📘 Week 10 – Binary Logistic Regression, Odds & Risk

🔷 1. What is Binary Logistic Regression?

- Used when the **outcome is binary** (e.g. yes/no, 1/0, depressed/not depressed).
- Predicts the **probability** of an event occurring.
- Expressed as a model using **log-odds (logit)**:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots$$

🔷 2. Key Terms

Term	Definition
Odds	Ratio of probability an event **will** happen to it **won't** happen
Odds Ratio (OR)	How much more likely one group is to have the outcome compared to another
Risk	Probability of the event happening (between 0–1)
Relative Risk (RR)	Risk in one group ÷ risk in another group
Logit	The natural log of the odds (what the model estimates)

Odds

In health care, the odds describes the ratio of the number of people with the event to the number without.

Odds of 10-1 at the bookmakers ...

- ... means: the probability that the outcome will not happen is 10 times the probability that it will

Or odds of developing a disorder...

- odds of disorder A = the probability that disorder A **does** happen versus the probability that disorder A **does not** happen

- Other examples:
  - an odds of 0.01 is often written as 1:100,
  - odds of 0.33 as 1:3, and
  - odds of 3 as 3:1

Risk

Risk describes the probability with which a health outcome (usually an adverse event) will occur.
Risk is commonly expressed as a decimal number between 0 and 1

A new drug reduced cancer incidence by 50%
- In absolute terms, the new drug reduced cancer incidence from 2 in 1000 to 1 in 1000.

Relative risk ...

- is the probability of an adverse outcome in an exposure group versus its likelihood in an unexposed group. This statistic indicates whether exposure corresponds to increases, decreases, or no change in the probability of the adverse outcome.

- The exposed group has 0.6 times the risk of the outcome (or 40% less risk of the outcome) compared to the unexposed group.

- More examples
  - when the risk is 0.1, about 10 people out of every 100 will have the event;
  - when the risk is 0.5, about 50 people out of every 100 will have the event.
  - In a sample of 1000 people, these numbers are 100 and 500 respectively.

How would we calculate the odds and risk?

- Could use a **contingency table**

abuse	psychosis	no psychosis	total
exposed +	127	275	**402**
exposed -	187	1,081	**1,268**
total	**314**	**1,356**	**1,670**

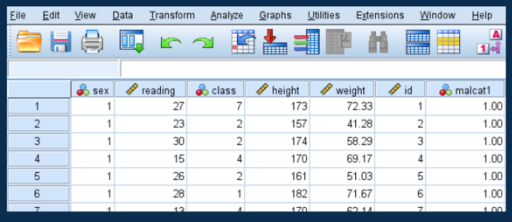
- A contingency table summarizes the frequency distribution of each of two categorical variables as well as the **association between two categorical variables**
- Each cell contains the frequency at which combination of its row and column categories occurred
- A contingency table allows us to check
  - How each of the potential explanatory variables are related to the dependent variable, one by one
  - That categories for explanatory variables are large enough (suggest at least 5 cases per category)
  - How many missing cases there are for each variable

🧪 3. When to Use

Situation	Method
Continuous outcome (e.g., weight)	Linear Regression
Binary outcome (e.g., low birth weight: Yes/No)	Logistic Regression
Compare binary outcome between groups	Chi-square or Logistic

SPSS Slide

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



	sex	reading	class	height	weight	id	malcat1
1	1	27	7	173	72.33	1	1.00
2	1	23	2	157	41.28	2	1.00
3	1	30	2	174	58.29	3	1.00
4	1	15	4	170	69.17	4	1.00
5	1	26	2	161	51.03	5	1.00
6	1	28	1	182	71.67	6	1.00

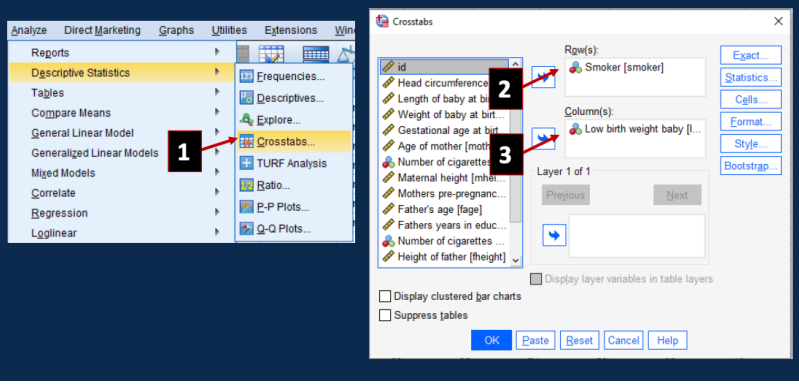
The dataset contains data from 42 babies, with respect to their
Specific body measurements at birth : headcircumf, length, weight (lbs)
Gestation: Gestational age at birth
Information about the baby's mother: smoker, motherage, mnocig, mheight, mppwgt
Information about the baby's father: fage, fedyrs, fnocig, fheight
lowbwt: Low birthweight Baby 0 = No, 1 = Yes
Mage35: 0=under 35, 1=Over 35

SPSS Slide: 'how to'

The next question is: Are the proportions of low weight babies different from mothers who smoked through pregnancy compared to those who did not smoke through pregnancy?

Step 1: Create a contingency table

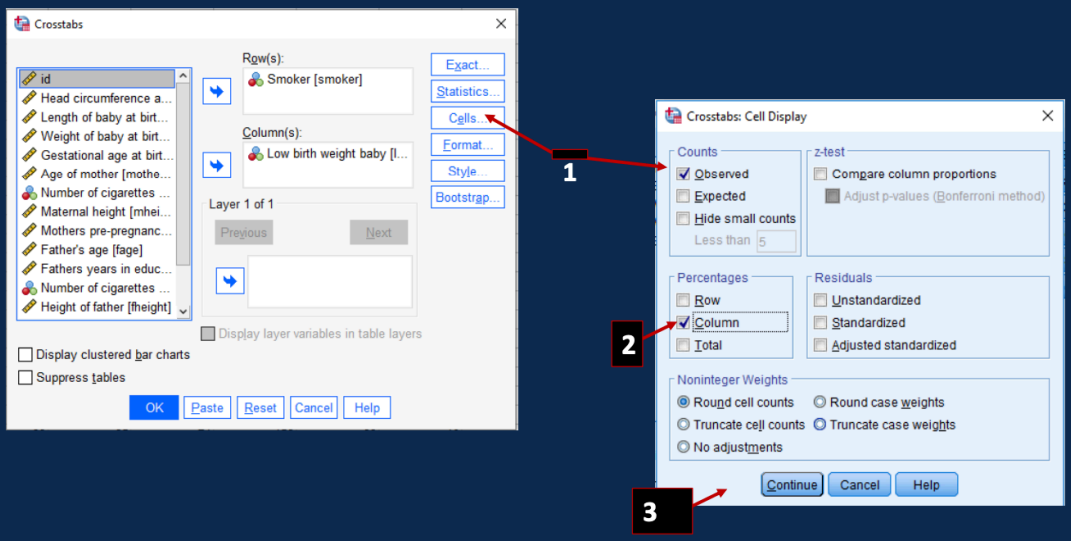
Analyse -> Descriptive Statistics-> Crosstabs



Add the variable of interest (our outcome) (Low Birth Wgt) in to the 'columns box'.
Add the second variable interest (Smoker) in the 'rows box'

SPSS Slide: 'how to'

Step 1: Choose the most appropriate 'Percentages'



Output and Interpretations

Smoker * Low birth weight baby Crosstabulation

			Low birth weight baby		
			No	Yes	Total
Smoker	Non-smoker	Count	15	5	20
		% within Smoker	75.0%	25.0%	100.0%
	Smoker	Count	9	13	22
		% within Smoker	40.9%	59.1%	100.0%
Total		Count	24	18	42
		% within Smoker	57.1%	42.9%	100.0%

Among those babies who were low birth weight, the proportion of those whose mothers smoked during pregnancy was higher than the proportion of whose mothers did not smoke during pregnancy (59.1% versus 25.0%, respectively).

