Jumping into Statistics: Introduction to Study Design and Statistical Analysis for Medical Research Using JMP Pro Statistical Software

FALL 2022/SPRING 2023

DR. TERRIE VASILOPOULOS, DR. CYNDI GARVAN,
& DR. PENNY REYNOLDS

Meet the Instructors

TERRIE VASILOPOULOS, PHD

Research Associate Professor in Anesthesiology and Orthopaedics and Rehabilitation

CYNTHIA GARVAN, PHD

Research Professor in Anesthesiology

PENNY REYNOLDS, PHD

Research Assistant Professor in Anesthesiology and Veterinary Medicine

Course Objectives

- Review fundamentals of study design and research methodology
- Understand how to choose best statistical test for your research question
- Practice basic statistical analysis use JMP Pro Software

Course Topics

- Life Cycle of Research and Asking a Good Research Question
- Choosing the Right Study Design for Your Research
- Clinical Trial Design
- Populations, Samples, and Hypothesis Testing in Medical Research
- Introduction to Data Types
- Best Practices in Data Collection and Database Management: Getting Started with SAS JMP Pro
- Summarizing and Visualizing Data
- Statistical Methods and How to Choose Them
- Risk Assessment Methods
- Introduction to Regression and Correlation
- Time-to-Event (Survival) Analysis
- Methods for Clinical Diagnostic Testing

Introduction to Regression and Correlation

1/25/2023

Learning Objectives

Participants will be able to:

1) Distinguish between correlation and regression, as well as between different types of correlation and regression methods.

2) Interpret correlation and regression coefficients.

3) Conduct regression and correlational analyses in SAS JMP Pro.

Why is this topic important?

Correlation is THE way of understanding the linear relationship between two numeric/ordinal variables.

Regression allows us to:

1) **Understand the relationship between a single numeric response (i.e., outcome) variable and a set of predictor variables.**
2) **Make a predictive model which quantifies uncertainty of our prediction.**

Overview

Correlation and Regression

1) Correlation
2) Regression
3) Multiple Regression
4) Relationship between Correlation and Regression
5) R^2: The Coefficient of Determination
6) Assumptions of Regression

1. Correlation

Correlation

Correlation is the measure of the strength and direction of *linear association* between two approximately normally distributed and independent measurements.

Correlation is not causation, nor does it imply a causal relationship.

The correlation coefficient ***r***

- ranges from –1 to +1.
- can never be greater than 1 or less than -1
- has no units of measurement

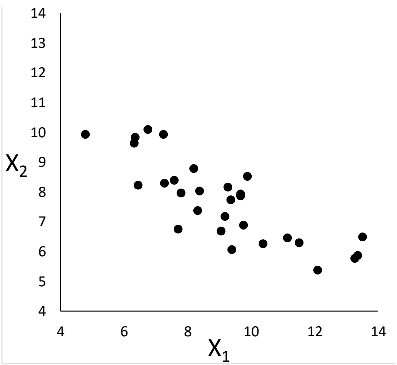
Strength of the correlation – effect size

The absolute value of ***r*** *(*|*r*| or –*r*, +*r*) is a *rough* measure of the strength and the "noisiness" of the relationship:

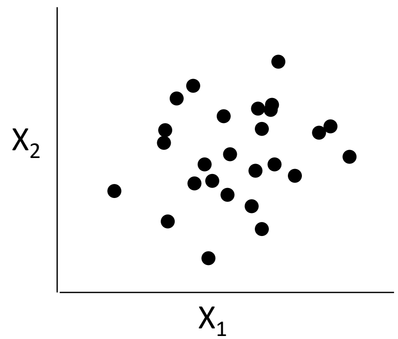
None or very weak	$|r| < 0.3$
Weak	$0.3 < |r| < 0.5$
Moderate	$0.5 < |r| < 0.7$
Strong	$|r| > 0.7$

Always report the actual value of the correlation coefficient. Do not merely describe correlation as low, moderate, or strong without numbers, and without scientifically justifying these categorizations in the context of the study.

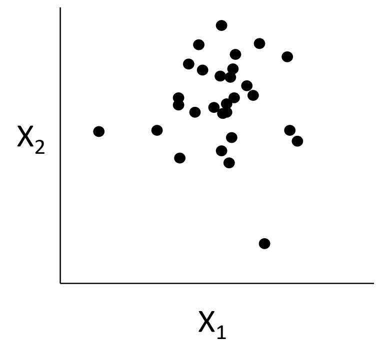
Checks to assess strength and appropriateness of correlation



Strong negative correlation
r = –0.82



Moderate positive correlation
r = 0.48



No correlation
r = –0.07

Scatterplots are indispensable for simple visual assessments of the data to determine

- if the assumption of linear association is appropriate,
- and to show the relationship between two variables.

Anscombe’s Quartet: The Importance of Graphs

Anscombe’s Quartet are four graphs constructed in 1973 by the statistician Francis Anscombe to demonstrate

- the importance of graphing data before analyzing it
- the effect of outliers and other influential observations on statistical properties.

RULE Always plot the data before ANY analysis.

