



**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

• 03/08/2020

**Module Title:** Introduction to Statistics

**Session Title:** Simple Linear Regression

---

**Topic title: Correlation and Linear Regression**



# Learning Outcomes

- Understand the difference between an **independent** and **dependent** variable
- Understand the **parameters** of simple linear regression (SLR)
- Interpret the **intercept** and **slope** parameters from a regression equation
- Use the simple linear regression (SLR) parameters to predict future observations
- Understand how to introduce a dummy categorical variable

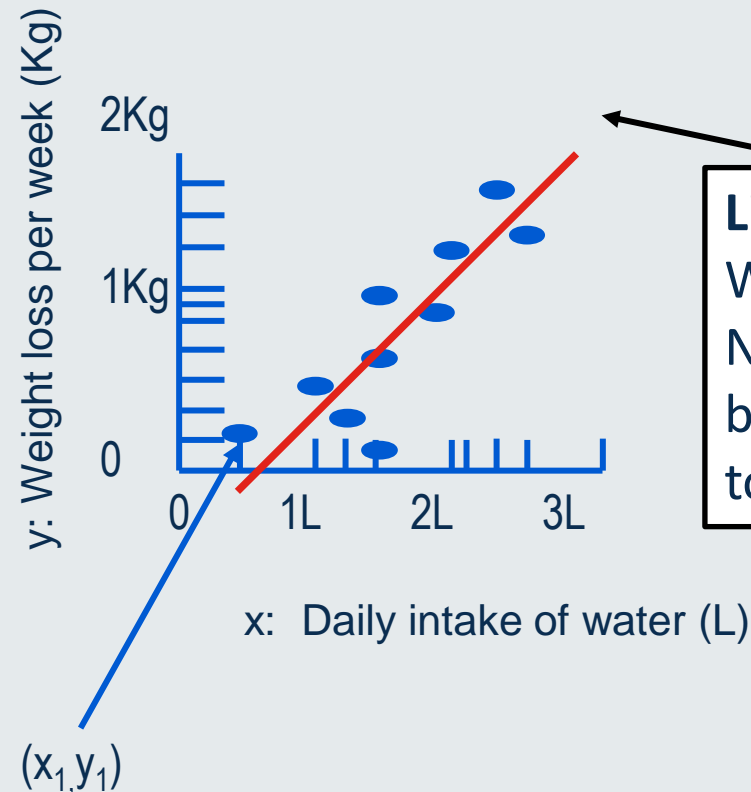
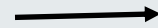


# Previously on 'Introduction to Statistics'

10 people were studied for the Hypothesis 'The higher the intake of water, the higher the weight loss'.

- 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  $x$  and  $y$  being continuous is called a **scatterplot**

	x	y
$(x_1, y_1)$	0.5	0.10
$(x_2, y_2)$	1.0	0.30
$(x_3, y_3)$	1.2	0.40
...	...	...



**Linear relationship:**  
We can draw a line.  
Not perfect fit,  
but the line is “close”  
to the points

# Previously on 'Introduction to Statistics'

We need an objective measure of strength of a linear relationship

**Correlation** is a statistical concept that refers to how close two variables are to having a linear relationship with each other, or in other words, the strength of their linear relationship. Correlation is a method to quantify the **Direction** and **Magnitude**, of linear association between two continuous variables.

Range of correlation coefficients	Degree of Correlation
0.80 to 1.00	Very strong positive
0.60 to 0.79	Strong positive
0.40 to 0.59	Moderate positive
0.20 to 0.39	Weak positive
0.00 to 0.19	Very weak positive - none
-0.19 to 0.00	Very weak negative - none
-0.39 to -0.20	Weak negative
-0.59 to -0.40	Moderate negative
-0.79 to -0.60	Strong negative
-1.00 to -0.80	Very strong negative

## Direction of effect

The co-efficient is positive or negative

## Magnitude of effect

The magnitude of the correlation coefficient ranges from -1 to 1, the close to  $\pm 1$  the stronger the effect

There are two types of correlation coefficients

- Pearson's Correlation Coefficient (normally distributed data)
- Spearman's Correlation Coefficient (skewed or ordinal data)

# Simple Linear Regression

---

In statistical modelling, a regression model is a set of statistical processes for estimating the relationships among variables. These models describe the relationship between variables by fitting a line to the observed data. The relationship is expressed as an equation.

In this session, we will focus on cases where there is a linear relationship between one continuous outcome and one predictor, for which the equation will look like:

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

This model is known as the **simple linear regression model**.

# Simple Linear Regression

---

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

- $x$  is called the **independent variable, predictor, explanatory** or **covariate** (continuous or categorical)
- $y$  is called the **dependent variable, outcome or response**.  $y$  'depends on'  $x$  (always continuous)
- The **intercept**  $\beta_0$  is the value that  $y$  takes when  $x$  is zero.
  - If the intercept is zero then  $y$  increases in proportion to  $x$  (i.e. double  $x$  then  $y$  doubles  $y=x$ )
- The **slope**  $\beta_1$  determines the change in  $y$  when  $x$  changes by one unit.
  - It is the amount that the dependent variable will increase (or decrease) for each unit increase in the independent variable
- $\varepsilon$  is called the **residual** (distance between the points and the line).
- $\beta_0$  and  $\beta_1$  are together known as **regression coefficients**.



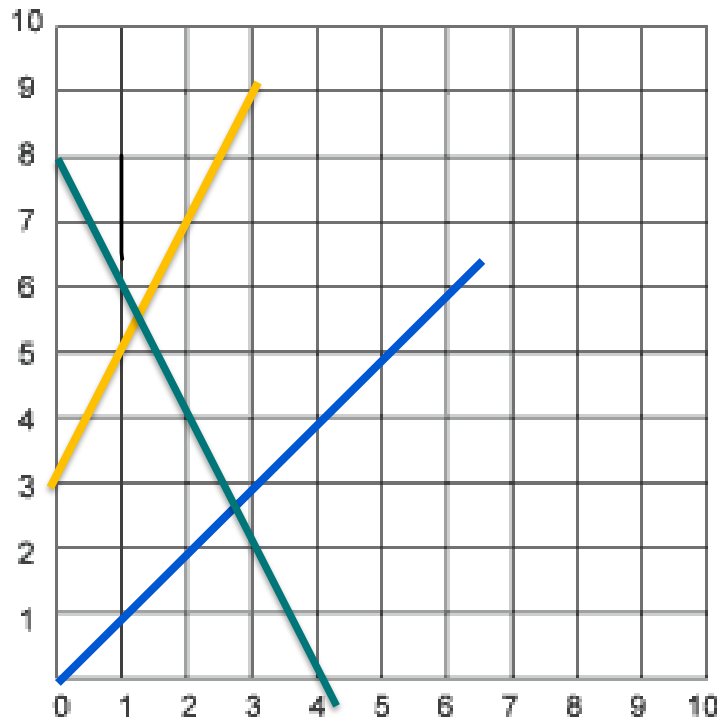
$$y = \beta_0 + \beta_1 x + \varepsilon$$

x	y
1	1
2	2
3	3
...	...
7	7

$$y = x$$

x	y
0	3
1	5
2	7
...	...
7	17

$$y = 3 + 2x$$



x	y
0	8
1	6
2	4
...	...
7	-6

$$y = 8 - 2x$$

$\beta_0$  represents where the line intercepts the y axis.

$\beta_1$  represent the slope of the line as x increases by one unit how much does y increase or decrease

# Example

You are a social researcher interested in the relationship between income and happiness. You survey 500 people whose incomes range from £15k to £75k and ask them to rank their happiness on a scale from 1 to 10.

