Bivariate Analysis

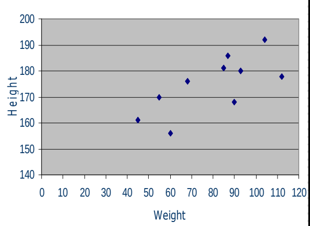
		Variable 1		
		2 LEVELS	>2 LEVELS	CONTINUOUS
Variable 2	2 LEVELS	X^2 chi square test	X^2 chi square test	t-test
	>2 LEVELS	X^2 chi square test	X^2 chi square test	ANOVA (F-test)
	CONTINUOUS	t-test	ANOVA (F-test)	-Correlation -Simple linear Regression

Correlation

- Used when you measure two continuous variables.
- Examples: Association between weight & height.
  Association between age & blood pressure

Correlation

Weight (Kg)	Height (cm)
55	170
93	180
90	168
60	156
112	178
45	161
85	181
104	192
68	176
87	186



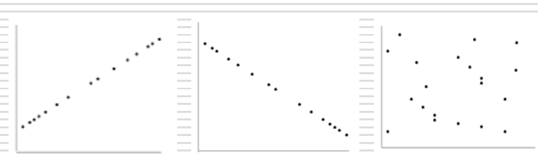
Pearson's Correlation Coefficient

- Correlation is measured by Pearson's Correlation Coefficient.
- A measure of the **linear association** between two variables that have been measured on a continuous scale.
- Pearson's correlation coefficient is denoted by r.
- A correlation coefficient is a number ranges between -1 and +1.

Pearson's Correlation Coefficient

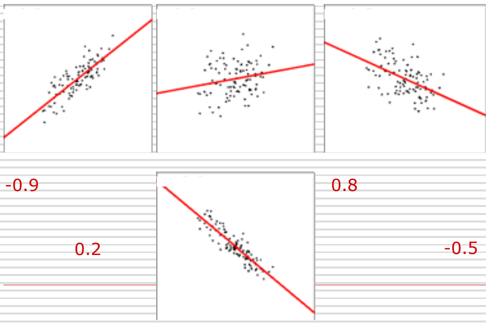
- If r = 1 ➔ perfect positive linear relationship between the two variables.
- If r = -1 ➔ perfect negative linear relationship between the two variables.
- If r = 0 ➔ No linear relationship between the two variables.

Pearson's Correlation Coefficient



r= +1   r= -1   r= 0

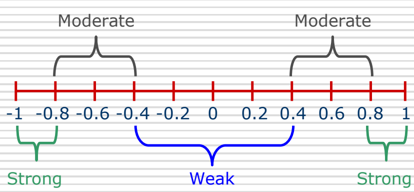
Pearson's Correlation Coefficient



Pearson's Correlation Coefficient

http://noppa5.pc.helsinki.fi/koe/corr/cor7.html

Pearson's Correlation Coefficient



Pearson's Correlation Coefficient

Example 1:

- **Research question:** Is there a linear relationship between the weight and height of students?
- H_o: there is no linear relationship between weight & height of students in the population ($p = 0$)
- H_a: there is a linear relationship between weight & height of students in the population ($p \neq 0$)
- **Statistical test**: Pearson correlation coefficient (R)

Pearson's Correlation Coefficient

Example 1: SPSS Output

Correlations

		weight	height
weight	Pearson Correlation	1	.651**
	Sig. (2-tailed)		.000
	N	1975	1954
height	Pearson Correlation	.651**	1
	Sig. (2-tailed)	.000	
	N	1954	1971

**. Correlation is significant at the 0.01 level

r coefficient

P-Value

Pearson's Correlation Coefficient

Example 1: SPSS Output

Correlations

		weight	height
weight	Pearson Correlation	1	.651**
	Sig. (2-tailed)		.000
	N	1975	1954
height	Pearson Correlation	.651**	1
	Sig. (2-tailed)	.000	
	N	1954	1971

**. Correlation is significant at the 0.01 level

- Value of statistical test: 0.651
- P-value: 0.000

Pearson's Correlation Coefficient

Example 1: SPSS Output

Correlations

		weight	height
weight	Pearson Correlation	1	.651**
	Sig. (2-tailed)		.000
	N	1975	1954
height	Pearson Correlation	.651**	1
	Sig. (2-tailed)	.000	
	N	1954	1971

**. Correlation is significant at the 0.01 level

- **Conclusion:** At significance level of 0.05, we reject null hypothesis and conclude that in the population there is significant linear relationship between the weight and height of students.

