Zahra Abdulla

Department: Biostatistics and Health Informatics

Institute of Psychiatry, Psychology and Neuroscience

08/2020

Module Title: Introduction to Statistics

Session Title: Prediction, Goodness of Fit and Classification

Topic title: Binary Logistic Regression

After working through this session, you should be able to:

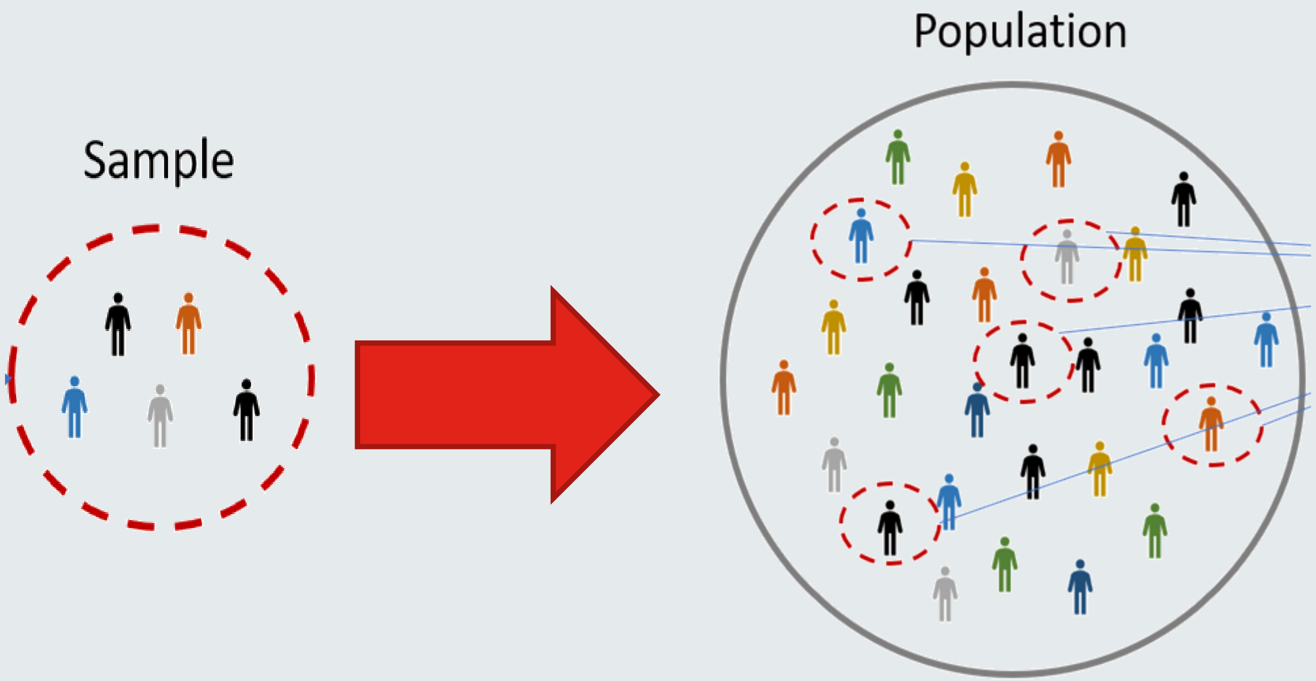
- Make predictions and describe these as probabilities
- Assess the goodness of fit of the model.

Prediction

- A logistic regression model can be used to make predictions
- The prediction is the value of the linear predictor
- We need to obtain the odds of the person experiencing an event - exponentiate the linear predictor.
- To get the probability you rearrange the odds equation.

Why is prediction important?

- Because we're modelling!
- We want to make predictions about what would happen in the general population at a given point



The logistic transformation: a recap

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

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)

$$L = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i$$

This is called the Linear Predictor

$$exp(L) \qquad = e^L$$

This is the Odds of an event

$$\hat{\pi} = \frac{odds}{1+odds} \qquad \hat{\pi} = \frac{exp(L)}{1+exp(L)} = \frac{1}{1+exp(-L)}$$

This is the Estimated Probability of an event

Estimating the probability of an event

What is the probability of a person starting smoking, if when they were born cigarettes cost £2?