Your independent variable (income) and dependent variable (happiness) are both quantitative, so you can do a regression analysis to see if there is a linear relationship between them.

We can ask

- How strong the relationship is between two variables (e.g. the relationship between income and happiness).
- The value of the dependent variable at a certain value of the independent variable (e.g. the amount of happiness at a certain level of income).

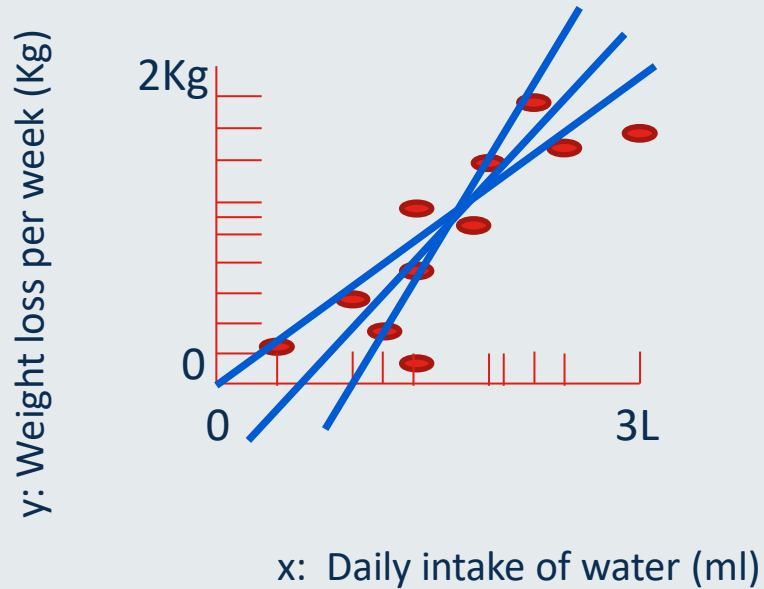




# Estimation

**Independent variable ( $x$ ):** Daily intake of water

**Dependent variable ( $y$ ):** Weight loss per week



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

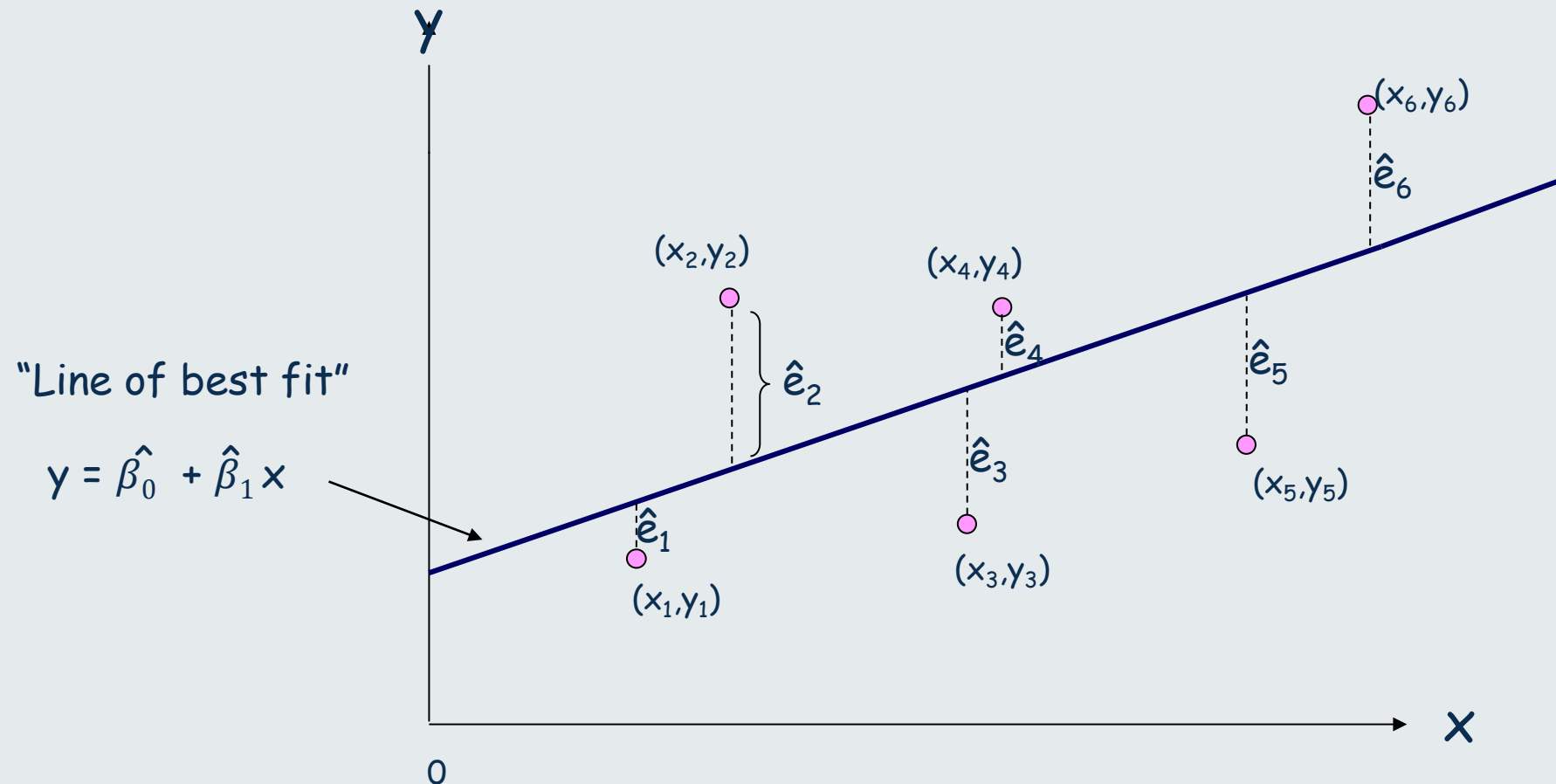
?

How can we find the best line fitting the cloud of points and therefore find the best estimates for  $\beta_0$  and  $\beta_1$ ?



# Estimation

- The best **linear regression line** is the closest to all data points, i.e. the line that makes the **residual**  $\varepsilon$  as small as possible.
- **Ordinary Least Squares (OLS)** – Is one method that can be used to estimate the regression line that **minimises the squared residuals** ( $\varepsilon_1^2 + \varepsilon_2^2 + \dots + \varepsilon_n^2$ ) to give us the estimates for  $\widehat{\beta}_0$  and  $\widehat{\beta}_1$ .



# Simple Linear Regression Model

---

## When to use it

- To measure to what extent there is a linear relationship between two variables

## Hypotheses:

- $H_0$ : There is no linear association e.g. the slope  $\beta_1$  in the population equals to 0
- $H_a$ : There is a linear association e.g. the slope  $\beta_1$  in the population does not equal to 0

## Assumptions:

- There is a linear relationship between the dependent and independent variable
- Residuals (or “errors”)  $\varepsilon$  are independent of one another: the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among observations
- Residuals follow a Normal distribution, with mean 0 and constant Standard Deviation  $\sigma$
- Homogeneity of variance (homoscedasticity): the size of the error in our prediction doesn’t change significantly across the values of the independent variable.

