 15 109

Exp % 60% 8%

Test of Homogeneity (M-H) : pvalue : 0.0209197

Crude OR for brazil : 10.82 [3.58-32.65]

MH OR for brazil adjusted for nut\_bar\_a : 6.41 [2.15-19.17]

Adjusted/crude relative change : -40.71 %

The effect of brazil nuts on illness is modifed by the consumtion of Nut bar Brand A. It seems that the exposure to brazil nuts has no effect on those that were also exposed to Nut bar Brand A and there is a clear association between brazil nuts when the consumption of Nut bar Brand A is not present.

Perform a stratified analysis using logistic regression to check for interactions.

You can obtain the same ORs shown before using logistic regression in a stratified manner:

logistic case brazil if nut\_bar\_a ==0

logistic case brazil if nut\_bar\_a ==1

or by adding an interaction term to the model.

You could manually generate the interaction term as a new variable (nutbara\_brazil)defined as the product of nut\_bar\_Aandbrazil*.* This variable equals one ifnut\_bar\_Aandbrazilare present at the same time. Otherwise it is zero. This interaction term can then be inserted into the model.

generate nutbara\_brazil = nut\_bar\_a\* brazil

logistic case nut\_bar\_a brazil nutbara\_brazil

In Stata, by typing the following, you don’t need to generate the interaction manually.

logistic case nut\_bar\_a##brazil

Logistic regression Number of obs = 139

LR chi2(3) = 45.53

Prob > chi2 = 0.0000

Log likelihood = -44.220186 Pseudo R2 = 0.3399

----------------------------------------------------------------------------------

case | Odds Ratio Std. Err. z P>|z| [95% Conf. Interval]

-----------------+----------------------------------------------------------------

1.nut\_bar\_a | 50 45.91659 4.26 0.000 8.265793 302.4513

1.brazil | 16.66667 10.52628 4.45 0.000 4.833368 57.47085

|

nut\_bar\_a#brazil |

1 1 | .05 .0664371 -2.25 0.024 .0036978 .6760788

|

\_cons | .06 .025219 -6.69 0.000 .0263255 .1367496

----------------------------------------------------------------------------------

The model is:

odds = exp(α + β1Χ1 + β2Χ2 + β3Χ3) = cons \* exp(β1nut\_bar\_a + β2brazil + β3nut\_bar\_a\*brazil)

odds = cons \* exp(β1)nut\_bar\_a \* exp(β2)brazil \* exp(β3 )nut\_bar\_a\*brazil

odds = 0.0519 \* 125.12nut\_bar\_a \* 0.99brazil \* 0.32nut\_bar\_a\*brazil

* The coefficient that accompanies the interaction term cannot be interpret alone, but the p value tells us that interaction term can remain in the model.
* The odds of illness among those who consumed Nut bar Brand A but did not consume brazil nuts was 50 times higher compared to those who consumed neither Nut bar Brand A nor brazil nuts.
* The odds of illness among those who consumed brazil nuts but did not consume Nut bar Brand A was 16 times higher compared to those who consumed neither Nut bar Brand A nor brazil nuts.
* The odds of illness among those consumed Nut bar Brand A and brazil nuts was 41.67 times (50\*16.66667\*0.05=41.666675) higher compared to those who Nut bar Brand A nor brazil nuts.

This result can be obtained with thelincomcommand executed after the logistic command.

lincom 1.nut\_bar\_a + 1.brazil + 1.nut\_bar\_a #1.brazil

( 1) [case]1.nut\_bar\_a + [case]1.brazil + [case]1.nut\_bar\_a#1.brazil = 0

------------------------------------------------------------------------------

case | Odds Ratio Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

(1) | 41.66667 39.0127 3.98 0.000 6.64974 261.0795

------------------------------------------------------------------------------

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Nut bar Brand A** | **Brazil nuts** | **Cases** | **Controls** | **OR** |
| 1 | 1 | 5 | 2 | 41.66667 |
| 0 | 1 | 9 | 9 | 16.66667 |
| 1 | 0 | 6 | 2 | 50 |
| 0 | 0 | 6 | 100 | Reference |

Table showing cases and controls according to level of exposure to Nut bar Brand A and brazil nuts consumption.

