Topic materials:
Dr Raquel Iniesta

Narration and contribution:
Zahra Abdula

Improvements:
Nick Beckley-Hoelscher
Kim Goldsmith
Sabine Landau

Institute of Psychiatry, Psychology and Neuroscience

KING'S College LONDON

Module Title: Introduction to Statistics

Session Title: Multiple Linear Regression Model

Topic title: Multiple regression with several explanatory variables: Adjusting for confounders

Learning Outcomes

- To extend a simple linear regression to a **multiple linear regression model**.
- Understand the difference between regression coefficients from a simple linear regression model and **partial regression coefficients**.
- Statistically evaluate associations between multiple independent variables and a dependent variable.
- Interpret the output from fitting a multiple linear regression in a statistical software.

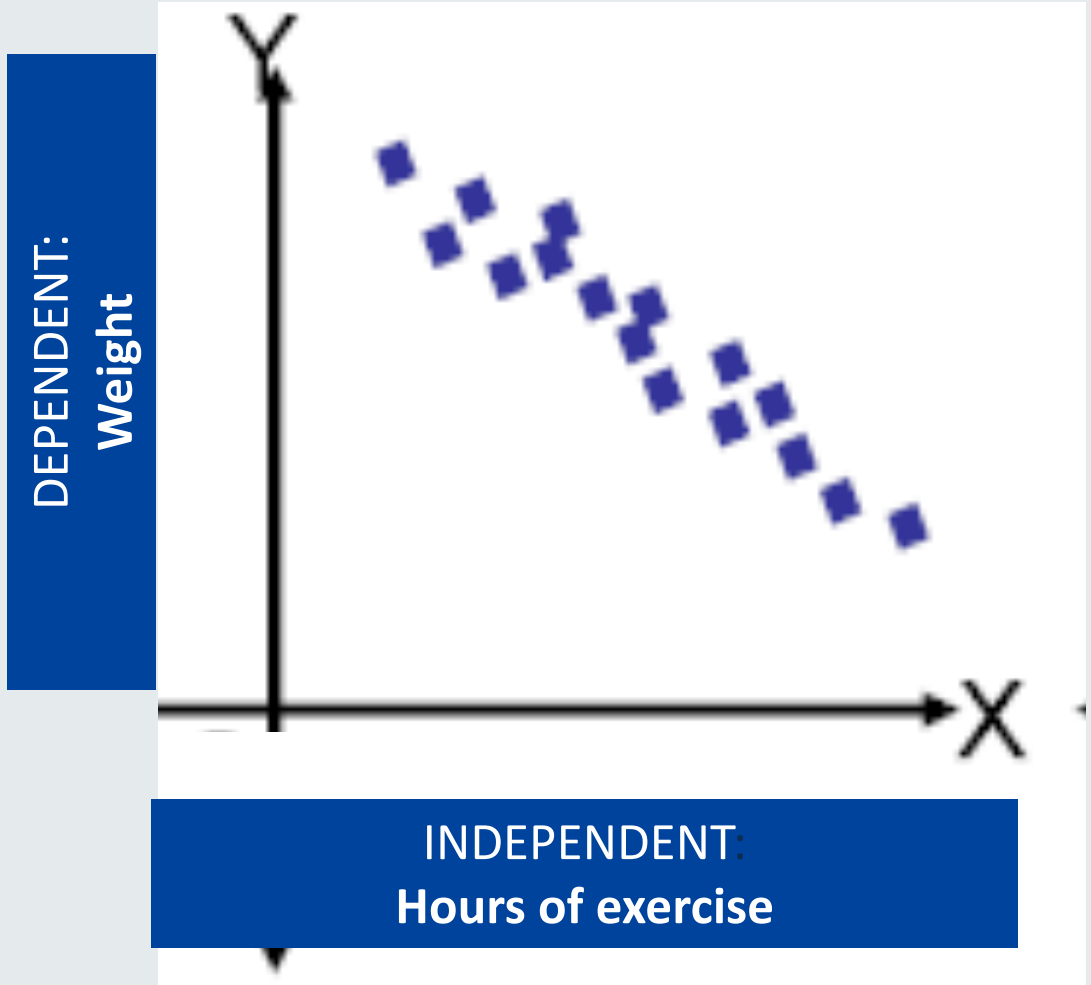
Previously on ‘Introduction to Statistics’

16 people were observed to see if the weight of a person, related to the hours of exercise they conducted. The following hypothesis was investigated:

Hypothesis ‘The higher the number of hours of exercise the lower the weight’.

Plotting the data is essential to understand and visually assess the relationship between pairs of continuous variables

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



Previously on ‘Introduction to Statistics’

Questions:

Q1: How strong is the linear relationship? Understand the direction and magnitude of the linear relationship

A1: Correlation Coefficient (Pearson) ***r*=-0.85**

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

Q2: Can the relationship between variables be described by fitting a line to the observed data?

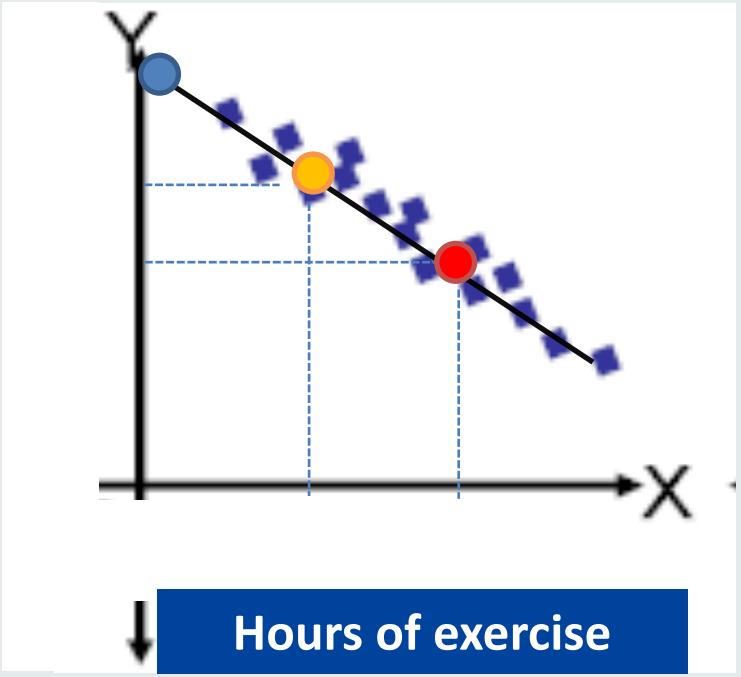
A2: Yes, because there is a **linear relationship**. The relationship is expressed as an equation

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

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

X	Y
0	70
1	65
2	60

$$y = 70 - 5x$$



β_0=70; β_1=-5;

Previously on ‘Introduction to Statistics’

Interpretation

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

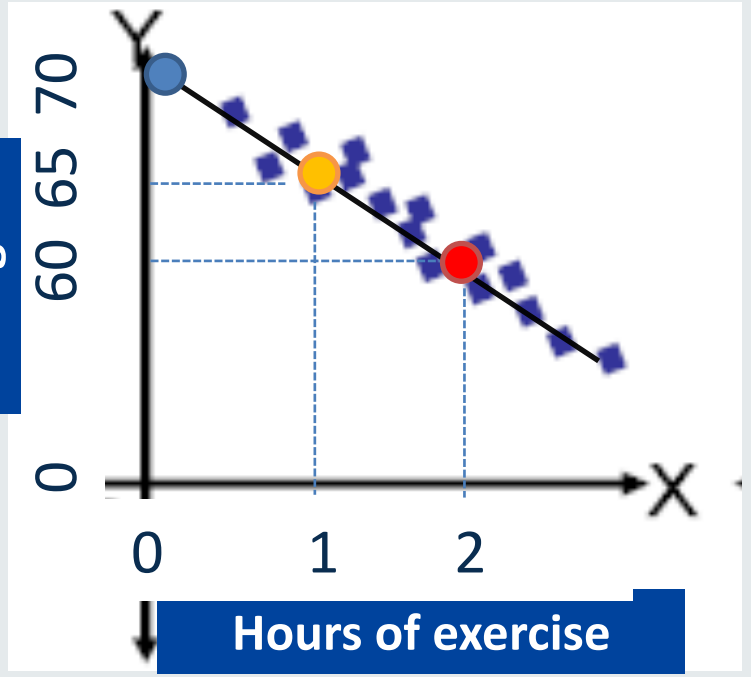
Linear regression model:

- To measure to what extent there is a linear relationship between two variables
- A rule that predicts weight given the hours of exercise.

β_0=70; β_1=-5;

$y = 70 - 5x$

X	Y
0	70
1	65
2	60

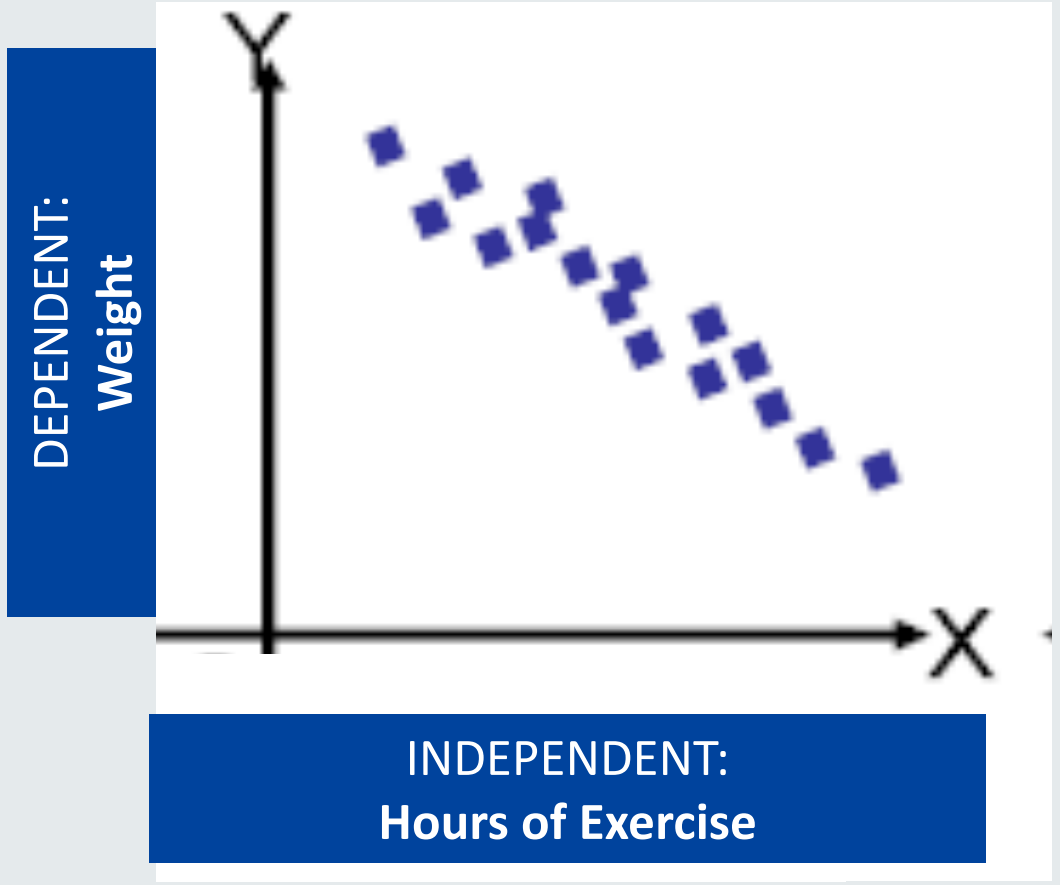


From Simple Linear Regression to Multiple Linear Regression

Using correlation and simple linear regression we found that individuals weight (y) depended on the hours of exercise (x). Specifically, that each extra hour of exercise reduces the average weight by 5 Kg.

$$y = 70 - 5x;\ r\text{=-0.85}$$

But is weight just related to exercise? Or could it also depend on **diet**, **water intake**, **age**, **gender**, ... ?