Smoker * Low birth weight baby Crosstabulation

			Low birth weight baby		
			No	Yes	Total
Smoker	Non-smoker	Count	15	5	20
		% within Low birth weight baby	62.5%	27.8%	47.6%
	Smoker	Count	9	13	22
		% within Low birth weight baby	37.5%	72.2%	52.4%
Total		Count	24	18	42
		% within Low birth weight baby	100.0%	100.0%	100.0%

Among those mothers who smoked during pregnancy there was a higher proportion who had a baby of low birthweight compared to a baby of normal birthweight (72.2% versus 37.5%, respectively).

Pearson's chi-square test

When to use

To test if, according to the current data, the proportions in the population of babies being born of low-birth-weight changes based on mothers smoking status during pregnancy

Hypotheses:

H_0: there is no association between the mother's smoking status and baby's birth weight
H_a: there is an association between the mother's smoking status and baby's birth weight

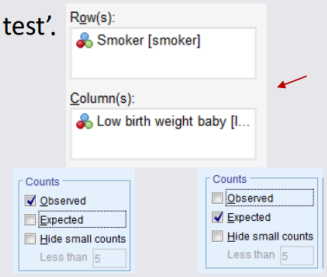
Assumptions:

- The observations are randomly and independently drawn
- The number of cells with expected frequencies less than 5, are less than 20%
- The minimum expected frequency is at the very least 1.
- The observations are not paired

Analyse -> Descriptive Statistics-> Crosstabs->Statistics->Chi-sqare

Output and Interpretations

Computations: 'Pearson's chi-square test'.



Smoker * Low birth weight baby Crosstabulation

Count

		Low birth weight baby		
		No	Yes	Total
Smoker	Non-smoker	15	5	20
	Smoker	9	13	22
Total		24	18	42

Smoker * Low birth weight baby Crosstabulation

Expected Count

		Low birth weight baby		
		No	Yes	Total
Smoker	Non-smoker	11.4	8.6	20.0
	Smoker	12.6	9.4	22.0
Total		24.0	18.0	42.0

$$\sum \frac{(O-E)^2}{E} = \frac{(15-11.4)^2}{11.4} + \frac{(5-8.6)^2}{8.6} + \frac{(9-12.6)^2}{12.6} + \frac{(13-9.4)^2}{9.4} = 5.05$$

Output and Interpretation Slide

Smoker * Low birth weight baby Crosstabulation

			Low birth weight baby	
			No	Yes	Total
Smoker	Non-smoker	Count	15	5	20
		% within Low birth weight baby	62.5%	27.8%	47.6%
	Smoker	Count	9	13	22
		% within Low birth weight baby	37.5%	72.2%	52.4%
Total		Count	24	18	42
		% within Low birth weight baby	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	4.972[a]	1	.026	.033	.027	
Continuity Correction[b]	3.677	1	.055			
Likelihood Ratio	5.104	1	.024	.033	.027	
Fisher's Exact Test				.033	.027	
Linear-by-Linear Association	4.853[c]	1	.028	.033	.027	.022
N of Valid Cases	42					

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 8.57.

b. Computed only for a 2x2 table

c. The standardized statistic is 2.203.

Among those mothers who smoked during pregnancy there was a higher proportion who had a baby of low birthweight compared to a baby of normal birthweight (72.2% versus 37.5%, respectively). This difference was statistically significant (Pearson χ^2=4.972, df=1, p=0.026).

Therefore, we conclude that mothers who smoked during pregnancy tend to have babies born with low birthweight than women who did not smoke during pregnancy. The variables 'Smoker' and 'Lowbirthwgt' are associated.

Quantifying the risk

Compare risk of smoking between babies classed as having a low birthweight and those who are not

Smoker * Low birth weight baby Crosstabulation

Count

		Low birth weight baby		
		No	Yes	Total
Smoker	Non-smoker	15	5	20
	Smoker	9	13	22
Total		24	18	42

The risk of an outcome is the number of times the outcome of interest occurs divided by the total number of possible outcomes.

For example: In the above study out of 22 mothers who smoked during pregnancy, there were 13 babies who were born with low birthweight. So, we get the risk of smoking during pregnancy and having a low birthweight baby by the following calculation:

Risk = 13 ÷ 22 = 0·59

Calculating the Risk Ratio (Relative Risk)

If we want to compare the effects of smoking during pregnancy and not smoking during pregnancy we could calculate the risk of having a baby of low birth weight for each group:

Smoker * Low birth weight baby Crosstabulation

Count

		Low birth weight baby		
		No	Yes	Total
Smoker	Non-smoker	15	5	20
	Smoker	9	13	22
Total		24	18	42

Risk of having baby of low birth weight in smokers = **13 ÷ 22 = 0·59**

Risk of having baby of low birth weight in non-smokers = **5 ÷ 20 = 0·25**

We can compare the risk for each of the groups using the risk ratio.

(Risk when smoker) ÷ (Risk when non-smoker)=
0·59 ÷ 0·25 = 2.36

So, the risk of having a low birthweight baby when the mother smoked through pregnancy is 2.36 times that of when the mother did not smoke during pregnancy.

Interpreting the Risk

Relative Risk = 1: The risk ratio equals one when the numerator and denominator are equal.

- This equivalence occurs when the probability of the event occurring in the exposure group equals the likelihood of it happening in the unexposed group.
- E.g: There is no association between mothers smoking status and a baby being born with a low birth weight.

Relative Risk > 1: The numerator is greater than the denominator in the risk ratio.

- Therefore, the event's probability is greater in the exposed group than in the unexposed group.
- E.g. If the RR = 1.4, the smoking status corresponds to a 40% greater probability of a mother having a child with low birthweight.

Relative Risk < 1: The numerator is less than the denominator in the risk ratio.

- Consequently, the probability of the event is lower for the exposed group than for the unexposed group.
- E.g. If the RR = 0.4, the smoking status corresponds to a 60% lower probability of a mother having a child with low birthweight.

Calculating the Odds

The odds in favour of a particular outcome is the number of times the outcome occurs divided by the number of times it doesn't occur.

