Model Assumptions

Course developed by
Deborah H. Glueck and Keith E. Muller

Slides developed by Jessica R. Shaw, Keith E. Muller, Albert D. Ritzhaupt and Deborah H. Glueck

 1

Learning Objectives

Define the multivariate linear model.

Define the mixed linear model.

Define a residual.

Describe how model assumptions impact power analysis.

2

Learning Objectives

Describe the assumptions of the multivariate linear model.

Define reversible mixed linear models.

Describe the assumptions of reversible mixed models.

3

THREE COMMON STATISTICAL MODELS

4

Recall, three common statistical models are used for multilevel and longitudinal studies

Three common models:

1. The univariate linear model
2. The multivariate linear model
3. The mixed linear model

We will not discuss how to fit the models for data analysis.

5

The univariate linear model is used to analyze experiments with a single outcome measured at only one point in time

Univariate linear models can only be applied to uncorrelated data, as found in studies with no clustering and a single outcome measure.

6

The multivariate linear model is used to analyze experiments with multiple outcomes or hypotheses

Multivariate outcomes may be multiple time points, multiple outcome variables measured at one time point, or multiple outcome variables measured at multiple time points.

7

The mixed linear model is a more flexible model for analyzing experiments with multiple outcomes

The mixed model is more flexible because it makes fewer assumptions than the multivariate model.

We will discuss model assumptions in the next lecture.

8

UNDERSTANDING CORRELATION, RESIDUALS, AND VARIANCE

9

Three interrelated terms are used to describe model assumptions

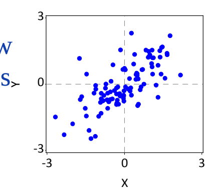
1. Correlation
2. Residual
3. Variance

10

Correlation measures the linear relationship between two variables

Correlation "indicates the strength and direction of the relationship between two random variables."

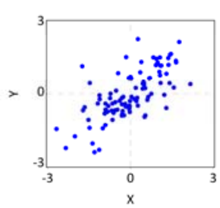
A correlation quantifies how well a straight line describes a scattergram:



Rosner, 2010, p. 127

11

Correlation of 0.70



12

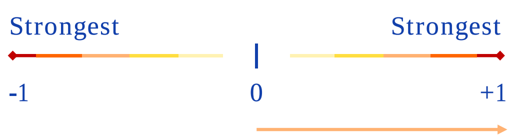
Correlation measures the linear relationship between two variables

Squared correlation, r^2, equals the proportion of variance in Y predicted by X.

The prediction uses only a straight line.

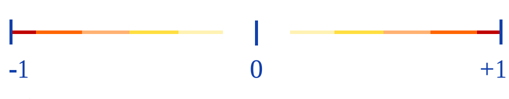
Numbers further from zero represent stronger correlation

Positive correlation (between 0 and 1) indicates that variables change in the same direction.



Correlation (r) ranges from -1 to 1

Negative correlation (between -1 and 0) indicates that two variables change in opposite directions.



A correlation of 0 indicates that two variables are unrelated

Variables with correlation equal to zero have rates and directions of change that are uncorrelated, or unrelated.

16

Variables that are independent always have correlation equal to zero

Mathematically speaking, variables with correlation equal to zero are not necessarily independent.

For the purposes of this class, we will focus on cases where correlation of zero indicates that two variables are independent.

17

The univariate, multivariate and mixed models have a common structure

A model is a statement about a population:

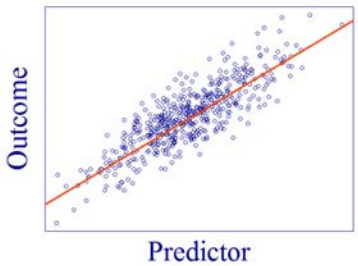
response = prediction + error.

Data analysis uses a sample of data to find estimates of the parts of the model.

We observe the response values.

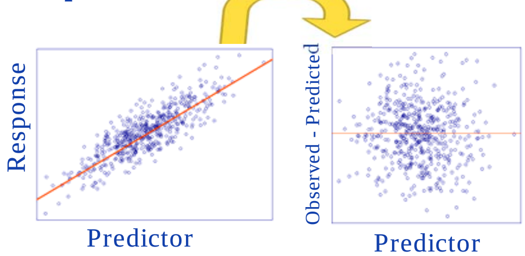
18

Recall, data analysis gives estimates called predicted values



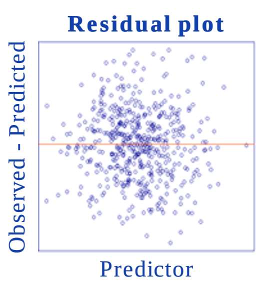
19

Graphing differences between observed and predicted values shows how well the estimated model fits the sample data



20

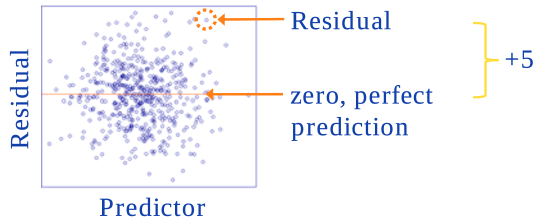
Graphing the differences gives a residual plot, frequently used to check model assumptions