From Simple Linear Regression to Multiple Linear Regression

Simple linear regression

$$y = 70 - 5x + \varepsilon$$

Where: ***y*=weight;**
***x*=exercise;**

Multiple linear regression

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

Where: ***y*=weight;**

x_1=exercise;

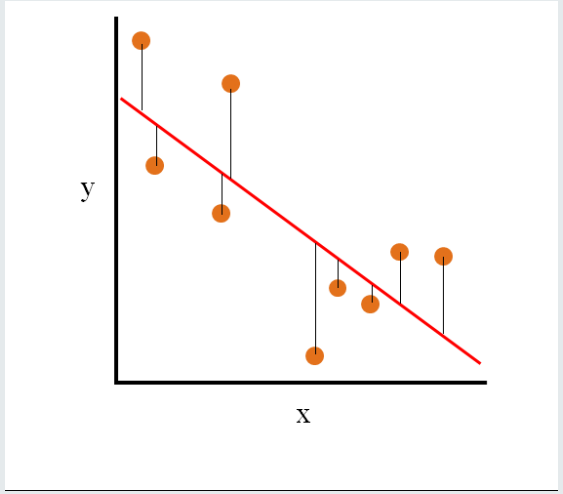
x_2=diet;

From Simple Linear Regression to Multiple Linear Regression

Simple linear regression

$$y = 70 - 5x + \varepsilon$$

Where: ***y*=weight; *x*=exercise;**



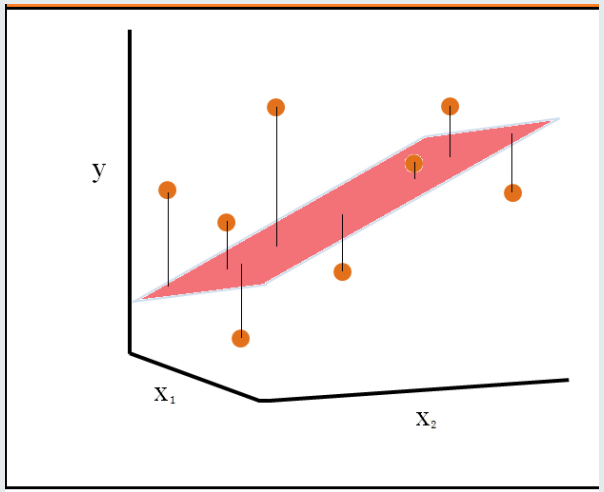
A **simple regression model** (one independent variable) fits a regression **line**

$y = \beta_0 + \beta_1 x_1$

Multiple linear regression

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

Where: ***y*=weight; x_1=exercise; x_2=diet;**



A **multiple regression model** with two explanatory variables fits a **regression plane**

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

Multiple Linear Regression Model

When to use it

- Method for studying the relationship between one dependent variable (e.g. weight) and two or more independent variables simultaneously (e.g. exercise, diet, water intake, age, gender...) to understand how a dependent variable can be explained by a set of other variables.

- We aim to answer:
  - Whether and how **several facts** are related with **one other fact**?
  - Whether and how a **set of independent variables** are related with a **dependent variable**?

- **E.g.** Understanding the factors that determine weight, to create clinical guides to advise patients on kind of diet, water intake, etc. for them to keep a healthy weight.

Multiple Linear Regression Model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon$$

- $\mathbf{x}_i$ are the **independent variables, predictors, explanatory or covariates** (continuous or categorical)
- $\mathbf{y}$ is called the **dependent variable, outcome or response**. y '**depends on**' $\mathbf{x}_i$. It is when $\mathbf{y}$ is continuous that we can use a linear model. If not ($\mathbf{y}$ is categorical) we have to use a different type of model.
- The **intercept** $\boldsymbol{\beta_0}$ is the value that y takes when $\mathbf{x}_i$ is zero.
- $\boldsymbol{\beta_i}$'s are the **partial regression coefficients.**

  $\boldsymbol{\beta_i}$ represents the change in **average *y* for one unit change in** $\mathbf{x}_i$ (holding (adjusting for) all other x's fixed)

  E.g. $\boldsymbol{\beta_1}$ Is the amount that the dependent variable $\mathbf{y}$ will increase (or decrease) for each unit increase in the independent variable $\boldsymbol{x_1}$ while **holding** all other variables $x_2, \ldots, x_n$ **constant**.
- $\boldsymbol{\varepsilon}$ is called the **residual** (distance between the points and the plane).

All you need are the regression coefficients β_i

Hypotheses:

Each partial regression coefficient will be tested for linear association while holding all other variables in the regression equation constant

E.g. Test $\boldsymbol{\beta_1}$ to check if variable $\boldsymbol{x_1}$ is significantly associated with the outcome $\boldsymbol{y}$ while holding all other variables $\boldsymbol{x_2}, \ldots, \boldsymbol{x_n}$ constant.

H_0: Holding (Adjusting for) all other variables constant, there is no linear association between $\boldsymbol{y}$ and $\boldsymbol{x_1}$
e.g. the slope β_1 in the population equals 0. **H_0: $\boldsymbol{\beta_1}$=0**

H_a: Holding (Adjusting for) all other variables constant, there is a linear association between $\boldsymbol{y}$ and $\boldsymbol{x_1}$
e.g. the slope β_1 in the population does not equal 0. H_a : **$\boldsymbol{\beta_1}\neq 0$**

If *p*<0.05, we reject the null $\boldsymbol{\beta_1 = 0}$ and conclude that x_1 is significantly associated with *y* at a population level.

Example

According to the researchers, in the population from which our data came, they believe there is a relationship between weight, frequency of exercise per week and the frequency of vegetables eaten per day.

$y = 72 - 4x_1 - 2x_2 + \varepsilon$		**p-value**
Slope for x_1 (β_1)	-4	0.01
Slope for x_2 (β_2)	-2	0.03

y = weight;
x_1= frequency of exercise per week;
x_2= frequency of vegetables per day;

a) Is the frequency of exercise associated with weight?

Yes, because the p-value for the hypothesis test for β_1= 0 is less than 0.05 (i.e. p = 0.01).

Then we can conclude that β_1 is significantly different than 0 at a population level.

The variable x_1 has a significant effect on y while holding x_2 constant.

In other words:

The variable x_1 is associated with y while holding x_2 constant, or while adjusting for x_2.

Example

$y = 72 - 4x_1 - 2x_2 + \varepsilon$		p-value
Slope for x_1 (β_1)	-4	0.01
Slope for x_2 (β_2)	-2	0.03

y = weight;
x_1= frequency of exercise per week;
x_2= frequency of vegetables per day;

b) How can we interpret the regression equation?

- A person exercising once a week (x_1=1) and eating vegetables twice a day **(x_2=2)** will have a weight of

  $y = 72 - (4 \times 1) - (2 \times 2)$
  y = 64kg

- A person exercising twice a week (x_1=2) and eating vegetables twice a day **(x_2=2)** will have a weight of

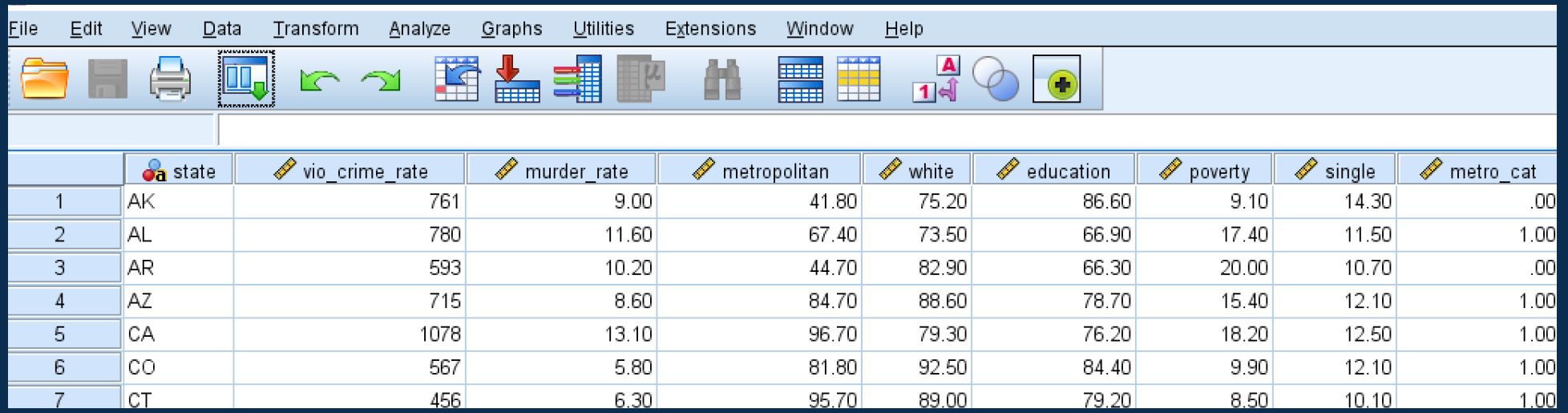
  $y = 72 - (4 \times 2) - (2 \times 2)$
  y = 60kg

Held constant

In other words: one added exercise session a week decreases the weight by 4kg if you eat vegetables with the same daily frequency (which is the interpretation of β_1= -4)

SPSS Slide

Download the data that we are going to use during the lecture. The dataset is the **lecture_7_data.sav**.



	state	vio_crime_rate	murder_rate	metropolitan	white	education	poverty	single	metro_cat
1	AK	761	9.00	41.80	75.20	86.60	9.10	14.30	.00
2	AL	780	11.60	67.40	73.50	66.90	17.40	11.50	1.00
3	AR	593	10.20	44.70	82.90	66.30	20.00	10.70	.00
4	AZ	715	8.60	84.70	88.60	78.70	15.40	12.10	1.00
5	CA	1078	13.10	96.70	79.30	76.20	18.20	12.50	1.00
6	CO	567	5.80	81.80	92.50	84.40	9.90	12.10	1.00
7	CT	456	6.30	95.70	89.00	79.20	8.50	10.10	1.00

The dataset contains data from 51 US states, measuring the crime rates and background measures for each State with respect to their

- **violent crime**: per 100,000 population
- **murder**: per 100,000 population
- **poverty**: percent below the poverty line
- **single**: percentage of lone parents

Questions

- What multiple facts may be related with the risk of someone committing a **crime?**
- A researcher suggested (had a theory) that both poverty and education have an effect on committing a crime?
- What is the **joint effect of poverty and education** on crime**?**

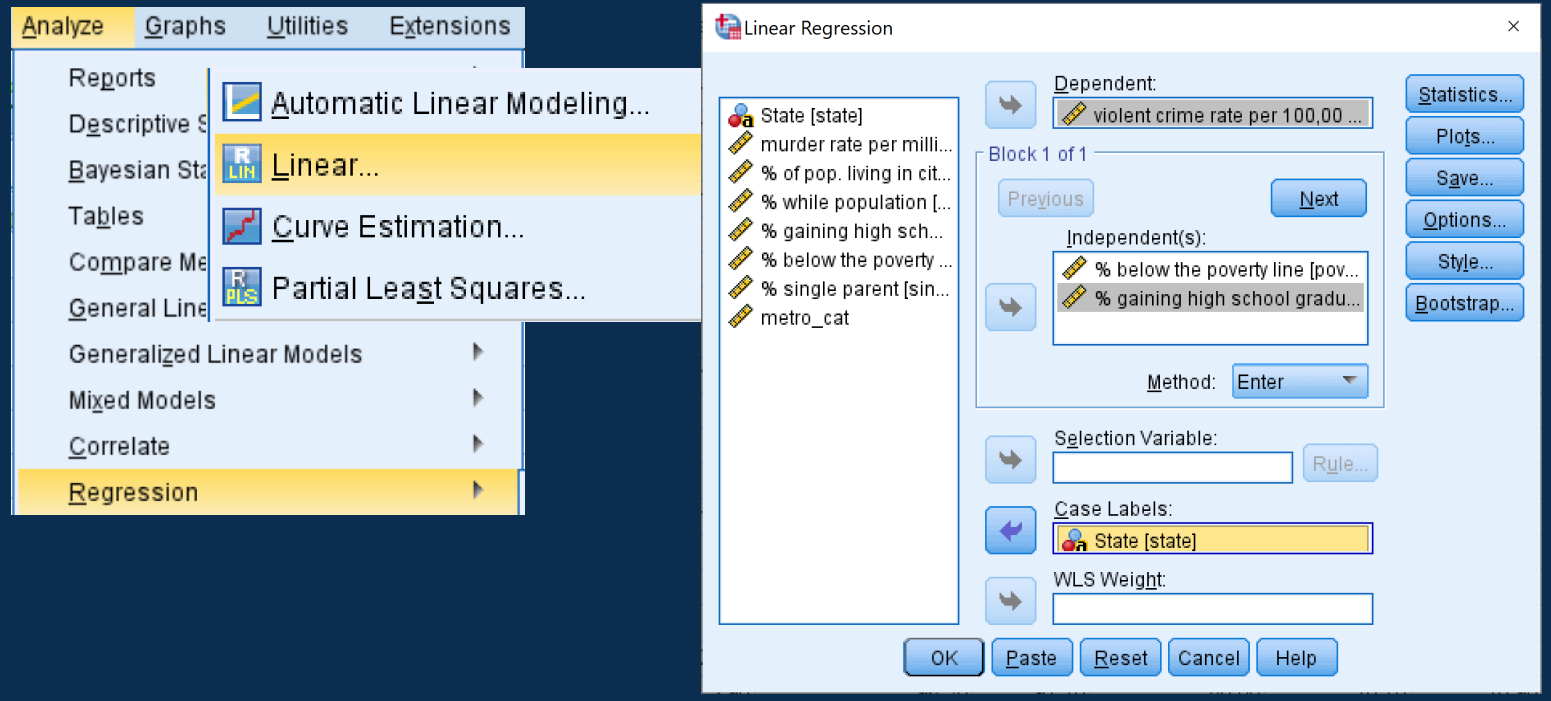
Facts like poverty or education are encoded in the form of **variables**