If we want to compare the effects of smoking and Not smoking during pregnancy we could calculate the odds for each group:

Smoker * Low birth weight baby Crosstabulation

Count

		Low birth weight baby		
		No	Yes	Total
Smoker	Non-smoker	15	5	20
	Smoker	9	13	22
Total		24	18	42

Odds for having a baby of low birthweight when mother is a smoker = **13 ÷ 9 = 1.44**

Odds for having a baby of low birthweight when mother is a non-smoker = **5 ÷ 15 = 0.33**

We can compare the odds using the odds ratio. The odds ratio for having a baby of low birthweight when mother smokes during pregnancy compare to a mother who did not smoke during pregnancy

(Odds when smoker) ÷ (Odds when non-smoker) = $1.44\dot{}$ ÷ $0.3\dot{3}$ = 4.33

- So the odds of having a low birthweight baby when the mother smoked during pregnancy is about 4.36 times larger than the odds for mothers who did not smoke during pregnancy.

Risk: yes/total=

Odds: yes/no=

Interpreting the odds

OR = 1

Odds of 1 mean the outcome occurs at the same rate in both groups

- Exposure does not affect the odds of outcome
- E.g There is no difference in the odds of low birth rate between smokers and non-smokers.

OR < 1

Odds of less than 1 mean the outcome occurs less often in the first group than the second group

- Indicates that the exposure is associated with a decreased risk of developing the disease
- E.g if the odds ratio = 0.339 then the odds of a non-smoker having a low birth weight baby is a third (33.9%) of smokers.

OR >1

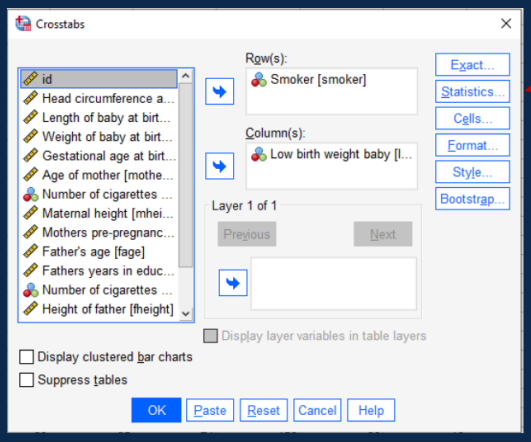
Odds of less than 1 mean the outcome occurs more often in the first group than the second group

- Indicates that the exposure is associated with an increased risk of developing the disease
- E.g. if the odds ratio = 1.5 then the odds of smokers having a low birth weight baby is 1.5 times that of the odds of non-smokers.

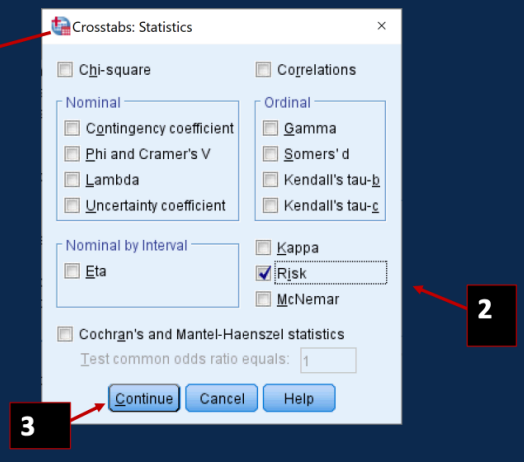
SPSS Slide: 'how to'

Calculate the risk of having a baby of low birthweight if mothers smoked during pregnancy versus if they did not.

Step 1: Create a contingency table



Step2: Click on '**Statistics**' and Choose **"Risk"**



Output and Interpretation

Risk Estimate

	Value	95% Confidence Interval Lower	Upper
Odds Ratio for Smoker (Non-smoker / Smoker)	4.333	1.156	16.248
For cohort Low birth weight baby = No	1.833	1.045	3.217
For cohort Low birth weight baby = Yes	.423	.184	.975
N of Valid Cases	42		

Odds Ratio

Risk Ratio (Relative Risk)

The risk ratio given here is 0.423, it is the risk of having a low birthweight baby when the mother did not smoke through pregnancy. To understand the risk of a mother who smoked, we take the reciprocal 1/0.423 = 2.36. So, the risk of a mother who smokes to have a low-birth-weight baby is 2.36 times that of a non-smoker.

Risk ratios and Odds ratios

Case-control studies

- The risk ratio cannot be used in a case-control study, the odds ratio can be used. Risk ratios cannot be used in studies where selection of subjects is based on the outcome.

Rare outcomes

- When an outcome is rare the risk ratio and odds ratio will be approximately equal.

Clinical practitioners often prefer the risk ratio due to its more direct interpretation. Statisticians tend to prefer the odds ratio as it applies to a wide range of study designs, allowing comparison between different studies and meta-analysis based on many studies. It also forms the basis of **logistic regression.**

General linear model (linear regression)

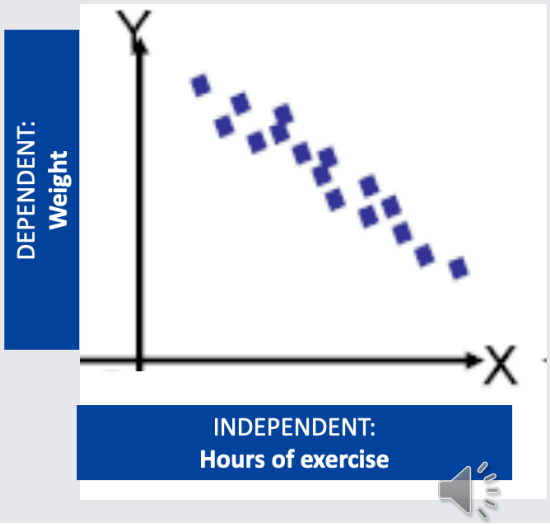
16 people were observed to see if the weight of a person is related to exercise:

Hypothesis 'The greater the number of hours of exercise, the lower the weight'.

The plot of data points (x,y) with x = **hours of exercise** and y = **weight** of a person where the data is continuous is called a **scatterplot.**

Correlation Coefficient (Pearson) ***r*=-0.85**

There is a strong, negative, linear association between hours of exercise and weight loss (*r*=-0.85)



General linear model (linear regression)

Interpretation

The relationship is expressed as a linear equation

$$y = \beta_0 + \beta_1 x$$

where β_0 is the y intercept = **70**
where β_1 is the slope of the line = **-5**

- β_0 = 70, When hours of exercise = 0, weight is 70kg.
- β_1 = -5, Each additional hour of exercise decreases weight by 5kg.

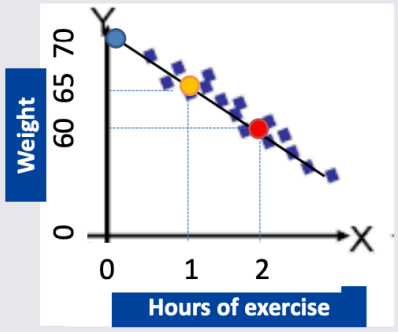
Linear regression model:

- To measure to what extent there is a linear relationship between two continuous variables, where the outcome variable (dependent variable) is continuous.
- A rule that predicts the dependent variable given the independent variable

β_0=70; β_1=-5;

$y = 70 - 5x$

X	Y
0	70
1	65
2	60



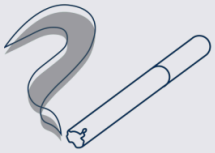
Some Scenarios

- Are clients with high scores on a personality test more likely to respond to psychotherapy than are clients with low scores?
- Do children have a better chance of surviving a severe illness than do adults?
- Do income, socio economic status and education distinguish persons who are depressed from persons who are not depressed.