# Formulae – for the curious

The slope is estimated as

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

The intercept is estimated as

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} \quad \left[ \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \right]$$

Test Statistic for the hypothesis test

$$t = \frac{\widehat{\beta}_1}{\widehat{se}(\widehat{\beta}_1)}, \text{ df} = n - 1$$



# SPSS Slide

Download the data that we are going to use during the lecture. The dataset is the **lecture\_6a\_data.sav**.

	sex	reading	class	height	weight	id	malcat1
1	1	27	7	173	72.33	1	1.00
2	1	23	2	157	41.28	2	1.00
3	1	30	2	174	58.29	3	1.00
4	1	15	4	170	69.17	4	1.00
5	1	26	2	161	51.03	5	1.00
6	1	28	1	182	71.67	6	1.00
7	1	13	4	170	62.14	7	1.00

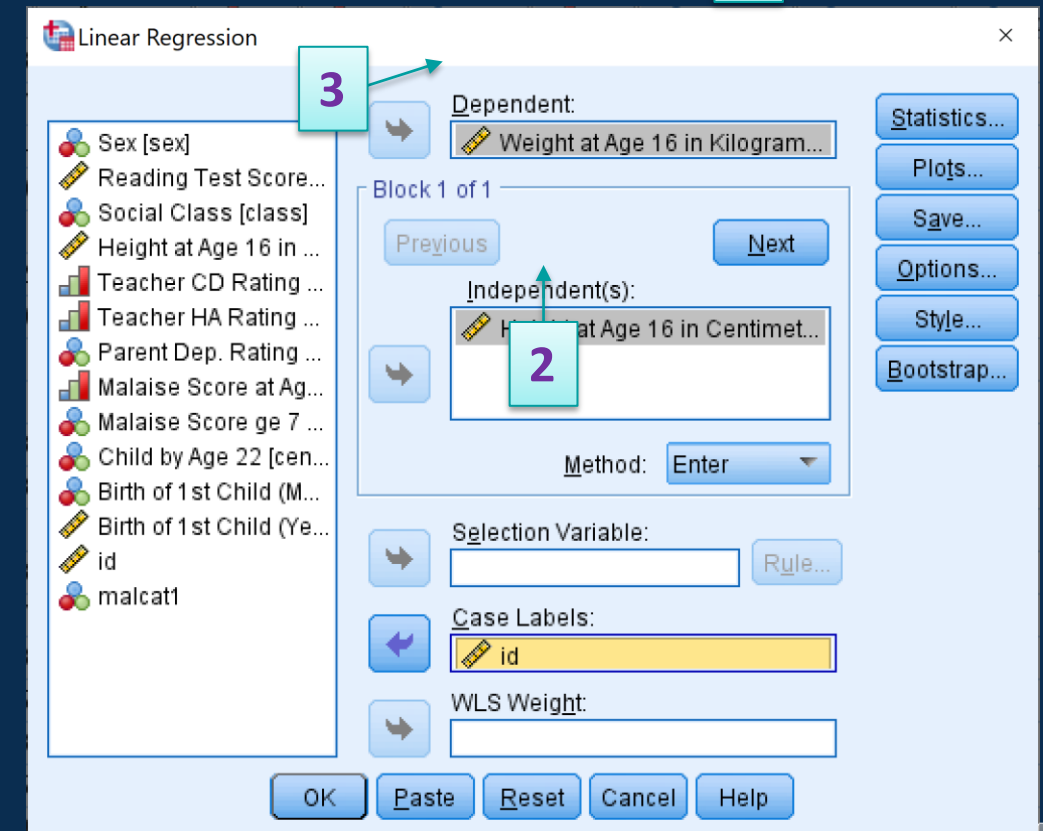
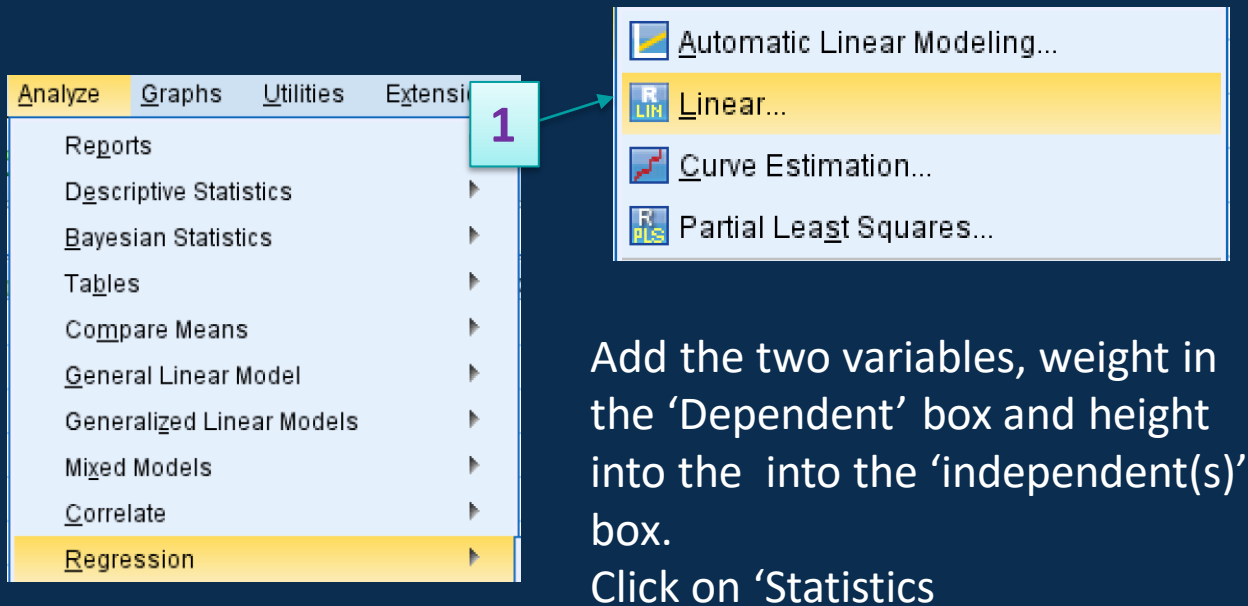
The dataset contains data from 1000 individuals, from the National Child Development Study (NCDS) with respect to their

- **sex**: gender of child (1=male, 2=female)
- **height** : height in cm at age 16
- **weight**: weight in kg at age 16
- **reading**: reading score
- **malcat1**: incidence of malaise at 22 years (0=yes, 1 = No)

# SPSS Slide: 'how to'

According to the researchers, in the population from which our data came, they believe there is a relationship between weight and height of the 16 year old children

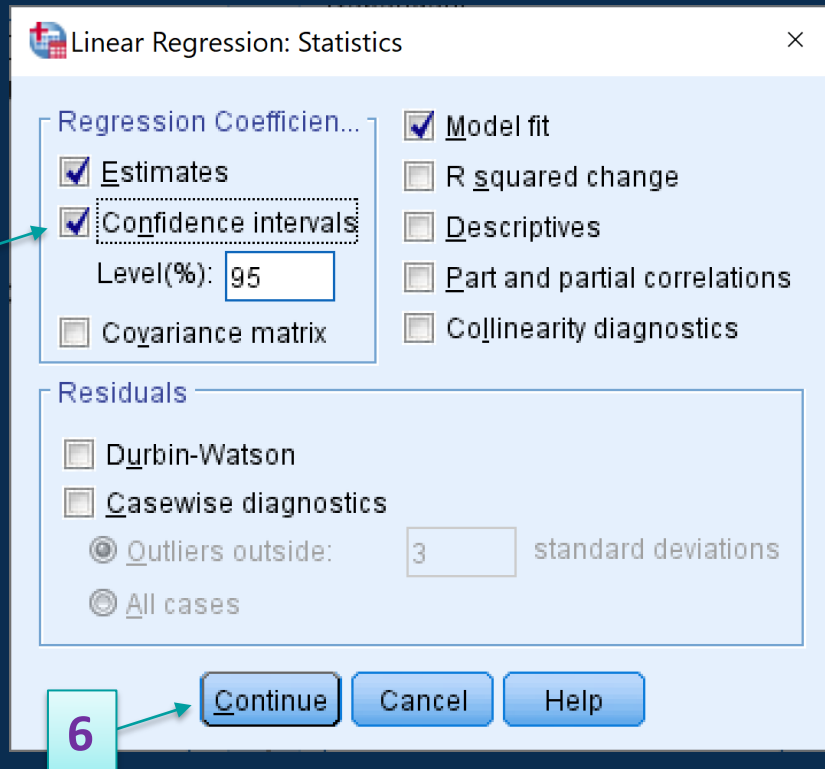
Step 1: Compute a Linear regression model for dependent variable 'weight' and independent variable 'height' from NCDS data Use 'Analyse' -> 'Regression' -> 'Linear'