Pearson's Correlation Coefficient

Example 2: SPSS Output

Correlations

		weight	age
weight	Pearson Correlation	1	.155**
	Sig. (2-tailed)		.000
	N	1975	1814
age	Pearson Correlation	.155**	1
	Sig. (2-tailed)	.000	
	N	1814	1846

**. Correlation is significant at the 0.01 level

- **Research question:** Is there a linear relationship between the age and weight of students?

Pearson's Correlation Coefficient

Example 2: SPSS Output

Correlations

		weight	age
weight	Pearson Correlation	1	.155**
	Sig. (2-tailed)		.000
	N	1975	1814
age	Pearson Correlation	.155**	1
	Sig. (2-tailed)	.000	
	N	1814	1846

**. Correlation is significant at the 0.01 level

- H_o: $p = 0$; No linear relationship between weight & age in the population
- H_a: $p \neq 0$; There is linear relationship between weight & age in the population

Pearson's Correlation Coefficient

Example 2: SPSS Output

Correlations

		weight	age
weight	Pearson Correlation	1	.155**
	Sig. (2-tailed)		.000
	N	1975	1814
age	Pearson Correlation	.155**	1
	Sig. (2-tailed)	.000	
	N	1814	1846

**. Correlation is significant at the 0.01 level

- Value of statistical test: 0.155
- P-value: 0.000

Pearson's Correlation Coefficient

Example 2: SPSS Output

Correlations

		weight	age
weight	Pearson Correlation	1	.155**
	Sig. (2-tailed)		.000
	N	1975	1814
age	Pearson Correlation	.155**	1
	Sig. (2-tailed)	.000	
	N	1814	1846

**. Correlation is significant at the 0.01 level

- **Conclusion:** At significance level of 0.05, we reject null hypothesis and conclude that in the population there is a significant linear relationship between the weight and age of students.

Pearson's Correlation Coefficient

Example 3: SPSS Output

Correlations

		age	height
age	Pearson Correlation	1	.084**
	Sig. (2-tailed)		.000
	N	1846	1812
height	Pearson Correlation	.084**	1
	Sig. (2-tailed)	.000	
	N	1812	1971

**. Correlation is significant at the 0.01 level

- **Research question:** Is there a linear relationship between the age and height of students?

Pearson's Correlation Coefficient

Example 3: SPSS Output

Correlations

		age	height
age	Pearson Correlation	1	.084**
	Sig. (2-tailed)		.000
	N	1846	1812
height	Pearson Correlation	.084**	1
	Sig. (2-tailed)	.000	
	N	1812	1971

**. Correlation is significant at the 0.01 level

- H_o: *p* = 0 ; No linear relationship between height & age in the population
- H_a: *p* ≠ 0 ; There is linear relationship between height & age in the population

Pearson's Correlation Coefficient

Example 3: SPSS Output

Correlations

		age	height
age	Pearson Correlation	1	.084**
	Sig. (2-tailed)		.000
	N	1846	1812
height	Pearson Correlation	.084**	1
	Sig. (2-tailed)	.000	
	N	1812	1971

**. Correlation is significant at the 0.01 level

- Value of statistical test: 0.084
- P-value: 0.000

Pearson's Correlation Coefficient

Example 3: SPSS Output

Correlations

		age	height
age	Pearson Correlation	1	.084**
	Sig. (2-tailed)		.000
	N	1846	1812
height	Pearson Correlation	.084**	1
	Sig. (2-tailed)	.000	
	N	1812	1971

**. Correlation is significant at the 0.01 level

- **Conclusion:** At significance level of 0.05, we reject null hypothesis and conclude that in the population there is a significant linear relationship between the height and age of students.

SPSS command for r

Example 1

- **Analyze**
  - **Correlate**
    - **Bivariate**
      - select **height** and **weight** and put it in the "variables" box.

In-class questions

T (True) or F (False):

In studying whether there is an association between gender and weight, the investigator found out that r= 0.90 and p-value<0.001 and concludes that there is a strong significant correlation between gender and weight.

In-class questions

T (True) or F (False):

The correlation between obesity and number of cigarettes smoked was r=0.012 and the p-value= 0.856. Based on these results we conclude that there isn't any association between obesity and number of cigarette smoked.

Simple Linear Regression

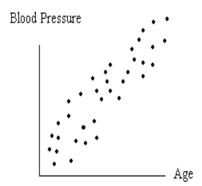
- Used to explain observed variation in the data
- For example, we measure blood pressure in a sample of patients and observe:

I=Pt#	1	2	3	4	5	6	7
Y= BP	85	105	90	85	110	70	115

Simple Linear Regression

- In order to explain why BP of individual patients are different, we try to associate the differences in PB with differences in other relevant patient characteristics (variables).
- Example: Can variation in blood pressure be explained by age?