SPSS Slide: ‘how to’

Researchers believe, in the population from which our data came, the % below the poverty line and % gaining a high school graduation have and effect on the Violent Crime rate

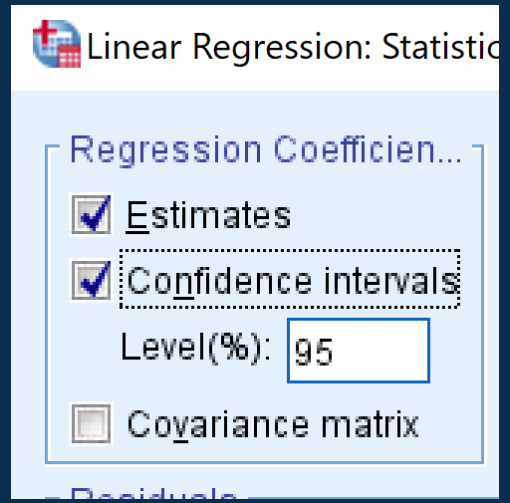
Step 1) Computing a multiple linear regression model for dependent variable ‘**crime**’ and independent variables ‘**poverty**’ and ‘**education**’

Use **Analyse -> Regression -> Linear**



Put ‘**crime**’ in ‘dependent’, and ‘**poverty**’ and ‘**education**’ in ‘independent’.

Click **Statistics**, select ‘**Confidence intervals**’.



Output and Interpretation Slide

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	345.852	1026.638		.337	.738	-1719.478	2411.181
	% below the poverty line	23.927	14.763	.347	1.621	.112	-5.774	53.627
	% gaining high school graduation	-1.502	11.239	-.029	-.134	.894	-24.112	21.109

a. Dependent Variable: violent crime rate per 100,000 population

β_0 β_1 β_2

$$y = 345.852 + 23.927\,x_1 - 1.502\,x_2$$

The intercept (β_0), is the extrapolated Violent Crime Rate at 0% below the poverty line and 0% of high school education

The estimated slope coefficient (β_1), suggests a 1% increase in poverty is associated with a 23.927 increase in Violent crime rate per 100 000 holding % of education constant (or adjusting for % of education).

The estimated slope coefficient (β_2), suggests a 1% increase in education is associated with a 1.502 decrease in Violent crime rate per 100 000 holding % poverty constant (or adjusting for % of poverty).

Output and Interpretation Slide

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	345.852	1026.638		.337	.738	-1719.478	2411.181
	% below the poverty line	23.927	14.763	.347	1.621	.112	-5.774	53.627
	% gaining high school graduation	-1.502	11.239	-.029	-.134	.894	-24.112	21.109

a. Dependent Variable: violent crime rate per 100,000 population

Based on the multiple regression, poverty (x_1) has a partial regression coefficient β_1 of 23.927, with a 95% CI [-5.774, 53.627]

Given the hypothesis test for β_1: $\begin{cases} H_0: \beta_1 = 0 \\ H_a: \beta_1 \neq 0 \end{cases}$ gives p=0.112 the result is not significant.

We conclude that poverty is not statistically significantly associated with crime when the poverty-crime relationship is adjusted for education (or education is held constant). We cannot generalise that poverty is associated with crime in the population (β_1= 23.927, t=1.621, p=0.112, 95%CI (-5.774, 53.627))

Knowledge Check

A clinical trial aims to compare the efficacy of two different antidepressant drugs (escitalopram and nortriptyline). A multiple regression model was considered with the dependent variable $\boldsymbol{y}$ being the "percentage of improvement in depression severity after being treated with an antidepressant drug".

- The treatment variable ($\boldsymbol{x_1}$) was coded as 0 or 1 (0 = escitalopram and 1 = nortriptyline).
- The patient severity at the start of the trial ($\boldsymbol{x_2}$) was also considered as an explanatory variable.
- $\boldsymbol{x_2}$ ranged from 0 to 100, being 0 the minimum severity and 100 the maximum.

The estimated multiple linear regression model was:

$$\boldsymbol{y = 40 - 10x_1 - 2x_2 + \varepsilon}$$

with p-value=0.02 for β_1 , p-value=0.01 for β_2

Q1: Given the model, which drug is more effective, escitalopram, or nortriptyline?
Q2: TRUE or FALSE: Patients who are more severe at baseline, improve less under any treatment.

Knowledge Check Solutions – Q1

$\boldsymbol{y}$ = percentage of improvement in depression severity after being treated with an antidepressant drug

$$\boldsymbol{x_1} = \begin{cases} 0 \text{ escitalopram} \\ 1 \text{ nortriptyline} \end{cases}$$

$\boldsymbol{x_2}$ = The patient severity at the start of the trial;

$\boldsymbol{x_2} \in [0,100]$

		p-value
Slope for $\boldsymbol{x_1}$ $(\boldsymbol{\beta_1})$	-10	0.02
Slope for $\boldsymbol{x_2}$ $(\boldsymbol{\beta_2})$	-2	0.01

$$\boldsymbol{y = 40 - 10x_1 - 2x_2 + \varepsilon}$$

Q1: Which drug is more effective, escitalopram, or nortriptyline?

$\boldsymbol{x_1}$ is significantly associated with $\boldsymbol{y}$ (p-value 0.02) so we conclude there is association between treatment type and depression severity.

When holding baseline severity constant the percentage of improvement was higher for escitalopram, so escitalopram was more effective.

We can further see this by calculating predicted outcome values for people with a fixed baseline severity level ($\boldsymbol{x_2}$), for example $\boldsymbol{x_2 = 15}$.

If treated with escitalopram, $\boldsymbol{x_1}$=0. Then $\boldsymbol{y = 40 - (10 \times 0) - (2 \times 15)}$; $\boldsymbol{y = 10\%}$

If treated with nortriptyline, $\boldsymbol{x_1}$=1. Then $\boldsymbol{y = 40 - (10 \times 1) - (2 \times 15)}$; $\boldsymbol{y = 0\%}$

Knowledge Check Solutions – Q2

$\boldsymbol{y}$ = percentage of improvement in depression severity after being treated with an antidepressant drug

$$\boldsymbol{x_1} = \begin{cases} 0 \text{ escitalopram} \\ 1 \text{ nortriptyline} \end{cases}$$

$\boldsymbol{x_2}$ = The patient severity at starting the trial;

$\boldsymbol{x_2} \in [0,100]$

		p-value
Slope for $\boldsymbol{x_1}$ $(\boldsymbol{\beta_1})$	-10	0.02
Slope for $\boldsymbol{x_2}$ $(\boldsymbol{\beta_2})$	-2	0.01

$$\boldsymbol{y = 40 - 10x_1 - 2x_2 + \varepsilon}$$

Q2: TRUE or FALSE: Patients that are more severe when starting with any treatment, improve less.

TRUE.

$\boldsymbol{x_2}$ is significantly associated with $\boldsymbol{y}$ (p-value 0.01) we conclude there is association between depression severity at the start of the trial and depression severity after treatment.

In the model, β_2 = -2 ; For every one point increase in baseline severity score, the % improvement in depression decreases by 2.

References

Agresti, A., & Finlay, B. (2009).
Statistical Methods for the Social Sciences (4th ed.). New Jersey, NJ: Prentice Hall Inc.

Douglas, C., Montgomery, D. C., Peck, E. A., & Vining, G. G. (2006).
Introduction to Linear Regression Analysis. New York, NY: Wiley.

Thank you

Please contact your module leader or the course lecturer of your programme, or visit the module's forum for any questions you may have.

If you have comments on the materials (spotted typos or missing points) please contact Dr Iniesta:

Raquel Iniesta, PhD
Department of Biostatistics and Health Informatics
IoPPN, King's College London, SE5 8AF, London, UK
raquel.iniesta@kcl.ac.uk

For any other comments or remarks on the module structure, please contact one of the three module leaders of the Biostatistics and Health Informatics department:

Zahra Abdula: **zahra.abdulla@kcl.ac.uk**

Raquel Iniesta: **raquel.iniesta@kcl.ac.uk**

Silia Vitoratou: **silia.vitoratou@kcl.ac.uk**

Topic materials:
Dr Raquel Iniesta

Narration and contribution:
Zahra Abdula

Improvements:
Nick Beckley-Hoelscher
Kim Goldsmith
Sabine Landau

Institute of Psychiatry, Psychology and Neuroscience
Department of Biostatistics and Health Informatics

KING'S College LONDON

Module Title: Introduction to Statistics

Session Title: Confounding

Topic title: Multiple regression with several explanatory variables: Adjusting for confounders

Learning Outcomes

After listening to this session you should be able to:-

- Understand what is meant by “confounding” or “confounders”.
- Be aware that confounding is a theory and always involves at least three variables, an exposure-outcome relationship of interest and a third variable that is thought to be cause of both the exposure and the outcome.
- Understand the statistical problem caused by the existence of confounding.
- Know how to use multiple linear regression models to adjust for potential confounding variables.

Confounding Variables

Confounding: A situation in which the association between an explanatory variable (e.g exercise $\boldsymbol{x_1}$) and outcome (e.g. weight $\boldsymbol{y}$) is distorted by the presence of another variable (e.g. hours of free time $\mathbf{x_2}$).

Theory:

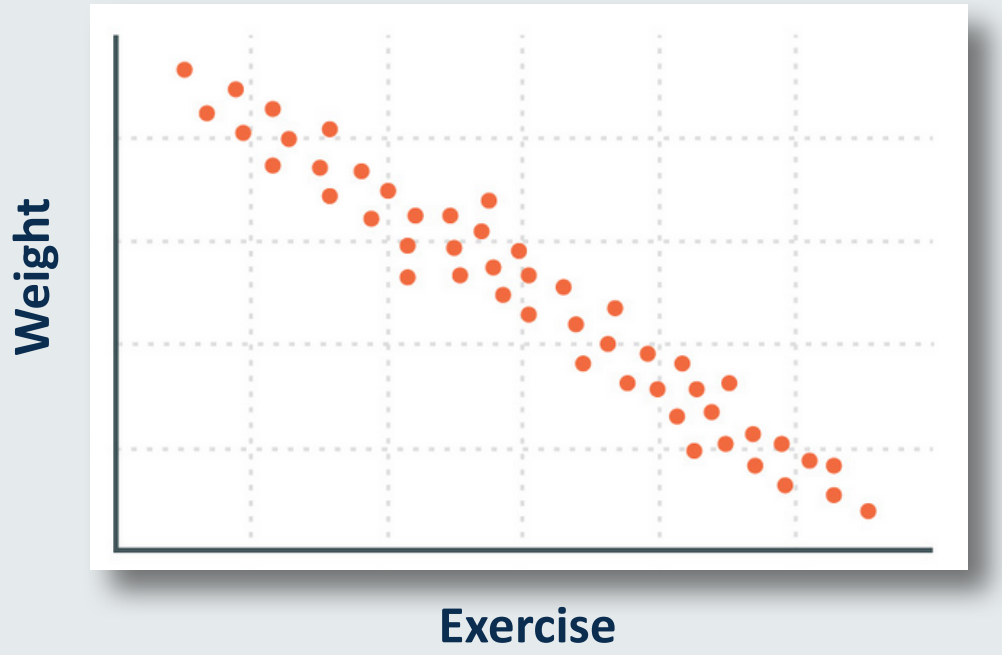
hours of free time causes **more exercise**
hours of free time also affects **weight**

i.e. hours of free time is a **common cause** of the exposure (exercise) and outcome (weight) of interest.

What happens if we only test the relationship between exercise ($\boldsymbol{x_1}$) and weight ($\boldsymbol{y}$) in a simple linear regression model with only exercise as an independent variable?