We know

$$\log\left(\frac{\pi}{1-\pi}\right) = L, \quad \text{where } L = 3.69 - 0.07x$$

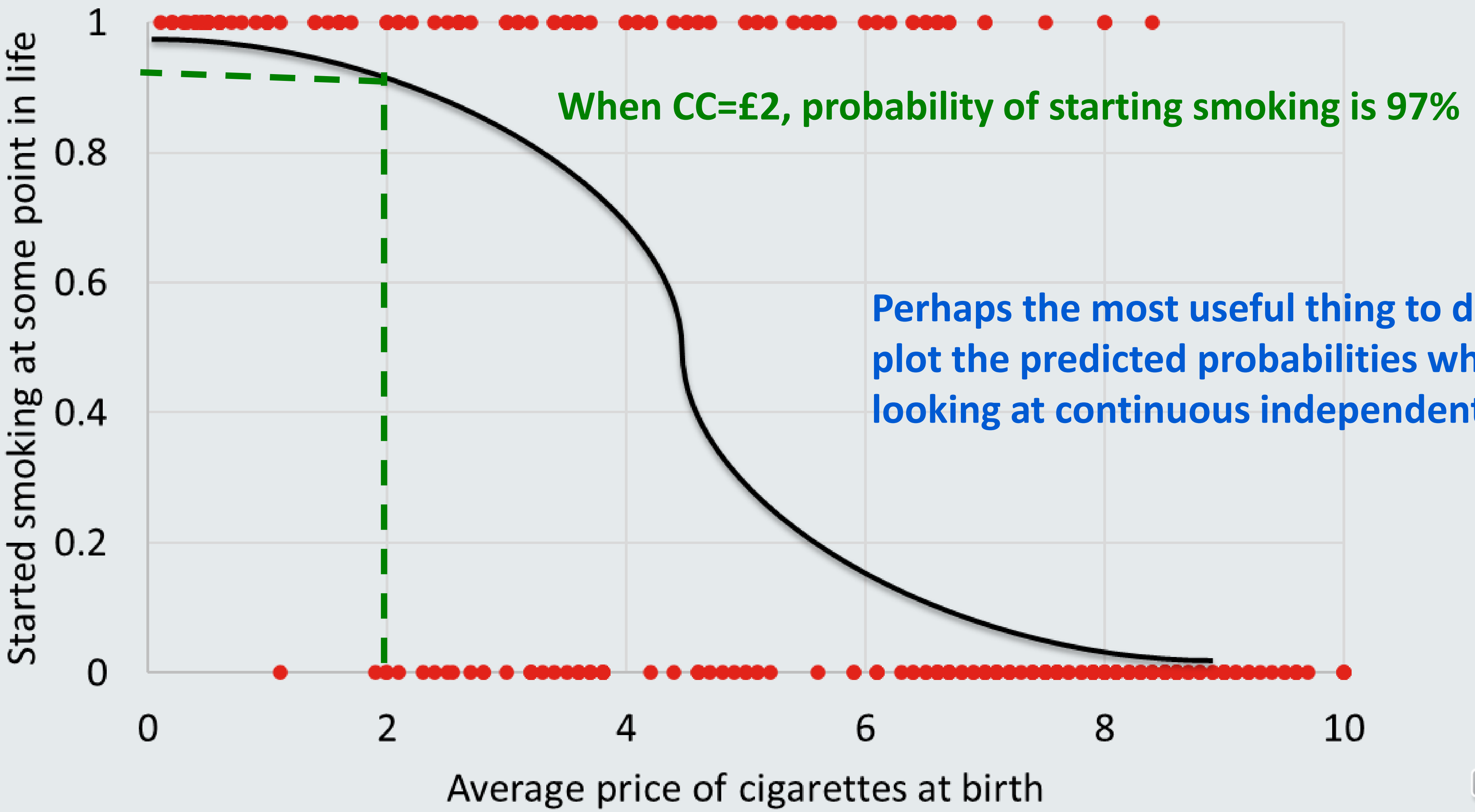
To calculate the probability of starting smoking, as per the conditions above

