



Topic materials:
Dr Raquel Iniesta



Narration and contribution:
Zahra Abdula

Improvements:
Nick Beckley-Hoelscher
Kim Goldsmith
Sabine Landau

Institute of Psychiatry, Psychology and Neuroscience

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.



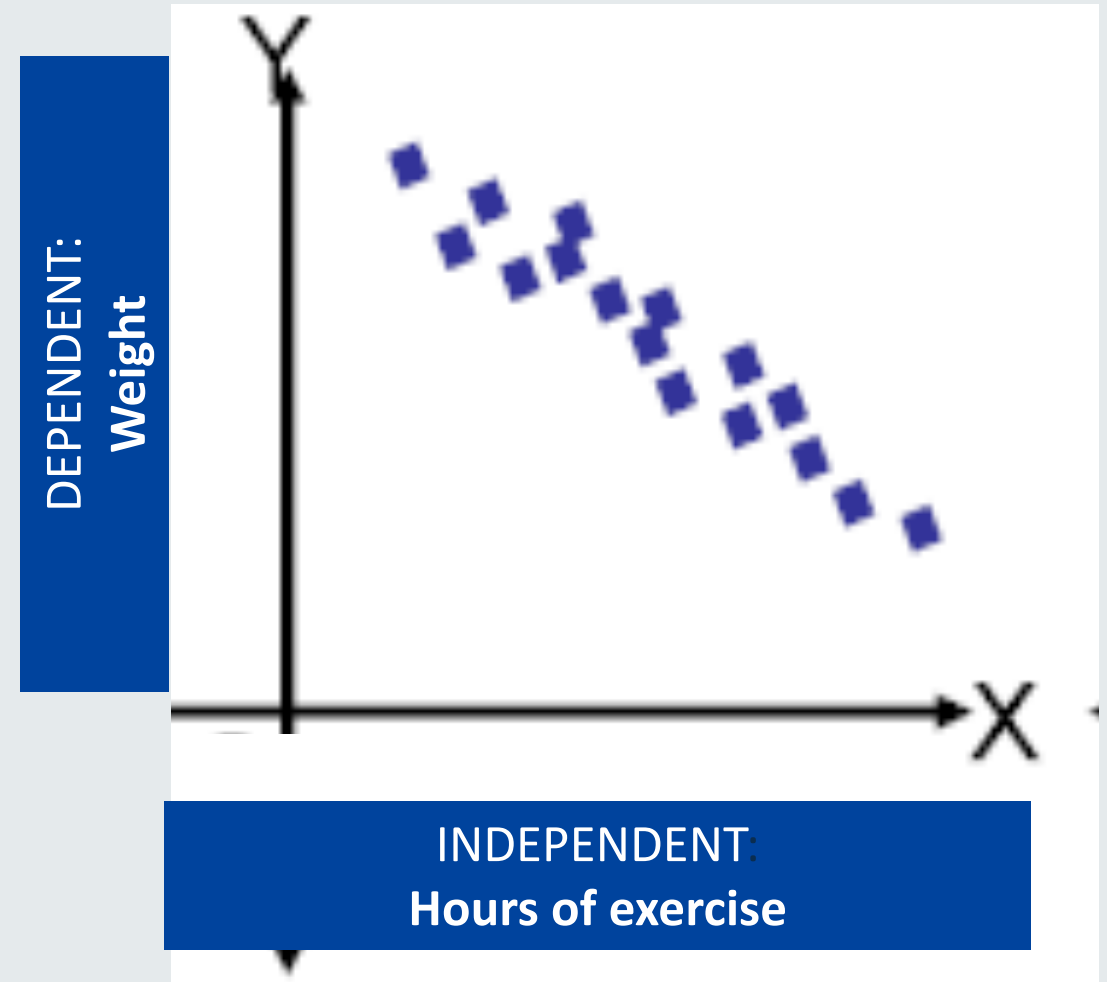
Previously on 'Introduction to Statistics'

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 $x = \text{hours of exercise}$ and $y = \text{weight}$ 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

