Institute of Psychiatry, Psychology and Neuroscience

08/2020

Zahra Abdulla

Department: Biostatistics and Health Informatics

Module Title: Introduction to Statistics

Session Title: Multiple Independent Variables

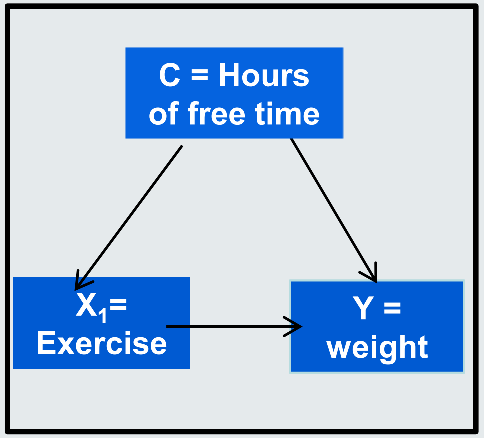
Topic title: Binary Logistic Regression

After working through this session you should be able to:

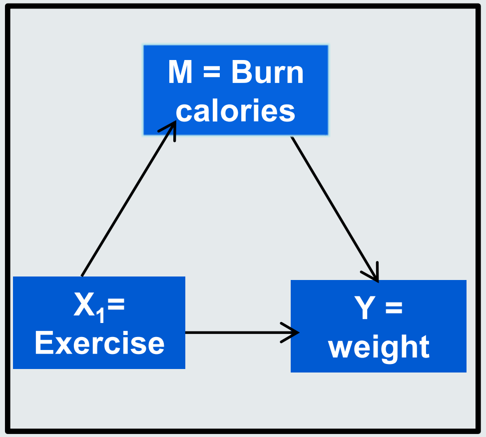
- Interpret a binary logistic regression model with multiple independent variables.
- Run a binary logistic regression analysis with multiple predictors in a software package.

Dealing with third variables

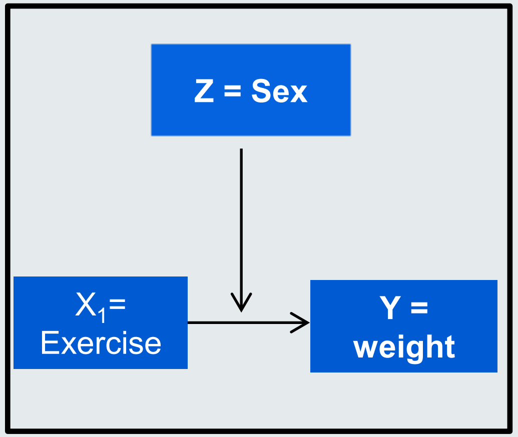
Both confounder, mediator and moderator, are third variables that explain a part (or most) of the association between an independent and dependent variable.



A **confounder (C)** has a common effect on the independent and dependent variables. A confounder **is extrinsic to the causal** pathway.



A **mediator (M)** is caused by the independent variable which in turn causes the dependent variable. A mediator **is in the causal** pathway



A **moderator (Z)** modifies the effect of an independent variable on a dependent variable. The association varies depending on the values/levels of Z

The logistic transformation: Multiple predictors

Just as we would be able to develop a Multiple Linear Regression model we are able to build a Binary logistic regression with multiple independent variables. This includes investigating

- Confounding Variables
- Mediators
- Effect Modifiers or Interaction Terms.