Does the interaction term improve the fit of the model?

First the parameters of both models have to be stored to be compared by the likelihood ratio test. Then the test is applied.

logistic case nut\_bar\_a brazil

Logistic regression Number of obs = 139

LR chi2(2) = 40.77

Prob > chi2 = 0.0000

Log likelihood = -46.60236 Pseudo R2 = 0.3043

------------------------------------------------------------------------------

case | Odds Ratio Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

nut\_bar\_a | 17.16436 12.23122 3.99 0.000 4.246848 69.3727

brazil | 9.468999 5.399341 3.94 0.000 3.09698 28.95141

\_cons | .075868 .0280695 -6.97 0.000 .0367394 .1566696

------------------------------------------------------------------------------

Note: \_cons estimates baseline odds.

estimates store model1

logistic case nut\_bar\_a##brazil

Logistic regression Number of obs = 139

LR chi2(3) = 45.53

Prob > chi2 = 0.0000

Log likelihood = -44.220186 Pseudo R2 = 0.3399

----------------------------------------------------------------------------------

case | Odds Ratio Std. Err. z P>|z| [95% Conf. Interval]

-----------------+----------------------------------------------------------------

1.nut\_bar\_a | 50 45.91659 4.26 0.000 8.265793 302.4513

1.brazil | 16.66667 10.52628 4.45 0.000 4.833368 57.47085

|

nut\_bar\_a#brazil |

1 1 | .05 .0664371 -2.25 0.024 .0036978 .6760788

|

\_cons | .06 .025219 -6.69 0.000 .0263255 .1367496

----------------------------------------------------------------------------------

Note: \_cons estimates baseline odds.

estimates store model2

lrtest model1 model2

Likelihood-ratio test LR chi2(1) = 4.76

(Assumption: model1 nested in model2) Prob > chi2 = 0.0291

The likelihood ratio test tells us that the model with the interaction has a better fit of the data.

Is the model with the interaction a better model?

Use glm commads and save the model statistic

glm case nut\_bar\_a brazil, link(logit) family(binomial) eform nolog

Generalized linear models No. of obs = 139

Optimization : ML Residual df = 136

Scale parameter = 1

Deviance = 93.20472006 (1/df) Deviance = .6853288

Pearson = 137.9440168 (1/df) Pearson = 1.014294

Variance function: V(u) = u\*(1-u) [Bernoulli]

Link function : g(u) = ln(u/(1-u)) [Logit]

AIC = .713703

Log likelihood = -46.60236003 BIC = -577.8837

------------------------------------------------------------------------------

| OIM

case | Odds Ratio Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

nut\_bar\_a | 17.16436 12.23122 3.99 0.000 4.246847 69.37269

brazil | 9.468998 5.39934 3.94 0.000 3.096979 28.95141

\_cons | .075868 .0280695 -6.97 0.000 .0367394 .1566697

------------------------------------------------------------------------------

Note: \_cons estimates baseline odds.

estat ic

Akaike's information criterion and Bayesian information criterion

-----------------------------------------------------------------------------

Model | Obs ll(null) ll(model) df AIC BIC

-------------+---------------------------------------------------------------

. | 139 . -46.60236 3 99.20472 108.0081

-----------------------------------------------------------------------------

Note: N=Obs used in calculating BIC

glm case nut\_bar\_a##brazil, link(logit) family(binomial) eform nolog

Generalized linear models No. of obs = 139

Optimization : ML Residual df = 135

Scale parameter = 1

Deviance = 88.44037218 (1/df) Deviance = .6551139

Pearson = 138.9999777 (1/df) Pearson = 1.029629

Variance function: V(u) = u\*(1-u) [Bernoulli]

Link function : g(u) = ln(u/(1-u)) [Logit]