Can we use linear regression to answer these questions?

Generalised linear model (logistic regression)

Not all data are suitable for general linear models (linear regression)



What happens when we have other types of data e.g., binary data?

An example: Imagine we wanted to predict whether a person starts smoking or not based on the price of cigarettes at the time they were born

→ Here, we have a **binary dependent variable: starts smoking (yes, no)**
→ And a **numerical continuous independent variable: price of cigarettes**
- *As the independent variable is continuous, we **can't** use cross-tabs.*

We want to know the **probability** that any given person will start smoking or not, at each price

And hence the **proportion** of people that will start smoking at each price on average

Examples of binary Outcomes

Outcomes in Psychology and Psychiatry are often binary:
- Illness (Schizophrenia, Autism,..)
- Passing some threshold (Depression, Anxiety, Obese, ….)
- Recurrence of psychosis
- Hospitalization
- Survival
- Hospital discharge
- Relapse to alcohol use

Often you need to define a timeframe:
- Depressive symptoms within the last year

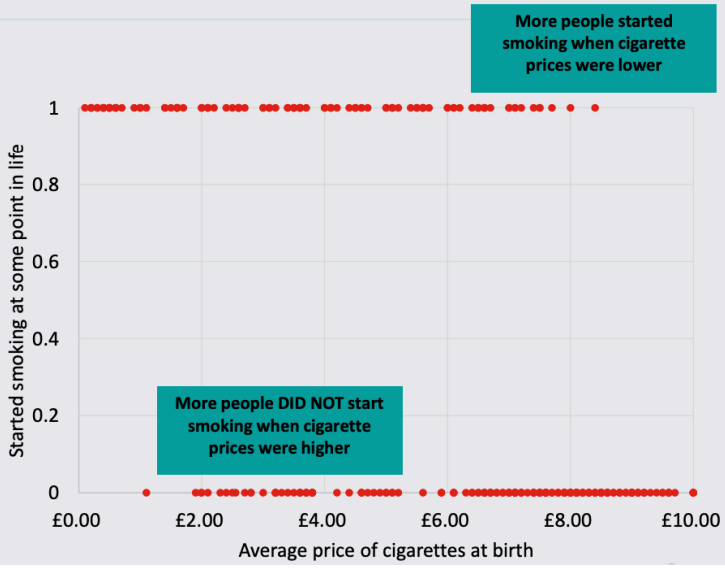
Do not dichotomize if not necessary (Loss of information)

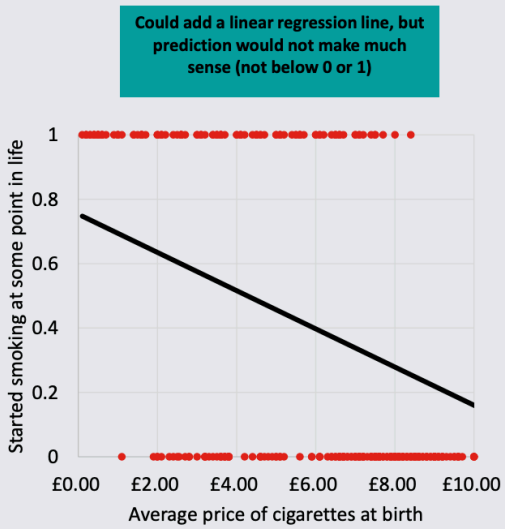
What is wrong with Simple Linear Regression?

We want to predict a probability; this can only vary between zero and 1

But our simple linear regression may predict values that are below zero or above 1

This is a scatter plot of 400 people who answered a survey about their smoking behaviours, plotted against the average price of cigarettes at the time they were born



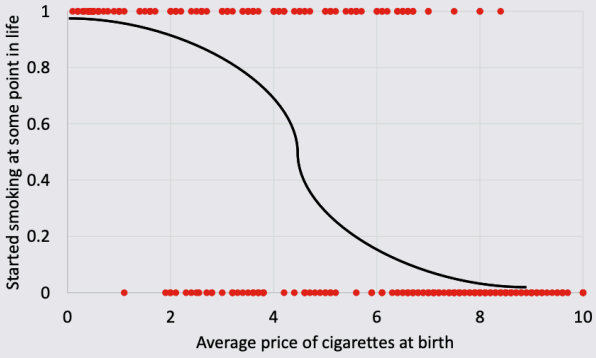


For linear regression we assumed that the population distribution was **normally distributed** around the mean, for each value of the X variable.

That's not going to be the case if we've got a **binary response**. The distribution around the mean is going to be quite different.

Non-linear relationship

- We assume there is a non-linear S-shaped (or sigmoid) relationship between cigarette price and starting smoking.
- Here is a more realistic representation of the relationship between the probability of cigarette price and starting smoking:



- Lower the price = more likely to start smoking
- There is never a probability of 0 or 100%

The Link Function

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

Linear regression describes the **linear** relationship between the outcome y and the predictor variable(s) x_i in a general linear model, where ϵ describes the random component (error) which is assumed to be normal distributed.

Generalised Linear Models (GLM) extend the ordinary regression model and allow the response variable (dependent, outcome) y to have an error distribution other than the normal distribution.

In a **logistic** regression, we relate x_i and the mean outcome at x_i (μ) by way of a **function**, known as **link function g(μ)**:

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

a link function will connect a model's outcome to its predictors in a linear way, so that we can model a linear relationship between the left- and right-hand side of the equation.

Logistic Regression: The Link Function

The link function in **logistic regression** is called the **Logit** link (used when data are binary):

$$\mathbf{g(\mu)} = ln\left(\frac{\pi}{1-\pi}\right)$$

When we have a binary outcome, the errors will follow a binomial distribution, where the mean of outcome y is represented by the probability (proportion) π of an event bounded by 0 and 1, as a function of the predictor variables. The logit link function will transform the data into a logit scale so that we can model a linear relationship between the left and right hand side of the equation.

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

Natural log.

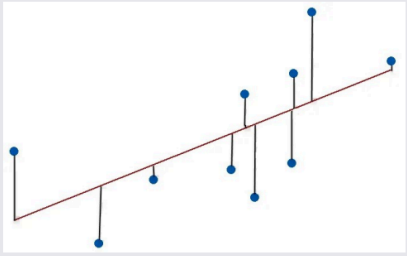
This is just the odds, the probability that expected outcome **does** happen divided by the probability that expected outcome **does not** happen

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.

Fitting this model (1)

- With SLR we tried to **minimize the squares of the residuals**, to get the best fitting line.



