Institute of Psychiatry, Psychology and Neuroscience

KING'S College LONDON

Zahra Abdulla

Department: Biostatistics and Health Informatics

Module Title: Introduction to Statistics

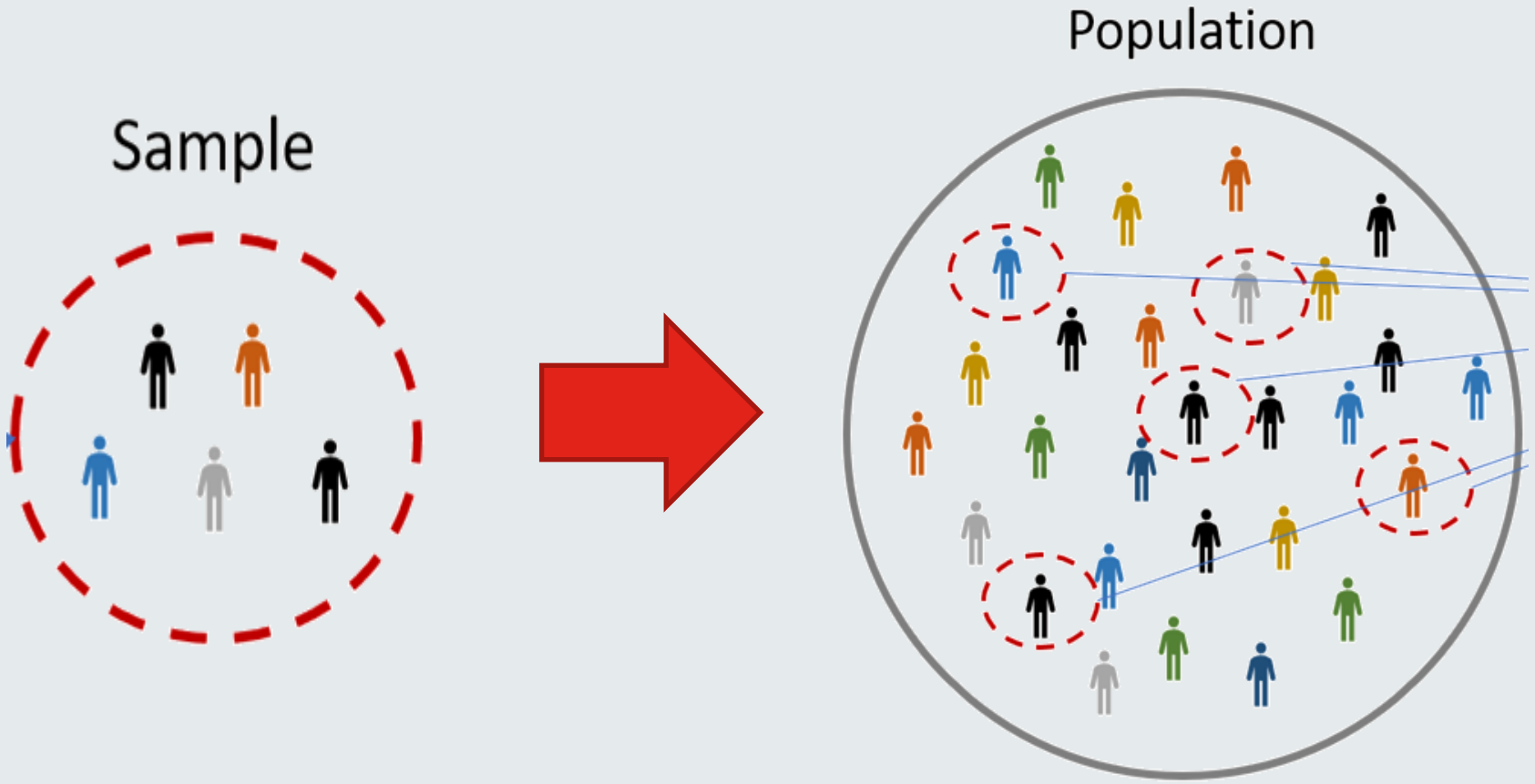
Session Title: Binary Logistic Regression

Topic title: Binary Logistic Regression

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

- Understand what modelling is and why we do it
- Recognise when a binary logistic regression analysis is suitable
- Run a binary logistic regression analysis in a software package

What is modelling? Recap



Why is statistical modelling important

- Example: Investigating the effect of a new app for treating depression

- In a randomised trial we observe reduced depression scores in the group that used the app

Could the difference have occurred by chance?