Anscombes' data

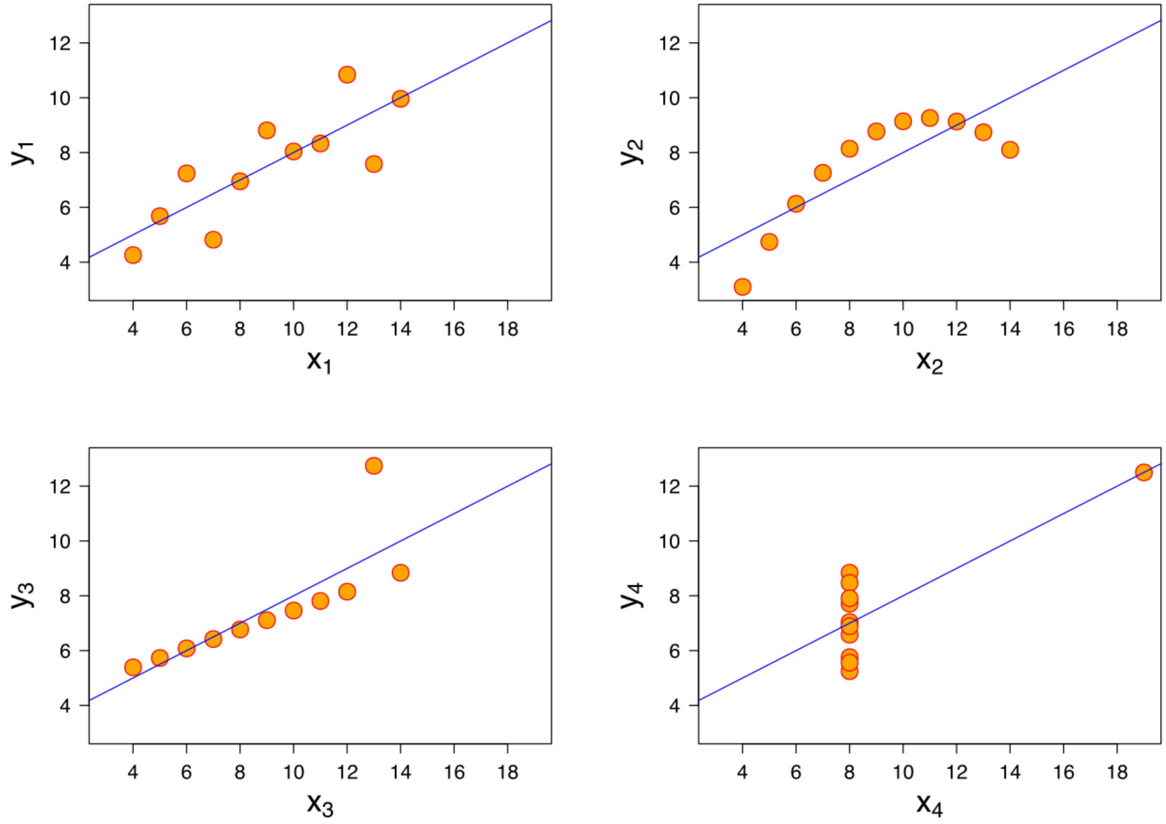
All four datasets have *identical summary statistics* (mean, SD, correlation coefficient r, intercept, slope).

However, scatterplot graphs of these data show that the behaviour of each dataset is quite different.

	Dataset 1		Dataset 2		Dataset 3		Dataset 4	
	X_1	Y_1	X_2	Y_2	X_3	Y_3	X_4	Y_4
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.76	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	8	5.56
	12	10.84	12	9.13	12	8.15	8	7.91
	7	4.82	7	7.26	7	6.42	8	6.89
	5	5.68	5	4.74	5	5.73	19	12.5
Mean	**9.0**	**7.5**	**9.0**	**7.5**	**9.0**	**7.5**	**9.0**	**7.5**
SD	**3.3**	**2.0**	**3.3**	**2.0**	**3.3**	**2.0**	**3.3**	**2.0**
Correlation r	**0.82**		**0.82**		**0.82**		**0.82**	
Intercept	**3**		**3**		**3**		**3**	
Slope	**0.5**		**0.5**		**0.5**		**0.5**	

Anscombe's quartet

This graphic represents the four datasets defined by Francis Anscombe
The descriptive statistics (mean, variance, correlation and regression line) ***are the same.***
The correlation of each X and Y pair is ***r = 0.82***.



Reference: Anscombe, Francis J. (1973) Graphs in statistical analysis. American Statistician, 27, 17–21.

Always plot your data. If you are reading a journal article, look for the data plots. If data plots are not presented, question how results could be affected by irregular patterns in the data.

KEY POINT

Pearson versus Spearman Correlation

Pearson

- Linear relationship of two continuous variables
- Distribution of each variable is normal

Spearman

- Linear relationship of two variables, either of which could be of the continuous or ordinal data type
- Based on data ranks
- Distribution of each variable is not assumed to be normal

2. Regression

Regression

Linear regression models a straight-line relationship between a *response*, or *output*, variable, and one or more *predictor, input,* or *explanatory*, variables.

Simple linear regression models the relationship between a single response and a single predictor variable

Multiple linear regression models the relationship between a single response and two or more predictor variables.

Linear regression

- **quantifies** the functional relationship between the response Y and explanatory variables X,
- **predicts** or forecasts future values of the response variable Y for given values of the explanatory variables X.