$$\hat{\pi} = \frac{exp(L)}{1+exp(L)}$$

$$\hat{\pi} = \frac{e^{3.69-0.07x}}{1+e^{3.69-0.07x}}$$

$$\hat{\pi} = \frac{e^{3.69-0.07\times2}}{1+e^{3.69-0.07\times2}} = 0.97$$

Thinking about prediction



Perhaps the most useful thing to do is to plot the predicted probabilities when looking at continuous independent variable

Estimating probabilities

What is the probability of a mother whose pre-pregnancy weight is 110 LLbs and a smoker of having a baby of low birth weight?

The **Linear Predictor (L)** is given by

$$L = 3.898 + 1.575 \times Smoker - 0.040 \times \text{Mppwgt}$$

$$L = 3.898 + (1.575 \times 1) - (0.040 \times 110)$$

$$L = 1.073$$

The **Probability (P)** is given by

$$P = \frac{\exp(L)}{1 + \exp(L)}$$

$$P = \frac{\exp(1.073)}{1 + \exp(1.073)}$$

$$P = \frac{2.924}{3.924}$$

$$P = 0.745$$

Interpretation

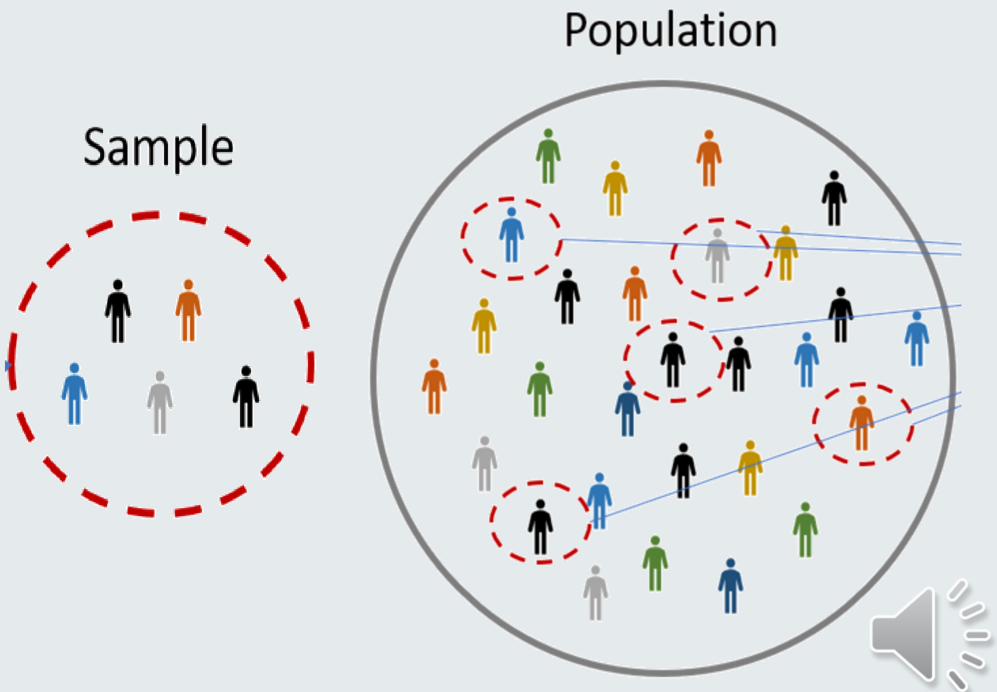
The probability of a baby born with a low birth weight is 74.5%

Goodness of fit

- Goodness-of-Fit tests help determine if observed data aligns with what is expected in the actual population.

- More specifically, it is used to test if sample data fits a distribution from a certain population (e.g., a population with a normal distribution)

- Remember, we're still modelling…



Goodness of fit

Here we will discuss two ways of assessing goodness of fit:

1. Classification analysis
2. Hosmer and Lemeshow test

Classification Analysis

One way of assessing goodness of fit is to use a **classification table.**

This allows us to evaluate **predictive accuracy** of the logistic regression model.