- Modelling allows us to calculate the **association between variables (e.g., odds ratios)** as well as **uncertainty about the association (e.g., confidence intervals and p values)**

General linear model (linear regression)

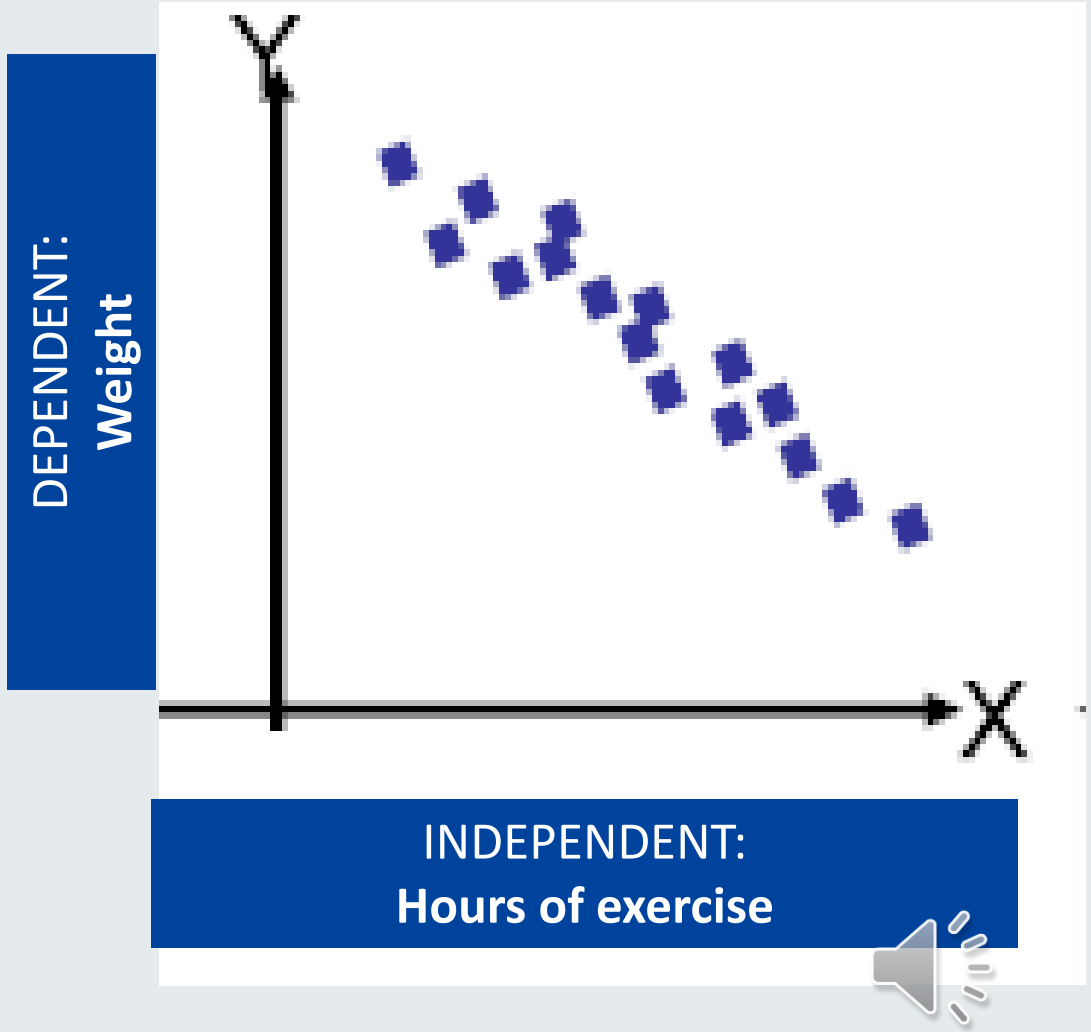
16 people were observed to see if the weight of a person is related to exercise:

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.