21

A residual is the difference between a single observed outcome value and the model predicted response value

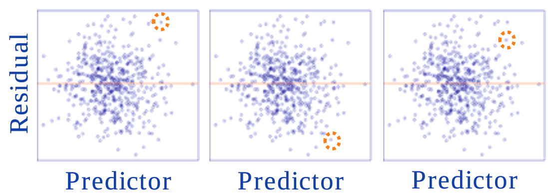
Residual = observed value – predicted value



22

Residuals vary in magnitude and may be positive or negative

Residual value		
+5	-5	+3



23

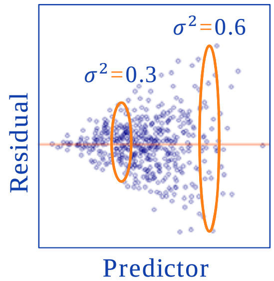
Variance is a summary measure used to describe the spread of many residuals

The variance of the residuals summarizes the degree to which the values observed in an experiment differ from the values predicted by the estimated statistical model.

24

Variance quantifies the spread of values in a set of observations

A greater range in observed values usually corresponds to greater variance in the residuals.



25

UNIVARIATE MODEL ASSUMPTIONS

26

Violating model assumptions may invalidate a data or power analysis

Studies discussed in this course employ one of three key statistical linear models:

1. The univariate model
2. The multivariate model
3. The mixed model

27

We will now discuss the assumptions of the univariate, multivariate and mixed linear models

The assumptions of the three models overlap, with some variation.

28

The univariate model has five assumptions

1. Independence
2. Homogeneity
3. Linearity
4. Existence
5. Normal (Gaussian) errors

29

1. INDEPENDENCE

30

The independence assumption requires that the values for each ISU be independent of values for every other ISU

If the independence assumption is met, then the correlation between independent sampling units (ISUs) is equal to zero.

31

The correlation between ISUs must be equal to zero, but observations within an ISU may still be correlated

Examples of independent units include unrelated people or separate schools located in different cities.

32

2. UNIVARIATE HOMOGENEITY

33

The homogeneity assumption requires that data values vary in predictable and consistent ways

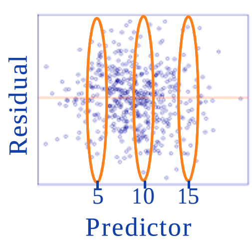
The specific requirements of the homogeneity assumption depend on whether a study has one outcome, or many.

Muller and Fetterman, 2002

34

The univariate homogeneity assumption requires a constant variance of the errors for all values of the predictor

Residual variance can estimate error variance.



Predictor	Variance of residuals
5	0.6
10	0.6
15	0.6

Muller and Fetterman, 2002

35

Residual analysis is outside the scope of the course

Residuals are homogenous only for balanced designs.

In contrast to the variance of the hypothetical errors, the variance of residuals typically depends on the predictor value.

Data analysts account for that and use modified residuals (Muller and Fetterman, Chapter 3).

Muller and Fetterman, 2002

36

3. LINEARITY

37

The linearity assumption mandates a linear relationship between predictors and response variables

The linearity assumption means that the observed change in the response variable is approximately the same for each one unit increase in a predictor.

Muller and Fetterman, 2002

38

4. EXISTENCE

39

The existence assumption requires that observations are real values with a variance less than infinity

Any finite set of values has finite variance.

Concepts of infinity play a useful role in statistical and mathematical theories across the sciences, from astrophysics to zoology (black holes; catastrophe theory in psychology).

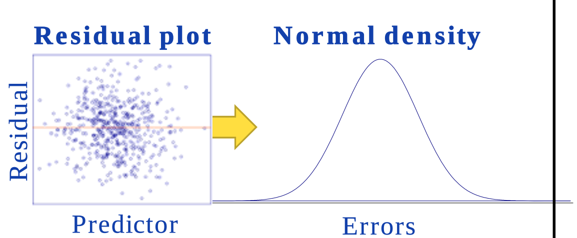
Muller and Fetterman, 2002

40

5. NORMALITY

41

The normality assumption requires that the errors follow a normal distribution (residuals estimate errors)



42

Multivariate model assumptions parallel the univariate model, with some elaborations

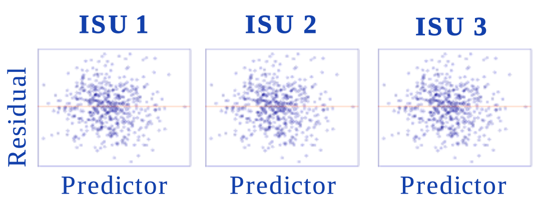
1. Independence
2. Multivariate homogeneity
3. Linearity
4. Existence
5. Multivariate normal (Gaussian) errors
6. Complete data

