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

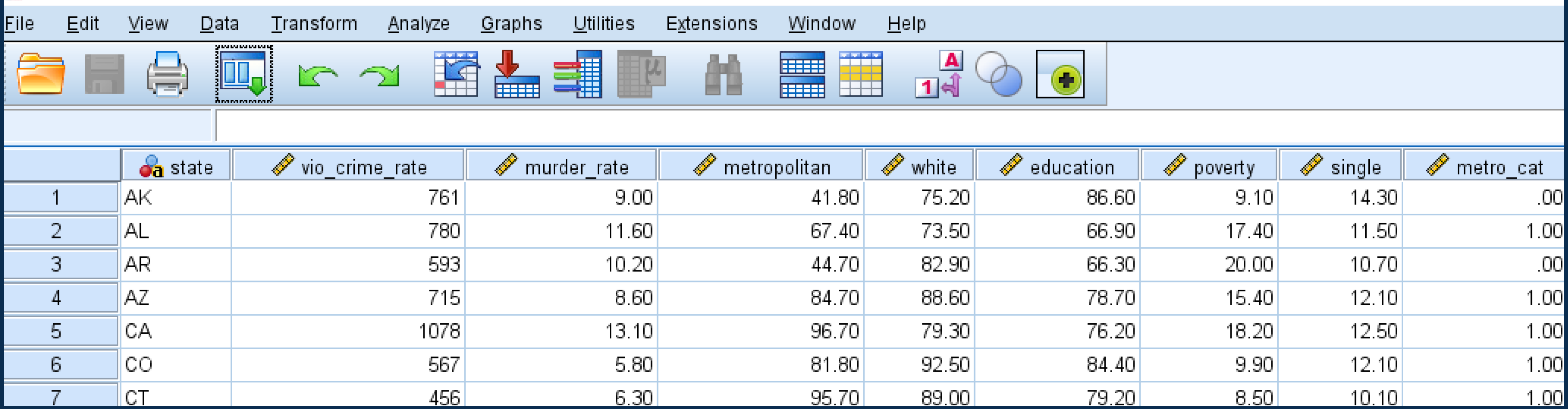
   e.g. repeated measurements collected on people over time are **not** independent

   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

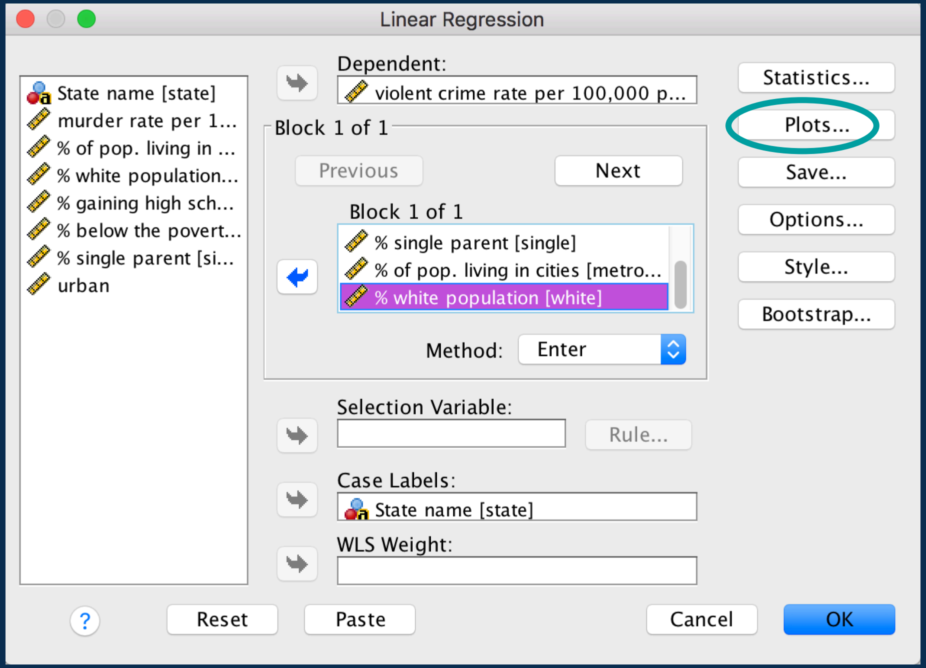
SPSS Slide: ‘how to’

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.