Bivariate Continuous Data

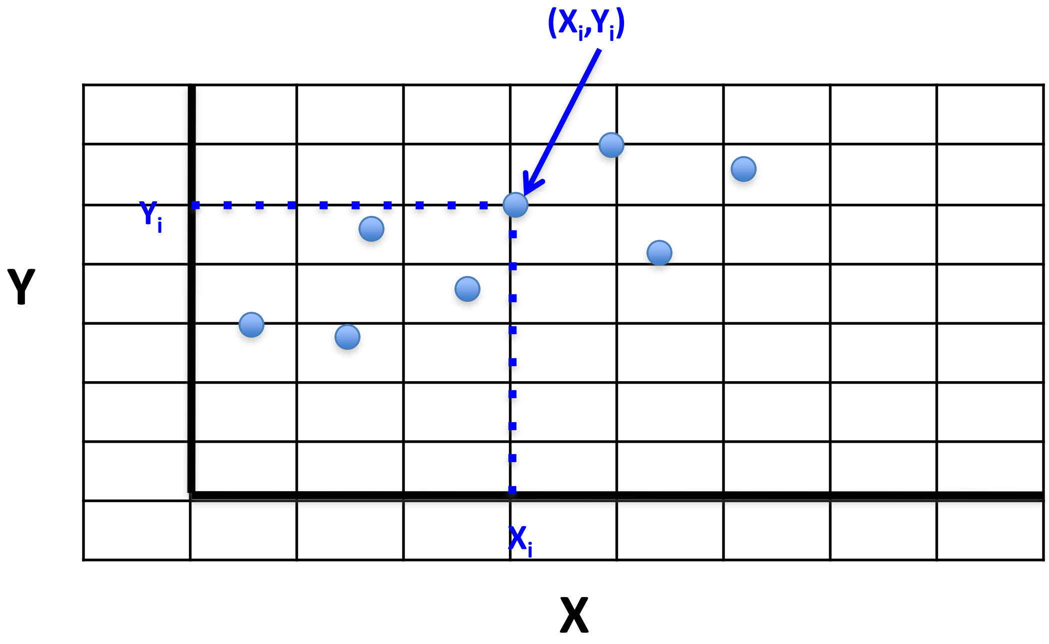
Suppose a researcher is interested in testing if there is a relationship between continuous measures (X and Y). For example, let X= age and Y=time to recover from surgery. The data collected in this study looks like:

ID	X	Y
100	X_1	Y_1
101	X_2	Y_2
102	X_3	Y_3
103	X_4	Y_4

The X_i and Y_i are both continuous data.

Scatter plot of Bivariate Numerical Data

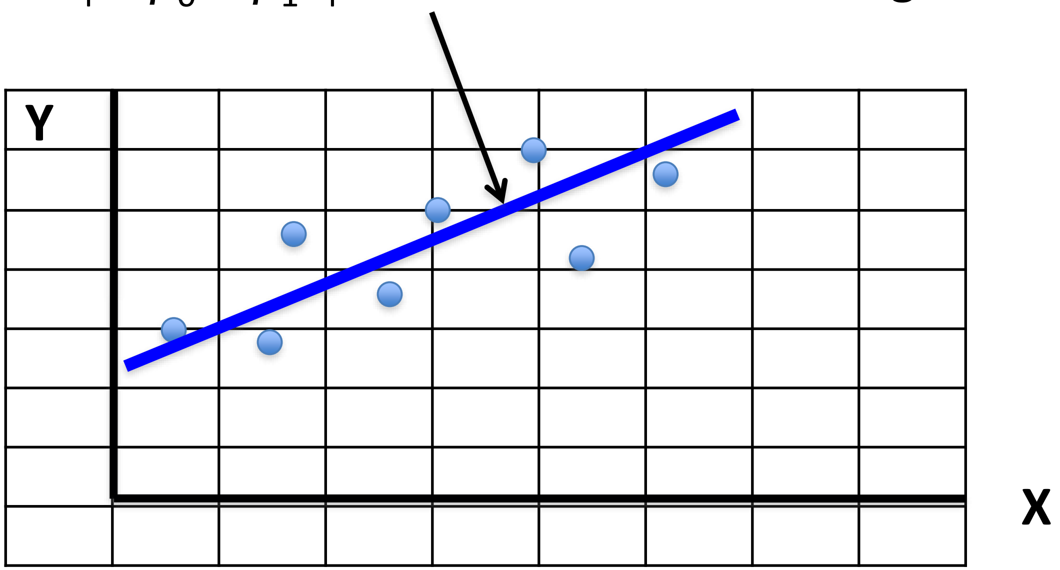
This scatter plot shows the graph of 8 observations of Y for a given values of X. There are 8 (X_i, Y_i) pairs.