- This doesn't really make sense here (remember the errors won't be normally distributed as there's only two values).
- We use something called **maximum likelihood** to estimate the coefficients of the linear predictors

Fitting this model (2)

- **Maximum likelihood** is an **iterative process** that estimates the best fitted equation.
- The coefficients maximise the probability (likelihood) of obtaining the actual group membership for cases in the sample (e.g. depressed)
- Coefficients are known as **Maximum Likelihood parameters**

An example…

Variable	Coefficient value	Standard error	*p*-value
Cigarette price	-0.07	0.01	0.00
Intercept	3.69	0.72	0.00

- In OLS linear regression, a change of one unit on the X variable meant that the Y variable would increase by the coefficient for X.
- That's not what the coefficient associated with X in our logistic regression means.
  - It's clear that cigarette price has a negative (and statistically significant) effect on starting smoking – i.e., as cigarette price increases the probability of starting smoking decreases.
  - But what does the -0.07 actually mean?

In logistic regression an increase in *X* of 1 unit will decrease our log (odds) by 0.07.

The anti-log (e^x) of -0.07 gives us the odds ratio for price

Binary Logistic Regression

When to use

To test, according to the current data, if in the population there is an association between babies being born of low birth weight and mothers' smoking status during pregnancy

Hypotheses:

H_0: there is no association between the mother's smoking status and baby's birth weight
H_a: there is an association between the mother's smoking status and baby's birth weight

Assumptions:

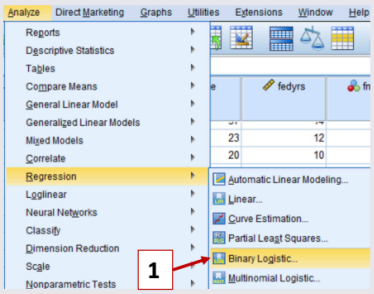
- Binary dependent variable which has a **Bernoulli (binomial)** Distribution
- Continuous variables have a linear effect on the log-odds scale (Is linearly related to the predictor variables only after transforming into the **logit** scale)
- Observations are independent
- Adequate Sample size
- Absence of multicollinearity
- No outliers

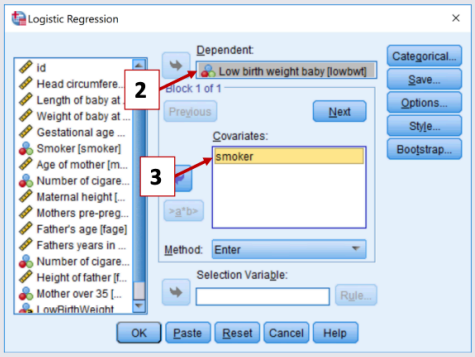
SPSS slide: 'how to'

Is there an association between having a baby of low birth weight with mothers who smoked through pregnancy? Use the Lecture_10_data.sav

Step 1: Use the appropriate test, here: '**Binary Logistic Regression**'.

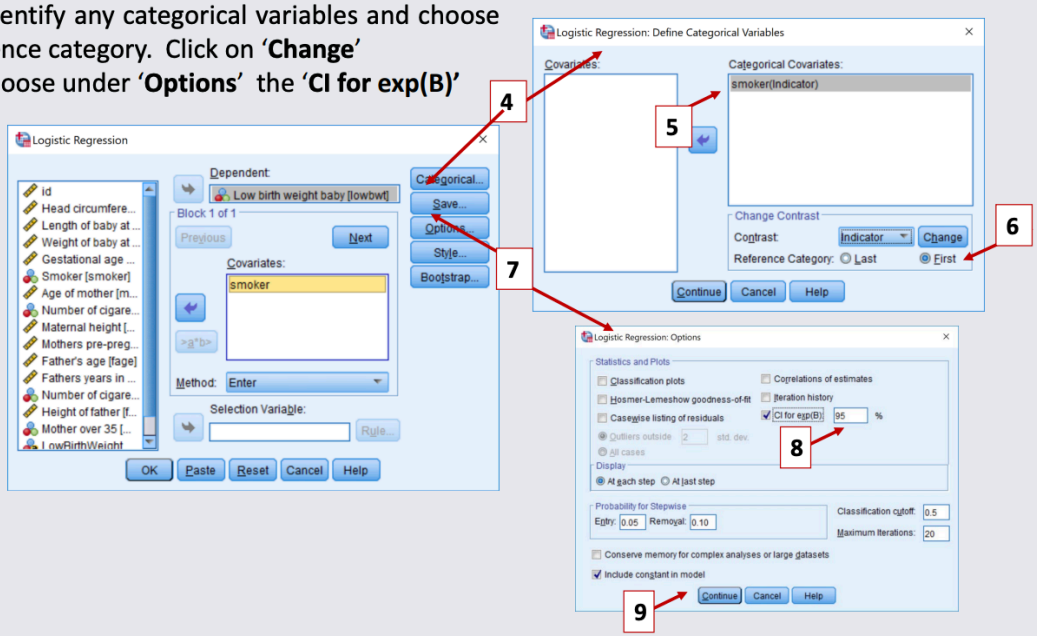
Analyse -> Regression> Binary Logistic





SPSS slide: 'how to'

Step 2: Identify any categorical variables and choose the reference category. Click on '**Change**'
Step 3: choose under '**Options**' the '**CI for exp(B)**'



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)	1.466	.674	4.729	1	.030	4.333	1.156	16.248
	Constant	-1.099	.516	4.526	1	.033	.333		

a. Variable(s) entered on step 1: Smoker.

Regression Equation

$$\ln\frac{p}{1-p} = -1.099 + 1.466 smoker$$

Odds ratio for the effect of mothers who smoked during pregnancy on low-birth-weight **Exp(β) = 4.333**. There is significant evidence (p=.030) of an association between mothers smoking status during pregnancy and a baby being born at a low birth weight. Mothers who smoke during pregnancy have a **4.33 times larger** odds of having a baby born with low birth weight compared to a mother who did not smoke during pregnancy **95%CI 1.156 to 16.248, p=0.030.**

📊 4. How to Run Binary Logistic Regression in SPSS

Example: Predicting Low Birth Weight (lowbwt) from Smoking (smoker)

1. **Go to**: `Analyze > Regression > Binary Logistic`
2. Dependent = `lowbwt`
3. Independent(s) = `smoker`, `mppwgt` (optional: multiple predictors)
4. Under **Categorical**, define reference groups (e.g., 0 = non-smoker)
5. Under **Options**:
    - Tick **CI for exp(B)** (to get odds ratios)
    - Tick **Classification Plot**
    - Tick **Hosmer-Lemeshow Test** (for goodness of fit)

📈 5. Interpretation of Output

- **B**: log-odds coefficient
- **Exp(B)**: odds ratio
    - > 1 → higher odds
    - <1 → lower odds
- **p-value (Sig.)**: test of whether predictor is significant
- **95% CI for Exp(B)**: confidence range for the odds ratio

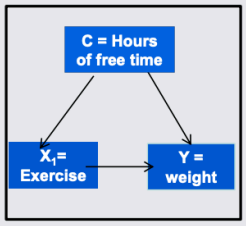
Example Output:

> $$\ln\left(\frac{p}{1 - p}\right) = 3.898 + 1.575(smoker) - 0.040(mppwgt)]$$

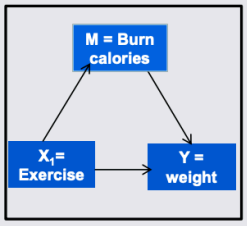
- Smoking increases odds of low birthweight (OR = 4.83, p = 0.026)
- Each 1lb increase in weight reduces odds by 4% (OR = 0.961, p = 0.077)

Dealing with third variables

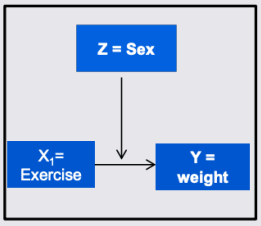
Both confounder, mediator and moderator, are third variables that explain a part (or most) of the association between an independent and dependent variable.