Linear Regression

State name [state]
murder rate per 1...
% of pop. living in ...
% white population...
% gaining high sch...
% below the povert...
% single parent [si...
urban

Dependent: violent crime rate per 100,000 p...

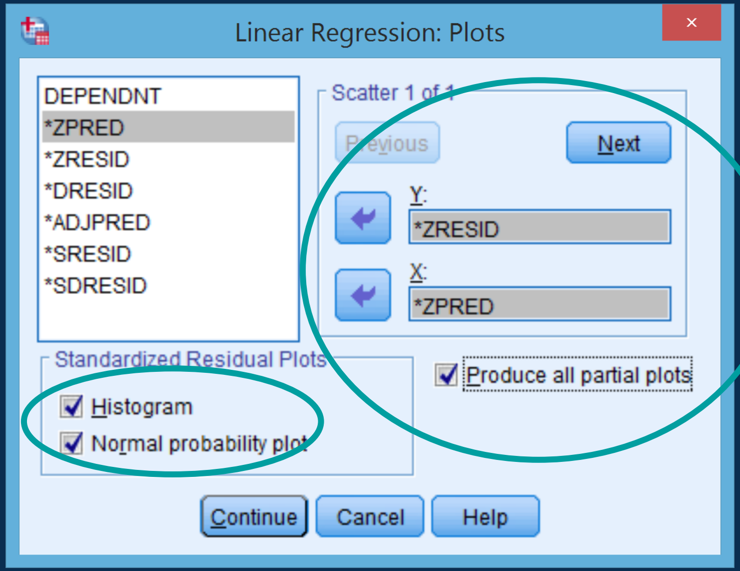
Block 1 of 1
% single parent [single]
% of pop. living in cities [metro...
% white population [white]

Method: Enter

Selection Variable:

Case Labels: State name [state]

WLS Weight:



Linear Regression: Plots

DEPENDNT
*ZPRED
*ZRESID
*DRESID
*ADJPRED
*SRESID
*SDRESID

Scatter 1 of 1

Y: *ZRESID

X: *ZPRED

Standardized Residual Plots

☑ Histogram

☑ Normal probability plot

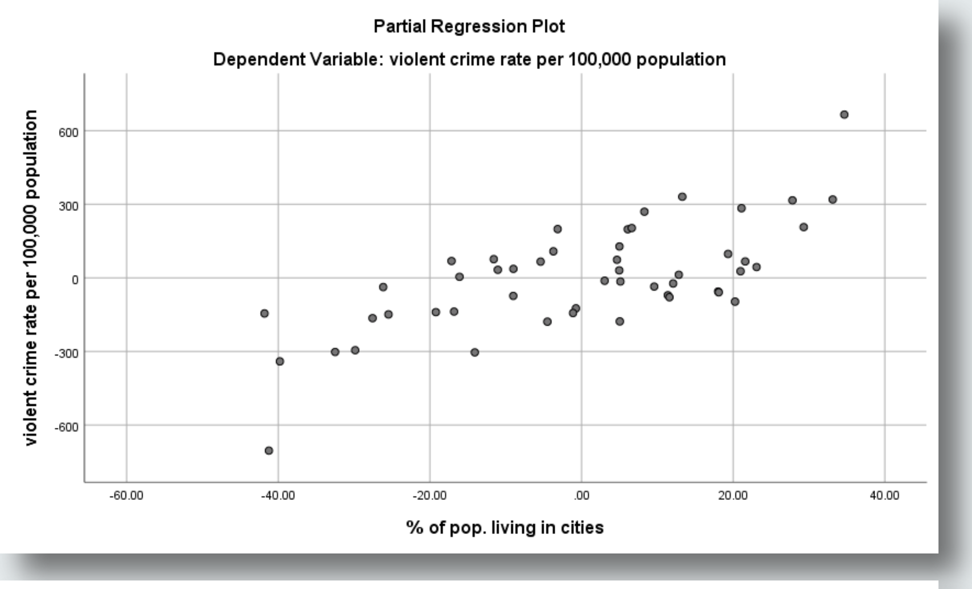
☑ Produce all partial plots

Output and Interpretation Slide – Assessing Linearity

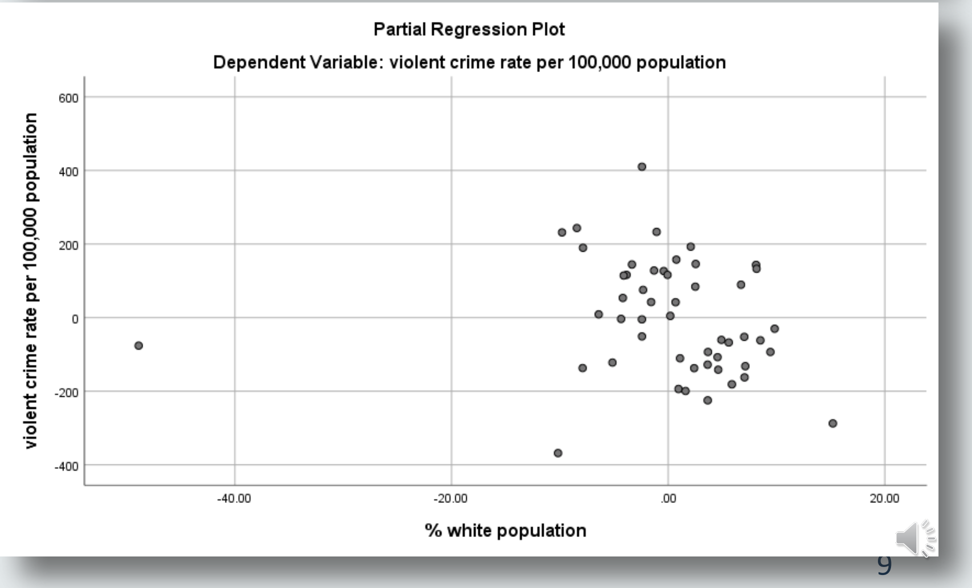
Assumption #1

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



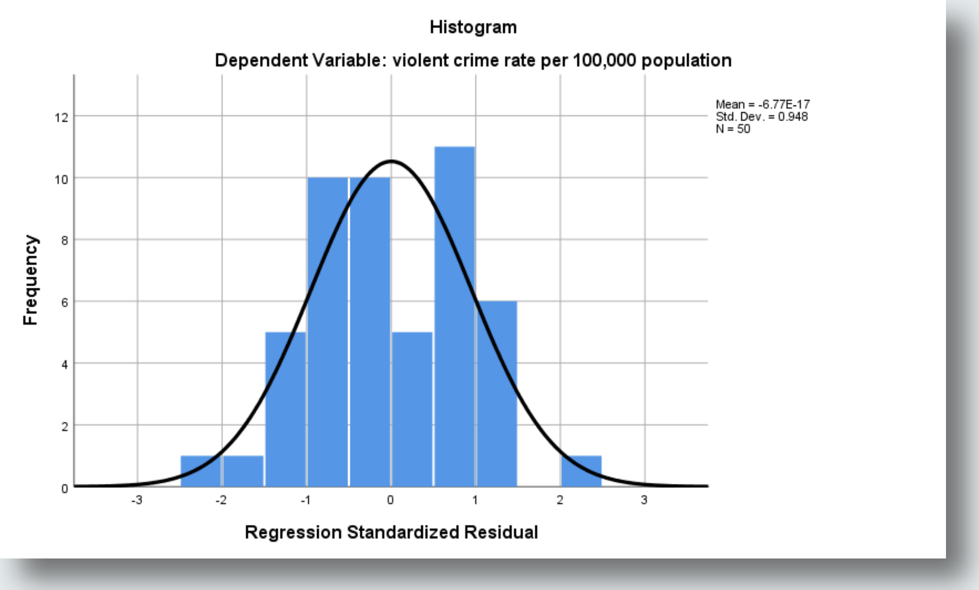
Bottom plot suggests there is not a linear relationship between the independent variable **white** and the outcome variable **crime** – **the linearity assumption is not met** for the **white** variable