Simple linear regression result: unadjusted relationship

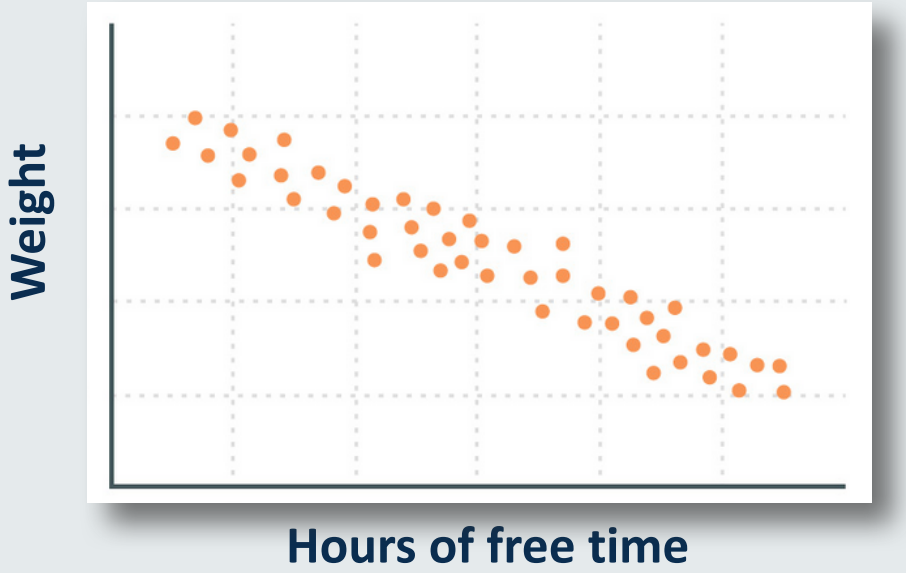
What happens if we only assess the relationship between exercise (x_1) and weight (y) in a simple linear regression model with only exercise as an independent variable?



$y = 70 - 5x_1 + \varepsilon$; p=0.01 for β_1

Simple linear regression results: possible relationship between hours of free time and outcome of interest

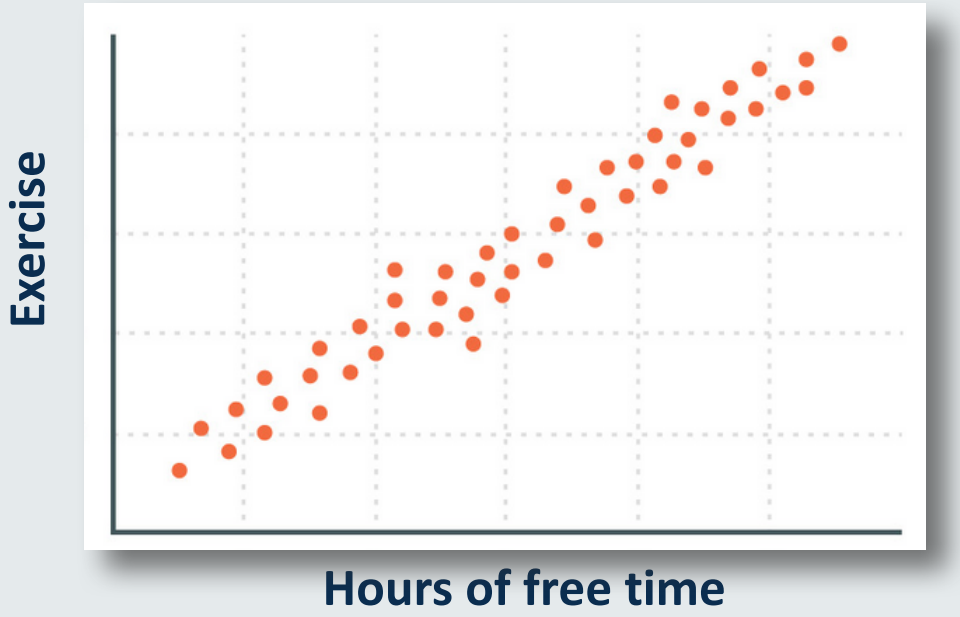
What happens if we only assess the relationship between hours for free time (x_2) and weight (y) in a simple linear regression model with only hours free time as an independent variable?



$y = 69 - x_2 + \varepsilon$; p=0.001 for β_2

Simple linear regression: possible relationship between hours of free time and exposure of interest

Note that exercise is also associated with hours of free time ... so our two potential independent variables are also associated



$Y = 0.5x + \varepsilon$; p=0.015

Confounding Problem

It appears:

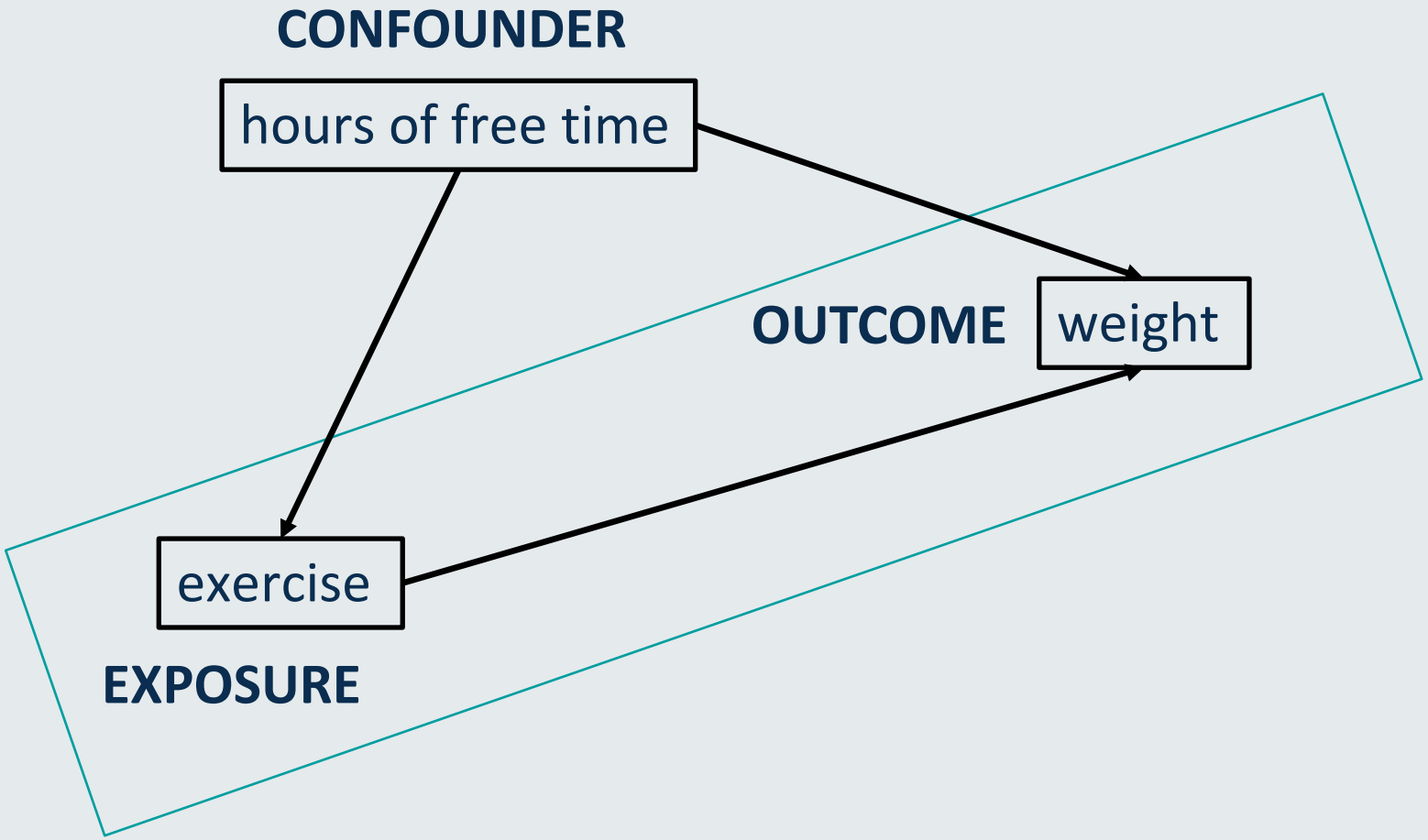
- Weight might go down with more free time.
- Those with more free time might exercise more.

So even if more exercise did not cause any weight loss, we might still observe an association between exercise and weight.

In the presence of such a common cause, we **cannot attribute all the observed association** between the exposure and the outcome **to the exposure causing the outcome**.

This is known as **confounding**; or in other words free time is a **confounder** of the effect of exercise on weight.

How Does Confounding Work?



Confounding Variables: Explanation

In an experiment, the **independent variable** is typically thought to **cause** your dependent variable.

Example: If you are researching whether lack of exercise leads to weight gain:
Exercise is your independent variable and weight gain is your dependent variable.

Confounding variables are any other variables that cause both your dependent and your main independent variable of interest.

They are like extra independent variables that are having a **hidden** effect on your dependent variables while being related to the independent variable of interest.

If not taken into account, confounding variables will **introduce bias** in the estimation of β_1.

Multiple Linear Regression Model: Confounding

We can formulate the model in terms of confounding

The researcher's ultimate goal is to be able to estimate the effect of an independent variable (or exposure) on a dependent variable (or outcome) while adjusting for other variables that distort this relationship (the confounders).

If we have an independent variable we are interested in – we want to get an **adjusted estimate** of the association between this independent and the dependent variable.

Using multiple linear regression allows us to hold all other independent variables constant allowing us to get an estimate of the effect of the independent variable of interest while adjusting for other variables in the model which are hypothesized to be confounders.

Multiple Regression Model: Adjust for Confounding

$$\boldsymbol{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon}$$

- Multiple regression framework is a **natural** and **practical** way of **adjusting for confounders**.

- To **adjust the association** between exercise (x_1) and weight (y) for hours of free time (x_2), **all we need to do is include the confounder** hours of free time (x_2) in the regression model as an **additional predictor**, which will automatically adjust the y and x_1 association for x_2.

- The coefficient of x_1 (i.e., β_1) in this case will represent the **adjusted association** between weight and exercise, controlling for the effect of the number of hours of free time.

Adjusting for One Confounder

$y = 72 - 3x_1 - x_2 + \varepsilon$

		p-value
Slope for x_1 (β_1)	-3	0.04
Slope for x_2 (β_2)	-1	0.01

Where:

y=weight;
x_1=frequency of exercise per week;
x_2=hours of free time per week;

The effect of exercise on weight **adjusted for** hours of free time is β_1=-3.
This effect is statistically significant (p=0.04).
We can infer the association for the whole population.

As I increase number of weekly exercise sessions by 1, I am decreasing my weight by 3kg, keeping the hours of free time per week fixed.

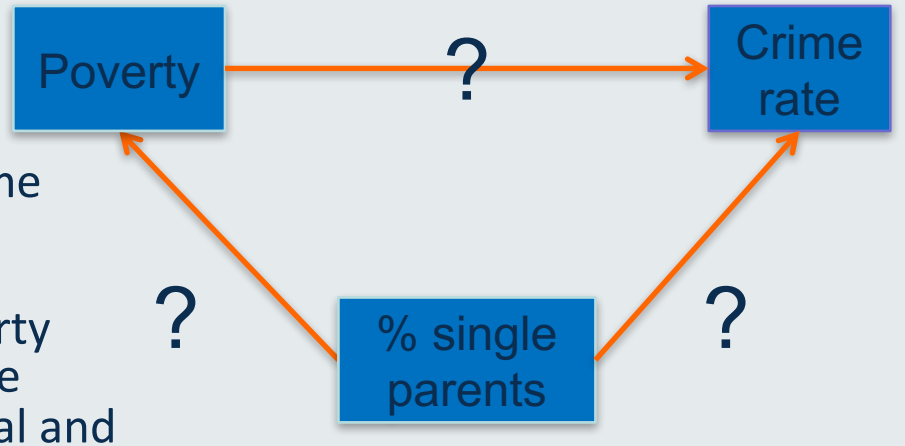
Dealing with Multiple Confounders

- Multiple regression model can deal with any number of confounders (within reason).
- All confounders are adjusted for by including them simultaneously as additional predictors.
- However, the sample size can restrict the number of variables that can be included in a regression model.
- A rule of thumb is to ensure that there are more than 10 observations (data points) per each independent variable.
- A sample of size 100, for example, will allow us to consider up to 10 independent variables.