Independent or Predictor variables can be numerical or categorical

$$\boxed{ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i}$$

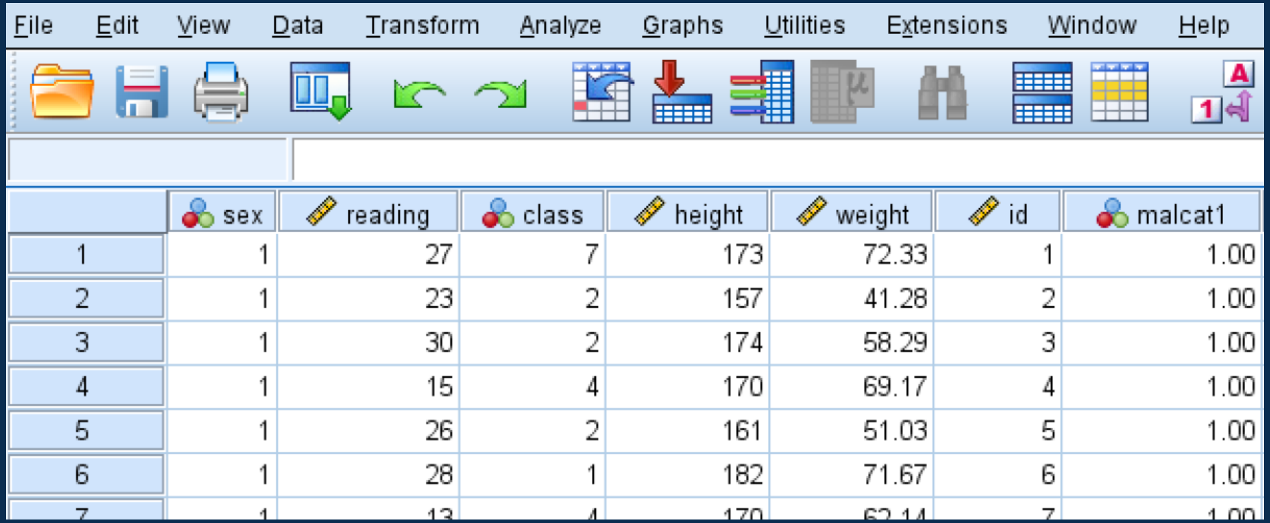
This is just the *odds.*

The (adjusted) odds ratio is the estimated change in odds for a unit change in x1 (holding x2 x3,...xi constant)

For variables coded as binary or dummy variables 'one unit' usually means a comparison between the group of interest and a reference group.

SPSS Slide

Download the data that we are going to use during the lecture. The dataset is the **lecture_10_data.sav**.



The dataset contains data from 42 babies, with respect to their

Specific body measurements at birth : headcircumf, length, weight (lbs)