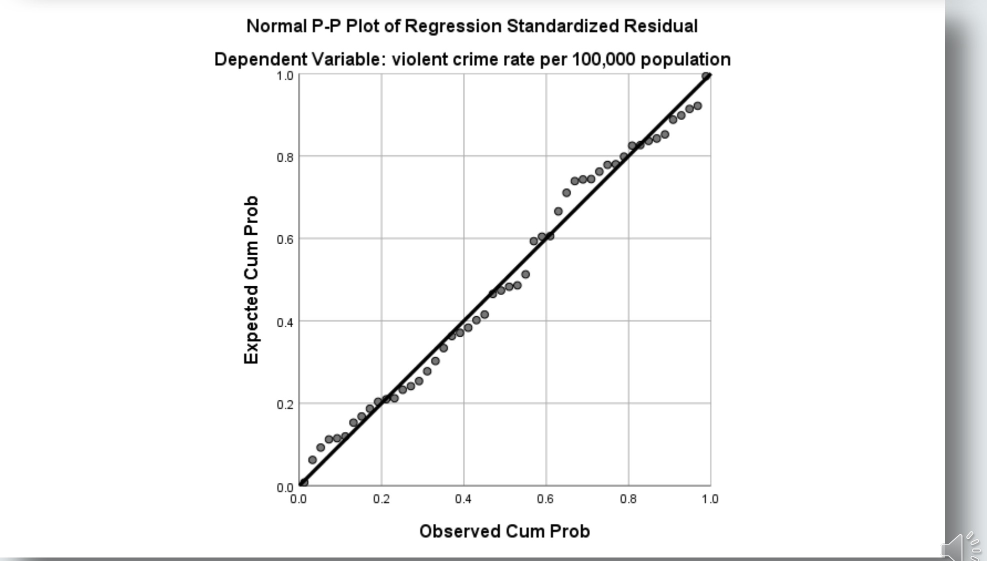
SPSS Interpretation Slide – Plots of Residuals for Assessing Normality

Assumption #2

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



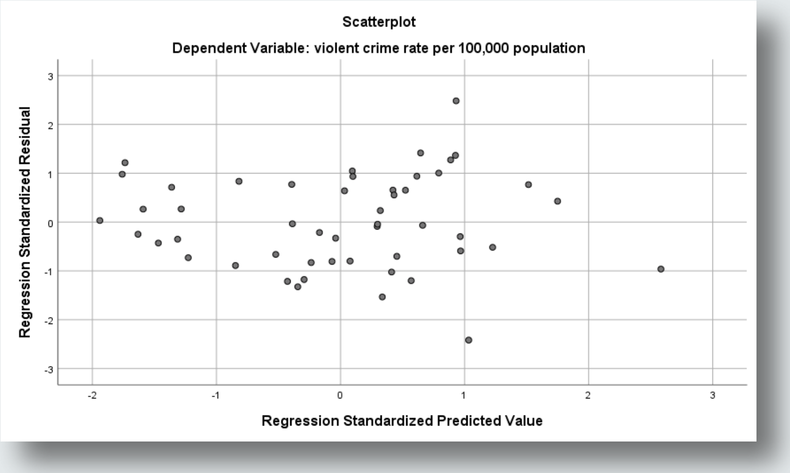
Normal P-P plot – gives similar information to the histogram; here we want to see that the points lie more or less close to the diagonal reference line, which they do.