# SPSS Slide: 'how to'

According to the researchers, in the population from which our data came, they believe there is a relationship between weight and height of the 16 year old children

Step 1: Compute a Linear regression model for dependent variable 'weight' and independent variable 'height' from NCDS data



In the Statistics tab.  
Check the 'Estimates'  
Check the 'Confidence Intervals'  
Click on 'Continue'  
Click on 'OK'

# Output and Interpretation Slide

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.520 <sup>a</sup>	.270	.270	8.25311

a. Predictors: (Constant), Height at Age 16 in Centimeters  
b. Dependent Variable: Weight at Age 16 in Kilograms

This table provides the R and R<sup>2</sup> values. The R value represents the simple correlation and is 0.520 which indicates a moderate degree of correlation.

The R<sup>2</sup> value indicates how much of the total variation in the dependent variable, weight, can be explained by the independent variable, height. In this case, 27.0% can be explained.

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	25172.852	1	25172.852	369.570	.000 <sup>b</sup>
	Residual	67977.581	998	68.114		
	Total	93150.434	999			

a. Dependent Variable: Weight at Age 16 in Kilograms  
b. Predictors: (Constant), Height at Age 16 in Centimeters

The ANOVA table, reports how well the regression equation fits the data (i.e., predicts the dependent variable). This table indicates that the regression model predicts the dependent variable significantly well ( $p < 0.001$ ).

This indicates the statistical significance of the regression model that was run and overall, the regression model statistically significantly predicts the outcome variable (i.e., it is a good fit for the data).





# Output and Interpretation

Coefficients <sup>a</sup>							
Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound Upper Bound
1	(Constant)	-46.764	5.413		-8.639	.000	-57.386 -36.142
	Height at Age 16 in Centimeters	.626	.033	.520	19.224	.000	.562 .689

a. Dependent Variable: Weight at Age 16 in Kilograms

$\beta_0$

$\beta_1$

$SE(\beta_1)$

The estimated slope coefficient ( $\beta_1$ ), suggests a 1cm increase in height is associated with a 0.626kg increase in weight. The units of the slope is kg/cm.

The intercept ( $\beta_0$ ), is the extrapolated weight for a 16 year old of zero height.

In addition to getting point estimation for  $\beta_1$ , it is possible to calculate a confidence interval for the slope parameter  
The confidence interval formula is:

$$95\% \text{ CI} = [\beta_1 - 1.96 \times SE(\beta_1), \beta_1 + 1.96 \times SE(\beta_1)]$$

E.g. for the NCDS data, a CI for  $\beta_1$  can be derived as follows:

Lower limit:  $0.626 - 1.96 \times 0.033 = 0.562$

Upper limit:  $0.626 + 1.96 \times 0.033 = 0.689$   
= [0.562, 0.689]

# Output and Interpretation

Coefficients <sup>a</sup>								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-46.764	5.413		-8.639	.000	-57.386	-36.142
	Height at Age 16 in Centimeters	.626	.033	.520	19.224	.000	.562	.689

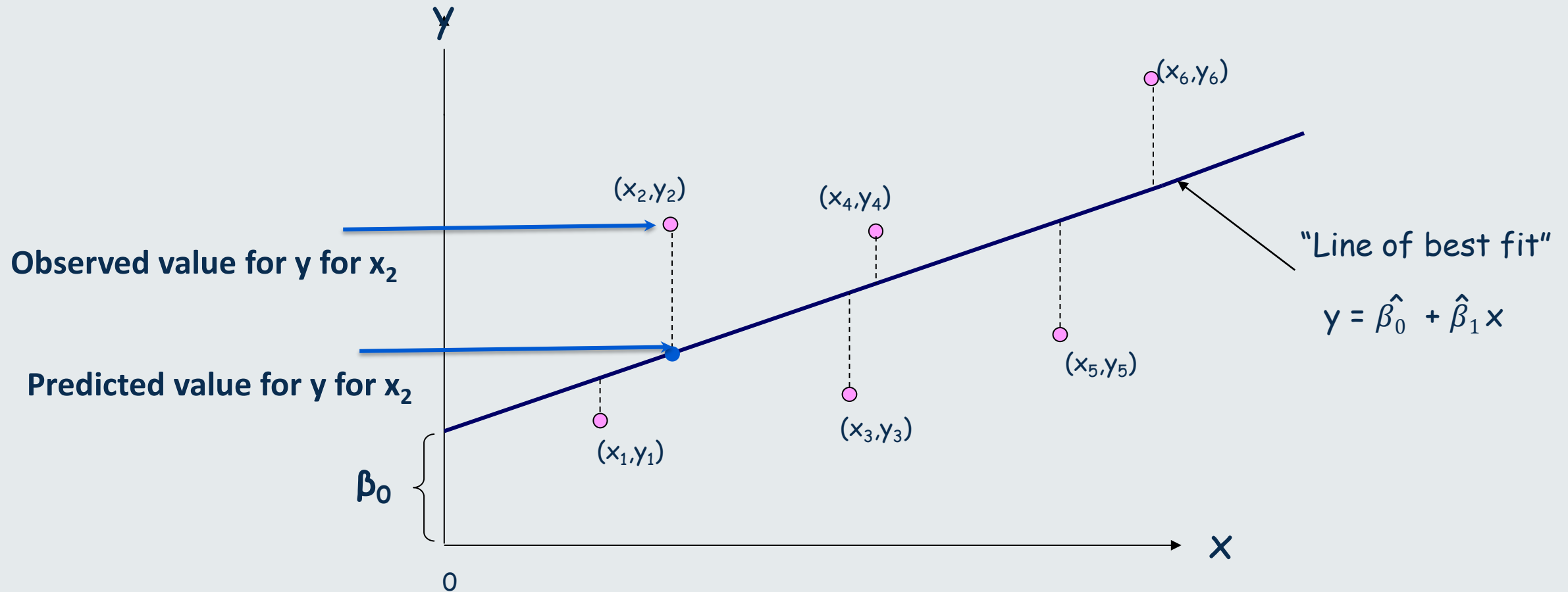
a. Dependent Variable: Weight at Age 16 in Kilograms

We found a significant relationship between weight and height of 16 year olds with a 1cm increase in height associated with a 0.626kg increase in weight ( $\beta_1=0.626$ ,  $t=19.224$ ,  $p<0.001$  95%CI (0.562, 0.689))

# Prediction

Regression models are used to predict new cases.

The predicted value  $\hat{y}$  for a new observation  $x$  is its corresponding value on the regression line.



# Prediction

---

We can use the regression equation to predict the weight for new case, added to the sample:  
If  $x=186\text{cm}$  for a given 16 years old new case, and knowing that  $y=-46.764 + 0.626 x$ ,  
What would be the child's weight?