Classification tables are useful because they provide information that allow us to consider goodness of fit in different ways e.g., specificity and sensitivity (we will come back to these).

They are built on regression models used to predict **probability** of an outcome. When we use classification tables we identify a **threshold probability,** beyond which, an outcome is expected.

> For example, if we want to identify a threshold probability, beyond which, a healthcare worker is encouraged to remove a breathing tube from an intensive care patient – we could do this based on a regression model in which we predict the probability of success, when removing a breathing tube, under different conditions.

An example with birth weight

Classification Table[a]

			Predicted		
			Low birth weight baby		Percentage Correct
	Observed		No	Yes	
Step 1	Low birth weight baby	No	15	9	62.5
		Yes	5	13	72.2
	Overall Percentage				66.7

a. The cut value is .500

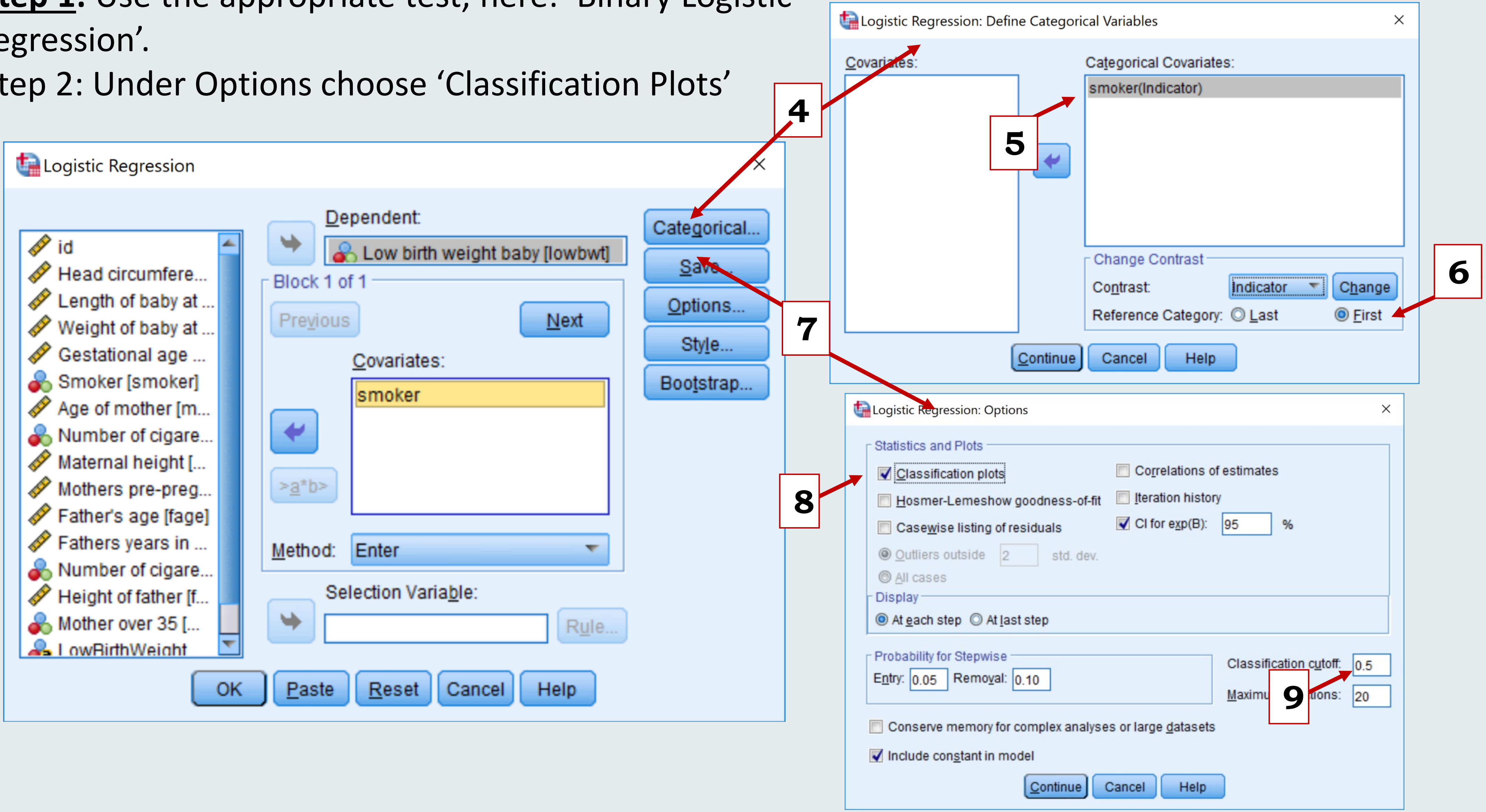
So, following our regression model, the observed values for the DV and the predicted values are cross-classified.