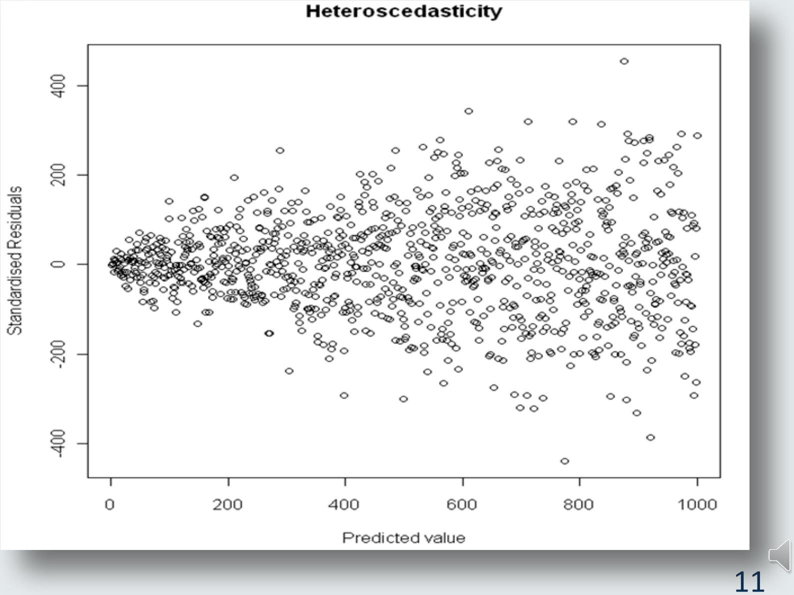
SPSS Interpretation Slide – Assessing Variance Homogeneity

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

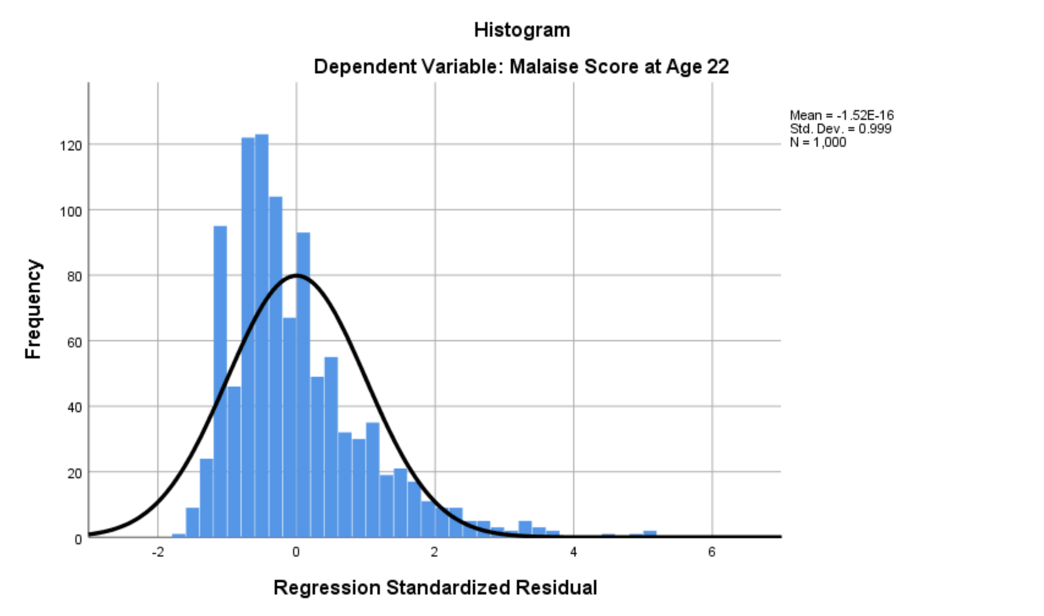
Use the Lecture_6a_data.sav NCDS dataset,