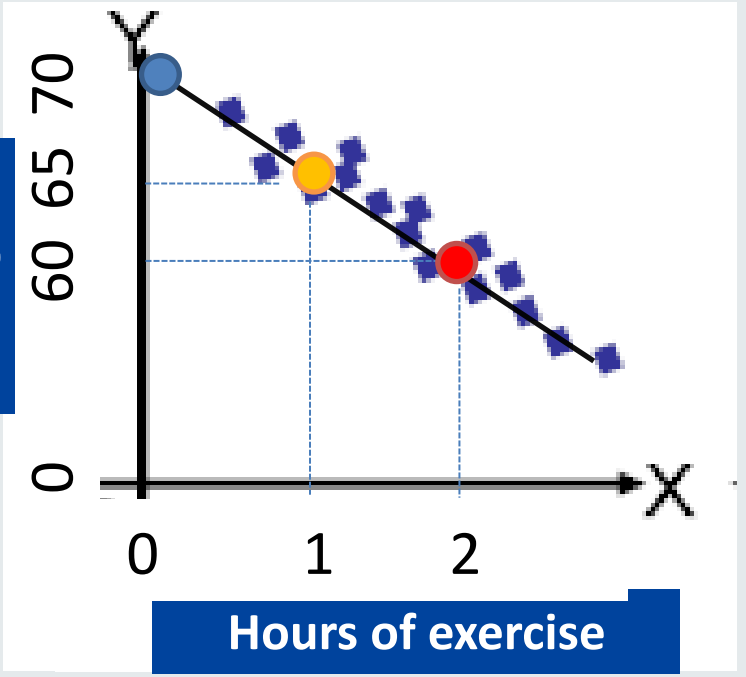
Linear regression model:

- 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



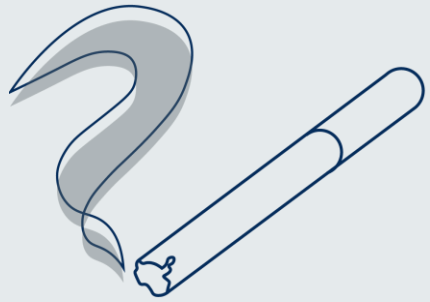
Some Scenarios

- 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?

General*ised* 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

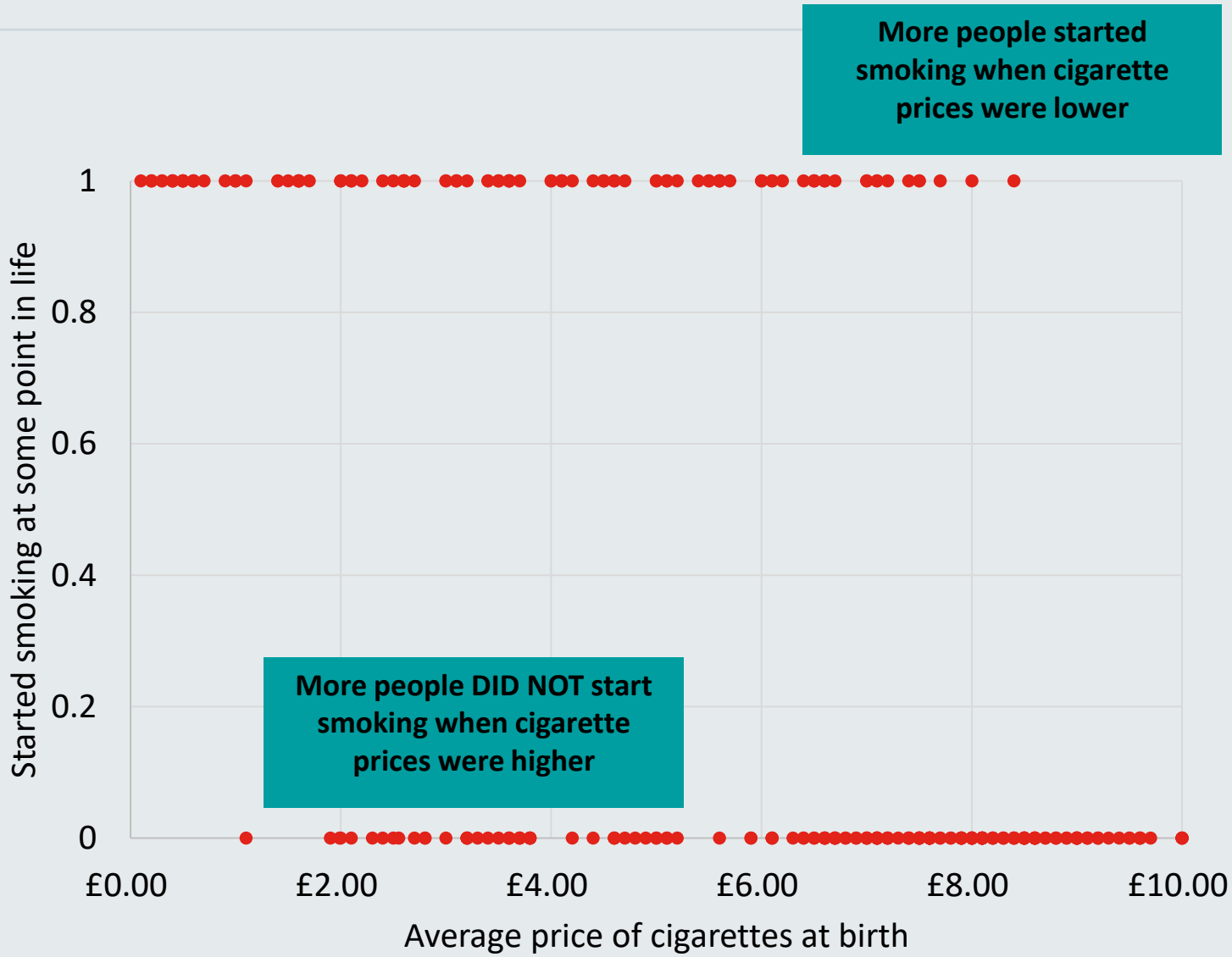
Do not dichotomize if not necessary (Loss of information)