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

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

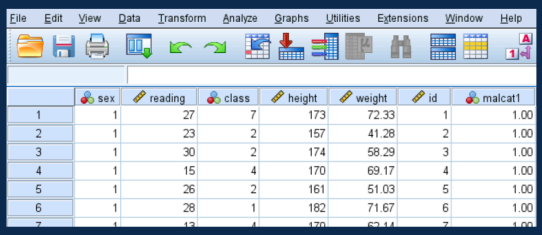
This is just the *odds*.

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**.



	sex	reading	class	height	weight	id	malcat1
1	1	27	7	173	72.33	1	1.00
2	1	23	2	157	41.28	2	1.00
3	1	30	2	174	58.29	3	1.00
4	1	15	4	170	69.17	4	1.00
5	1	26	2	161	51.03	5	1.00
6	1	28	1	182	71.67	6	1.00

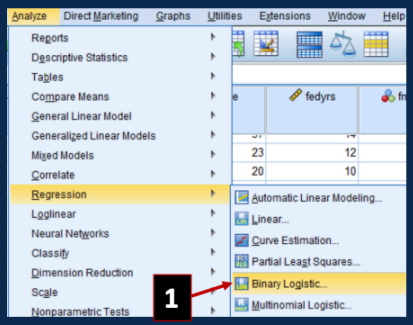
The dataset contains data from 42 babies, with respect to their
Specific body measurements at birth : headcircumf, length, weight (lbs)
Gestation: Gestational age at birth
Information about the baby's mother: smoker, motherage, mnocig, mheight, mppwgt
Information about the baby's father: fage, fedyrs, fnocig, fheight
lowbwt: Low birthweight Baby 0 = No, 1 = Yes
Mage35: 0=under 35, 1=Over 35

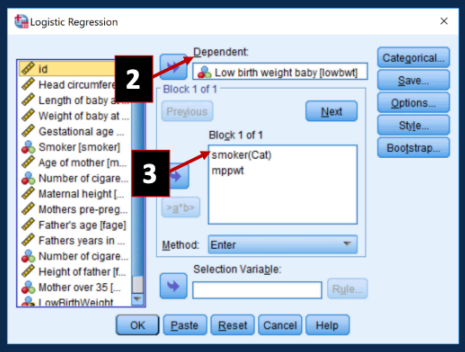
SPSS slide: 'how to'

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

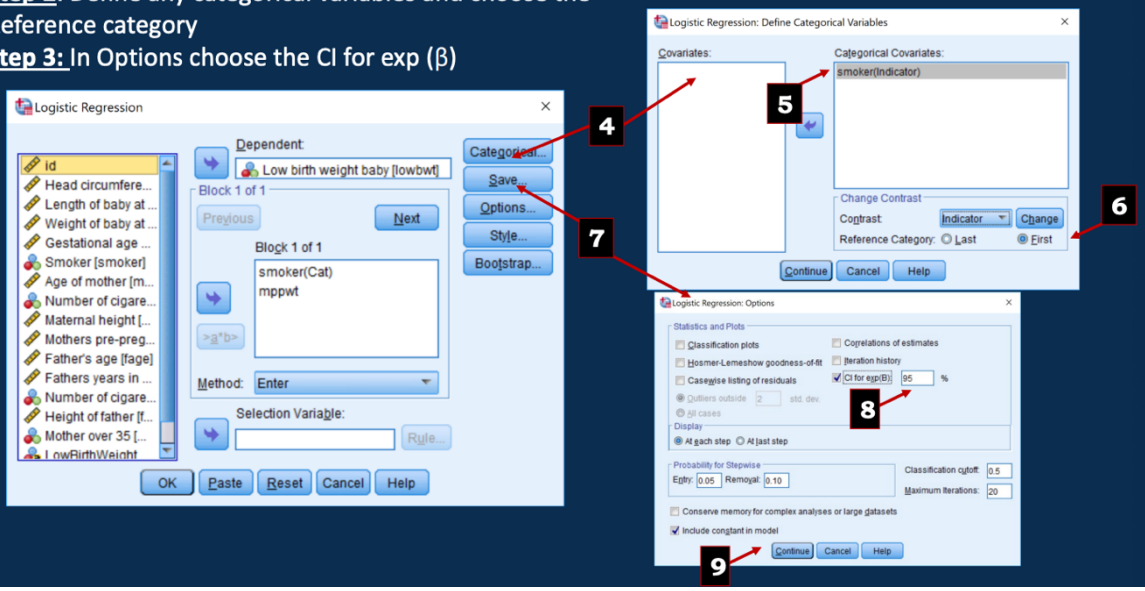




SPSS slide: 'how to'

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)**

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

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.

Classification Table[a]

			Predicted		
			Low birth weight baby		Percentage Correct
	Observed		No	Yes	
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

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

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**

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.

The logistic transformation: a recap

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

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**

$exp(L) \quad = e^L$ — **This is the Odds of an event**

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

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

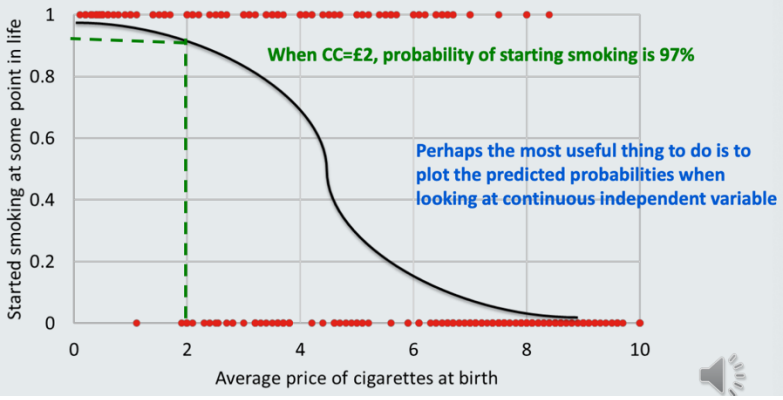
To calculate the probability of starting smoking, as per the conditions above

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

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

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

Thinking about prediction



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]