We can then **classify individuals** by saying that all individuals with a predicted value higher than a certain threshold probability are positive i.e. will have babies with a low birth weight.

- For every individual we use the linear predictor to estimate their probability of having a binary outcome (e.g., babies of low birth weight)
- Based on some cut-off probability we classify them as positive or negative
- Cross tabulate the predicted values versus the true values

SPSS slide: 'how to'

Step 1: Use the appropriate test, here: 'Binary Logistic regression'.
Step 2: Under Options choose 'Classification Plots'



Classification Table

Based on a cut-off of 0.5, 62.5% of those without low birth weight are correctly predicted to be negative and 72.2% of those with babies with low birth weight is correctly predicted to be positive.

Classification Table[a]

Observed			Predicted: Low birth weight baby No	Predicted: Low birth weight baby Yes	Percentage Correct
Step 1	Low birth weight baby	No	15	9	62.5
		Yes	5	13	72.2
	Overall Percentage				66.7

a. The cut value is .500

"The cut value is .500". This means that if the probability of a case being classified into the "yes" category is greater than .500, then that particular case is classified into the "yes" category. Otherwise, the case is classified as in the "no" category (as mentioned previously).

Sensitivity and specificity

In order to choose a threshold probability to turn a probability model into a classification model we usually consider the quantities **sensitivity** and **specificity**

Sensitivity, which is the percentage of cases that had the observed characteristic (e.g., "yes" for baby with low birth weight) which were correctly predicted by the model (i.e., true positives).

Specificity, which is the percentage of cases that did not have the observed characteristic (e.g., "no" for baby with low birth weight) and were also correctly predicted as not having the observed characteristic (i.e., true negatives).

In an ideal world we would like to maximise both sensitivity and specificity, but there is often a trade-off

We select an optimal threshold by considering what degree of sensitivity and specificity are acceptable

Positive and negative predicted values

We can also use the classification table to look at **positive and negative predictive values**

Remember again we're still modelling...

The positive predictive value is the percentage of correctly predicted cases "with" the observed characteristic compared to the total number of cases predicted as having the characteristic.

The negative predictive value is the percentage of correctly predicted cases "without" the observed characteristic compared to the total number of cases predicted as not having the characteristic.

How can I calculate these!?

	Outcome successful	Outcome unsuccessful
Classification successful	True positives (TP)	False positives (FP)
Classification unsuccessful	False negatives (FN)	True negatives (TN)

The formulae for the various quantities are as follows:

Sensitivity = $\frac{TP}{(TP+FN)}$

Specificity = $\frac{TN}{(FP+TN)}$

Positive Predictive Value (PPV) = $\frac{TP}{(TP+FP)}$

Negative Predictive Value (NPV) = $\frac{TN}{(FN+TN)}$

Calculation

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	15	9	62.5
		Yes	5	13	72.2
	Overall Percentage				66.7

a. The cut value is .500

Percentage Accuracy in Classification (PAC) is the overall percentage of cases correctly classified by the model = (15 +13)/ (15+9+5+13) = **66.7**