What is wrong with Simple Linear Regression?

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

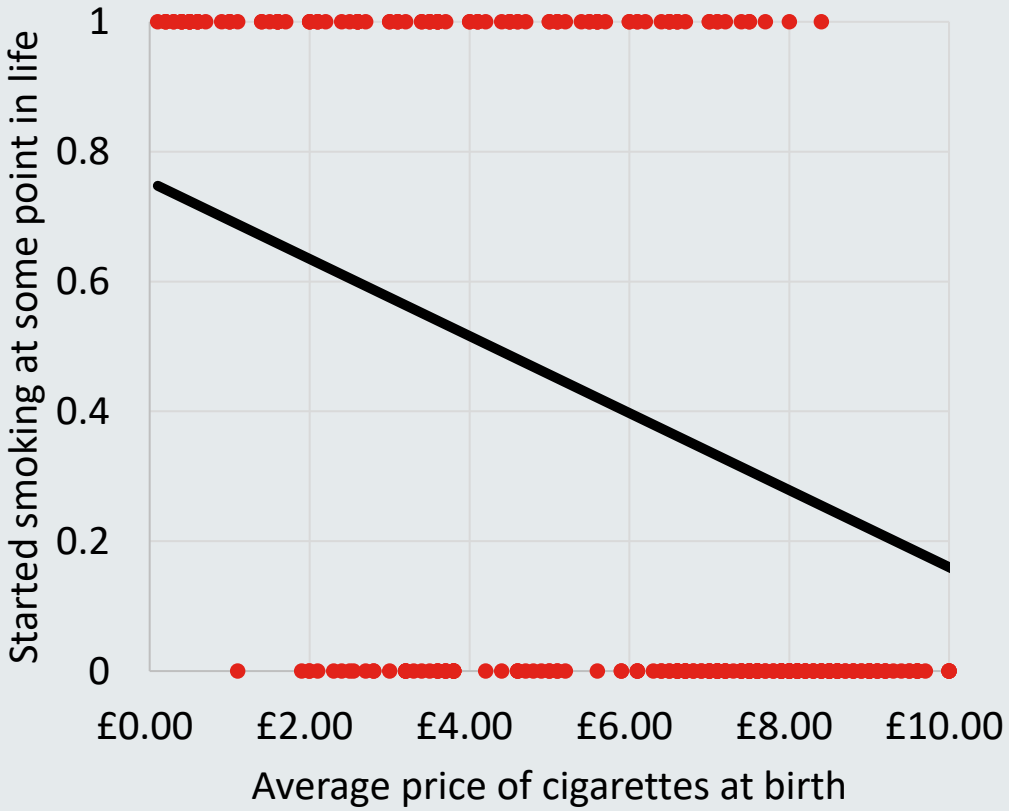
But our simple linear regression may predict values that are below zero or above 1