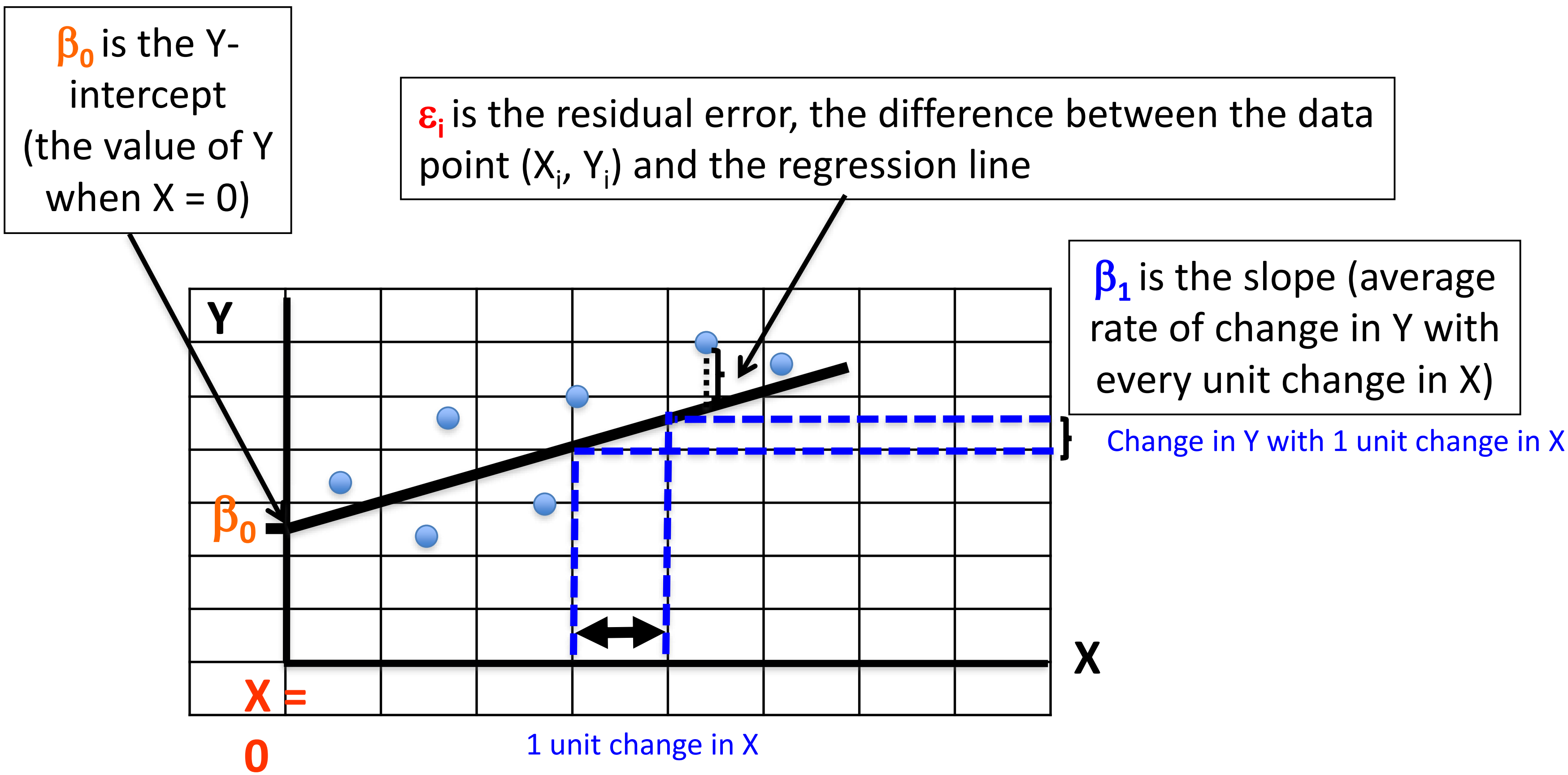
The regression model is given by:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$Y_i = \beta_0 + \beta_1 X_i$ is the *model* of the straight line, ε_i are the *residuals*



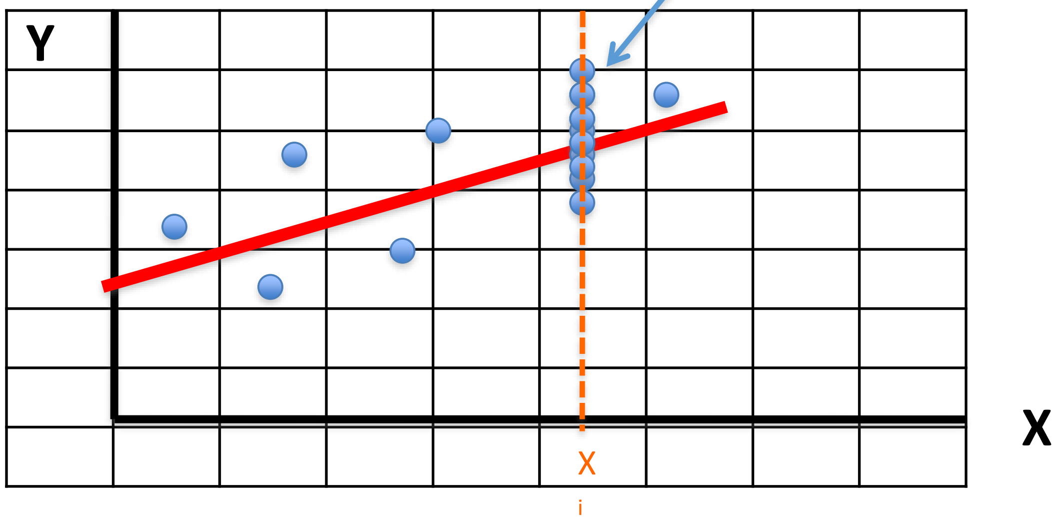
Understanding the Regression Model



Understanding the Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \varepsilon_i \sim N(0,\sigma^2)$$

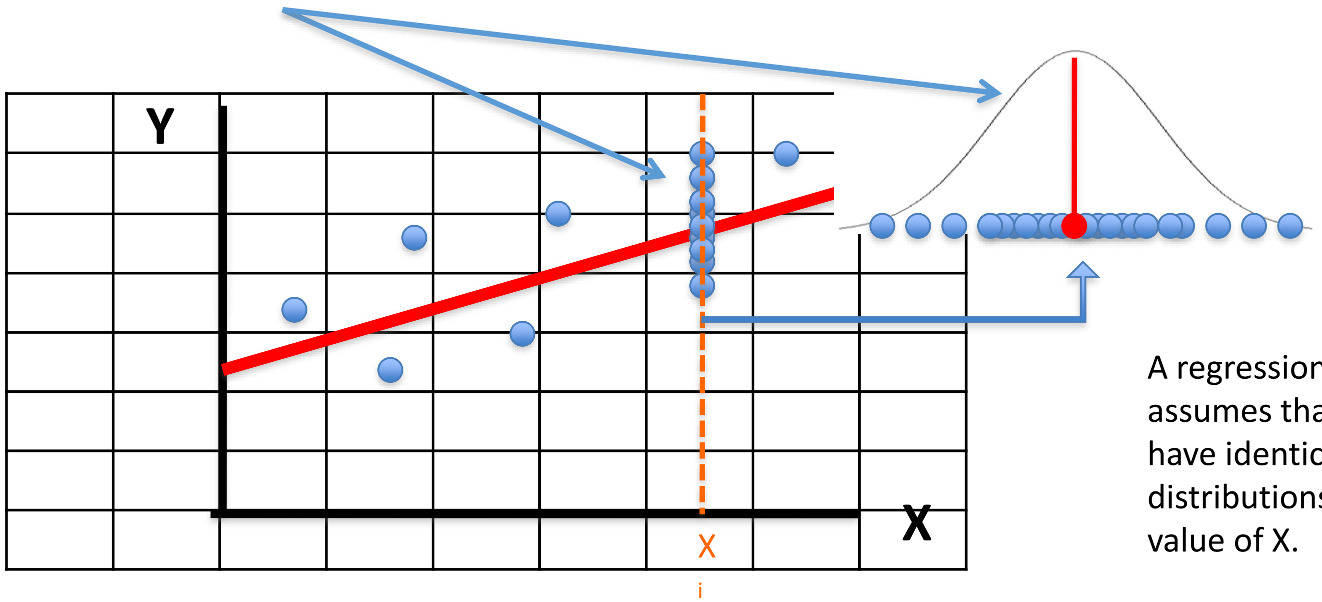
To understand what $\varepsilon_i \sim N(0,\sigma^2)$ means, imagine that you have a large amount of data with an X value = X_i