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

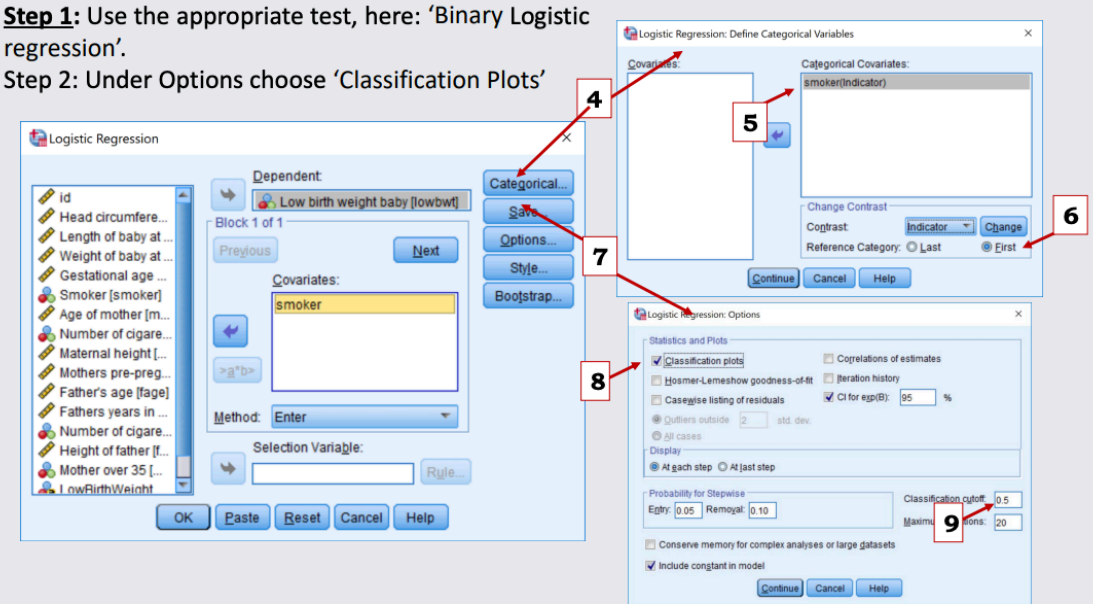
So, following our regression model, the observed values for the DV and the predicted values are cross-classified.

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]

			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

"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	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.

Outcome=observed (in the form)

Classification=predicted

🧩 1. Definitions

Term	What it means	Where to look
Outcome successful	The **actual observed outcome** is "Yes" (e.g. baby has low birth weight)	Look in the **"Observed" row** of the table
Outcome unsuccessful	The actual observed outcome is "No" (e.g. baby does **not** have low birth weight)	Also in the "Observed" row
Classification successful	The model correctly predicted the outcome	Look at diagonal cells (True Positives + True Negatives)
Classification unsuccessful	The model got the prediction wrong	Look at off-diagonal cells (False Positives + False Negatives)

🧪 2. How to interpret your table:

	Predicted No	**Predicted Yes**
Observed No	15 (✅ TN)	9 (❌ FP)
Observed Yes	5 (❌ FN)	13 (✅ TP)

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.

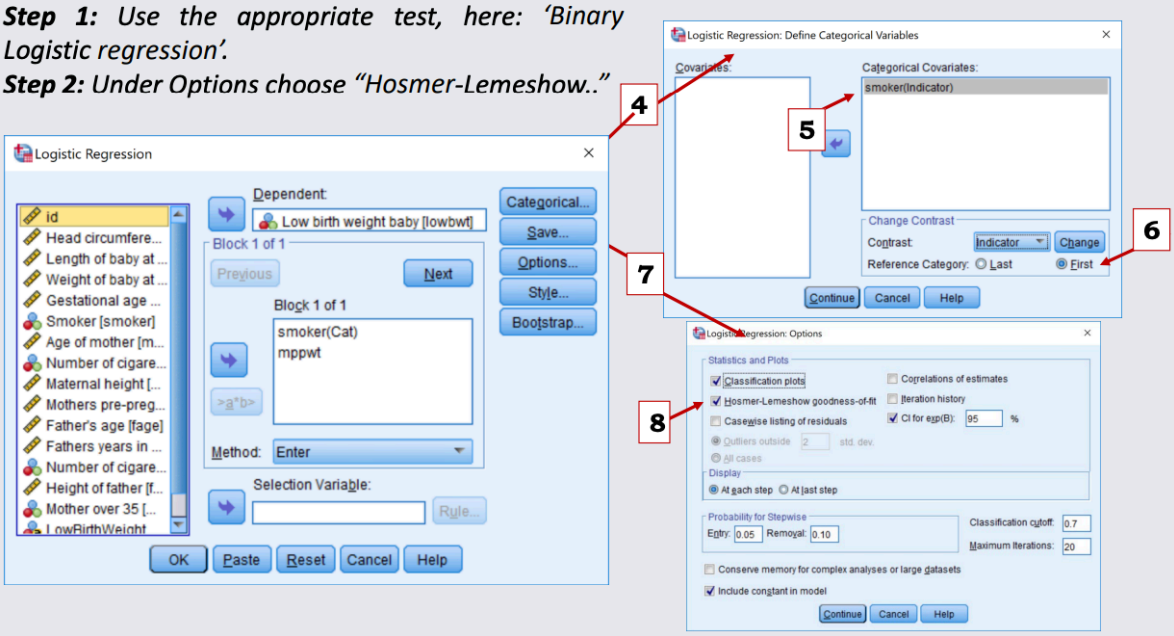
The data are divided into approximately ten groups defined by increasing order of estimated risk.

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

📋 7. Goodness of Fit

Method	Interpretation
Hosmer–Lemeshow Test	Non-significant = good fit
Classification Table	Accuracy of prediction (based on a cut-off like 0.5 or 0.7)
Sensitivity	True positive rate
Specificity	True negative rate
PPV / NPV	Predictive accuracy for yes/no cases

Quiz:

Question 2

Incorrect

Mark 0.00 out of 1.00

Flag question

The odds ratio in Binary logistic regression is:

a. The ratio of the odds after a unit change in the predictor variable

b. The ratio of the probability of an event happening to the probability of the event not happening. ×

c. The ratio of the probability of an event not happening to the probability of the event happening.

d. The probability of an event occurring.

Your answer is incorrect.

The odds ratio in a Binary Logistic regression is the ratio of the odds after a unit change in the predictor variable

The correct answer is:
The ratio of the odds after a unit change in the predictor variable

Topic 10 Skills check

Question 1

- **Binary Logistic regression is used when you want to:**
- Predict a dichotomous variable from continuous or categorical variables.

Question 2

- **The odds ratio in Binary logistic regression is:**
- The ratio of the odds after a unit change in the predictor variable

Question 3

- **Which of the following methods do we use to best fit the data in Logistic Regression?**
- Maximum Likelihood Estimation

Question 4

In a study to determine whether anxiety is associated with subsequent development of depression, the estimated relative risk for those with prior anxiety compared to those who never had anxiety was found to be 1.9. From this we can conclude:

Those with prior anxiety have a higher risk of developing depression than those who did not have prior anxiety

Question 5

There were 5842 men and women surveyed regarding whether they experience dizziness or not and if they used anti-depression medication more than twice in the past two weeks. The data are presented below. Compare the effects of using anti depression medication against not using anti-depression medication on dizziness using the odds ratio.