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



More problems with Simple Linear Regression

Could add a linear regression line, but prediction would not make much sense (not below 0 or 1)



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

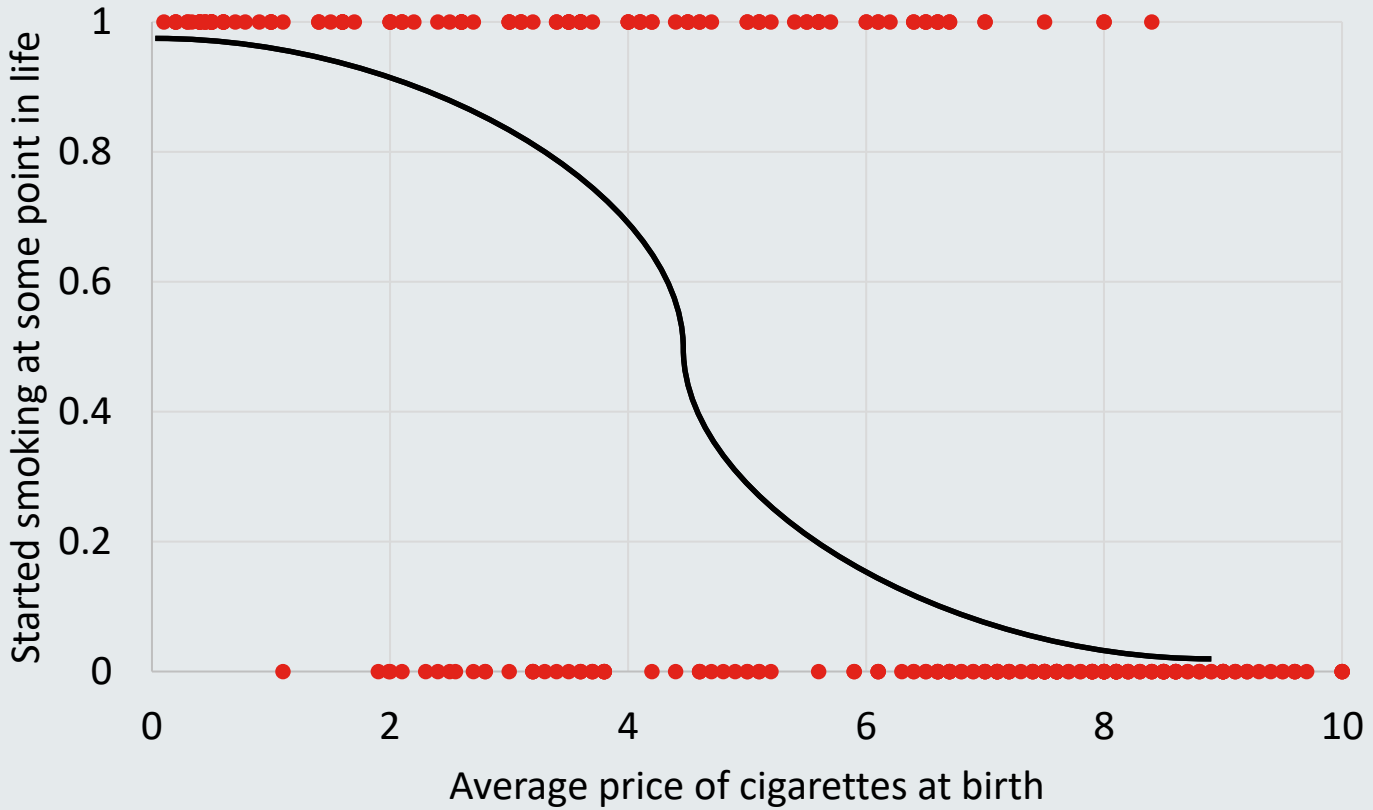
Linear relationship does not make sense for binary outcomes

We rather assume a **nonlinear** relationship as

- Output variable is limited to **[0,1],** some of our observations are outside this range
- Our goal is to separate best the two groups not to **minimize Mean Squared Error**.
- Linear regression would be **highly sensitive to influential cases**
- Assumptions of linear regression are violated (esp. homogeneity of variances) and hence inference is not valid

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):

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

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.