AIC = .6938156

Log likelihood = -44.22018609 BIC = -577.7136

----------------------------------------------------------------------------------

| OIM

case | Odds Ratio Std. Err. z P>|z| [95% Conf. Interval]

-----------------+----------------------------------------------------------------

1.nut\_bar\_a | 49.99998 45.91657 4.26 0.000 8.265791 302.4512

1.brazil | 16.66666 10.52627 4.45 0.000 4.833367 57.47084

|

nut\_bar\_a#brazil |

1 1 | .05 .0664371 -2.25 0.024 .0036978 .6760789

|

\_cons | .06 .025219 -6.69 0.000 .0263255 .1367496

----------------------------------------------------------------------------------

Note: \_cons estimates baseline odds.

estat ic

Akaike's information criterion and Bayesian information criterion

-----------------------------------------------------------------------------

Model | Obs ll(null) ll(model) df AIC BIC

-------------+---------------------------------------------------------------

. | 139 . -44.22019 4 96.44037 108.1783

-----------------------------------------------------------------------------

Note: N=Obs used in calculating BIC

The model with the lowest AIC is the one that includes the interaction.

# Optional – Binomial regression: dealing with RRs

The scenario presented was a case control study. Let’s imagine for the proposes of the exercise that it was a cohort study. Logistic regression only provides odds ratios. These can always be reported and they are not wrong … however, one may want to stick to risk ratios in the multivariable analysis of a cohort study. In this case, logistic regression is not appropriate.

**Question 3. If you wanted to use risk ratios, how would you account for the effect of the different exposures?**

Proposed steps for the analysis:

* Using **salm\_v1.4.csv,** start with the simplest model with one exposure variable only - Add one variable at a time and compare models.
* Add the interaction and interpret it.

The command for binomial regression is binreg and you need to indicate the option ,rr.

Using glm commands you need to indicate the options , family(binomial) link(log) and eform to display RR.

After the glm models, use the commands estimates store and estat ic to assess the contribution of the new variable and to identify the most parsimonious model.

Brief help:

glm case nut\_bar\_a, family(binomial) link(log) eform

estimates store model1

estat ic

glm case nut\_bar\_a brazil , family(binomial) link(log) eform

estimates store model2

estat ic

lrtest model1 model2

glm case nut\_bar\_a##brazil , family(binomial) link(log) eform

estimates store model3

estat ic

lrtest model2 model3

lincom 1.nut\_bar\_a + 1.brazil + 1.nut\_bar\_a #1.brazil, rr

# Real outbreak investigations…

During the analysis, exposures which were statistically significant in the univariate analysis (p<0.001) or had a OR > 20 were considered for inclusion into the multivariable model, using a backwards, stepwise approach. Each variable was considered in term, and its removal tested with a likelihood ratio test to assess whether it improved the fit of the model. Additionally, the ORs of the remaining variables were observed, and if any changed significantly (>50%), then the variable was considered a potential confounder and retained in the model. The multivariable logistic model was considered final when no further variables could be removed.

Table 6 contains the result from the final multivariable model (please note, these model results are from the real investigation and not from the altered dataset provided in this case study; as such, there may be discrepancies between the results displayed here and your outputs from the exercise).

**Table 6: Results from the final multivariable logistic regression model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Case (n, %) | OR (95% CI) | aOR (95% CI) | p-value |
| Nut Bar Brand B | 5 (20%) | 26.7 (4.0 – 525) | 53.1 (3.19 – 1760) | 0.009 |
| Nut Bar Brand A | 11 (42%) | 40.7 (9.8 – 280) | 23.8 (2.46 – 395) | 0.012 |
| Cabbage | 10 (39%) | 8.2 (2.8 – 24.7) | 14.9 (1.81 – 188) | 0.018 |
| Brazil nuts | 14 (46%) | 13.5 (4.9 – 39.3) | 6.79 (1.11 – 49.3) | 0.043 |
| Gender-male | 11 (42%) | 0.7 (0.3 – 1.7) | 0.15 (0.01 – 1.05) | 0.084 |
| Carrots | 17 (65%) | 7.8 (3.2 – 20.6) | 4.51 (0.84 – 28.7) | 0.086 |
| Cashew nuts | 10 (39%) | 6.6 (2.5 – 17.6) | 4.83 (0.77 – 34.2) | 0.094 |

