Bivariate Analysis

<table>
<thead>
<tr>
<th>Variable 1</th>
<th>Variable 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 LEVELS</td>
<td>2 LEVELS</td>
</tr>
<tr>
<td>$X^2$</td>
<td>$X^2$</td>
</tr>
<tr>
<td>chi square test</td>
<td>chi square test</td>
</tr>
<tr>
<td>t-test</td>
<td>t-test</td>
</tr>
<tr>
<td>&gt;2 LEVELS</td>
<td>&gt;2 LEVELS</td>
</tr>
<tr>
<td>$X^2$</td>
<td>$X^2$</td>
</tr>
<tr>
<td>chi square test</td>
<td>chi square test</td>
</tr>
<tr>
<td>ANOVA</td>
<td>ANOVA</td>
</tr>
<tr>
<td>(F-test)</td>
<td>(F-test)</td>
</tr>
<tr>
<td>CONTINUOUS</td>
<td>CONTINUOUS</td>
</tr>
<tr>
<td>t-test</td>
<td>t-test</td>
</tr>
<tr>
<td>ANOVA</td>
<td>ANOVA</td>
</tr>
<tr>
<td>(F-test)</td>
<td>(F-test)</td>
</tr>
<tr>
<td>Pearson's Correlation</td>
<td>Simple linear Regression</td>
</tr>
</tbody>
</table>

Correlation

- Used when you measure two continuous variables.

- Examples: Association between weight & height, Association between age & blood pressure.

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.

<table>
<thead>
<tr>
<th>Weight (Kg)</th>
<th>Height (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>55</td>
<td>170</td>
</tr>
<tr>
<td>93</td>
<td>180</td>
</tr>
<tr>
<td>90</td>
<td>168</td>
</tr>
<tr>
<td>60</td>
<td>156</td>
</tr>
<tr>
<td>112</td>
<td>178</td>
</tr>
<tr>
<td>45</td>
<td>161</td>
</tr>
<tr>
<td>85</td>
<td>181</td>
</tr>
<tr>
<td>104</td>
<td>192</td>
</tr>
<tr>
<td>68</td>
<td>176</td>
</tr>
<tr>
<td>87</td>
<td>186</td>
</tr>
</tbody>
</table>

Pearson's Correlation Coefficient
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.

http://noppa5.pc.helsinki.fi/koe/corr/cor7.html
Pearson's Correlation Coefficient

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

- $H_0$: there is no linear relationship between weight & height of students in the population ($p = 0$)
- $H_1$: there is a linear relationship between weight & height of students in the population ($p 
eq 0$)

**Statistical test:** Pearson correlation coefficient ($R$)

**Example 1:**

- **Value of statistical test:** $0.651$
- **P-value:** $0.000$
Correlation is significant at the 0.01 level

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

** 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

<table>
<thead>
<tr>
<th></th>
<th>weight</th>
<th>age</th>
<th>weight</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Correlation</td>
<td>1.155**</td>
<td>.000</td>
<td>1.197</td>
<td>1.814</td>
</tr>
<tr>
<td>N</td>
<td>1975</td>
<td>1814</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

** Correlation is significant at the 0.01 level.

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

Example 3: SPSS Output

<table>
<thead>
<tr>
<th></th>
<th>age</th>
<th>height</th>
<th>age</th>
<th>height</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Correlation</td>
<td>.084**</td>
<td>.000</td>
<td>.084</td>
<td>1.000</td>
</tr>
<tr>
<td>N</td>
<td>1846</td>
<td>1812</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

** Correlation is significant at the 0.01 level.

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

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

H₀: \( \rho = 0 \); No linear relationship between height & age in the population

H₁: \( \rho ≠ 0 \); There is a linear relationship between height & age in the population
Pearson’s Correlation Coefficient

Example 3: SPSS Output

<table>
<thead>
<tr>
<th></th>
<th>age</th>
<th>height</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>.084</td>
<td></td>
</tr>
<tr>
<td>Sig.</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>1846</td>
<td>1812</td>
</tr>
<tr>
<td></td>
<td>.084</td>
<td>1</td>
</tr>
<tr>
<td>Pearson</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>Sig.</td>
<td>1846</td>
<td>1971</td>
</tr>
</tbody>
</table>