Understanding the Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \varepsilon_i \sim N(0,\sigma^2)$$

$\varepsilon_i \sim N(0,\sigma^2)$ means that the Y data at a fixed X level (i.e., X = X_i) has a normal distribution centered at the regression line with a variance of σ^2.



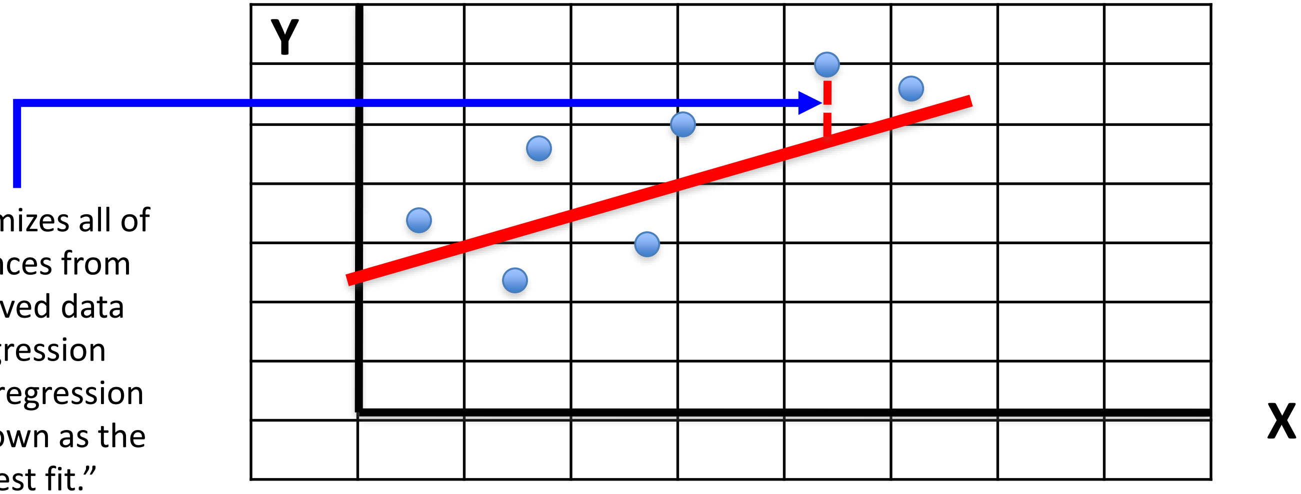
A regression model assumes that the data have identical normal distributions for every value of X.

Understanding the Regression Equation

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \varepsilon_i \sim N(0,\sigma^2)$$

To find statistics to estimate β_0 and β_1, calculus is used to minimize the residual errors. This estimation method is known as the principle of least squared errors or ordinary least squares (OLS).

OLS minimizes all of the distances from the observed data to the regression line. The regression line is known as the line of "best fit."



Formulas

Resource Slide

Residual = difference between the observed value of Y and the value of Y which falls on the regression line (i.e., the predicted value of Y).

$$\varepsilon_i = \left(Y_i - \hat{Y}_i\right)$$

$$\hat{\beta}_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Slope of the regression line.

Intercept of the regression line.

Resource Slide

Notation

	Symbol
Observed value of response variable	Y
Explanatory variable	X
Population parameter for intercept	$\boldsymbol{\beta_0}$
Population parameter for slope (i.e., regression coefficient)	$\boldsymbol{\beta_1}$
Statistic which is estimate of intercept	$\widehat{\boldsymbol{\beta}_0}$ **or** $\mathbf{b_0}$
Statistic which is estimate of slope (i.e., regression coefficient)	$\widehat{\boldsymbol{\beta}_1}$ **or** $\mathbf{b_1}$
Predicted value of Y from the regression line (also called Y-hat)	$\hat{Y}$
Residual (distance between Y and $\hat{Y}$)	$\boldsymbol{\varepsilon_i}$
Variance of data around each fixed value of X	σ^2

Simple Linear Regression – one predictor

The regression model of the form:

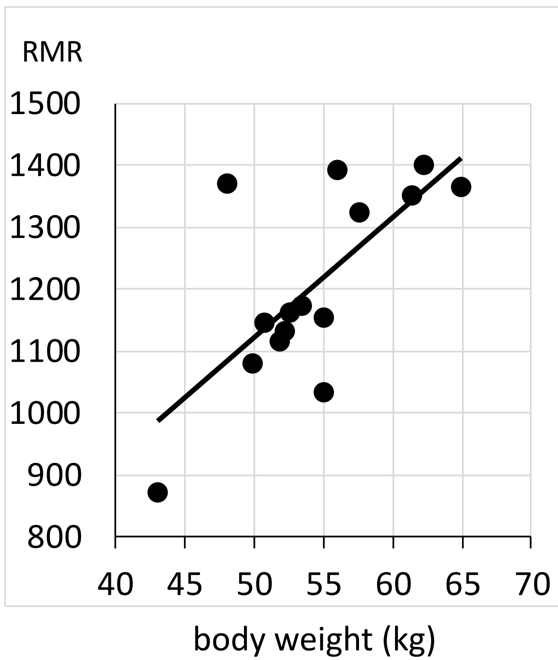
$$Y_i = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} X_i + \boldsymbol{\varepsilon_i} \qquad \boldsymbol{\varepsilon_i} \sim N(0,\sigma^2)$$

is called a **simple** linear regression because there is one "X" variable and the X variable has the continuous data type. It is called a simple **linear** regression because the parameters $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ are linear (i.e., not squared or cubed, etc.).

Example of Simple Regression: Resting metabolic rate and weight

BW kg	RMR
49.9	1079
50.8	1146
51.8	1115
52.6	1161
57.6	1325
61.4	1351
62.3	1402
64.9	1365
43.1	870
48.1	1372
52.2	1132
53.5	1172
55.0	1034
55.0	1155
56.0	1392

Data on resting metabolic rate (RMR, kcal/24 h) and body weight (kg) were obtained for 15 women.



The regression of RMR on body weight (BW) is

$$RMR = 143.1 + 19.6 \cdot BW$$

$N = 15$

Example: Resting metabolic rate and weight

The estimates for the coefficients are :

	Estimate	SE	t Stat	P-value
Intercept	143.1	293.8	0.487	0.634
Slope	19.6	5.4	3.631	0.003

The intercept is not significantly different from zero

The slope is significantly different from zero

RMR increases by almost 20 kcal/day for every 1 kg increase in body weight

The ANOVA is:

	df	SS	MS	F	P-value
Regression	1	172595.29	172595.29	13.19	0.003
Residual	13	170163.65	13089.51		
Total	14	342758.93			

The overall F-test is statistically significant

3. Multiple Regression

Multiple Regression

Multiple regression is an extension of simple linear regression. It is used when we want to model a continuous response or outcome variable (i.e., the Y variable) in terms of two or more predictor variables (X variables).

Rule of thumb: Multiple regression should have at least 10 observations per variable.

Multiple linear regression

Multiple linear regression can be used for several purposes:

1. To predict Y from several X variables.
2. To adjust data. You are most interested in the effect of one particular X variable and therefore need to isolate its effects from other X variables. (This is also called ANCOVA, **AN**alysis of **COVA**riance). In this model the control variables are also called ***covariates***.
3. To assess interactions among multiple variables to determine if the effect of one depends on one or more of the other X variables influence.
4. To model nonlinear data.

The equation for a multiple regression model is given below:

$$\mathbf{Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} \ldots + \beta_p X_{pi} + \varepsilon_i} \text{ where } \boldsymbol{\varepsilon_i} \sim N(0,\sigma^2) \quad i = 1, \ldots, n$$

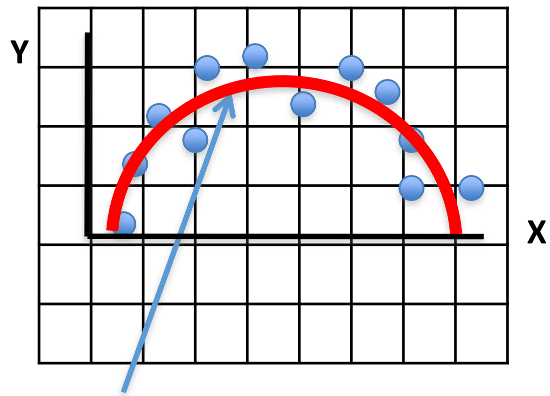
Each X is a variable and can be of any data type.

Each β is a regression coefficient. If the test of hypothesis concludes that β is not zero then we have evidence to support a relationship between this X and Y.

The ε_i's are the errors. They measure the departure from the actual (observed) Y data values and the "fitted" Y values (the Y values predicted by the model).

There are "n" subjects

Multiple regression and curvilinear relationships



Curvilinear relationship between X and Y

We may be able to model a curvilinear relationship between X and Y by squaring the predictor :

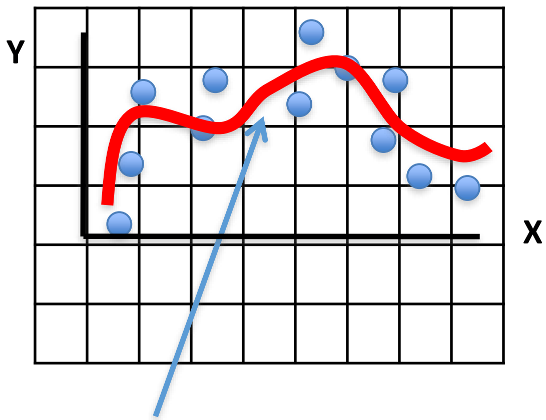
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 (X_{1i})^2 + \varepsilon_i$$

This relationship is quadratic so it is modeled by a quadratic equation

Example: Y = bank account balance, X = age

Model $\mathbf{(\$)_i = \beta_0 + \beta_1(age)_i + \beta_2(age_i)^2 + \varepsilon_i}$

Multiple regression and curvilinear relationships



Curvilinear relationship between X and Y

Sometimes there is a more complex curvilinear relationship between X and Y. We can model this relationship by including an "X-square" term, "X-cubed" term, etc. in the model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 (X_{1i})^2 + \beta_2 (X_{1i})^3 + \varepsilon_i$$

Multiple regression gives us a powerful tool for the type of data analysis needed to address many complex research questions.