Gestation: Gestational age at birth

Information about the baby's mother: smoker, motherage, mnocig, mheight, mppwgt

Information about the baby's father: fage, fedyrs, fnocig, fheight

lowbwt: Low birthweight Baby 0 = No, 1 = Yes

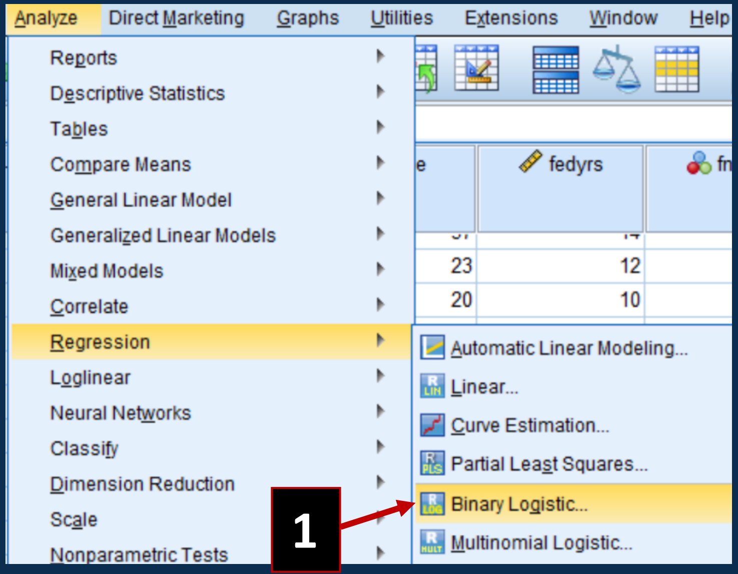
Mage35: 0=under 35, 1=Over 35

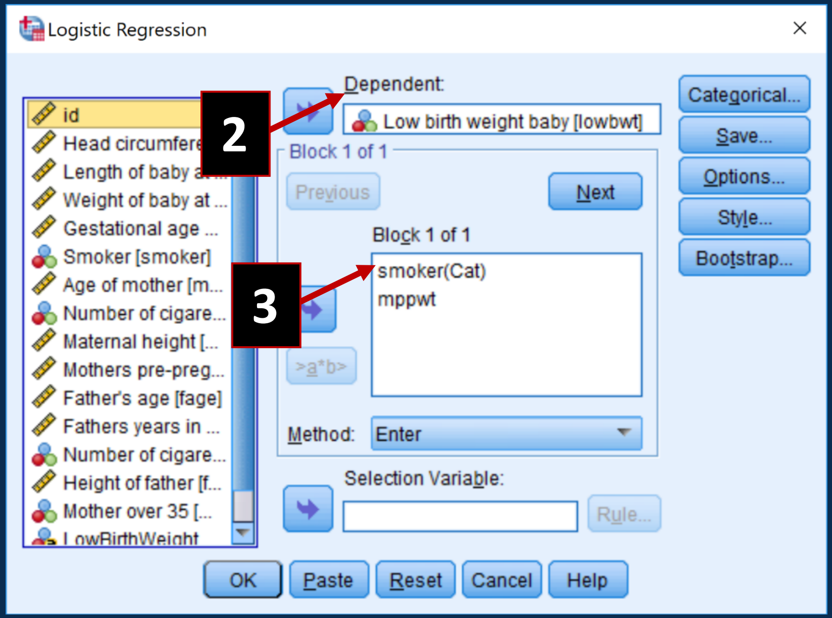
SPSS slide: ‘how to’

Is there an association between having a baby of low birth weight with mothers who smoked through pregnancy adjusting for mother's weight pre-pregnancy?

Step 1: Use the appropriate test, here: ‘Binary Logistic Regression’.