Sensitivity =13/(13+5) = **72.2 %**

Specificity =15/(9+15) = **62.5** %

Positive Predictive Value (PPV) = 13/(13+9) = **29.1%**
Negative Predictive Value (NPV) = 15/(5+15) = **75%**

Interpretation

Classification Table[a]

			Predicted		
			Low birth weight baby		Percentage Correct
Observed			No	Yes	
Step 1	Low birth weight baby	No	15	9	62.5
		Yes	5	13	72.2
	Overall Percentage				66.7

a. The cut value is .500

Overall, the model correctly classified 66.7% of the cases. Sensitivity, 72.2% is high compared to specificity, which is 62.5%. The positive predictive value, computed for low-birth-weight baby, is 29.1%; the negative predictive value, computed for no low-birth-weight baby, is 75%. The low PPV may be indicative that the model is not a good predictor of low birth weight, as only 29.1% of cases predicted to have a baby of low birthweight had babies of low birthweight.

Hosmer and Lemeshow Goodness of fit

Another way of assessing goodness of fit is (i.e., is our model any good?) is to use a **Hosmer and Lemeshow** test.

This is a **statistical test for goodness of fit** for the logistic regression model.

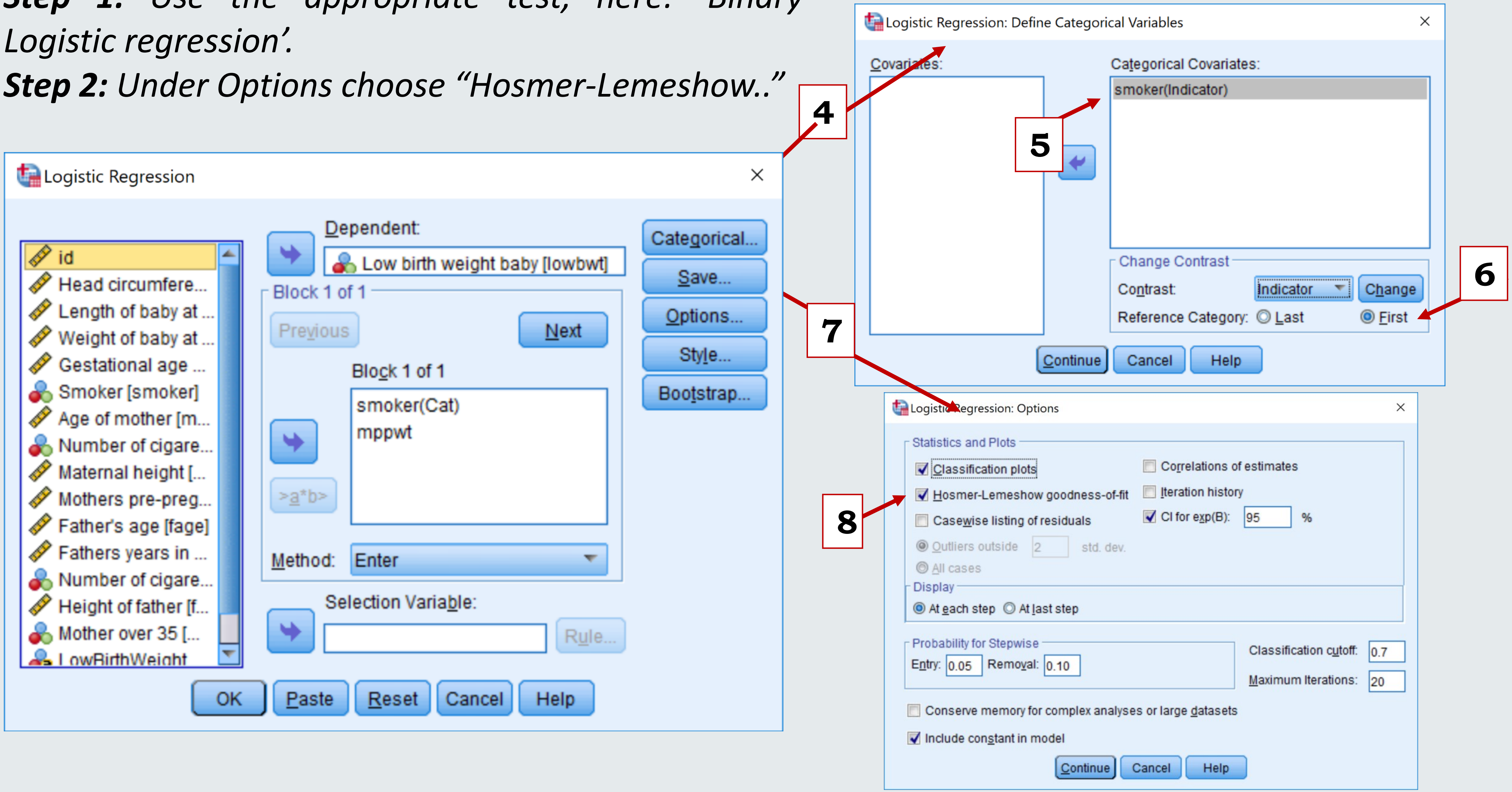
The data are divided into approximately ten groups defined by increasing order of estimated risk.