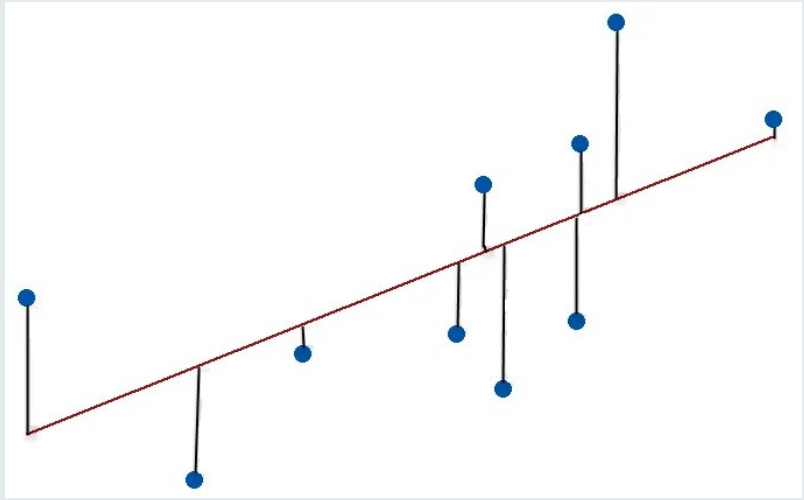
This is just the odds,
the probability that expected outcome **does** happen divided by the probability that expected outcome **does not** happen

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

Logistic Regression: The Logistic Model

π — This is the **Probability** of an event

$\frac{\pi}{1-\pi}$ — This is the **Odds** of an event

Model: $ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x$ — This is called the **Logit**

$L = \alpha + \beta x$ — This is called the **Linear Predictor**

The model is **linear** in the **logit** but **non-linear** in the probability π. The data are fitted using **logits** and then one **transforms** the **fitted** parameters to **probabilities** afterwards.

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

$\hat{\pi} = \frac{odds}{1+odds}$ — This is the **Estimated Probability** of an event

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

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:

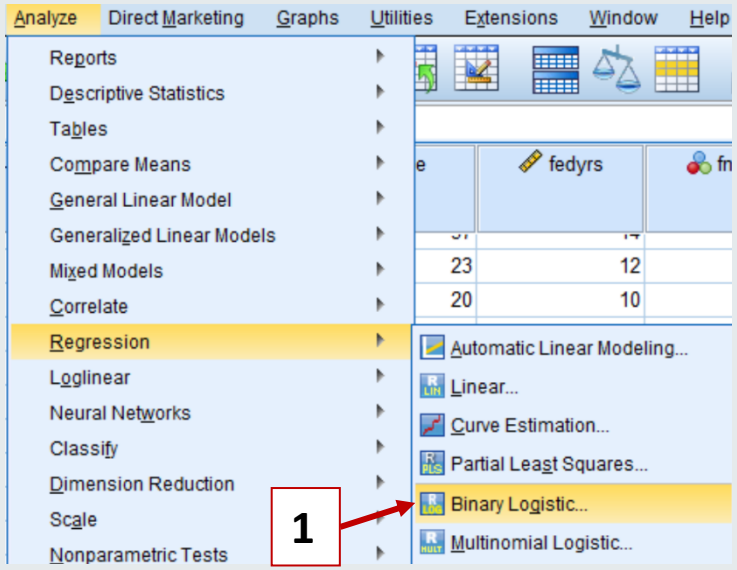
- Binary dependent variable which has a **Bernoulli (binomial)** Distribution
- Continuous variables have a linear effect on the log-odds scale (Is linearly related to the predictor variables only after transforming into the **logit** scale)
- Observations are independent
- Adequate Sample size
- Absence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