We can estimate:

$$\begin{aligned}y &= -46.764 + 0.626 x, \\y &= -46.764 + 0.626 \times 186 \\y &= 69.672 \text{ Kg}\end{aligned}$$

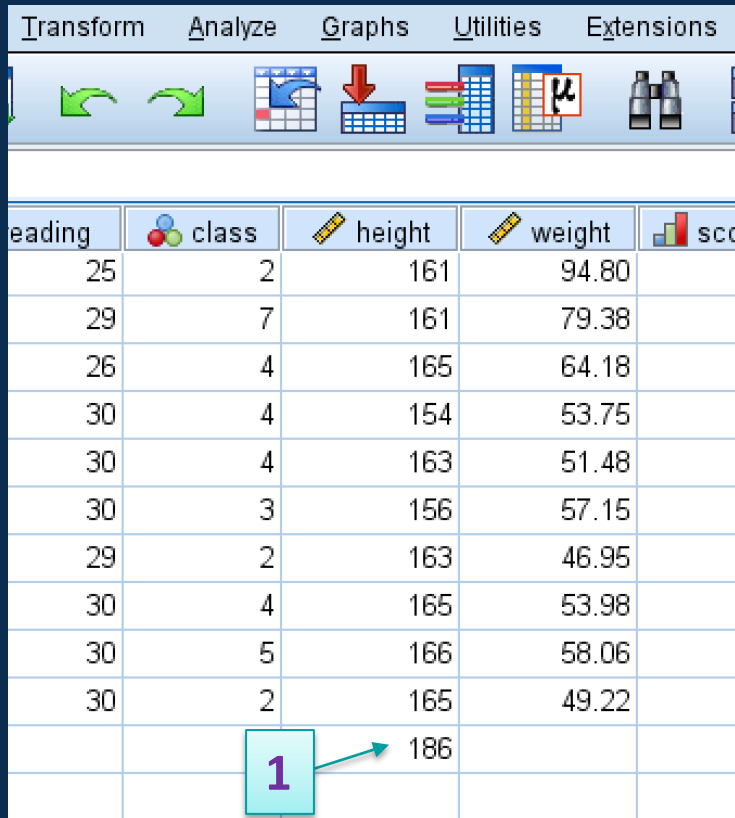
**The model predicts a weight of 69.672 Kg for a 16-years old child that is 186 cm tall**

In addition to getting predicted values of weight for any given height, it is possible to calculate a confidence interval for that prediction.

# SPSS Slide: 'how to'

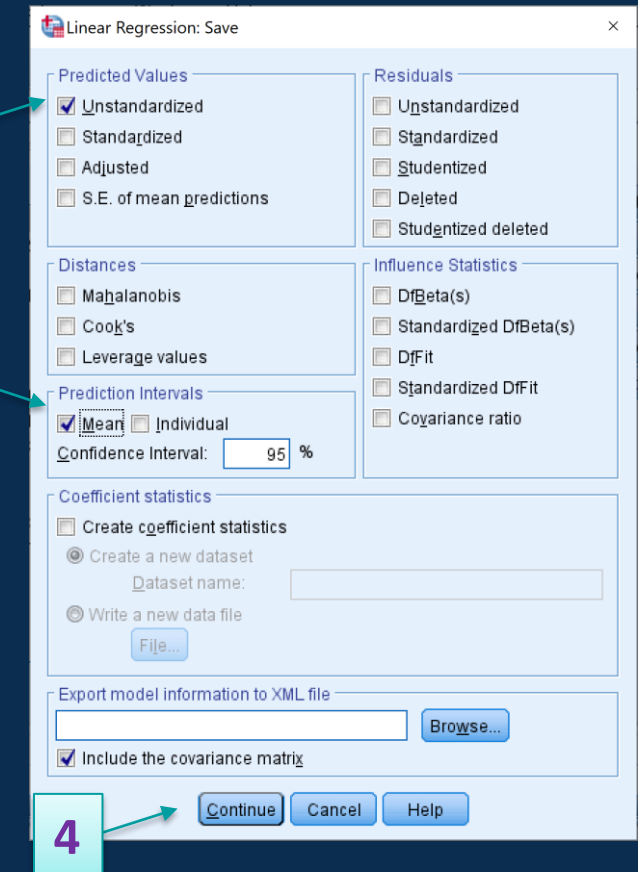
If  $x=186\text{cm}$  for a given 16 years old new case, and knowing that  $y=-46.764 + 0.626 x$ ,  
What would be expect the child's weight to be?

Step 1: Add the x-values at which you want to predict y to the y-variable (here height) in the data. Use height= 186 cm



leading	class	height	weight	scd
25	2	161	94.80	2
29	7	161	79.38	4
26	4	165	64.18	2
30	4	154	53.75	1
30	4	163	51.48	3
30	3	156	57.15	1
29	2	163	46.95	4
30	4	165	53.98	3
30	5	166	58.06	2
30	2	165	49.22	1
		186		

Step 2) Use **Analyse -> Regression -> Linear**  
Step 2) Put 'weight' in dependent, and  
'height' in independent.  
Click '**Save**', select 'Prediction values'  
'Unstandardised' and 'Prediction intervals'  
'mean'.  
Click on 'Continue'  
Click on 'OK'



Linear Regression: Save

**Predicted Values**

- ☒ Unstandardized
- ☐ Standardized
- ☐ Adjusted
- ☐ S.E. of mean predictions

**Residuals**

- ☐ Unstandardized
- ☐ Standardized
- ☐ Studentized
- ☐ Deleted
- ☐ Studentized deleted

**Distances**

- ☐ Mahalanobis
- ☐ Cook's
- ☐ Leverage values

**Prediction Intervals**

- ☒ Mean ☐ Individual
- Confidence Interval: 95 %

**Influence Statistics**

- ☐ DfBeta(s)
- ☐ Standardized DfBeta(s)
- ☐ DfFit
- ☐ Standardized DfFit
- ☐ Covariance ratio




**Coefficient statistics**

- ☐ Create coefficient statistics
- ☒ Create a new dataset  
Dataset name:
- ☒ Write a new data file  
File:

**Export model information to XML file**

☒ Include the covariance matrix

# Output and Interpretation

 PRE_1	 LMCI_1	 UMCI_1
53.94261	53.33354	54.55167
53.94261	53.33354	54.55167
56.44463	55.92713	56.96213
49.56407	48.63380	50.49434
55.19362	54.64309	55.74414
50.81508	49.98842	51.64174
55.19362	54.64309	55.74414
56.44463	55.92713	56.96213
57.07013	56.55788	57.58238
56.44463	55.92713	56.96213
69.58024	68.21403	70.94645

The 'Data View' in SPSS you will see three new columns one for the predicted  $y$  (PRE\_1)  $\hat{y} = 69.58$  kg based on the value of 186cm height and the lower (LMCI\_1) and upper (UMCI\_1) confidence interval limits **95%CI (68.21, 70.95)**.

Prediction Intervals

☒ Mean ☐ Individual

Confidence Interval:  %

For instance, to predict the average weight of 16 year olds if the height is 186cm use the **confidence interval of the mean**.

To predict the weight of Jasmine, a 16 year old with weight 186cm then use the **confidence interval for the individual**.



# Categorical Predictors

What do we do if we have a predictor that is **categorical** ?

Focus on continuous outcome  $y$  = weight and categorical explanatory variable  $x$  = gender.