KEY POINT

Multiple Regression Example

Investigators wished to obtain a regression model of patient BMI (kg/m^2) as a function of waist circumference (WC, cm; X_1) and mid-upper arm circumference (MUAC, cm; X_2) in 86 female patients.

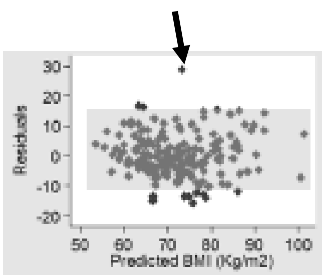
The model is $Y = b_0 + b_1X_1 + b_2X_2$

The regression is BMI = -5.94 + 0.18·WC + 0.59·MUAC

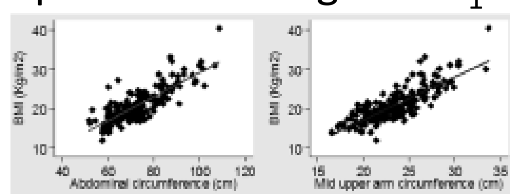
b_0= -5.94 (95%CI -8.10,-3.77); b_1 = 0.18 (95%CI 0.14, 0.22); b_2 = 0.59 (95%CI 0.45, 0.74)

Diagnostics

Residual plot shows outlier



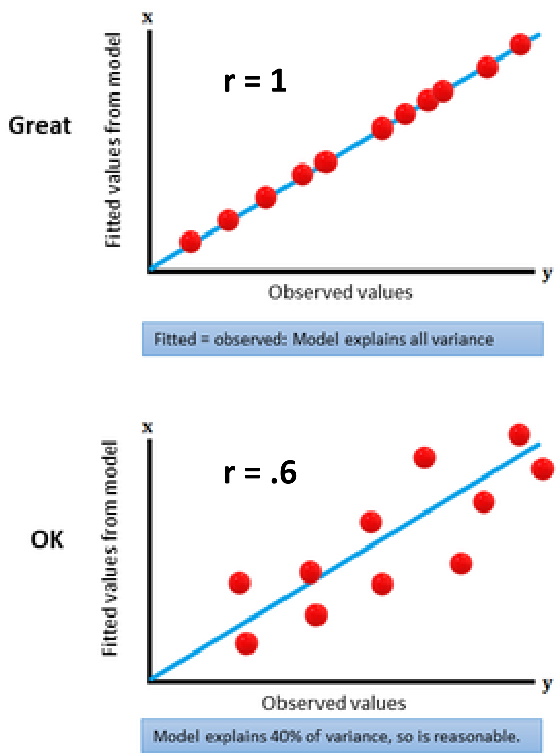
Linear plots of BMI against X_1 and X_2 are linear



Adapted from Bland http://www-users.york.ac.uk/~mb55/

4. Relationship between Correlation and Regression

Relationship between correlation and slope



Relationship between correlation and slope

There is a mathematical relationship between correlation (ρ) and slope of regression line ($\boldsymbol{\beta_1}$):

$$\beta_1 = \rho \, (\sigma_y / \sigma_x)$$

Where σ_y is the standard deviation of the Y data and σ_x is the standard deviation of the X data.

This relationship says that a change of one standard deviation in X corresponds to a change of ρ standard deviations in Y. When X and Y are perfectly correlated (i.e., $\rho = 1$ or $\rho = -1)$, then a change of one standard deviation in X corresponds to a change of one standard deviation in Y. As the correlation grows less strong, the predicted Y moves less in response to changes in X.

Note: A test of hypothesis about ρ is mathematically equivalent to a test of hypothesis about $\boldsymbol{\beta_1}$.

The correlation coefficient is mathematically equivalent to the slope in a regression model.

KEY POINT

5. R^2: The Coefficient of Determination

R^2: The Coefficient of Determination

The statistic R^2 is proportion of variation in Y explained by the linear regression model fitted to the data.

$$R^2 = 1 - \frac{\text{Unexplained variation}}{\text{Total variation}}$$

Example. $R^2 = 0.91$: 91% of the variation in Y can be explained by regression on X.

In the case of simple linear regression (one X), R^2 is equal to the correlation coefficient squared.

6. Assumptions of Regression

Assumptions for Validity of Regression Analysis

Linearity - the relationships between the predictors and the outcome variable should be linear.

Normality - the errors should be normally distributed.

Homogeneity of variance (homoscedasticity) - the error variance should be constant.

Independence - the errors associated with one observation are not correlated with the errors of any other observation.

Errors in variables - predictor variables are measured without error.