The results can be interpreted as follows: Nut Bar Brand A, Nut Bar Brand B, brazil nuts and cabbage were identified as independent risk factors for infection by multivariable logistic regression. No significant interactions were identified in this original dataset (as such, none were included in the final model). Despite multiple nuts being statistically significantly associated with infection of the outbreak strain during univariate analysis, only Brazil nuts and two nut bars (and cabbage) remained significant following multivariable analysis – showing the power of multivariable analysis in gaining further insight into the vehicle(s) of infection!

These results were immediately reported back to the incident management team. On the 6th August, PHE visited the premise of Producer A (which produced both implicated nut bars) focusing on gathering information on the batches of nuts used in nut bars over the period of interest. Samples of brazil nuts were obtained for testing. In total, 119 x25g samples of brazil nuts were tested from 19 consignments of 6 batches. *Salmonella* was isolated from samples belonging to two batches (5 positive samples from batch one and 1 positive sample from batch two). Whole genome sequencing was conducted on all isolates and samples from batch one were found to genetically identical to the outbreak strain! An additional *Salmonella* serovar was detected in an isolate from batch two which was found to genetically identical to a strain of *Salmonella* Anatum that had been isolated from a human isolate in June 2020.

# Conclusion

As of 20 October 2020, there had been 105 cases identified as part of this outbreak, from England, Wales, Northern Ireland and Scotland. Additionally, 19 cases had been identified outside of the UK in France, Luxembourg, Netherlands and Canada. The UK’s case-control study, combined with case interviews in the UK, France and Luxembourg indicated Brazil nuts and nut bars as likely vehicles of infection.

One batch of Brazil nuts from Bolivia, sampled at the Producer A’s premise, tested positive for *S.* Typhimurium matching the outbreak strain. These Brazil nuts were used for the production of certain batches of Brand A and Brand B nut products (nut bars and other products such as cereals containing nuts). The Brazil nuts were also used in other nut products manufacturers by different companies in the UK and one in Austria. It is likely that the vehicle of infection was contaminated Brazil nuts, however the true root cause of how the contamination of the brazil nuts occurred was not possible with the data that was available.

Extensive recalls and withdrawals of nut products were implemented from August 2020. As of October 2020, the outbreak appeared to be controlled, with the low likelihood of further cases occurring due to the long shelf life of many of the nut products.

# Further information and reading

A rapid outbreak assessment was published by ECDC in October 2020, which can be found at: <https://www.ecdc.europa.eu/en/publications-data/salmonella-typhimurium-multi-country-outbreak-brazil-nuts>

Full details on the outbreak investigation will be made available shortly in a peer-review publication.

## About Public Health England

Public Health England exists to protect and improve the nation’s health and wellbeing, and reduce health inequalities. We do this through world-class science, knowledge and intelligence, advocacy, partnerships and the delivery of specialist public health services. We are an executive agency of the Department of Health, and are a distinct delivery organisation with operational autonomy to advise and support government, local authorities and the NHS in a professionally independent manner.

Public Health England  
Wellington House

133-155 Waterloo Road  
London SE1 8UG  
Tel: 020 7654 8000  
[www.gov.uk/phe](http://www.gov.uk/phe)   
Twitter: [@PHE\_uk](https://twitter.com/PHE_uk)

Facebook: [www.facebook.com/PublicHealthEngland](http://www.facebook.com/PublicHealthEngland)

© Crown copyright 2020  
You may re-use this information (excluding logos) free of charge in any format or medium, under the terms of the Open Government Licence v3.0. To view this licence, visit [OGL](https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/) or email [psi@nationalarchives.gsi.gov.uk](mailto:psi@nationalarchives.gsi.gov.uk). Where we have identified any third party copyright information you will need to obtain permission from the copyright holders concerned.