Knowledge Check – Confounding

Use the lecture_7_data.sav dataset on Keats

We want to look at the relationship between poverty and crime rate and consider % single parents in the State as a confounder of the poverty – crime rate relationship.



Q1: Fit a simple linear regression model for the relationship between poverty (independent) and crime rate (dependent). Write down the estimate of the regression coefficient, β_1, for poverty, as well as the 95% confidence interval and p-value for β_1.

Q2: To check whether the data is consistent with the theory that % single parents is a confounder, we assess whether it is associated with both our independent and dependent variables of interest. Fit two simple linear regression equations: poverty (dependent) and % single parents (independent), and crime rate (dependent) and % single parents (independent). Write down the estimates of the regression coefficients, the β_1 (and 95% confidence intervals and p-values) for % single parents from both equations.

Q3: Does it seem that % single parents is likely to be a confounder of the poverty – crime rate association? Why or why not? What would you need to do to account for confounding by % single parents in assessing the poverty – crime rate association?

Knowledge Check Solutions

Q1:

B_1 for poverty and crime rate is 25.452, 95% CI 6.833 to 44.072, $p = 0.008$

Q2:

B_1 for % single parents and poverty is 1.250, 95% CI 0.489 to 2.012, $p = 0.02$

B_1 for % single parents and crime rate is 130.110, 95% CI 85.809 to 174.411, $p < 0.001$

Q3:

As % single parents is associated both with our independent (poverty) and dependent (crime rate) variables of interest, the data set confirms that it acts as a confounder of the poverty – crime rate relationship.

To deal with this, we would want to include this variable in a multiple regression model where both poverty and % single parents were independent variables to get an estimate of this relationship **adjusted for confounding**.

References

Agresti, A., & Finlay, B. (2009).
Statistical Methods for the Social Sciences (4th ed.). New Jersey, NJ: Prentice Hall Inc.

Douglas, C., Montgomery, D. C., Peck, E. A., & Vining, G. G. (2006).
Introduction to Linear Regression Analysis. New York, NY: Wiley.

Thank you

Please contact your module leader or the course lecturer of your programme, or visit the module's forum for any questions you may have.

If you have comments on the materials (spotted typos or missing points) please contact Dr Iniesta:

Raquel Iniesta, PhD
Department of Biostatistics and Health Informatics
IoPPN, King's College London, SE5 8AF, London, UK
raquel.iniesta@kcl.ac.uk

For any other comments or remarks on the module structure, please contact one of the three module leaders of the Biostatistics and Health Informatics department:

Zahra Abdula: **zahra.abdulla@kcl.ac.uk**

Raquel Iniesta: **raquel.iniesta@kcl.ac.uk**

Silia Vitoratou: **silia.vitoratou@kcl.ac.uk**

Topic materials:
Dr Raquel Iniesta
Department of Biostatistics and Health Informatics

Narration and contribution:
Zahra Abdula

Improvements:
Nick Beckley-Hoelscher
Kim Goldsmith
Sabine Landau

Institute of Psychiatry, Psychology and Neuroscience

KING'S College LONDON

Module Title: Introduction to Statistics

Session Title: Prediction and Model Fit

Topic title: Multiple regression with several explanatory variables: Adjusting for confounders

Learning Outcomes

After working through this session you should be able to:

- Use the multiple linear regression model as a tool for prediction.
- Use multiple linear regression models to obtain predicted values of dependent variables given a regression equation and values of the independent variables.
- Assess the fit of your model / quality of your prediction model.
- Understand the difference between the standard coefficient of determination R^2 and its adjusted version R^2_{adj}.

Multiple Linear Regression Model: Prediction

We can formulate the model in terms of prediction

The researcher's ultimate goal is to be able to predict the value for a **dependent variable** given **a set of other variables**.

Independent variables can also be known as
Explanatory variables and also as
Predictor variables

A multiple linear regression model can help us find the **factors useful** for the clinician to **predict…**

E.g. weight that a person can reach if he/she does not follow recommendations on habits like diet, water, exercise.

Example: Using the Model to Predict

$y = 72 - 4x_1 - 2x_2 + \varepsilon$

		p-value
Slope for x_1 (β_1)	-4	0.01
Slope for x_2 (β_2)	-2	0.03

Where:

y=weight;
x_1=frequency of exercise per week;
x_2=frequency of vegetables per day;

Use the model to predict the weight for a person who exercises 3 times a week and normally has vegetables 2 times a day, i.e.

x_1=3
x_2=2

$y = 72 - (4 \times 3) - (2 \times 2)$
$y = 72 - 12 - 4$
$\hat{y} = 56\text{kg}$

The model predicts a weight of 56kg for a person who does physical activity 3 times a week and normally has vegetables 2 times a day.

R^2 – The Coefficient of Determination

- The **coefficient of determination**, denoted R^2 and pronounced R-Squared, is a statistical measure of how well the regression line/hyperplane approximates the real data points.
- It is also known as a measure of **goodness of fit:** The goodness of fit of any statistical model describes how well it fits a set of observations.
- $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$ where SS = sum of squares, **res** = residuals (or errors) and **tot** = total
- R^2 ranges from 0 to 1.
  - R^2 of 0 indicates poor fit; the regression line would be perfectly horizontal.
  - R^2 of 1 indicates perfect fit; the regression line/hyperplane fit exactly to all data points.

R^2 – continued

- R^2 measures the fit of the model both in simple and multiple linear regression.

- In **simple linear regression $R^2 = r^2$**, where r is the Pearson correlation.

- In a context of regression where we are assessing **associations between variables**, R^2 is often interpreted as the proportion of the variance in the dependent variable that is "explained" by the independent variables in the model.
  - In our earlier example, this would be the proportion of variance in the weight that is explained by frequency of exercise and hours of free time.
  - R^2 of 0 indicates that none of the variance in y is explained.
  - R^2 of 1 indicates that 100% of the variance in y is explained.

- In a context of **prediction analysis**, R^2 is often interpreted as how well the model will be able to predict values of Y based on observed values for the independent variables x_i; with higher values of R^2 indicating better prediction.

What R^2 Does Not Indicate

R^2 **does not indicate** whether:

- the independent variables are a **cause** of the changes in the dependent variable;
  - (we can only say the variables are associated, not that one causes the other)
- the correct type of regression was used;
- the most appropriate set of independent variables have been chosen;
- there are enough data points to make a solid conclusion.

Adjusted R^2 as a Measure for Model Selection

Adjusted R^2 (denoted R^2_{adj}) is a modified version of R^2 that adjusts for the **number of independent variables *p*** in the model:

$$\boldsymbol{R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}}$$

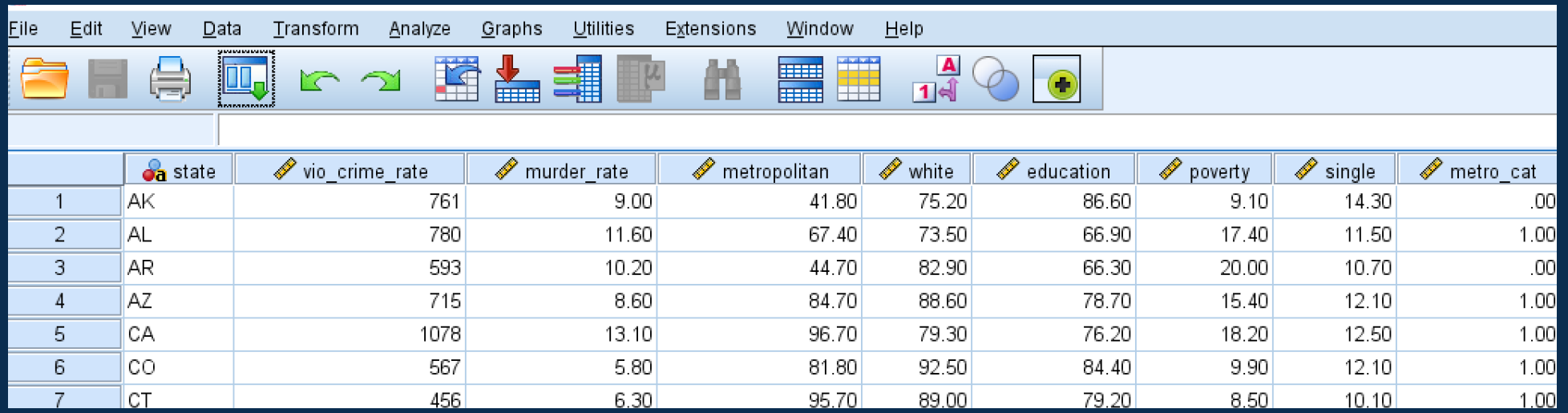
R^2_{adj} takes account of the phenomenon whereby R^2 **increases** every time an **extra independent variable** is added regardless of whether this added variable adds substantially to the explanation of dependent variable variance.

R^2_{adj} increases only when the increase in R^2 (due to the inclusion of a new independent variable) is more than one would expect to see by chance.

$R^2_{\text{adj}}$is considered to be a **better indicator for model selection:** between different models, the one with **higher** R^2_{adj} is the one that better fits the data, and should be selected.

SPSS Slide

Download the data that we are going to use during the lecture. The dataset is the **lecture_7_data.sav**.



	state	vio_crime_rate	murder_rate	metropolitan	white	education	poverty	single	metro_cat
1	AK	761	9.00	41.80	75.20	86.60	9.10	14.30	.00
2	AL	780	11.60	67.40	73.50	66.90	17.40	11.50	1.00
3	AR	593	10.20	44.70	82.90	66.30	20.00	10.70	.00
4	AZ	715	8.60	84.70	88.60	78.70	15.40	12.10	1.00
5	CA	1078	13.10	96.70	79.30	76.20	18.20	12.50	1.00
6	CO	567	5.80	81.80	92.50	84.40	9.90	12.10	1.00
7	CT	456	6.30	95.70	89.00	79.20	8.50	10.10	1.00

The dataset contains data from 51 US states, measuring the crime rates and background measures for each state with respect to their

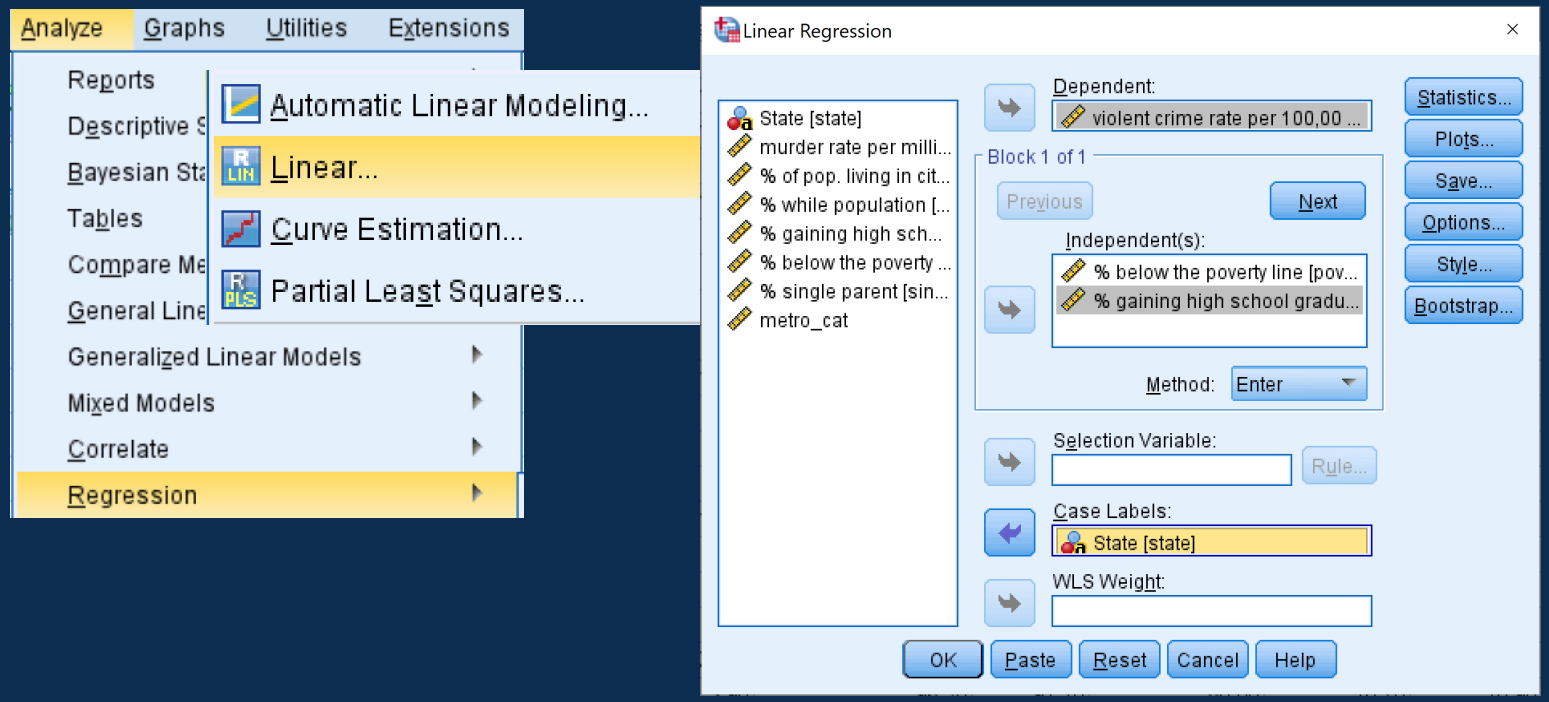
- **violent crime**: per 100,000 population
- **murder**: per 100,000 population
- **poverty**: percent below the poverty line
- **single**: percentage of lone parents