43

2. MULTIVARIATE HOMOGENEITY

44

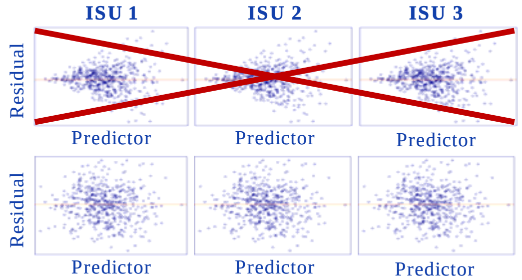
The multivariate homogeneity assumption requires that errors (estimated by residuals) from each ISU have the same pattern of correlation and variance



Muller and Fetterman, 2002

45

With the multivariate homogeneity assumption, the error variance does not depend on predictor values



46

6. COMPLETE DATA

47

The complete data assumption in the multivariate model requires that none of the measurements be missing for any ISU

Some missing (Not usable)

$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$
$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$
$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	$y_{3,4}$
$y_{4,1}$	$y_{4,2}$	$y_{4,3}$	$y_{4,4}$
$y_{5,1}$	$y_{5,2}$	$y_{5,3}$	$y_{5,4}$
$y_{6,1}$	NA	$y_{6,3}$	$y_{6,4}$
$y_{7,1}$	$y_{7,2}$	$y_{7,3}$	$y_{7,4}$
$y_{8,1}$	$y_{8,2}$	$y_{8,3}$	$y_{8,4}$
NA	$y_{9,2}$	NA	$y_{9,4}$

None missing (Data Used)

$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$
$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$
$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	$y_{3,4}$
$y_{4,1}$	$y_{4,2}$	$y_{4,3}$	$y_{4,4}$
$y_{5,1}$	$y_{5,2}$	$y_{5,3}$	$y_{5,4}$
$y_{7,1}$	$y_{7,2}$	$y_{7,3}$	$y_{7,4}$
$y_{8,1}$	$y_{8,2}$	$y_{8,3}$	$y_{8,4}$

48

To justify the deletion of missing observations for the multivariate model, data must be missing at random

Data may only be deleted if missingness is not determined by the outcome or predictors.

49

MIXED MODEL ASSUMPTIONS

50

The mixed model assumptions parallel the multivariate model with elaborations and one exception

1. Independence (but N=1 can be ok)
2. Multivariate homogeneity
   (but can allow groupwise heterogeneity)
3. Linearity (more general form)
4. Existence
5. Normal (Gaussian) errors
6. ~~Complete data~~

51

Unlike the multivariate model, the mixed model can accommodate missing data

Original data

$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$
$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$
$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	$y_{3,4}$
$y_{4,1}$	$y_{4,2}$	$y_{4,3}$	$y_{4,4}$
$y_{5,1}$	$y_{5,2}$	$y_{5,3}$	$y_{5,4}$
$y_{6,1}$	NA	$y_{6,3}$	$y_{6,4}$
$y_{7,1}$	$y_{7,2}$	$y_{7,3}$	$y_{7,4}$
$y_{8,1}$	$y_{8,2}$	$y_{8,3}$	$y_{8,4}$
NA	$y_{9,2}$	NA	$y_{9,4}$

Data used

$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$
$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$
$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	$y_{3,4}$
$y_{4,1}$	$y_{4,2}$	$y_{4,3}$	$y_{4,4}$
$y_{5,1}$	$y_{5,2}$	$y_{5,3}$	$y_{5,4}$
$y_{6,1}$	NA	$y_{6,3}$	$y_{6,4}$
$y_{7,1}$	$y_{7,2}$	$y_{7,3}$	$y_{7,4}$
$y_{8,1}$	$y_{8,2}$	$y_{8,3}$	$y_{8,4}$
NA	$y_{9,2}$	NA	$y_{9,4}$

52

In general, mixed models have fewer restrictions than multivariate models

In addition to missing data, many mixed models can accommodate mistimed data and repeated covariates.

Kachman, 2015

53

In this course, we only consider ‘reversible’ mixed models

‘Reversible’ mixed models are models which are equivalent to multivariate models.

54

Mixed models are 'reversible' under the following three conditions

1. Each ISU contains the same number of units of observation.
2. Units of observation are measured at the same times, the same locations, or are measured for the same variables.
3. Predictor variables are only measured once.

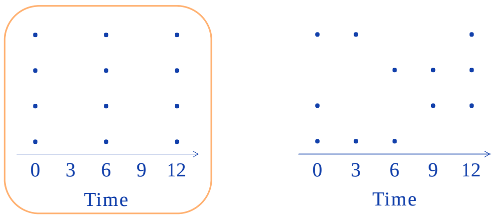
REVIEW OF LEARNING OBJECTIVES

Which study design has an equal number of observations for each ISU?

ISU	Observation 1	2	3
1	·	·	·
2	·		·
3	·	·	
4	·		·
5	·	·	
6		·	·
7	·	·	

ISU	Observation 1	2	3
1	.	.	.
2	.	.	.
3	.	.	.
4	.	.	.
5	.	.	.
6	.	.	.
7	.	.	.

Which study design has equal timing of measurements?



58