Analyse -> Regression> Binary Logistic

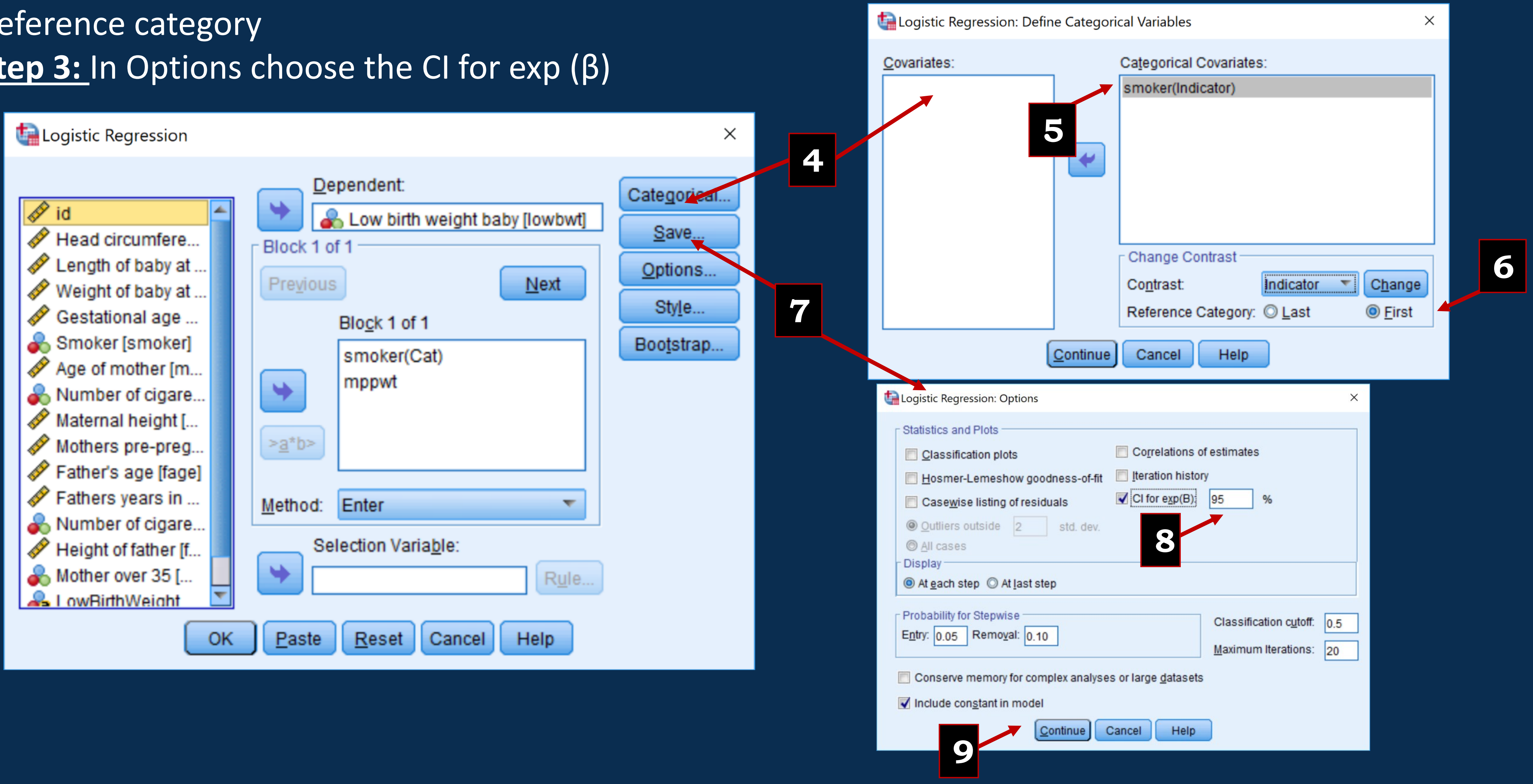




SPSS slide: ‘how to’

Step 2: Define any categorical variables and choose the Reference category

Step 3: In Options choose the CI for exp (β)



Output and Interpretation

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	8.573	2	.014
	Block	8.573	2	.014
	Model	8.573	2	.014

A p-value (sig) of less than 0.05 for block means that the final model is a significant improvement to the constant only model. **(chi-square=8.573, df=2, p=.014)**

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	48.791[a]	.185	.248

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Nagelkerke R^2 = 24.8% of the variation in lowbwt can be explained by the final model.

Classification Table[a]

Observed			Predicted: Low birth weight baby No	Predicted: Low birth weight baby Yes	Predicted: Percentage Correct
Step 1	Low birth weight baby	No	19	5	79.2
		Yes	7	11	61.1
	Overall Percentage				71.4

a. The cut value is .500

The correct classification rate has increased by **14.3% to 71.4%**

Output and Interpretation

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Smoker(1)	1.575	.709	4.936	1	.026	4.831	1.204	19.386
	Mothers pre-pregnancy weight (lbs)	-.040	.023	3.130	1	.077	.961	.919	1.004
	Constant	3.898	2.840	1.884	1	.170	49.306		

a. Variable(s) entered on step 1: Smoker, Mothers pre-pregnancy weight (lbs).

Regression Equation

$$\ln\frac{p}{1-p} = 3.898 + 1.575 smoker + \text{-}0.040 mppwt$$

Odds ratio for the effect of mothers who smoked during pregnancy on low birth weight **Exp(β) = 4.831** once adjusted for mothers pre-pregnancy wgt (lbs). Mothers who smoke during pregnancy have **a 4.831 times larger** odds of having a baby born with low birth weight compared to a mother who did not smoke during pregnancy adjusting for mother's pre-pregnancy weight. This was a significant association **95%CI 1.204 to 19.386, p=0.026.**

One lbs increase in mothers pre-pregnancy weight would lead to **a 4% reduction (exp(β) = 0.961)** in the odds of having a baby of low birth weight, if the mother is a non-smoker. **This is not a significant association 95% CI (0.919 to 1.004), p=0.077**

Reference Categories and Dummy Variables

- Categorical Independent **dichotomous** variables:
  - E.g. Gender defined at birth
  - One category is treated as a baseline, or reference category.
  - Reference Category is arbitrarily coded 0, comparison group coded 1

- Categorical independent variables with **more than two levels** need to be recoded into **dummy** variables
  - A **"dummy variable"** is a numerical variable used in regression analysis to represent subgroups of the sample in your study.
  - E.g. Variable X has three levels, create two new variables, each comparing one level to the baseline or reference category
  - Coding represents a contrast between categories.