SPSS Slide: ‘how to’

Researchers believe, in the population from which our data came, the % below the poverty line and % gaining a high school graduation have and effect on the Violent Crime rate

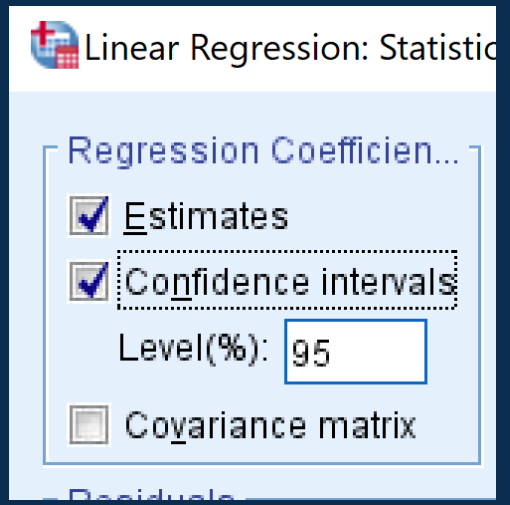
Step 1) Computing **R^2** for a multiple linear regression model with dependent variable ‘crime’ and independent variables ‘poverty’ and ‘education’ from practical_7_data.sav data

Use **Analyse -> Regression -> Linear**



Put ‘**crime**’ in ‘dependent’, and ‘**poverty**’ and ‘**education**’ in ‘independent’.

Click **Statistics**, select ‘**Confidence intervals**’.



SPSS Interpretation Slide

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	345.852	1026.638		.337	.738	-1719.478	2411.181
	% below the poverty line	23.927	14.763	.347	1.621	.112	-5.774	53.627
	% gaining high school graduation	-1.502	11.239	-.029	-.134	.894	-24.112	21.109

a. Dependent Variable: violent crime rate per 100,000 population

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.369[a]	.136	.100	280.763

a. Predictors: (Constant), % gaining high school graduation, % below the poverty line

The linear multiple regression model has an R^2_{adj} of 0.100. Poverty and education explained 10.0% of the variance in violent crime.

Knowledge Check – Prediction

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	3.813	.334		11.411	.000	3.158	4.469
	Reading Test Score	-.080	.013	-.184	-6.000	.000	-.106	-.054
	Sex	1.410	.174	.248	8.093	.000	1.068	1.752

a. Dependent Variable: Malaise Score at Age 22

This analysis was done using the Lecture_6_data (NCDS Data) dataset.

It shows the result of fitting a multiple linear regression model with malaise score at age 22 as the dependent variable, with reading and sex as independent variables (sex coded as 0 = male, 1 = female).

Q1: Write out the regression equation, both in terms of *Y* and *X* as well as using the variable names.

Q2: What is the predicted malaise score at age 22 for a female with a reading test score of 11?

Q3: What is the predicted malaise score at age 22 for a male with a reading test score of 28?

Knowledge Check Solutions - Prediction

Q1:

Regression equation:

$y = 3.813 – 0.08(x_1) + 1.410(x_2)$

Malaise score at age 22 = 3.813 – (0.08 x reading score) + (1.410 x sex)

Q2:

Malaise at age 22 = 3.813 – (0.08 x 11) + (1.410 x 1)

Malaise score at age 22 = 4.343

Q3:

Malaise at age 22 = 3.813 – (0.08 x 28) + (1.410 x 0)

Malaise score at age 22 = 1.573

Knowledge Check - R^2

The Psychosis department at the IoPPN is investigating whether quality of life in people diagnosed with schizophrenia depends on a series of demographic and clinical variables. They have asked us to help them choose among different models.

Q4: Which one should they keep as the best model?

Dependent variable:
Quality of Life (QoL) measured with QOLS scale (ranging from 16 to 112)
Independent variables:
Severity of illness, age, gender (1=female), marital status (1=married)

Model	*y*	β_0	Severity β_1 (p-value)	Age β_2 (p-value)	Gender β_3 (p-value)	Marital Status β_4 (p-value)	R^2_{adj}
I	QOLS	50	-3.4 (0.01)	-2.1 (0.10)	Not included	5.1 (0.001)	0.73
II	QOLS	47	Not included	-1.8 (0.07)	1.03 (0.13)	6.2 (0.002)	0.51
III	QOLS	56	-3.1 (0.02)	Not included	Not included	5.3 (0.001)	0.85

Knowledge Check Solutions – R^2

Q4: Which one should they keep as the best model and why?

The best model is the model III with Severity of illness and status as the independent variables. This is because we see from the adjusted R^2 that it explains 85% of the variability in quality of life. If we compare to model I we can see that adding age decreased the adjusted R^2 – this makes sense in combination with the fact that age doesn't seem to be a significant predictor of quality of life. Model II has a lower adjusted R^2 because it is missing the important severity predictor.

Model	y	β_0	Severity β_1 (p val)	Age β_2 (p val)	Gender β_3 (p val)	Marital Status β_4 (p val)	R^2_{adj}
I	QOLS	50	-3.4 (0.01)	-2.1 (0.10)	Not included	5.1 (0.001)	0.73
II	QOLS	47	Not included	-1.8 (0.07)	1.03 (0.13)	6.2 (0.002)	0.51
III	**QOLS**	**56**	**-3.1 (0.02)**	**Not included**	**Not included**	**5.3 (0.001)**	**0.85**

References

Agresti, A., & Finlay, B. (2009).
Statistical Methods for the Social Sciences (4th ed.). New Jersey, NJ: Prentice Hall Inc.

Douglas, C., Montgomery, D. C., Peck, E. A., & Vining, G. G. (2006).
Introduction to Linear Regression Analysis. New York, NY: Wiley.

Thank you

Please contact your module leader or the course lecturer of your programme, or visit the module's forum for any questions you may have.

If you have comments on the materials (spotted typos or missing points) please contact Dr Iniesta:

Raquel Iniesta, PhD
Department of Biostatistics and Health Informatics
IoPPN, King's College London, SE5 8AF, London, UK
raquel.iniesta@kcl.ac.uk

For any other comments or remarks on the module structure, please contact one of the three module leaders of the Biostatistics and Health Informatics department:

Zahra Abdula: **zahra.abdulla@kcl.ac.uk**

Raquel Iniesta: **raquel.iniesta@kcl.ac.uk**

Silia Vitoratou: **silia.vitoratou@kcl.ac.uk**

Topic materials:
Dr Raquel Iniesta

Narration and contribution:
Zahra Abdula

Improvements:
Nick Beckley-Hoelscher
Kim Goldsmith
Sabine Landau

Institute of Psychiatry, Psychology and Neuroscience
Biostatistics and Health Informatics

KING'S College LONDON

Module Title: Introduction to Statistics

Session Title: Checking Regression Model Assumptions

Topic title: Multiple regression with several explanatory variables: Adjusting for confounders

Learning Outcomes

After listening to this session you should be able to:

- List the assumptions that need to be met when fitting linear regression models.
- Know how to check using your data that these assumptions hold.
- Understand what a partial plot or a residual plot tells you.

Assumptions for Multiple Regression Inference

1. **The relationship** between the dependent (*Y*) and each continuous independent variable (*x* variables) is **linear.**

   Obtain **scatterplots** of:
   residuals of the dependent variable (*Y*) plotted against
   residuals of each independent variable (*x*) in turn
   when **both** variables are regressed **separately** on **the rest of the independent variables**.

   These plots will show the relationship between y and that specific x with **effects of other x's removed**.

   Partial residual plots show the **net relationship** where the influence of other variables is **partialled out**.

   At **least two independent variables** must be in the equation for a partial plot to be produced.

Assumptions for Multiple Regression Inference

2. **Residuals** or error terms ε should be approximately **normally distributed.**

   A common misconception about linear regression is that it assumes that the dependent variable *Y* is normally distributed. Actually, linear regression assumes normality for the residual errors ε, which represent variation in *Y* which is not explained by the predictors.

   We can plot a **histogram** of the **error terms** to see if the errors more or less follow a normal distribution

   Or we can use a **normal P-P plot**, which plots the data against a theoretical normal distribution, and check that the points more or less follow a **straight line**.

Assumptions for Multiple Regression Inference - continued

3. Homoscedasticity (stability in variance of residuals):

A **scatterplot** of **standardised residuals** ε and **standardised predicted** values shows no pattern.

This equates to the error terms having the same variance irrespective of the values of x (i.e., variance does not depend on x).

Also called “homoscedasticity”.

Assumptions for Multiple Regression Inference - continued

4. **Independent observations:**

   Note: this is an aspect of design rather than an assumption that can be tested

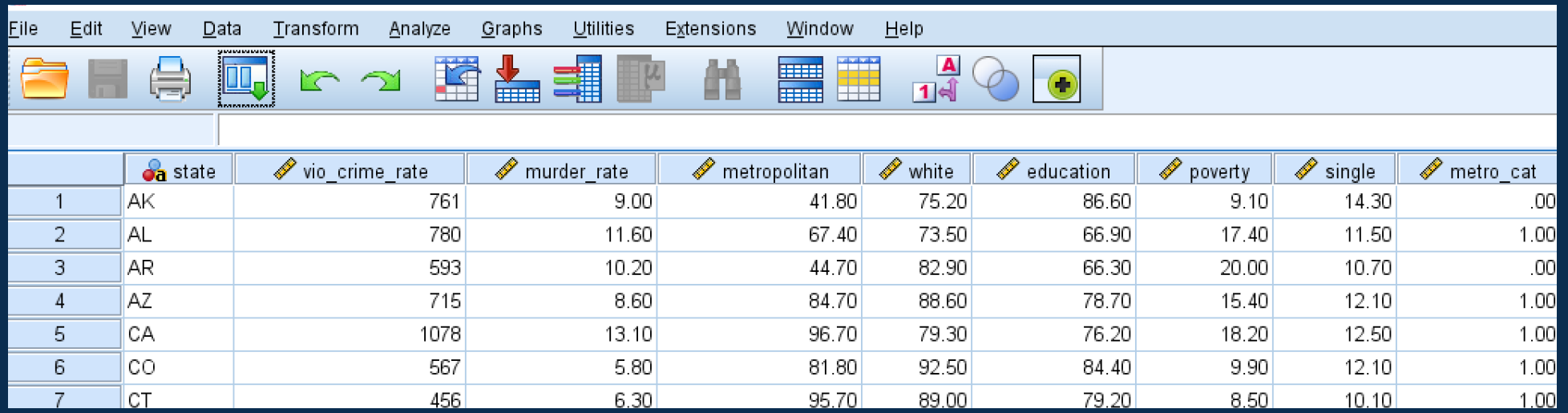
   e.g. repeated measurements collected on people over time are **not** independent

   e.g. measurements with a natural pairing of individuals (twins, couples), or matching are **not** independent

   In these latter situations, refer to earlier lectures for methods for dealing with paired data; regression models for analysing such data are beyond the scope of the course.