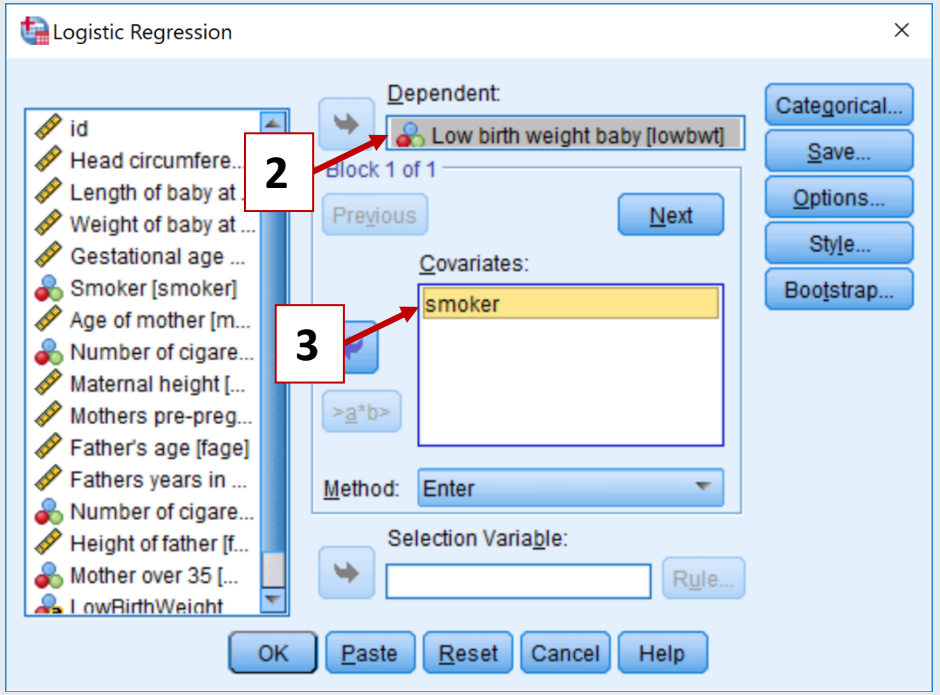
SPSS slide: ‘how to’