Simple Linear Regression



Questions:

1) What is the most appropriate mathematical Model to use? A straight line, parabola, etc…

2) Given a specific model, how do we determine the best fitting model?

Simple Linear Regression

Mathematical properties of a straight line

- $Y = B_0 + B_1X$
  - Y = dependent variable
  - X = independent variable
  - B_0= Y intercept
  - B_1= Slope
- The intercept B_0 is the value of Y when X=0.
- The slope B_1 is the amount of change in Y for each 1-unit change in X.

Simple Linear Regression

Estimation of a simple Linear Regression Model

- Optimal Regression line = $B_0 + B_1X$
- $Y = B_0 + B_1X$

Simple Linear Regression

Example 1:

- **Research Question:** Does height help to predict weight using a straight line model? Is there a linear relationship between weight and height? Does height explain a significant portion of the variation in the values of weight observed?
- Weight = $B_0 + B_1$ Height

Simple Linear Regression

- SPSS output: Example 1

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	height[a]	.	Enter

a. All requested variables entered.
b. Dependent Variable: weight

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.651[a]	.424	.423	10.878

a. Predictors: (Constant), height

Simple Linear Regression

- SPSS output (Continued): Example 1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	169820.3	1	169820.297	1435.130	.000[a]
	Residual	230982.0	1952	118.331		
	Total	400802.3	1953			

a. Predictors: (Constant), height
b. Dependent Variable: weight

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-95.246	4.226		-22.539	.000
	height	.940	.025	.651	37.883	.000

a. Dependent Variable: weight

Simple Linear Regression

- SPSS output (Continued): Example 1

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.651[a]	.424	.423	10.878

a. Predictors: (Constant), height

0.424 → Height explains 42.4% of the variation seen in weight

Simple Linear Regression

- SPSS output (Continued): Example 1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-95.246	4.226		-22.539	.000
	height	.940	.025	.651	37.883	.000

a. Dependent Variable: weight

Weight = B_0 + B_1 Height

-95.246 **0.940**

Weight = -95.246 + **0.94** Height
Increasing height by 1 unit (1 cm) increases weight by **0.94** Kg

Simple Linear Regression

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-95.246	4.226		-22.539	.000
	height	.940	.025	.651	37.883	.000

a. Dependent Variable: weight

- H_0: B_1=0
- H_a: $B_1 \neq 0$

- Because the p-value of the B_1 is < 0.05; then reject H_0 and conclude that height provides significant information for predicting weight.

In-class questions

Question 1:

In a simple linear regression model the predicted straight line was as follows:

Weight (Kg) = 3.5 – 1.32 (weekly hours of PA)
R^2= 0.22; p-value for the slope= 0.04

What is the dependent/ independent variable?
Dependent variable: Weight
Independent Variable: Weekly hours of PA

In-class questions

Question 1:

In a simple linear regression model the predicted straight line was as follows:

Weight (Kg) = 3.5 – 1.32 (weekly hours of PA)
R^2= 0.22; p-value for the slope= 0.04

Interpret the value of R^2

Number of weekly hours of PA explain 22% of the variation observed in weight

In-class questions

Question 1:

In a simple linear regression model the predicted straight line was as follows:

Weight (Kg) = 3.5 – 1.32 (weekly hours of PA)
R^2= 0.22; p-value for the slope= 0.04

What is the null hypothesis? Alternative?

H_0: $B_{\text{weekly hours of PA}}=0$
H_a: $B_{\text{weekly hours of PA}}\neq 0$

In-class questions

Question 1:

In a simple linear regression model the predicted straight line was as follows:

Weight (Kg) = 3.5 – 1.32 (weekly hours of PA)
R^2= 0.22; p-value for the slope= 0.04

Is the association between weight & weekly hours of PA positive or negative?

Negative

In-class questions

Question 1:

In a simple linear regression model the predicted straight line was as follows:

Weight (Kg) = 3.5 – 1.32 (weekly hours of PA)
R^2= 0.22; p-value for the slope= 0.04

What is the magnitude of this association?

1.32 => One hour increase of PA in a week decreases weight by 1.32 Kg.

In-class questions

Question 1:

In a simple linear regression model the predicted straight line was as follows:

Weight (Kg) = 3.5 – 1.32 (weekly hours of PA)
R^2= 0.22; p-value for the slope= 0.04

Is the association significant at a level of 0.05?

Because the p-value of the B_1 is < 0.05; then reject H_0 and conclude that weekly hours of PA provide significant information for predicting weight.