When  $x$  is categorical binary then:

- The regression line connects the mean response in one group with the mean response in the other.
- The slope coefficient simply measures the group difference in means (remember: slope measures predicted change in  $y$  when  $x$  changes by one unit=switches groups)

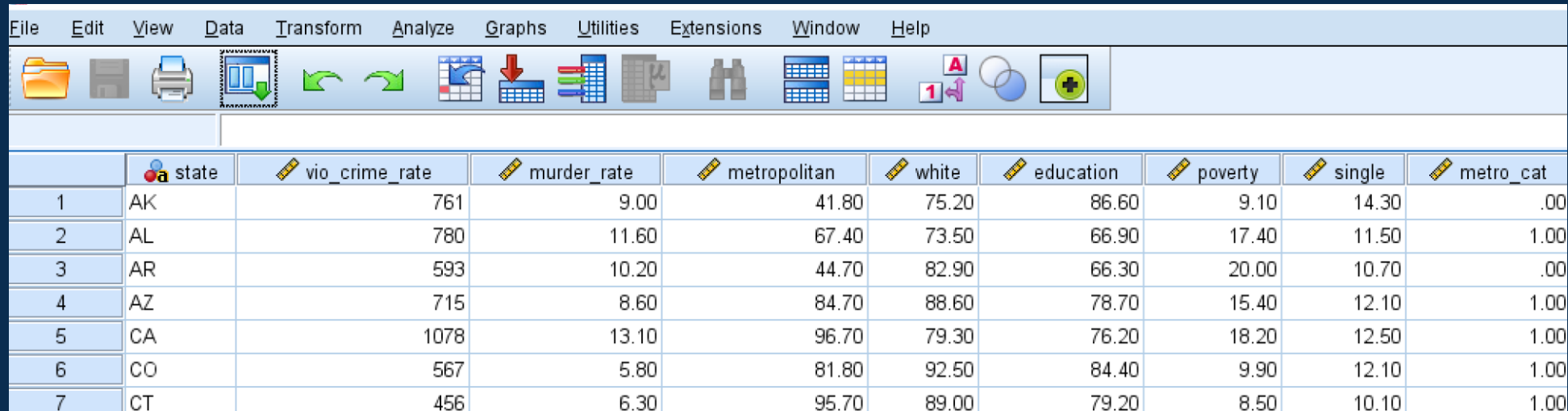
Coefficients <sup>a</sup>								
		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	64.124	.943		67.971	.000	62.273	65.975
	Sex	-4.607	.593	-.239	-7.763	.000	-5.772	-3.442

a. Dependent Variable: Weight at Age 16 in Kilograms

Represents the difference in means between males and females, as we change  $x$  by one unit (move from male to female), the weight changes by 4.607kg. **On average females weigh 4.607kg less than males ( $\beta_1 = -4.607$ .  $t = -7.763$ ,  $p < 0.001$ , 95% CI (-5.772, -3.442))**

# SPSS Slide

Download the data that we are going to use during the lecture. The dataset is the **lecture\_6b\_data.sav**.



The screenshot shows the SPSS software interface. The menu bar includes File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, Extensions, Window, and Help. The toolbar contains various icons for file operations, data manipulation, and analysis. The data view shows a table with 10 columns and 7 rows of data. The columns are labeled: state, vio\_crime\_rate, murder\_rate, metropolitan, white, education, poverty, single, and metro\_cat. The rows represent different US states: AK, AL, AR, AZ, CA, CO, and CT.

	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



# Categorical Predictors

What do we do if we have a predictor that has more than 2 **categories**?

Focus on continuous outcome  $y$  = Violent Crime and categorical explanatory variable  $x$  = Urbanicity.

state	urban	
AK	Low	
AR	Low	
IA	Low	
ID	Low	
KY	Low	
ME	Low	
AL	Medium	
GA	Medium	
KS	Medium	
MN	Medium	
MO	Medium	
NC	Medium	
AZ	High	
CA	High	
CO	High	
CT	High	
DE	High	

The variable **urban** is a categorical variable with three levels “Low”, “Medium” and “High”

- Categorical variables which are non binary cannot be included directly in a regression model.
- Need to be recoded into a set of dummy variables
- A dummy (indicator) variable is a binary (0,1) variable indicating a category of the predictor variable.
- A predictor with k levels can be coded as k dummy variables
- Only k-1 dummy variables are necessary to fully represent a categorical predictor.

# Categorical Predictors

US crime data. The variable urban is a categorical variable with three levels “Low”, “Medium” and “High”  
Let’s consider a linear regression for violent\_crime and urban

state	urban
AK	Low
AR	Low
IA	Low
ID	Low
KY	Low
ME	Low
AL	Medium
GA	Medium
KS	Medium
MN	Medium
MO	Medium
NC	Medium
AZ	High
CA	High
CO	High
CT	High
DE	High

The variable **urban** is a categorical variable with three levels “Low”, “Medium” and “High”

Dummy coding of **urban** ( $k=3$ )

	d1	d2	d3
	1	0	0
	1	0	0
	1	0	0
	1	0	0
	1	0	0
	1	0	0
	0	1	0
	0	1	0
	0	1	0
	0	1	0
	0	1	0
	0	1	0
	0	0	1
	0	0	1
	0	0	1
	0	0	1
	0	0	1

# SPSS Slide: 'how to'

Researchers believe there is a relationship between Violent Crime and the level of urbanicity in an area. The variable urban is a categorical variable with three levels “Low”, “Medium” and “High” and needs to be converted to dummy variables to include in the regression.





Step 1: Generating a dummy variable for “Low” urbanicity level in ‘urban’ variable from US crime dataset (We need to repeat this process to create a dummy variable for “Medium” level)

The image shows the SPSS 'Recode into Different Variables' dialog box with several numbered annotations:

- 1**: Points to the 'Transform' menu, specifically the 'Recode into Different Variables...' option.
- 2**: Points to the 'Numeric Variable -> Output Variable:' list, where 'urban --> ?' is entered.
- 3**: Points to the 'Output Variable' section, where the 'Name' is 'D1\_Low' and the 'Label' is 'Low urbanicity'.
- 4**: Points to the 'Old and New Values...' button.
- 5**: Points to the 'Old Value' section, where the 'Value' radio button is selected.
- 6**: Points to the 'New Value' section, where the 'Value' radio button is selected, and the 'Old --> New:' list shows the mapping: 1 --> 1, 2 --> 0, 3 --> 0.

Below the dialog box, the text reads: **Use 'Transform' -> 'Recode into Different Variables'**

# Output and Interpretation Slide

 urban	 D1_Low	 D2_Med	 D3_High
2.00	.00	1.00	.00
3.00	.00	.00	1.00
2.00	.00	1.00	.00
3.00	.00	.00	1.00
3.00	.00	.00	1.00
3.00	.00	.00	1.00
3.00	.00	.00	1.00
3.00	.00	.00	1.00
3.00	.00	.00	1.00
3.00	.00	.00	1.00
3.00	.00	.00	1.00
3.00	.00	.00	1.00
2.00	.00	1.00	.00
1.00	1.00	.00	.00

Only 2 dummy variables (e.g. d1 and d2) are needed to represent a variable with 3 levels.