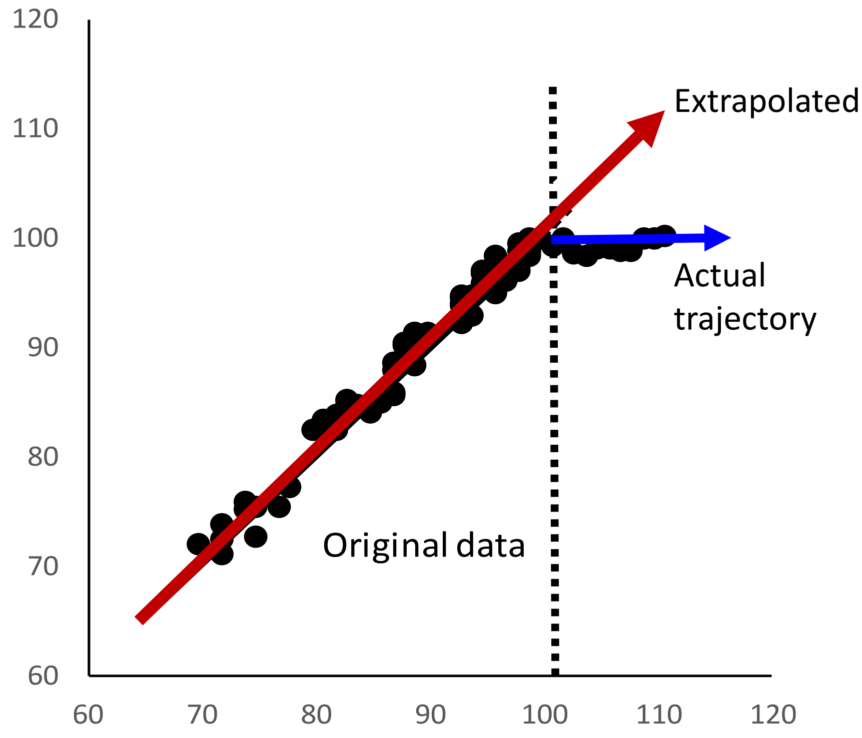
Model specification - the model should be properly specified (including all relevant variables, and excluding irrelevant variables)

No Collinearity - If two predictors are extremely highly correlated, estimates of model parameter can be biased.

Misuse of regression: extrapolation

Extrapolating a fitted regression beyond the range of the data used to obtain it can be extremely misleading if the relationship does not hold outside that range.

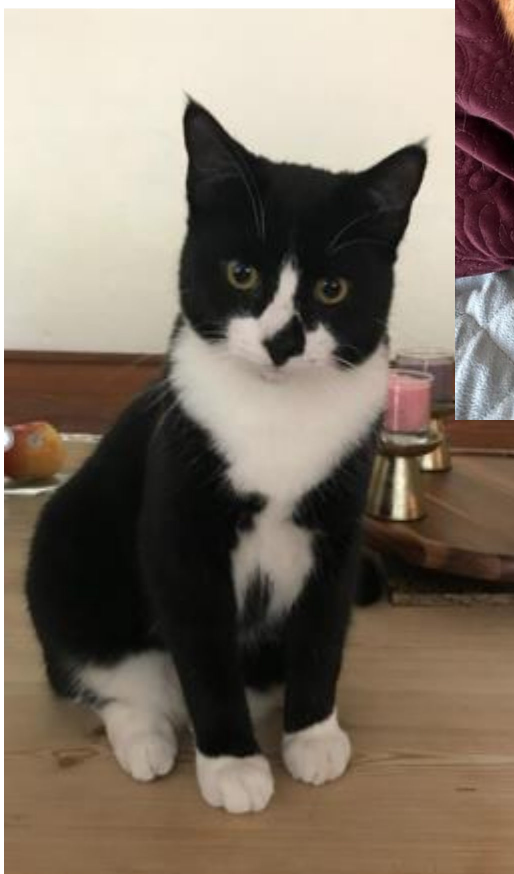
Extrapolation beyond the data range used to fit the regression model will result in seriously biased prediction if the relationship does not hold.



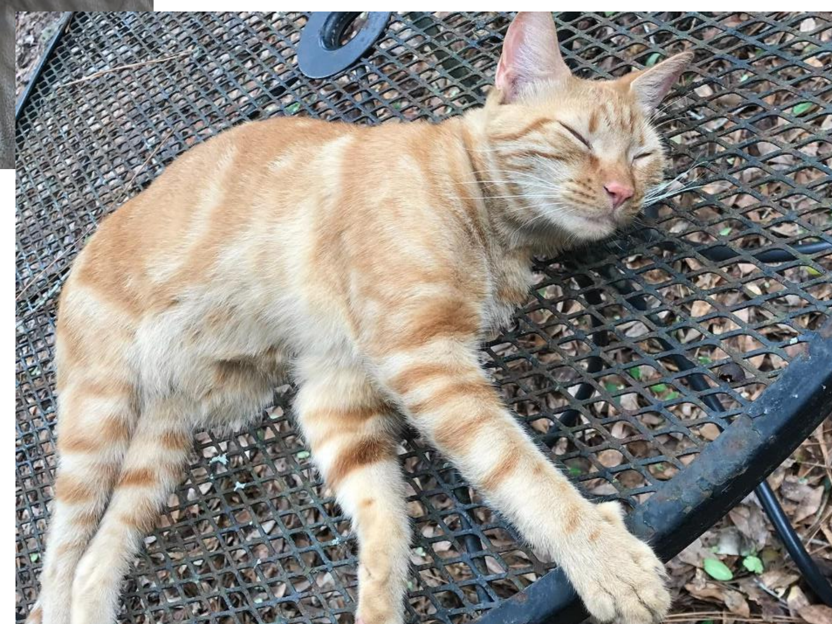
Summary Tips

- Practice Correlation and Regression in JMP!

THANK YOU!



- Stuart
- Mandalorian
- Willis (dog)
- Tigger



References

- Box GEP. Use and abuse of regression,Technometrics 1966; 8:4, 625-629.
- Draper NR, Smith H. Applied regression analysis. Wiley.
- Goldstein M, Wooff D. Bayes linear statistics, theory & methods. Wiley.
- Sedgwick P. Pearson’s correlation coefficient. BMJ 2012; 345:e4483
- Wei et al. Quantile regression methods for reference growth charts. Statistics in Medicine. 2005; 25(8):1369-1382.
- Zar JH Biostatistical analysis. Prentice Hall.

JMP Pro!

https://software.ufl.edu/