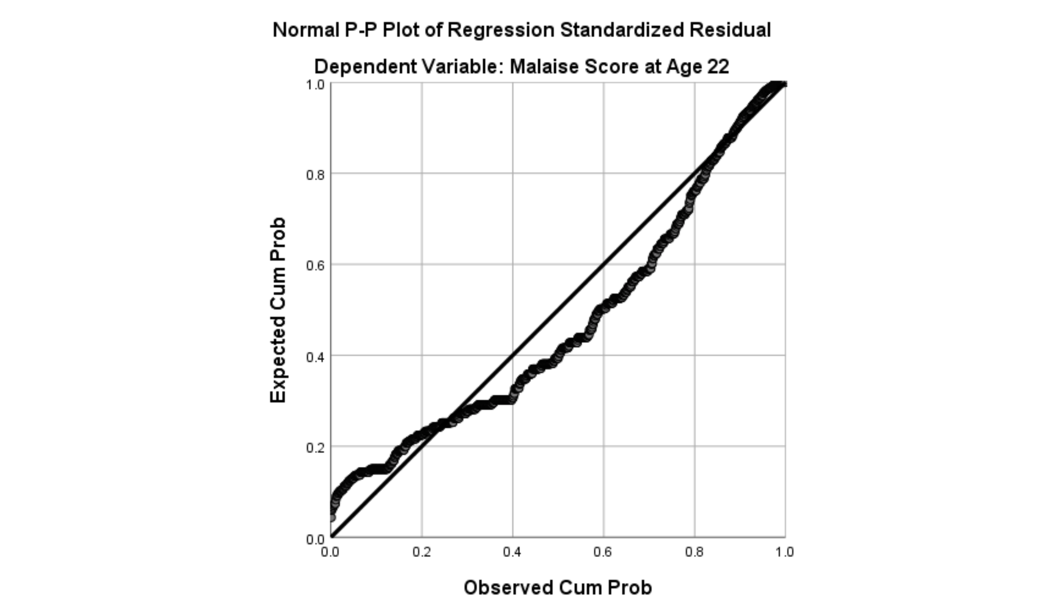
Researchers want to understand if the multiple regression model they have run on malaise score at age 22 as the dependent variable, with reading and sex as independent variables (sex coded as 0 = male 1 = female) meets the model assumptions.

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?