The observed and expected number of cases in each group is calculated and a **Chi-squared statistic** is produced.

You can only do this test with multiple predictors

SPSS slide: ‘how to’

Step 1: *Use the appropriate test, here: ‘Binary Logistic regression’.*
Step 2: *Under Options choose “Hosmer-Lemeshow..”*



Hosmer and Lemeshow Goodness of fit

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	7.199	8	.515

Null hypothesis: The model is consistent with the data. i.e. **a non-significant p-value indicates good fit.**

A large value of Chi-squared (with small p-value < 0.05) indicates poor fit and small Chi-squared values (with larger p-value closer to 1) indicate a good logistic regression model fit.

The Contingency Table for Hosmer and Lemeshow Test table shows the details of the test with observed and expected number of cases in each group

To conclude...

You should now be able to analyse data using binary logistic regressions

You should be able to run binary logistic regressions adjusting for covariates

You should understand goodness of fit

You should be able to make predictions based on data with dichotomous outcomes and continuous predictors

Knowledge Check

What is the probability of a mother whose pre pregnancy weight is 210lbs and a non-smoker of having a baby of low birth weight?

If we were to raise the cutoff to 0.70, how well is the model predicting babies with low birth weight?

Knowledge Check Solutions

What is the probability of an mother whose pre pregnancy weight is 210lbs and a non-smoker of having a baby of low birth weight?

$L = 3.898 + 1.575 \ \times Smoker \ - 0.040 \ \times$ Mppwgt

=3.898 + (1.575 x 0) – (0.040 x 210)

=-4.502

P = (exp(L))/(1+exp(L))

=($\boldsymbol{exp}$(-4.502))/($\mathbf{1}$+$\boldsymbol{exp}$(-4.502)

=(0.0112)/(1.0112)

=0.0112 =1.1%

Knowledge Check Solutions

If we were to raise the cutoff to 0.70, how well is the model predicting babies with low birth weight?

Classification Table[a]

			Predicted		
			Low birth weight baby		
	Observed		No	Yes	Percentage Correct
Step 1	Low birth weight baby	No	22	2	91.7
		Yes	14	4	22.2
	Overall Percentage				61.9

a. The cut value is .700

Based on a cut-off of 0.7, 91.7% of those without low birth weight are correctly predicted to be negative but only 22.2% of those with babies with low birth weight is correctly predicted to be positive.

References

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

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

Binomial Logistic Regression using SPSS Statistics, Laerd Statistics.
https://statistics.laerd.com/spss-tutorials/binomial-logistic-regression-using-spss-statistics.php:Accessed 05/01/22

Thank you

Contact details/for more information:

Zahra Abdulla

Zahra.abdulla@kcl.ac.uk

Dr. Silia Vitoratou

silia.vitoratou@kcl.ac.uk

Dr Raquel Iniesta

raquel.iniesta@kcl.ac.uk

Department of Biostatistics and Health Informatics (BHI)

IoPPN