Building Models

Which predictor variables should I include?

- Literature
- Researcher theory

- Iterative Multivariable Logistic Regression
  - Often have too many variables to legitimately include in the logistic regression model.
    - At least 50 times as many subjects as predictors

  - Used to find a good subset of variables
    - A subset that includes only **statistically** significant predictors and that results in good negative and positive predictive values (more about this in the next section).

- Forward, backward Stepwise regression

Model Building Strategies

The log likelihood (LL), the deviance (-2LL), or the likelihood ratio(LR) give an overall goodness of fit measurement for the model.

Forward Selection

- Variables are tested one at a time.
- First variable added has the smallest LR (and is statistically significant).
- Other variables added if their LR is also significant when adjusted for other variables in the model.
- Model Building stops.
  - All variables have been entered.
  - LR is non-significant for all variables not entered.

Backward Selection

- Start with all the predictors (significant and not significant).
- Variables are tested one at a time.
- First variable removed has a LR with the largest probability that is greater than alpha.
- Continue until only statistically significant variables remain.

Stepwise Selection

- Combination of forward and backward.
- Each variable is tested for entry to the model.
- When a predictor is entered, other variables are tested for removal.
- Continue until no more variables can be entered or removed.

Knowledge Check

Q1. The researcher was also interested to see if the length of gestation had a impact on low birth weight of babies alongside other factors already tested. Interpret these results.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	9.078	3	.028
	Block	9.078	3	.028
	Model	9.078	3	.028

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	48.286[a]	.194	.261

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Smoker(1)	1.557	.715	4.746	1	.029	4.746	1.169	19.271
	Mothers pre-pregnancy weight (lbs)	-.037	.023	2.496	1	.114	.964	.921	1.009
	Gestational age at birth (weeks)	-.100	.141	.497	1	.481	.905	.686	1.194
	Constant	7.326	5.701	1.651	1	.199	1519.126		

a. Variable(s) entered on step 1: Smoker, Mothers pre-pregnancy weight (lbs), Gestational age at birth (weeks).

Knowledge Check Solutions

Q1. The chi-square is significant (chi-square=9.078, df=3, p=0.028) so our new model is significantly better. Nagelkerke's R^2 suggests that the model explains roughly 26.1% of the variation in the outcome.

For every unit increase in the length of gestation, the odds of a mother having a lowbwt baby is decreased by 9.5%, 95% CI of the odds (0.686, 1.194) adjusting for Mothers smoking status and and mothers pre-pregnancy weight, this result was statistically non significant (Wald = 0.497, df=1, p=0.481)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	9.078	3	.028
	Block	9.078	3	.028
	Model	9.078	3	.028

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	48.286[a]	.194	.261

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Smoker(1)	1.557	.715	4.746	1	.029	4.746	1.169	19.271
	Mothers pre-pregnancy weight (lbs)	-.037	.023	2.496	1	.114	.964	.921	1.009
	Gestational age at birth (weeks)	-.100	.141	.497	1	.481	.905	.686	1.194
	Constant	7.326	5.701	1.651	1	.199	1519.126		

a. Variable(s) entered on step 1: Smoker, Mothers pre-pregnancy weight (lbs), Gestational age at birth (weeks).

References

Field, Andy. Discovering statistics using IBM SPSS statistics. Sage, 2013. (Chapter 19)

Agresti, Alan. Categorical data analysis. John Wiley & Sons, 2014.

Thank you

Contact details/for more information:

Zahra Abdulla

Zahra.abdulla@kcl.ac.uk

Department of Biostatistics and Health Informatics (BHI)

IoPPN