SPSS Slide

Download the data that we are going to use during the lecture. The dataset is the **lecture_7_data.sav**.



	state	vio_crime_rate	murder_rate	metropolitan	white	education	poverty	single	metro_cat
1	AK	761	9.00	41.80	75.20	86.60	9.10	14.30	.00
2	AL	780	11.60	67.40	73.50	66.90	17.40	11.50	1.00
3	AR	593	10.20	44.70	82.90	66.30	20.00	10.70	.00
4	AZ	715	8.60	84.70	88.60	78.70	15.40	12.10	1.00
5	CA	1078	13.10	96.70	79.30	76.20	18.20	12.50	1.00
6	CO	567	5.80	81.80	92.50	84.40	9.90	12.10	1.00
7	CT	456	6.30	95.70	89.00	79.20	8.50	10.10	1.00

The dataset contains data from 51 US states, measuring the crime rates and background measures for each state with respect to their

- **violent crime**: per 100,000 population
- **murder**: per 100,000 population
- **poverty**: percent below the poverty line
- **single**: percentage of lone parents

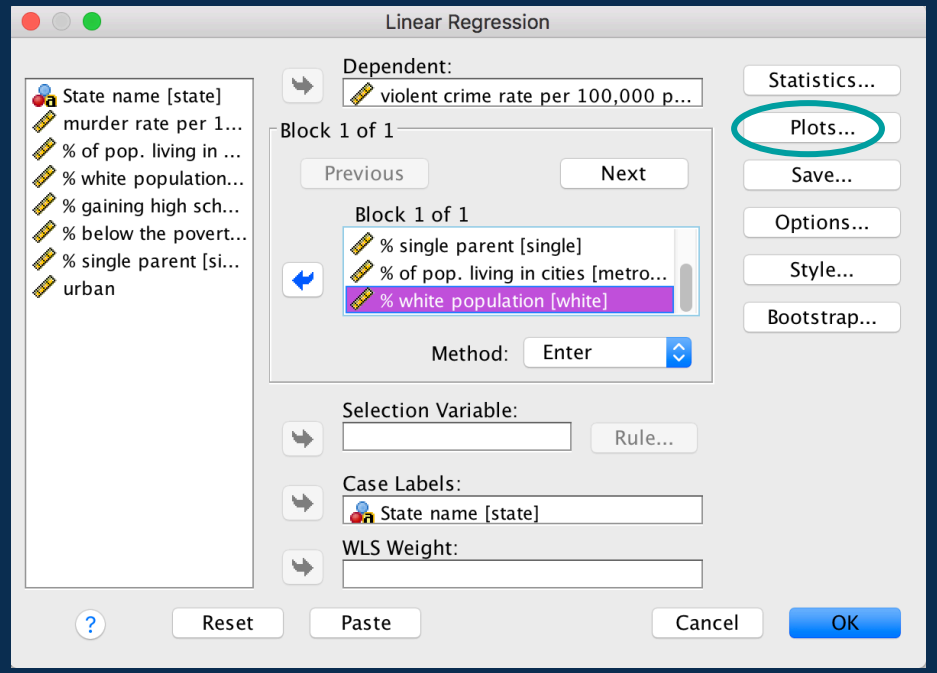
SPSS Slide: ‘how to’

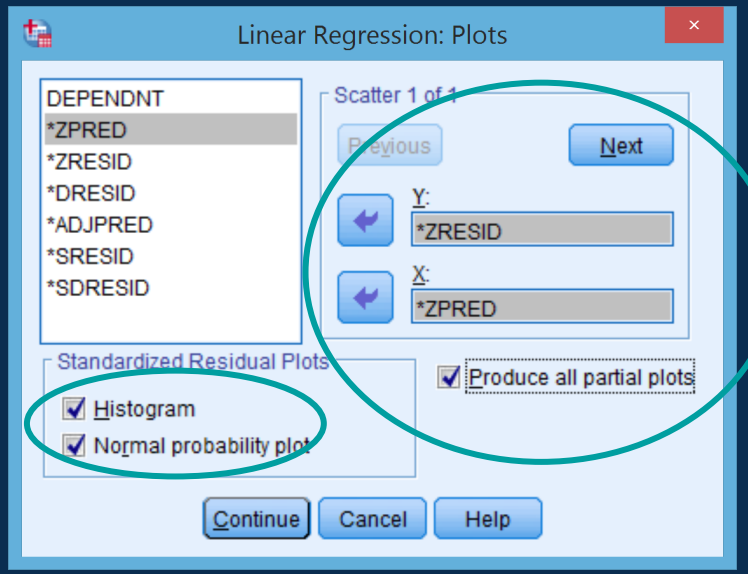
Assessing assumptions to make inference from a multiple linear regression model for crime rate from Lecture_7_data.sav **data.**

Use **Analyse -> Regression -> Linear**

Put ‘**crime**’ in **dependent**, and **poverty**, **education**, **single**, **metropol** and **white** in ‘**independent**’.

Click ‘**Plots**’, select ‘Histogram’, ‘Normal probability plot’, ‘Produce all partial plots’, put ZRESID (standardised residuals) in Y and ZPRED (standardised predicted values) in X.



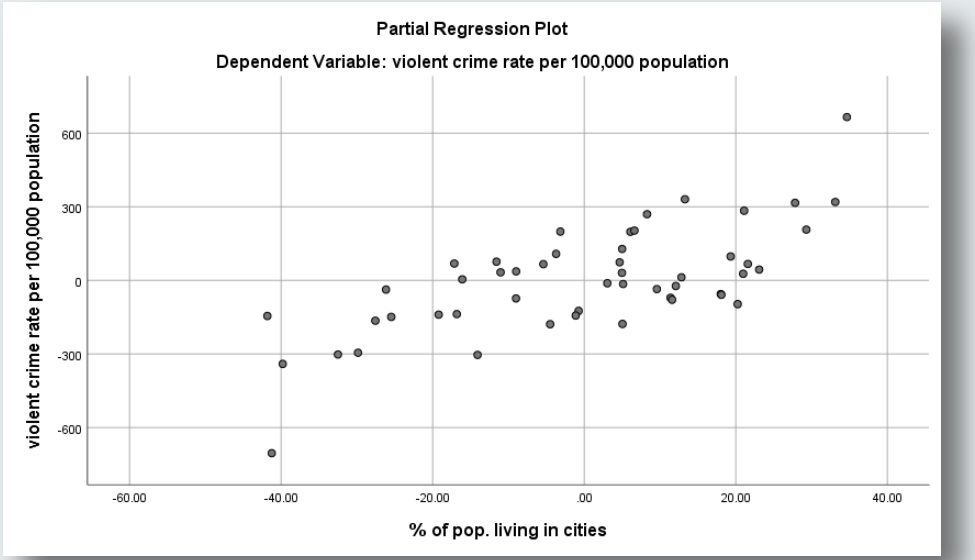


Output and Interpretation Slide – Assessing Linearity

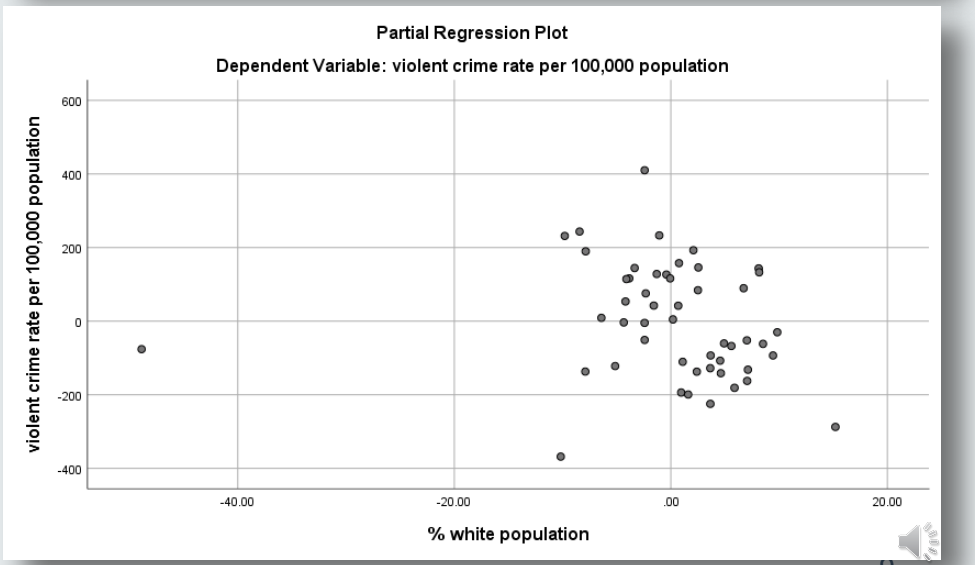
Assumption #1

Partial residual plots from the regression of **crime** on **poverty, edu, single, metropol**, and **white –** note only two of the five partial plots are shown:

Top plot suggests a linear relationship between the independent variable **metropol** and the outcome variable **crime** – the **linearity assumption is met** for the **metropol** variable



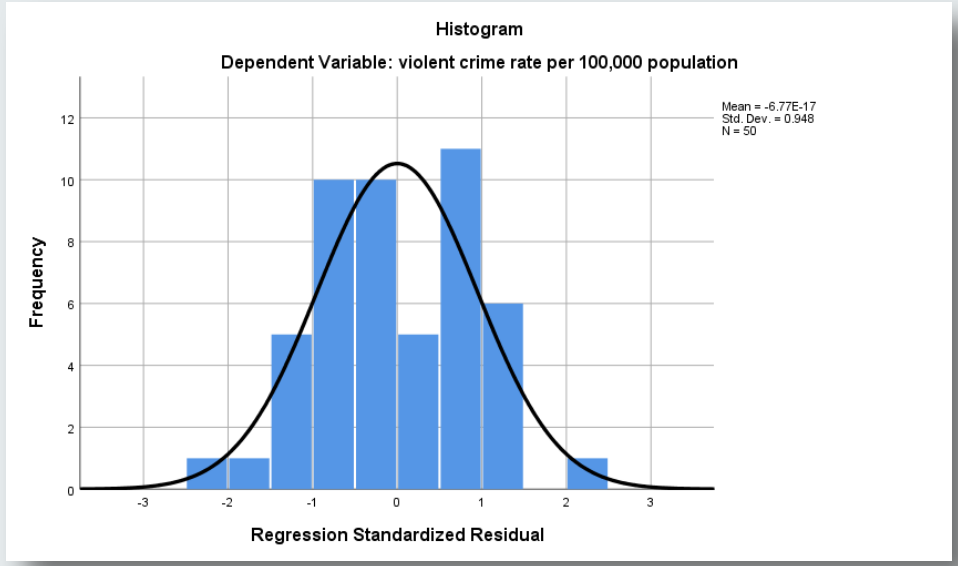
Bottom plot suggests there is not a linear relationship between the independent variable **white** and the outcome variable **crime** – **the linearity assumption is not met** for the **white** variable



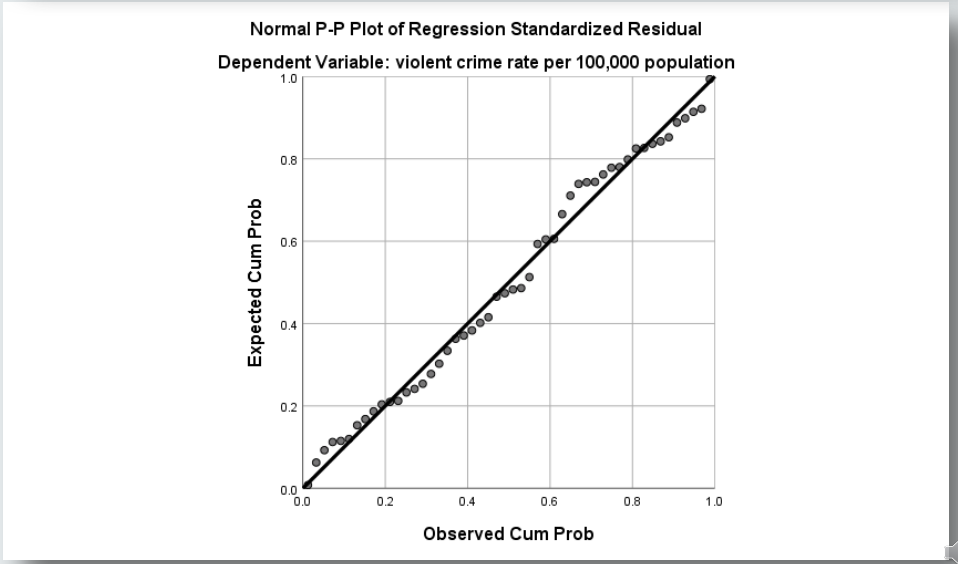
SPSS Interpretation Slide – Plots of Residuals for Assessing Normality

Assumption #2

Histogram – a gap at the right and possibly somewhat skewed, but errors/residuals look more or less normally distributed.