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

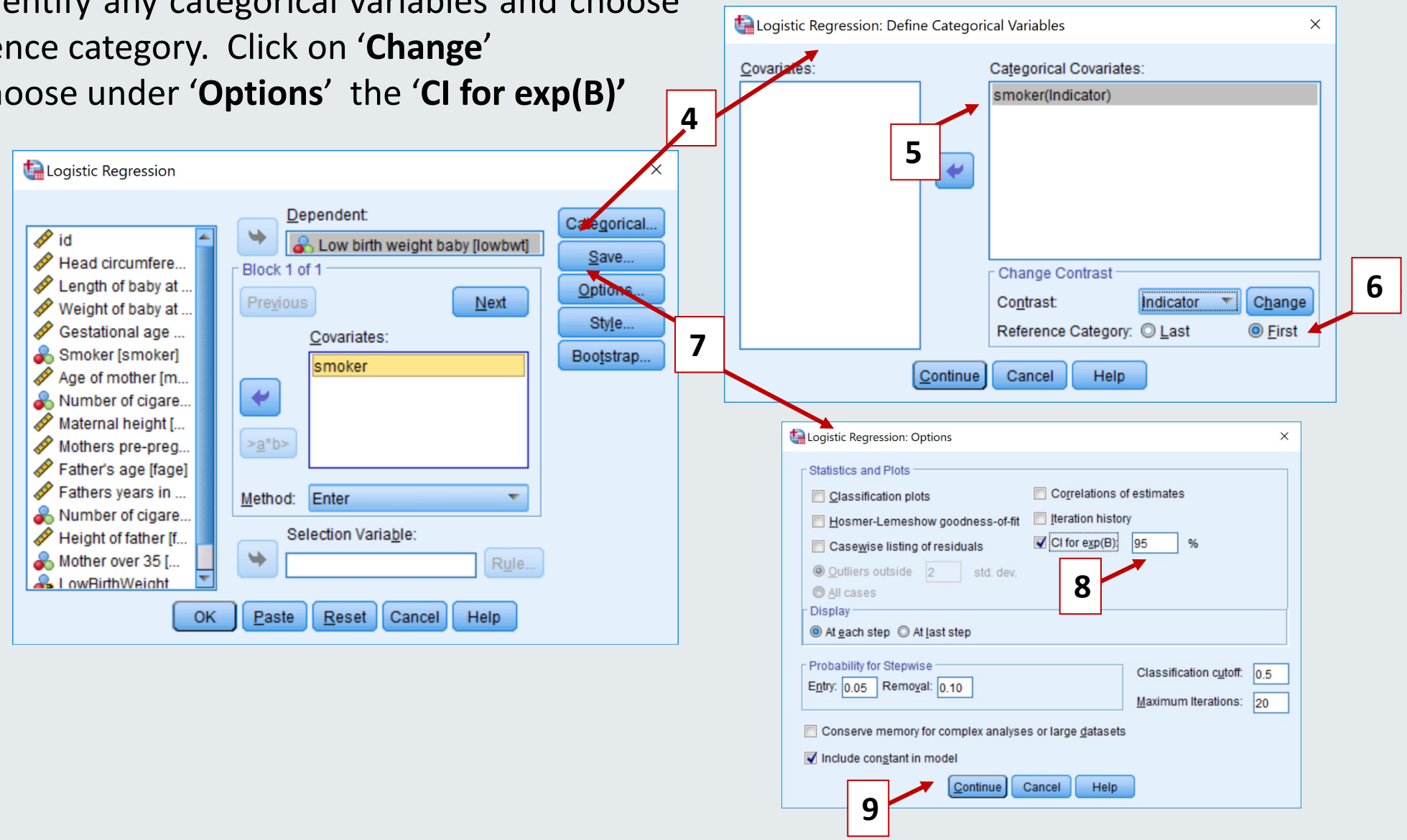




SPSS slide: ‘how to’

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

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

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.**

Thinking back to why this is important… - rerecord

- ***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.***

- We now have a measure of association (odds ratio): Exp(β) = 4.333

- As well as measures of uncertainty (confidence intervals and p values): 95%CI = 1.156 to 16.248, p=0.030

Knowledge Check

Q1:
Consider research by Wuensch & Poteat, published in the *Journal of Social Behavior and Personality* in 1998

College students (N = 315) were asked to pretend that they were serving on a university research committee regarding a complaint against animal research being conducted by a faculty member. The complaint included a description of the research in simple but emotional language.

In his defence, the researcher made a case about the benefits of his research and steps taken to ensure the animals did not experience pain.

After reading the case materials, each participant was asked to decide whether or not to withdraw the faculty members' authorization to conduct the research.

Knowledge Check

Q1 Cont. Let us first consider a simple (bivariate) logistic regression, using subjects' decisions as the dichotomous criterion variable and their gender as a dichotomous predictor variable.
gender coded with 0 = Female, 1 = Male, and decision with 0 = "Stop the Research" and 1 = "Continue the Research".
Write the regression equation and interpret the output below

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	399.913[a]	.078	.106

a. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	gender	1.217	.245	24.757	1	.000	3.376
	Constant	-.847	.154	30.152	1	.000	.429

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

Knowledge Check Solution

- We can interpret **Nagelkerke R^2** 10.6% of the variation in 'Stop the research' can be explained by the model including Gender.

- The exp(β) tells us that the model predicts that the odds of deciding to continue the research are 3.376 times higher for men than they are for women.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	399.913[a]	.078	.106

a. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	gender	1.217	.245	24.757	1	.000	3.376
	Constant	-.847	.154	30.152	1	.000	.429

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

References

J Scott Long, Sage, 1997 Regression Models for Categorical and Limited Dependent Variables

A Agresti, Wiley, 2002 An Introduction to Categorical Data Analysis 2nd ed by

Wuensch and Poteat example: http://core.ecu.edu/psyc/wuenschk/MV/Multreg/Logistic-SPSS.PDF

Thank you

Contact details/for more information:

Zahra Abdulla

Zahra.abdulla@kcl.ac.uk

Department of Biostatistics and Health Informatics (BHI)

IoPPN