The model will be:  $violent\_crime = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \varepsilon$

- $\beta_1$  will be the difference in mean between “Low” vs. “High” (the latter is called the “reference category”)

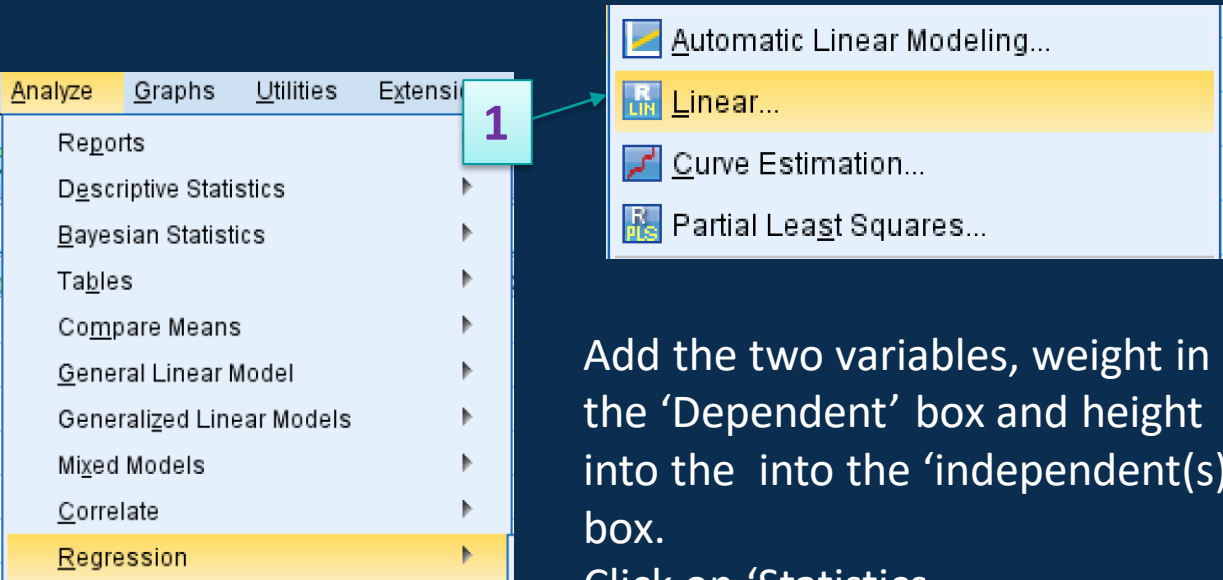
- $\beta_2$  will be the difference in mean between “Medium” vs. “High” (the latter is called the “reference category”)



# SPSS Slide: 'how to'

Researchers believe there is a relationship between Violent Crime and the level of urbanicity in an area.

Step 2: Compute a Linear regression model for dependent variable 'Violent Crime' and independent variable 'urban' using the dummy variables created