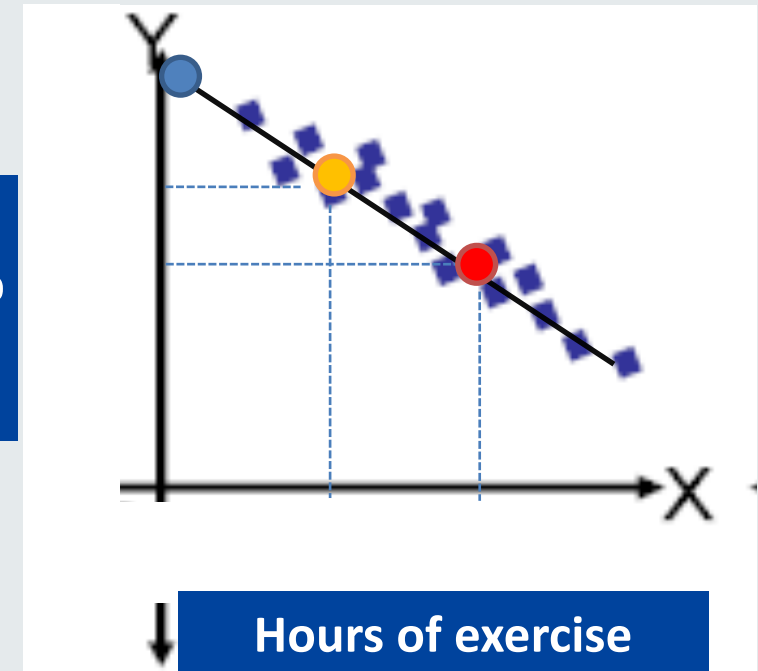
$$y = \beta_0 + \beta_1 x$$

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

Weight



$$\beta_0=70; \beta_1=-5;$$

Previously on 'Introduction to Statistics'

Interpretation

- $\beta_0 = 70$, When hours of exercise = 0, average weight is 70kg.
- $\beta_1 = -5$, Each additional hour of exercise decreases average weight by 5kg.

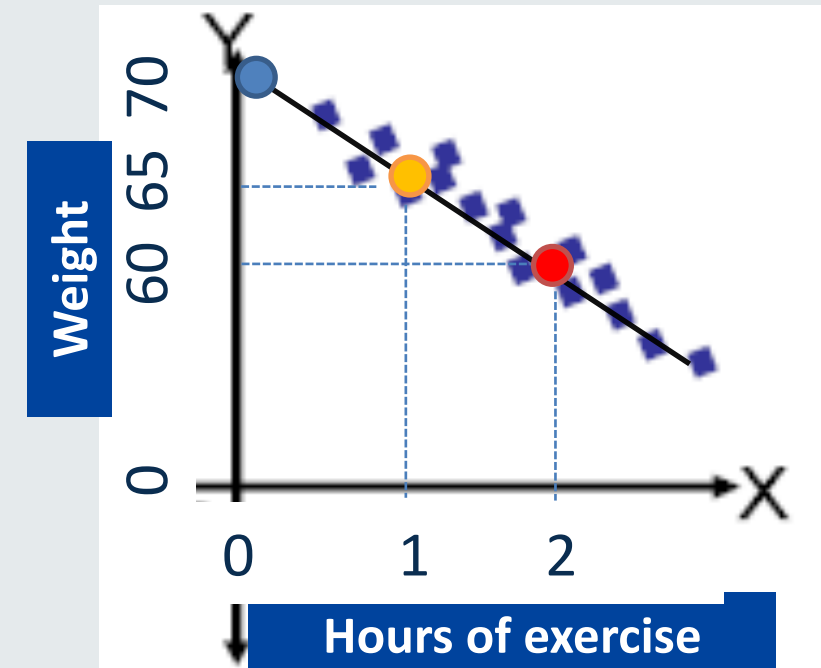
Linear regression model:

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

	X	Y
●	0	70
●	1	65
●	2	60

$$\beta_0=70; \beta_1=-5;$$

$$y = 70 - 5x$$

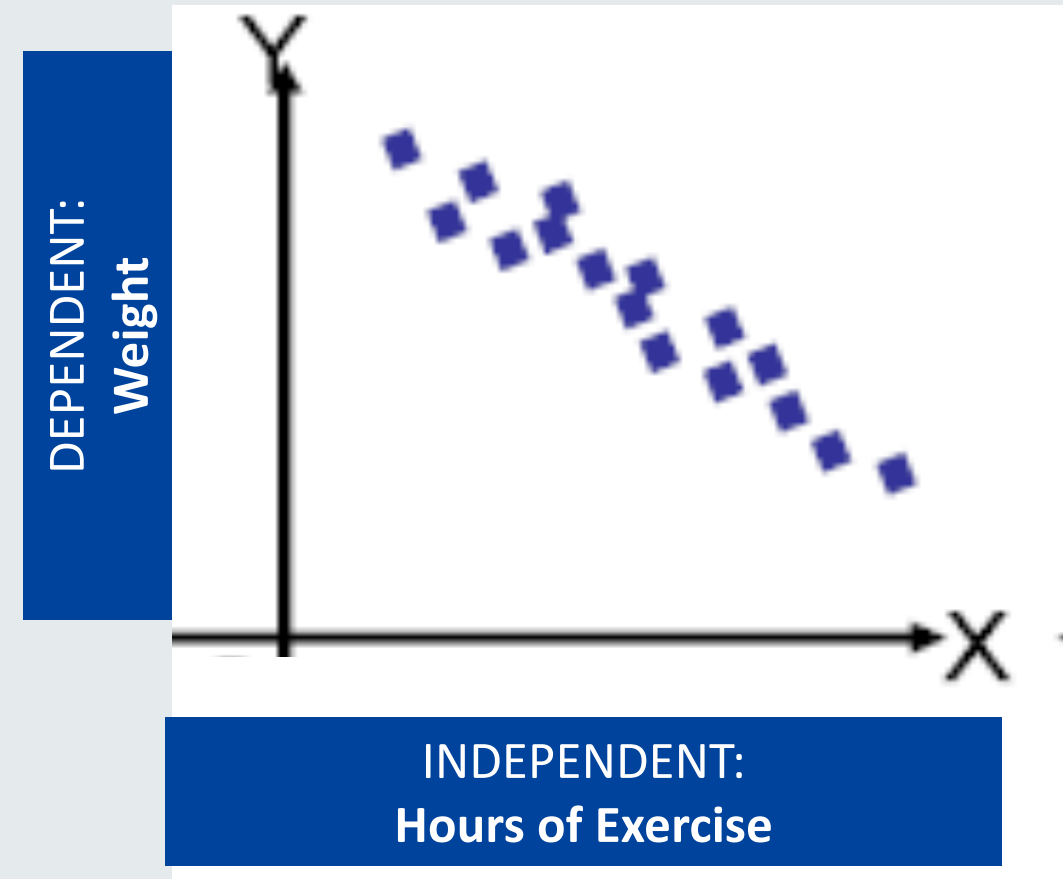


From Simple Linear Regression to Multiple Linear Regression

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.

$$y = 70 - 5x; r = -0.85$$

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;

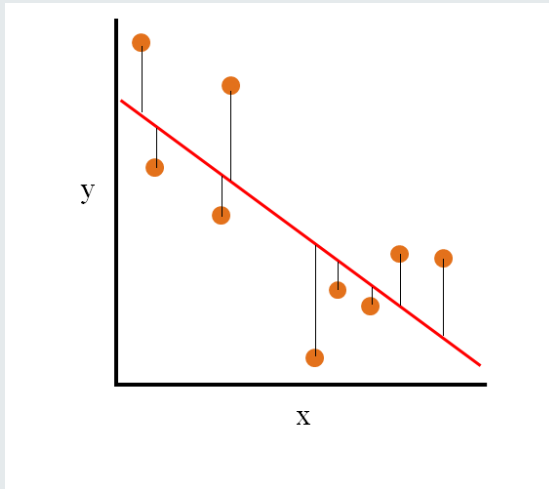
x_2 =diet;

From Simple Linear Regression to Multiple Linear Regression

Simple linear regression

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

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

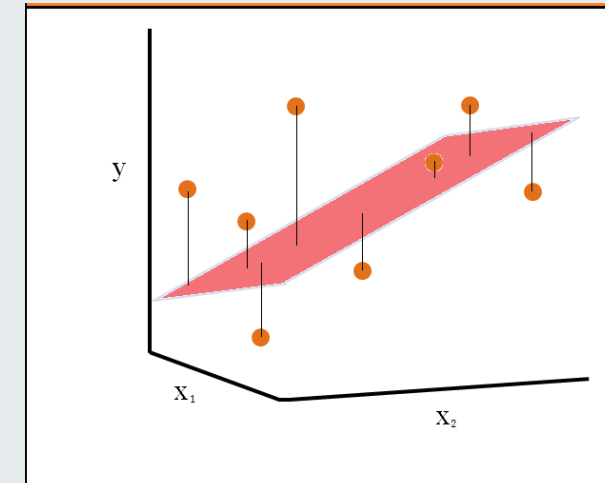


A **simple regression model** (one independent variable) fits a **regression line**
 $y = \beta_0 + \beta_1 x_1$

Multiple linear regression

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

Where: **y=weight**; **x₁=exercise**; **x₂=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** β_0 is the value that y takes when \mathbf{x}_i is zero.
- β_i 's are the **partial regression coefficients**.