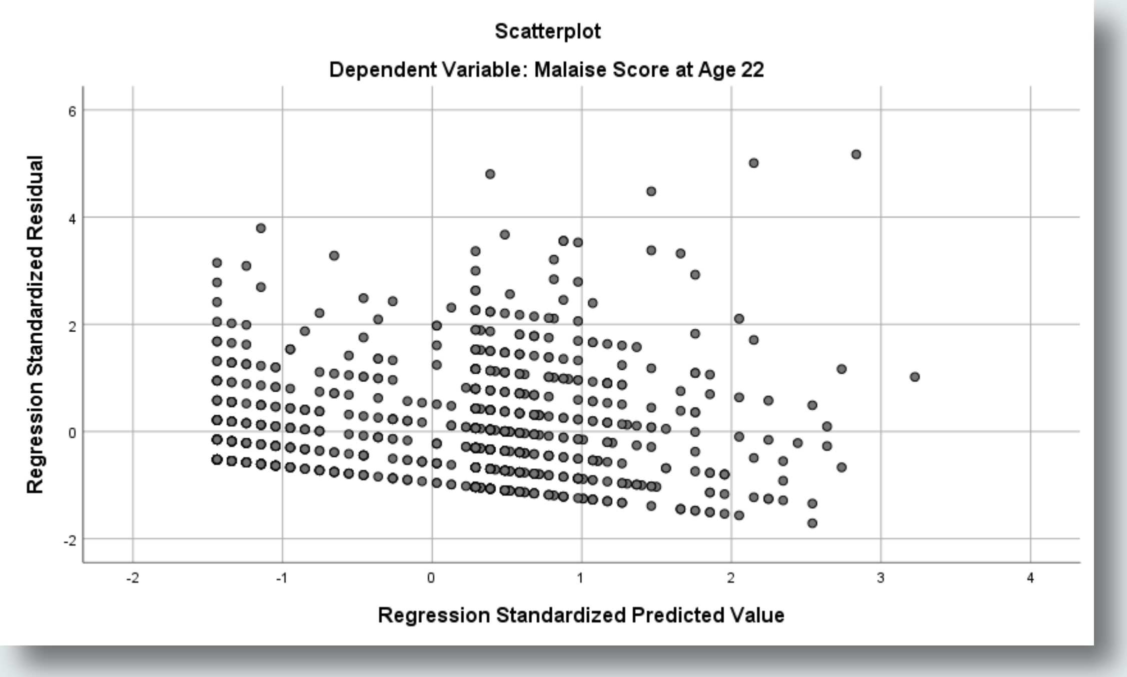




- The residuals seem to follow a skewed distribution with a longer tail to the right side of the histogram.
- The P-P plot also indicates a skewed distribution, with the points not falling along the straight reference line

Knowledge Check Solutions – Model Assumptions

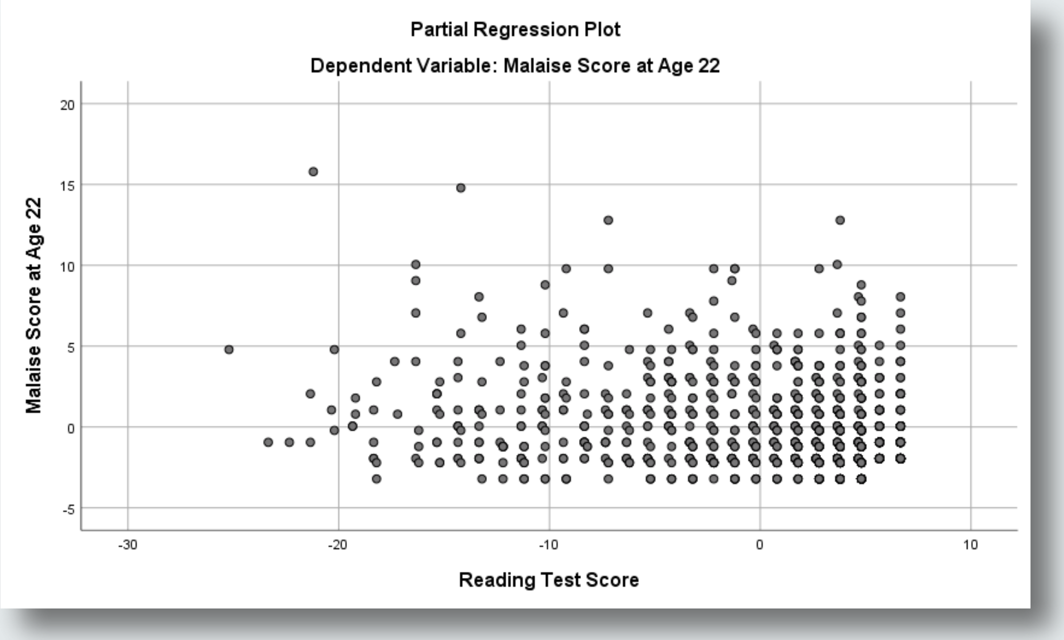
- Q: Fit this model and assess the normality of residuals and homoscedasticity assumptions. Can we make these assumptions for this model or not?



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**