1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

66

67

68

69

70

71

72

73

74

75

76

77

78

79

80

81

82

83

84

85

86

87

88

89

90

91

92

93

94

95

96

97

98

99

100

101

102

103

104

105

106

107

108

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

200

201

202

203

204

205

206

207

208

209

210

211

212

213

214

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229

230

231

232

233

234

235

236

237

238

239

240

241

242

243

244

245

246

247

248

249

250

251

252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

268

269

270

271

272

273

274

275

276

277

278

279

280

281

282

283

284

285

286

287

288

289

290

291

292

293

294

295

296

297

298

299

300

301

302

303

304

305

306

307

308

309

310

311

312

313

314

315

316

317

318

319

320

321

322

323

324

325

326

327

328

329

330

331

332

333

334

335

336

337

338

339

340

341

342

343

344

345

346

347

348

349

350

351

352

353

354

355

356

357

358

359

360

361

362

363

364

365

366

367

368

369

370

371

372

373

374

375

376

377

378

379

380

381

382

383

384

385

386

387

388

389

390

391

392

393

394

395

396

397

398

399

400

401

402

403

404

405

406

407

408

409

410

411

412

413

414

415

416

417

418

419

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

457

458

459

460

461

462

463

464

465

466

467

468

469

470

471

472

473

474

475

476

477

478

479

480

481

482

483

484

485

486

487

488

489

490

491

492

493

494

495

496

497

498

499

500

501

502

503

504

505

506

507

508

509

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

536

537

538

539

540

541

542

543

544

545

546

547

548

549

550

551

552

553

554

555

556

557

558

559

560

561

562

563

564

565

566

567

568

569

570

571

572

573

574

575

576

577

578

579

580

581

582

583

584

585

586

587

588

589

590

591

592

593

594

595

596

597

598

599

600

601

602

603

604

605

606

607

608

609

610

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

648

649

650

651

652

653

654

655

656

657

658

659

660

661

662

663

664

665

666

667

668

669

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

702

703

704

705

706

707

708

709

710

711

712

713

714

715

716

717

718

719

720

721

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

756

757

758

759

760

761

762

763

764

765

766

767

768

769

770

771

772

773

774

775

776

777

778

779

780

781

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

799

800

801

802

803

804

805

806

807

808

809

810

811

812

813

814

815

816

817

818

819

820

821

822

823

824

825

826

827

828

829

830

831

832

833

834

835

836

837

838

839

840

841

842

843

844

845

846

847

848

849

850

851

852

853

854

855

856

857

858

859

860

861

862

863

864

865

866

867

868

869

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917

918

919

920

921

922

923

924

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

940

941

942

943

944

945

946

947

948

949

950

951

952

953

954

955

956

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971

972

973

974

975

976

977

978

979

980

981

982

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

1000

1001

1002

1003

1004

1005

1006

1007

1008

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018

1019

1020

1021

1022

1023

1024

1025

1026

1027

1028

1029

1030

1031

1032

1033

1034

1035

1036

1037

1038

1039

1040

1041

1042

1043

1044

1045

1046

1047

1048

1049

1050

1051

1052

1053

1054

1055

1056

1057

1058

1059

1060

1061

1062

1063

1064

1065

1066

1067

1068

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

1080

1081

1082

1083

1084

1085

1086

1087

1088

1089

1090

1091

1092

1093

1094

1095

1096

1097

1098

1099

1100

1101

1102

1103

1104

1105

1106

1107

1108

1109

1110

1111

1112

1113

1114

1115

1116

1117

1118

1119

1120

1121

1122

1123

1124

1125

1126

1127

1128

1129

1130

1131

1132

1133

1134

1135

1136

1137

1138

1139

1140

1141

1142

1143

1144

1145

1146

1147

1148

1149

1150

1151

1152

1153

1154

1155

1156

1157

1158

1159

1160

1161

1162

1163

1164

1165

1166

1167

1168

1169

1170

1171

1172

1173

1174

1175

1176

1177

1178

1179

1180

1181

1182

1183

1184

1185

1186

1187

1188

1189

1190

1191

1192

1193

1194

1195

1196

1197

1198

1199

1200

1201

1202

1203

1204

1205

1206

1207

1208

1209

1210

1211

1212

1213

1214

1215

1216

1217

1218

1219

1220

1221

1222

1223

1224

1225

1226

1227

1228

1229

1230

1231

1232

1233

1234

1235

1236

1237

1238

1239

1240

1241

1242

1243

1244

1245

1246

1247

1248

1249

1250

1251

1252

1253

1254

1255

1256

1257

1258

1259

1260

1261

1262

1263

1264

1265

1266

1267

1268

1269

1270

1271

1272

1273

1274

1275

1276

1277

1278

1279

1280

1281

1282

1283

1284

1285

1286

1287

1288

1289

1290

1291

1292

1293

1294

1295

1296

1297

1298

1299

1300

1301

1302

1303

1304

1305

1306

1307

1308

1309

1310

1311

1312

1313

1314

1315

1316

1317

1318

1319

1320

1321

1322

1323

1324

1325

1326

1327

1328

1329

1330

1331

1332

1333

1334

1335

1336

1337

1338

1339

1340

1341

1342

1343

1344

1345

1346

1347

1348

1349

1350

1351

1352

1353

1354

1355

1356

1357

1358

1359

1360

1361

1362

1363

1364

1365

1366

1367

1368

1369

1370

1371

1372

1373

1374

1375

1376

1377

1378

1379

1380

1381

1382

1383

1384

1385

1386

1387

1388

1389

1390

1391

1392

1393

1394

1395

1396

1397

1398

1399

1400

1401

1402

1403

1404

1405

1406

1407

1408

1409

1410

1411

1412

1413

1414

1415

1416

1417

1418

1419

1420

1421

1422

1423

1424

1425

1426

1427

1428

1429

1430

1431

1432

1433

1434

1435

1436

1437

1438

1439

1440

1441

1442

1443

1444

1445

1446

1447

1448

1449

1450

1451

1452

1453

1454

1455

1456

1457

1458

1459

1460

1461

1462

1463

1464

1465

1466

1467

1468

1469

1470

1471

1472

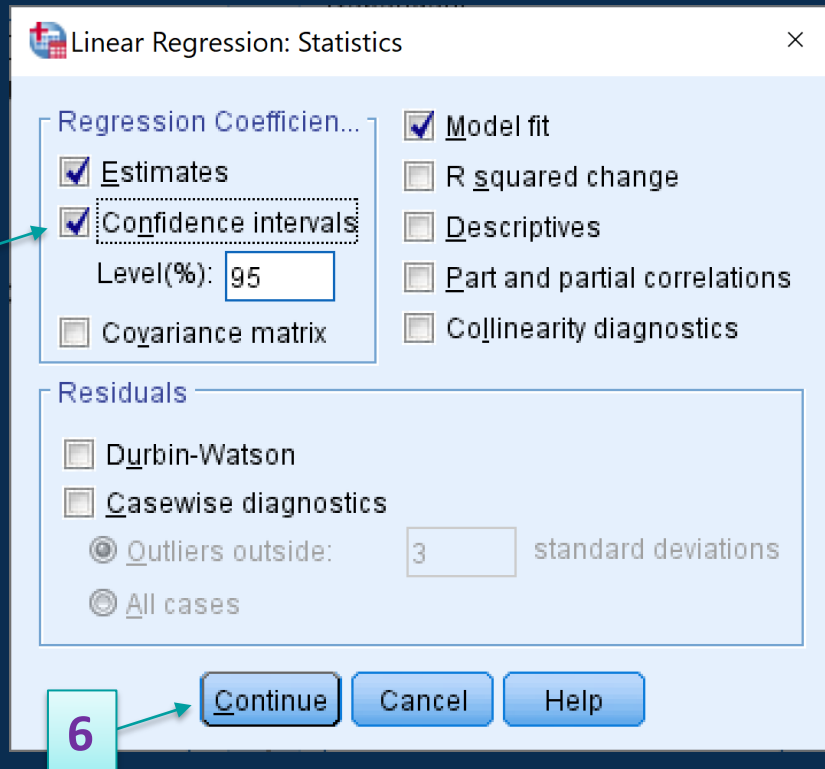
1473

1474

# SPSS Slide: 'how to'

Researchers believe there is a relationship between Violent Crime and the level of urbanicity in an area

Step 2: Compute a Linear regression model for dependent variable 'Violent Crime' and independent variable 'urban' using the dummy variables created



In the Statistics tab.  
Check the 'Estimates'  
Check the 'Confidence Intervals'  
Click on 'Continue'  
Click on 'OK'

# Output and Interpretation Slide

**Model Summary<sup>b</sup>**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.431 <sup>a</sup>	.186	.151	410.381

a. Predictors: (Constant), Medium urbanicity, Low urbanicity

b. Dependent Variable: violent crime rate per 100,00 population

**ANOVA<sup>a</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1808632.428	2	904316.214	5.370	.008 <sup>b</sup>
	Residual	7915378.052	47	168412.299		
	Total	9724010.480	49			

a. Dependent Variable: violent crime rate per 100,00 population

b. Predictors: (Constant), Medium urbanicity, Low urbanicity

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	749.281	72.546		10.328	.000	603.338	895.224
	Low urbanicity	-498.948	182.569	-.368	-2.733	.009	-866.230	-131.666
	Medium urbanicity	-324.531	138.915	-.314	-2.336	.024	-603.991	-45.071

a. Dependent Variable: violent crime rate per 100,00 population

There is a moderate degree of correlation between Violent Crime and Urbanicity  $r = 0.431$ . 18.6% of the variation in Violent crime can be explained by Urbanicity. the regression model statistically significantly predicts the outcome variable i.e., it is a good fit for the data.

On average low urbanised areas have 498.95 less cases of violent crime per 100 000 compared to high urbanised areas ( $\beta_1 = -498.948$ ,  $t = -2.733$ ,  $p < 0.009$ , 95% CI (-866.230, -131.666) , on average med urbanised areas have 324.53 less cases of violent crime per 100 000 compared to high urbanised areas ( $\beta_2 = -324.531$ ,  $t = -2.336$ ,  $p < 0.024$ , 95% CI (-603.991, -45.071)



# Knowledge Check

---

We examined the medical records of participants when they were between 65 and 70 years old, counting the number of health problems they had. Participants were given a questionnaire on how much they've smoked at different times in their life e.g number of cigarettes smoked per day between ages 20 and 50.

The following regression model describes the relationship between health problems (y) and smoking (x)

$$y' = 3.109 + 1.578x .$$

- Output from the analysis of the data showed  $r = 0.77$
  - Effects of both the intercept and slope show  $p = .045$  and  $p = 0.049$  respectively
1. Write an appropriate Null and Alternative hypothesis for these data.
  2. Interpret the coefficients of the regression.
  3. How many health problems will a participant be predicted to have if the number of cigarettes they smoke is 10 and 30. Calculate a confidence interval for the prediction given the s.e. is 0.435



# Knowledge Check Solutions

We examined the medical records of participants when they were between 65 and 70 years old, counting the number of health problems they had. Participants were given a questionnaire on how much they've smoked at different times in their life e.g number of cigarettes smoked per day between ages 20 and 50.

1. Write an appropriate Null and Alternative hypothesis for these data.  
H0: There is no linear association between health problems and amount of smoking e.g. the slope  $\beta_1$  in the population equals to 0  
Ha: There is a linear association between number of health problems and amount of smoking e.g. the slope  $\beta_1$  in the population does not equal to 0
2. Interpret the coefficients of the regression.  
 $\beta_0 = 3.109$  The intercept ( $\beta_0$ ), is the extrapolated number of health problems for a participant who does not smoke, this suggests that if a participant is a non-smoker they will have approx. 3 health problems.  
 $\beta_1 = 1.578$  The estimated slope coefficient ( $\beta_1$ ), suggests a increase of 1 cigarette smoked is associated with a 1.578 increase to number of health problems
3. How many health problems will a participant be predicted to have if the number of cigarettes they smoke is 10 and 30. Calculate a confidence interval for the prediction given the s.e. is 0.435

10 cigarettes will lead to  $3.109 + (10 \times 1.578) = 18.89$  health problems  
95% CI  $(18.89 \pm 1.96 \times 0.435) = (18.04, 19.74)$

30 cigarettes will lead to  $3.109 + (30 \times 1.578) = 50.45$  health problems  
95% CI  $(50.45 \pm 1.96 \times 0.435) = (49.60, 51.30)$

# References

---

**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

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

Chapter 8: Bivariate correlation and regression



# Thank you

## Contact details/for more information:

Zahra Abdulla

Department of Biostatistics and Health Informatics (BHI)

IoPPN

+44 (0)20 7848 0847

Zahra.abdulla@kcl.ac.uk

[www.kcl.ac.uk/xxxx](http://www.kcl.ac.uk/xxxx)

© 2020 King's College London. All rights reserved