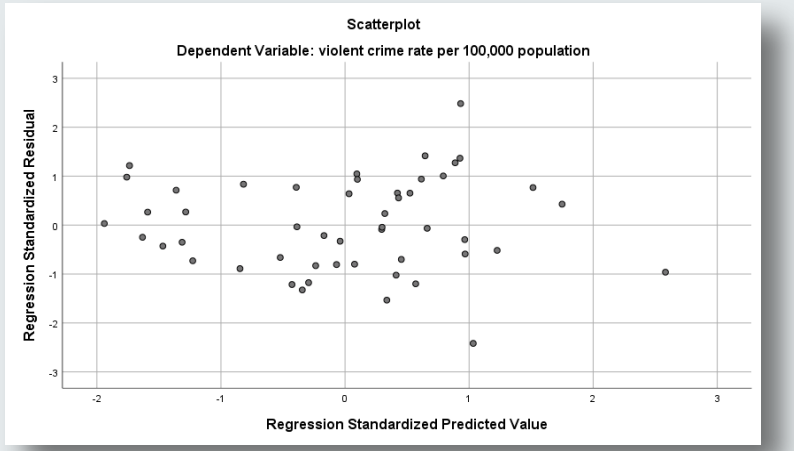
Normal P-P plot – gives similar information to the histogram; here we want to see that the points lie more or less close to the diagonal reference line, which they do.



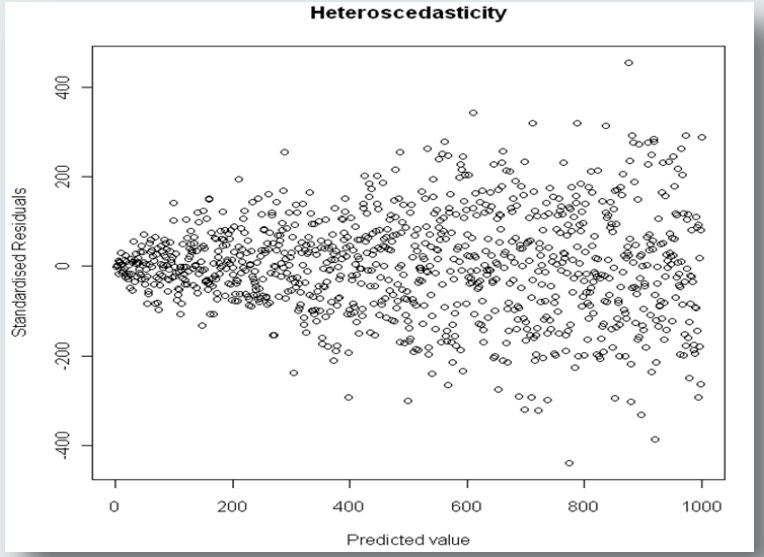
SPSS Interpretation Slide – Assessing Variance Homogeneity

Assumption #3

Top plot = **homoscedastic** = meets assumption.
There is no obvious trend, the residuals **scatter randomly** above and below zero. The scatter around the horizontal zero line is roughly constant, suggesting a constant variance. We can assume homogeneity and make inferences from our regression model.



Bottom plot = **heteroscedastic** = does not meet assumption.
Here the residual variance **increases** with the size of the predicted value. Homoscedasticity cannot be assumed, we cannot make inferences from our model, and would need to use approaches beyond the scope of this course. This is an example of heteroscedasticity.



Knowledge Check - Model Assumptions

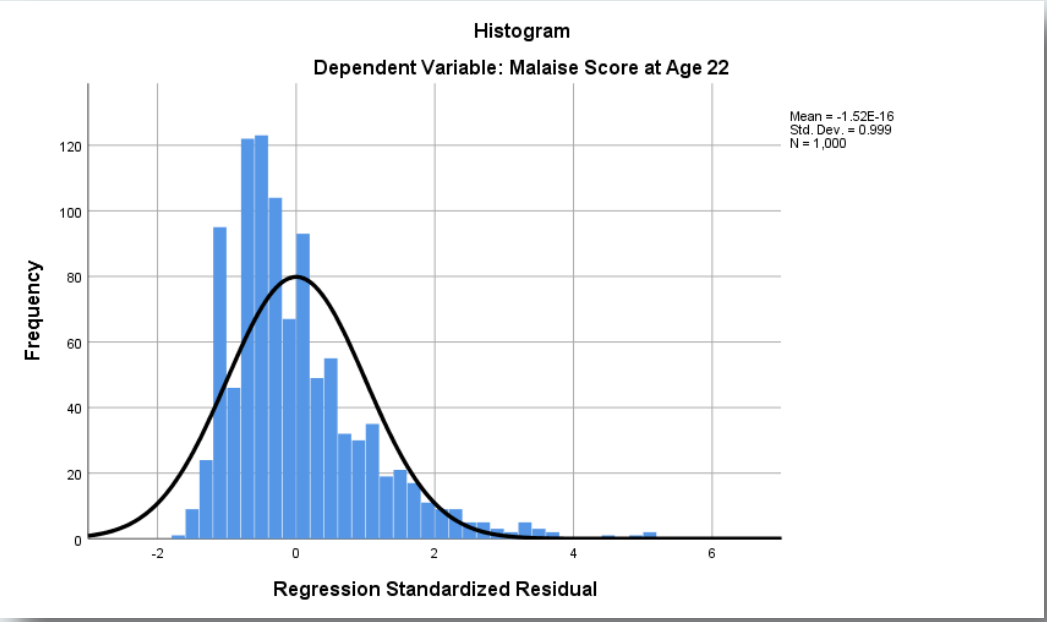
Use the Lecture_6a_data.sav NCDS dataset,

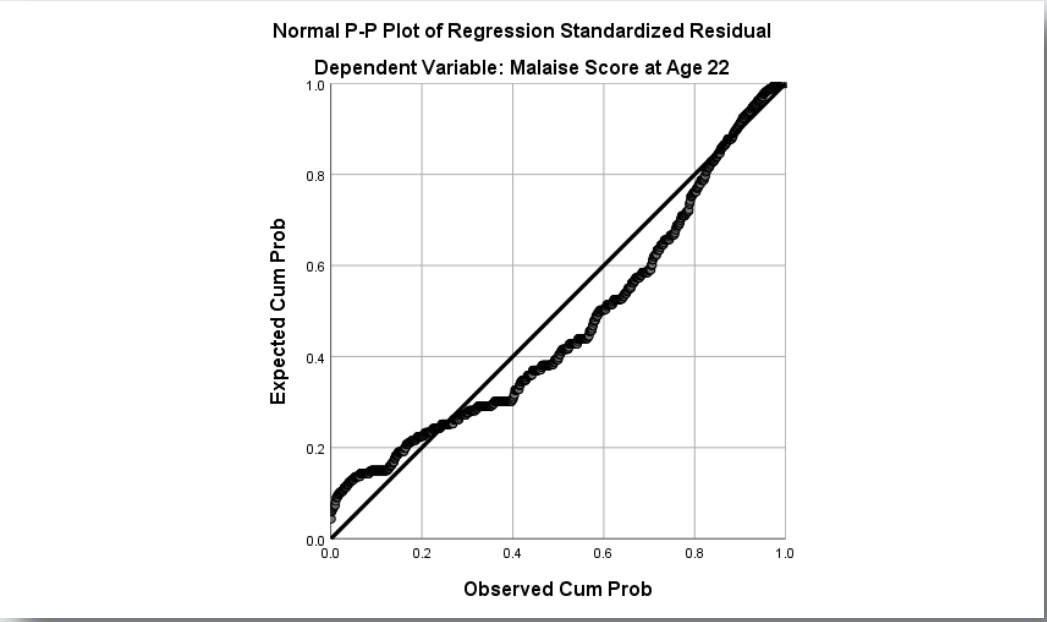
Researchers want to understand if the multiple regression model they have run on malaise score at age 22 as the dependent variable, with reading and sex as independent variables (sex coded as 0 = male 1 = female) meets the model assumptions.

Q: Fit this model and assess the linearity, normality of residuals and homoscedasticity assumptions. Can we make these assumptions for this model or not?

Knowledge Check Solutions – Model Assumptions

- Q: Fit this model and assess the normality of residuals and homoscedasticity assumptions. Can we make these assumptions for this model or not?

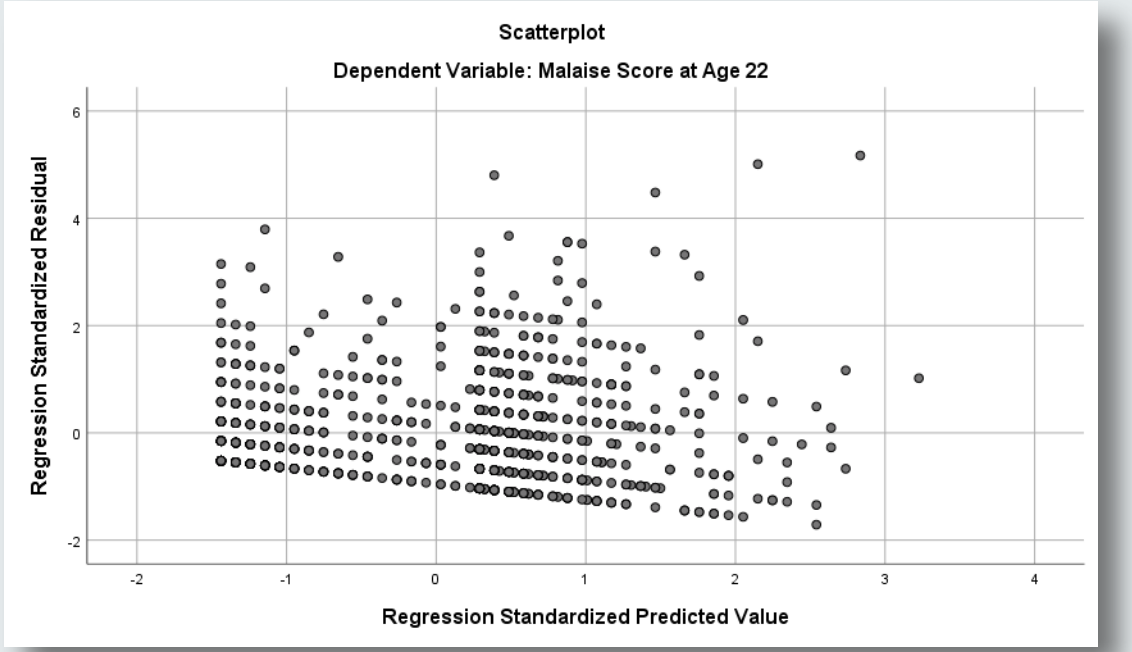




- The residuals seem to follow a skewed distribution with a longer tail to the right side of the histogram.
- The P-P plot also indicates a skewed distribution, with the points not falling along the straight reference line

Knowledge Check Solutions – Model Assumptions

- **Q: Fit this model and assess the normality of residuals and homoscedasticity assumptions. Can we make these assumptions for this model or not?**



The residual vs predicted value scatterplot shows some fanning out/increased variance, i.e. the errors do not have constant variance across the range of predicted values and are heteroscedastic.

Knowledge Check Solutions – Model Assumptions

- **Q: Fit this model and assess the normality of residuals and homoscedasticity assumptions. Can we make these assumptions for this model or not?**

- **The partial plot for the relationship between malaise and reading score does not form an approximate straight line/does not appear to be linear.**
- **The model assumptions don't appear to be fully met in this analysis.**

Suggested Reading

Field (2017) Discovering Statistics using SPSS, 5th Ed.

Chapter 8: Correlation

Chapter 9: The Linear Model (Regression)

Agresti and Finlay (2014) Statistical Methods for the Social Sciences, 4th Ed.

Chapter 9: Linear Regression and Correlation

Chapter 10: Introduction to Multivariate Relationships

Chapter 11: Multiple Regression and Correlation

Acock (2018) A Gentle Introduction to Stata, 6th Ed.

Chapter 8: Bivariate correlation and regression

Thank you

Please contact your module leader or the course lecturer of your programme, or visit the module's forum for any questions you may have.

If you have comments on the materials (spotted typos or missing points) please contact Dr Iniesta:

Raquel Iniesta, PhD
Department of Biostatistics and Health Informatics
IoPPN, King's College London, SE5 8AF, London, UK
raquel.iniesta@kcl.ac.uk

For any other comments or remarks on the module structure, please contact one of the three module leaders of the Biostatistics and Health Informatics department:

Zahra Abdula: **zahra.abdulla@kcl.ac.uk**

Raquel Iniesta: **raquel.iniesta@kcl.ac.uk**

Silia Vitoratou: **silia.vitoratou@kcl.ac.uk**