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

E.g. β_1 Is the amount that the dependent variable \mathbf{y} will increase (or decrease) for each unit increase in the independent variable \mathbf{x}_1 while **holding** all other variables $\mathbf{x}_2, \dots, \mathbf{x}_n$ **constant**.

- ε 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 β_1 to check if variable x_1 is significantly associated with the outcome y while holding all other variables x_2, \dots, x_n constant.

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

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

If $p < 0.05$, we reject the null $\beta_1 = 0$ 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 $\beta_1 = 0$ 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

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

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

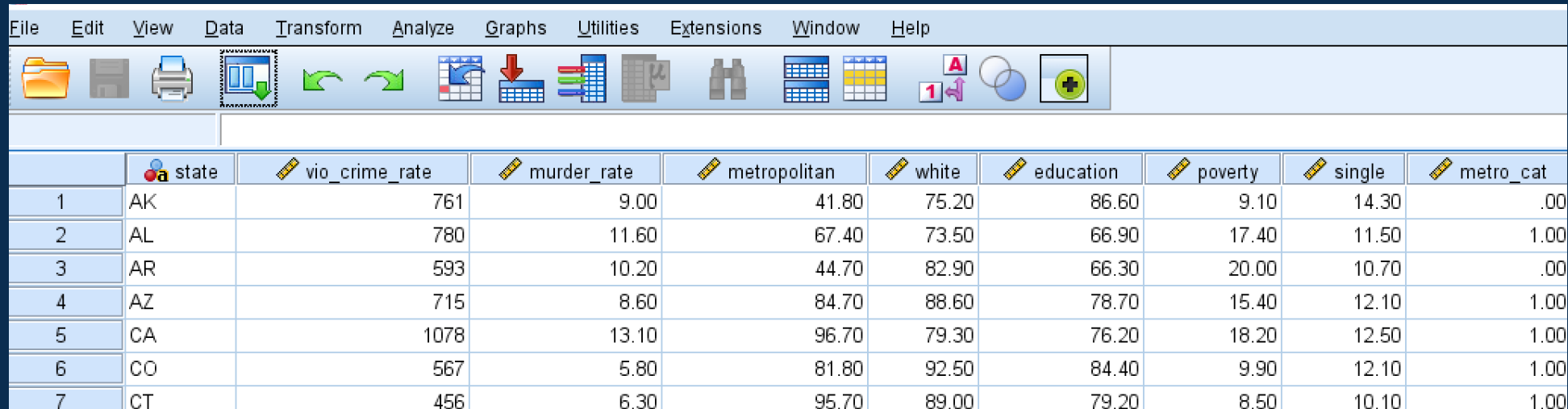
$$y = 72 - (4 \times 2) - (2 \times 2)$$
$$y = 60\text{kg}$$

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 $\beta_1 = -4$)

SPSS Slide

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



The screenshot shows the SPSS Data Editor window. The menu bar includes File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, Extensions, Window, and Help. The toolbar contains icons for file operations, data manipulation, and analysis. The data grid displays the following variables and values:

	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**

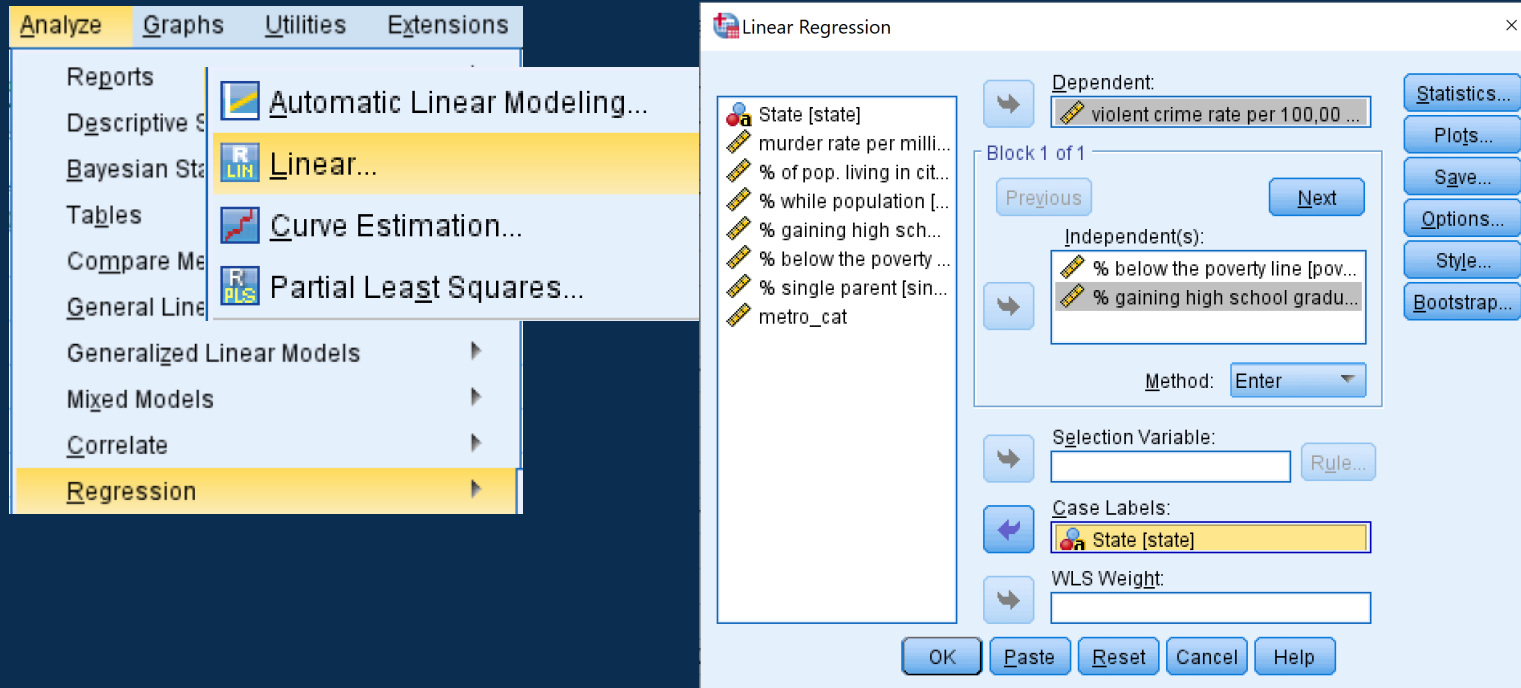


SPSS Slide: 'how to'

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

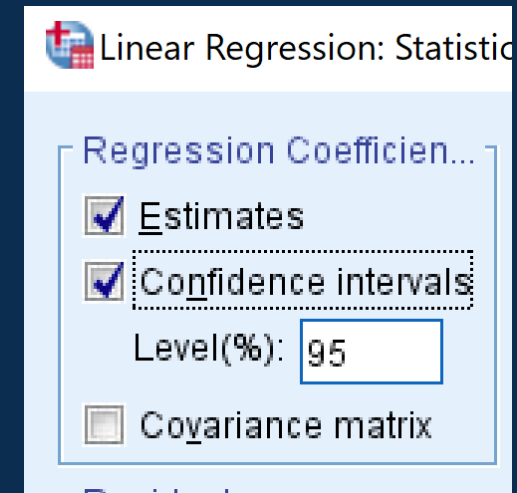
Step 1) Computing a multiple linear regression model for dependent variable 'crime' and independent variables 'poverty' and 'education'

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								
		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
Model		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

$$y = \beta_0 + \beta_1 x_1 - \beta_2 x_2$$
$$y = 345.852 + 23.927 x_1 - 1.502 x_2$$

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								
		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
Model		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 ($\beta_1 = 23.927$, $t=1.621$, $p=0.112$, 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 y being the “percentage of improvement in depression severity after being treated with an antidepressant drug”.

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

The estimated multiple linear regression model was:

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

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

y = percentage of improvement in depression severity after being treated with an antidepressant drug

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

x_2 = The patient severity at the start of the trial;

$$x_2 \in [0,100]$$

		p-value
Slope for x_1 (β_1)	-10	0.02
Slope for x_2 (β_2)	-2	0.01

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

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

x_1 is significantly associated with 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 (x_2), for example $x_2 = 15$.

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

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

Knowledge Check Solutions – Q2

y = percentage of improvement in depression severity after being treated with an antidepressant drug

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

x_2 = The patient severity at starting the trial;

$x_2 \in [0,100]$

		p-value
Slope for x_1 (β_1)	-10	0.02
Slope for x_2 (β_2)	-2	0.01

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

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

TRUE.

x_2 is significantly associated with 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, $\beta_2 = -2$; 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

© 2021 King's College London. All rights reserved