In-class questions

Question 2:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.407[a]	.166	.164	10.396

a. Predictors: (Constant), ISS - injury severity measure

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	.443	.747		.593	.554
	ISS - injury severity measure	.661	.066	.407	9.945	.000

a. Dependent Variable: Length of hospital stay

In-class questions

Question 2:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.407[a]	.166	.164	10.396

a. Predictors: (Constant), ISS - injury severity

Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	.443	.747		.593	.554
	ISS - injury severity mea	.661	.066	.407	9.945	.000

a. Dependent Variable: Length of hospital stay

What is the dependent/ independent variable?

Dependent variable: Length of hospital stay
Independent Variable: ISS- Injury severity score

In-class questions

Question 2:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.407[a]	.166	.164	10.396

a. Predictors: (Constant), ISS - injury severity

Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	.443	.747		.593	.554
	ISS - injury severity mea	.661	.066	.407	9.945	.000

a. Dependent Variable: Length of hospital stay

Interpret the value of R^2

ISS explains 40.7% of the variation observed in length of hospital stay.

In-class questions

Question 2:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.407[a]	.166	.164	10.396

a. Predictors: (Constant), ISS - injury severity

Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	.443	.747		.593	.554
	ISS - injury severity mea	.661	.066	.407	9.945	.000

a. Dependent Variable: Length of hospital stay

What is the null hypothesis? Alternative?

H_0: $B_{ISS}=0$
H_a: $B_{ISS}\neq 0$

In-class questions

Question 2:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.407[a]	.166	.164	10.396

a. Predictors: (Constant), ISS - injury severity

Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	.443	.747		.593	.554
	ISS - injury severity mea	.661	.066	.407	9.945	.000

a. Dependent Variable: Length of hospital stay

Is there a significant association between the dependent & the independent?

Because the p-value of the B_{ISS} is < 0.05; then reject H_0 and conclude that ISS provide significant information for predicting length of hospital stay.

In-class questions

Question 2:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.407[a]	.166	.164	10.396

a. Predictors: (Constant), ISS - injury severity

Coefficients[a]

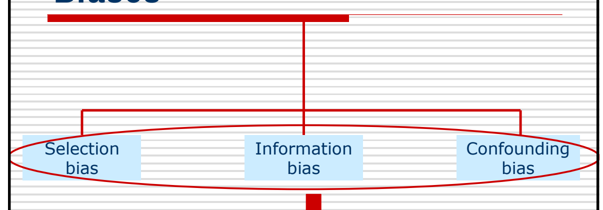
Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	.443	.747		.593	.554
	ISS - injury severity mea	.661	.066	.407	9.945	.000

a. Dependent Variable: Length of hospital stay

What is the magnitude of this association?

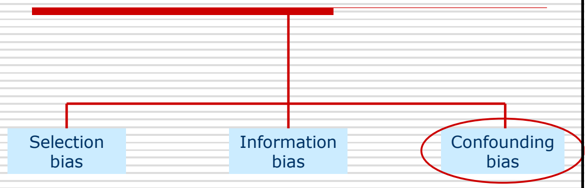
0.661 => Increasing ISS by 1 unit increases length of hospital stay by 0.661 days.

Biases



Bias is an error in an epidemiologic study that results in an incorrect estimation of the association between exposure and outcome.

Biases



Confounding Bias: Definition

Is present when the association between an exposure and an outcome is distorted by an extraneous third variable (referred to a confounding variable).

Confounding Bias: Example

Example : Study the association between coffee drinking and lung cancer

		LC	
		Yes	No
Coffee	Yes	**80**	**15**
	No	**20**	**85**

OR= (80x 85)/ (15 x 20)= 22

What would you conclude????

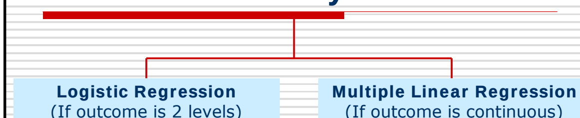
Confounding Bias: Minimize bias

- **Research Design:**
  - Use of randomized clinical trial
  - Restriction
- **Data Analysis:**
  - Multivariate statistical techniques

Bivariate Analysis

		Variable 1		
		2 LEVELS	**>2 LEVELS**	**CONTINUOUS**
Variable 2	**2 LEVELS**	X^2 chi square test	X^2 chi square test	t-test
	>2 LEVELS	X^2 chi square test	X^2 chi square test	ANOVA (F-test)
	CONTINUOUS	T-test	ANOVA (F-test)	-Correlation -Simple linear Regression

Multivariate analyses



Multivariate Analysis is used for adjusting for confounding variables.

Multivariate Analysis

WHY?

- To investigate the effect of more than one independent variable.
- Predict the outcome using various independent variables.
- Adjust for confounding variables

Multivariate analyses

- **Logistic Regression**
  (If outcome is 2 levels)
- **Multiple Linear Regression**
  (If outcome is continuous)