	Anti-depression medication used	Anti-depression medication not used	Total
Dizziness	443	774	1217
No Dizziness	1530	3095	4625
Total	1973	3869	5842

Question 5

- Odds of having dizziness when anti depression medication is taken is about 1.16 times larger than the odds when anti depression medication is not taken
- Odds of having dizziness when anti depression medication is taken is about 1.16 times larger than the odds when anti depression medication is not taken

- Odds of dizziness and AD = 443/1530 = 0.2895
- Odds of dizziness and no AD = 774/3095 = 0.25001
- Odds Ratio = 0.2895/0.25001 = 1.16

Question 6

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	3298.696	8	.000
	Block	3298.696	8	.000
	Model	3298.696	8	.000

Classification Table[a]

			Predicted		
			Free School Meal		
	Observed		no	yes	Percentage Correct
Step 1	Free School Meal	no	9373	263	97.3
		yes	1368	611	30.9
	Overall Percentage				86.0

Choose all that apply to this data representing cases of free school meals
The model fits the data better than the ‘baseline’ model
The model classifies 86% of cases correctly.
The model is more accurate when classifying those who are not eligible for a free school meal (Specificity).

Topic 10 Practical Knowledge Check Solutions

Data Information

- **Dataset HYPERADHERv3.sav**
- The data set consists of a sample of n = 60 hypertension patients. Several continuous and categorical variables have been recorded:
- Adherent (Participants adherence to their hypertension treatment, coded as 0 = Non adherence, 1 = Adherence)
- Age
- PHQ9 (Measurement of depression symptoms, higher the scores higher the experience of depression symptoms)
- Disease Years (Years with hypertension)
- Self-Compassion (Measurement of Self-Compassion, higher the scores higher the experience of compassion towards the self)
- Social Support (Experience of social support, coded as 0 = Low, 1 = high

Question 1

Researchers are trying to understand if level of social support impacted on a patients adherance to hypertension treatment. The researchers wanted to compare the effects of low social support and high social support on the non-adherance to hypertension treatment.

Calculate the risk of having non adherance to hypertension treatment for each group of social support and complete the interpretation below

Risk of being Non adherent when receiving low social support = 13/20 = 0.65

Risk of being Non adherent when receiving high social support = 13/40 = 0.325

Risk Ratio =0.65/0.325 = 2

The risk of having non adherance to hypertension treatment when patient received low social support is [2] times that of the patient receiving high social support

Calculate the odds of having non adherance to hypertension treatment for each group of social support and complete the interpretation below

Odds of being Non adherent when receiving low social support = 13/7 = 1.85714286

Odds of being Non adherent when receiving highsocial support = 13/27 = 0.48148148

Odds ratio = 1.85714286/0.48148148 = 3.857

The odds of having non adherance to hypertension treatment when the patient received low social support is about [3.86] times [larger] than the odds of patients receiving high social support

Of those patients who did not adhere to hypertension treatment there was a [higher] proportion who received low social support compared to those who received high social support ([65.0%] versus [32.5%]). This difference was statistically [significant] according to [Pearson's Chi Squared] test ([X2 = 5.735], df=1, p = [0.017])

Note

- It is important to know how your data is coded and what interpretation is the easiest to make and the most meaningful

- For example, I could work out the risk and odds for adherence of low compared to high and get a odds ratio of 0,259 this is a difficult number to interpret so I can use its inverse (reciprocal) to get the value comparing high to low e.g. 1/0.259 = 3.857 which is an easier number to interpret.

- Comparing adherence in high versus low social support is the same as comparing non-adherance in low versus high social support

Question 2

The investigator decides to further evaluate the association between social support and adherence among hypertension patients. Run the appropriate test and complete the inference paragraph below. With low social support being the reference category

The analysis results show that model was statistically [significant], [$\chi 2(1) = 5.763$], [$p = 0.016$]. The model explained [12.3%] [(Nagelkerke R2)] of variance in adherence. The correct classification rate has increased by [10]% to [66.7]%

Hypertension patients with [high] social support were [3.857] times more likely to indicate being adherent to treatment compared to those receiving [low] social support. This was a statistically significant result (Wald = [5.460], df = [1], p = [0.019], 95% CI [1.243 - 11.968]).

Note

- Notice for the Binary logistic regression we are looking at being adherent to treatment rather than non-adherent as asked for in Question 1, and comparing high social support to low social support rather than low social support in comparison to high in question 1) If I had changed the coding round in SPSS so adherent was 0 and non-adherent is 1 and if I changed the reference category to "high" then I would also be able to write the below

- Hypertension patients with [low] social support were [3.857] times more likely to indicate being non-adherent to treatment compared to those receiving [high] social support. This was a statistically significant result (Wald = [5.460], df = [1], p = [0.19], 95% CI [1.243 - 11.968]).

- It is important to know how your data is coded and what interpretation is the easiest to make

- For example, I could work out the risk and odds for adherence of low compared to high and get a odds ratio of 0,259 this is a difficult number to interpret so I can use its inverse to get the value comparing high to low e.g. 1/0.259 = 3.857 which is an easier number to interpret.

Question 3

The investigator decides to estimate the effect of social support on adherence after controlling for age, disease years, depression symptoms, and self-compassion. Run the appropriate test and complete the inference paragraph below.

The model fit for the adjusted analysis was [good], indicating that [Hosmer and Lemeshow] test result was [higher] than the 0.05 level of significance.

Hypertension patients with high social support were [7.491] times more likely to indicate being adherent after adjusting for age, disease years, depression symptoms, and self-compassion compared to patients with low social support. The effect of social support on adherence [was not] statistically significant in the adjusted model (Wald = [3.796], df= [1], p = [0.051], 95% CI [0.998 - 56.795]).

Model shows that an increase in age by one year is associated with an [increased] likelihood of adherence.

Note

Hypertension patients with **high** social support were [7.355] times more likely to indicate being **adherent** after adjusting for age, disease years, depression symptoms, and self-compassion compared to patients with **low** social support.

As before this could also be written as follows

Hypertension patients with **low** social support were [7.355] times more likely to indicate being **non-adherent** after adjusting for age, disease years, depression symptoms, and self-compassion compared to patients with **high** social support.

or

Hypertension patients with **low** social support were [**0.135**] times less likely to indicate being **adherent** after adjusting for age, disease years, depression symptoms, and self-compassion compared to patients with **high** social support. Where 0.135 is the **reciprocal** or **inverse** of 7.355 (1/7.355)

Question 4

The probability of a patient being adherent to hypertension treatment when they have high social support and are 58 years old, have a total score of 1 for depressive symptoms, 10 years of hypertension and a total score of 12 for self compassion is [99.8]%

Adherent log $\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)$= -4.292 + 2.014social support + 0.095age - 0.311depression+ 0.020hypertension + 0.259compassion

Adherent log $\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)$= -4.292 + (2.014 x 1) + (0.095 x 58) – (0.311 x 1) + (0.020 x10) + (0.259 x12)

= -4.292 + 2.013 + 5.51 – 0.311 + 0.2 + 3.108

=6.228

Probability = $\hat{\pi} = \frac{e^{6.228}}{1+e^{6.228}} = 0.998$