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

1. **Analyze**
2. **Correlate**
3. **Bivariate**
4. 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:

<table>
<thead>
<tr>
<th>Pt#</th>
<th>BP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>85</td>
</tr>
<tr>
<td>2</td>
<td>105</td>
</tr>
<tr>
<td>3</td>
<td>90</td>
</tr>
<tr>
<td>4</td>
<td>85</td>
</tr>
<tr>
<td>5</td>
<td>110</td>
</tr>
<tr>
<td>6</td>
<td>70</td>
</tr>
<tr>
<td>7</td>
<td>115</td>
</tr>
</tbody>
</table>

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?

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?

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 = $B_0 + B_1X$
- $Y = B_0 + B_1X$

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

Example 1:

<table>
<thead>
<tr>
<th>Model</th>
<th>Variables Entered/Removed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>height</td>
</tr>
</tbody>
</table>

Method
- All requested variables entered.
- Dependent Variable: weight

SPSS output: Example 1

Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>$R$</th>
<th>$R$ Square</th>
<th>Adjusted $R$ Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.451</td>
<td>.424</td>
<td>.423</td>
<td>10.879</td>
</tr>
</tbody>
</table>

$*$ Predictors (Constant), height

SPSS output (Continued): Example 1

ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>$F$</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>169820.3</td>
<td>1</td>
<td>169820.297</td>
<td>1435.130</td>
<td>.000</td>
</tr>
<tr>
<td>Residual</td>
<td>230982.0</td>
<td>1952</td>
<td>118.331</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>400802.3</td>
<td>1953</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$*$ Predictors (Constant), height

$P$ Dependent Variable: weight

Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>$t$</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$B$</td>
<td>$SE$</td>
<td>Beta</td>
<td>$t$</td>
</tr>
<tr>
<td>(Constant)</td>
<td>335.398</td>
<td>2.339</td>
<td>.940</td>
<td>1435.130</td>
</tr>
<tr>
<td>height</td>
<td>4.041</td>
<td>.029</td>
<td>.951</td>
<td>1435.130</td>
</tr>
</tbody>
</table>

$*$ Unstandardized Variable: weight
Simple Linear Regression

- SPSS output (Continued): Example 1

**Model Summary**

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.424</td>
<td>0.424</td>
<td>0.423</td>
<td>10.378</td>
</tr>
</tbody>
</table>

* Predictors: (Constant), height

0.424 Height explains 42.4% of the variation seen in weight

**Coefficients**

<table>
<thead>
<tr>
<th>Model</th>
<th>B</th>
<th>Std. Error</th>
<th>Unstandardized Coefficients</th>
<th>Beta</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-95.246</td>
<td>4.226</td>
<td>-22.539</td>
<td>.000</td>
<td>.940</td>
<td>37.883</td>
<td>.000</td>
</tr>
</tbody>
</table>

* Dependent Variable: weight

Weight = B0 + B1 Height

Weight = -95.246 + 0.94 Height

Increasing height by 1 unit (1 cm) increases weight by 0.94 Kg

---

Simple Linear Regression

**Question 1:**

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

\[
\text{Weight (Kg)} = 3.5 - 1.32 \text{ (weekly hours of PA)}
\]

\[
R^2 = 0.22; \ p\text{-value for the slope} = 0.04
\]

What is the dependent/independent variable?

- **Dependent variable:** Weight
- **Independent Variable:** Weekly hours of PA

---

In-class questions

Question 1:

- H0: B1=0
- H1: B1 ≠0

Because the p-value of the B1 is < 0.05; then reject H0 and conclude that height provides significant information for predicting weight.
Question 1:

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

\[ \text{Weight (Kg)} = 3.5 - 1.32 \text{ (weekly hours of PA)} \]

\[ R^2 = 0.22; \text{ p-value for the slope} = 0.04 \]

Interpret the value of \( R^2 \):

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

Question 1:

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

\[ \text{Weight (Kg)} = 3.5 - 1.32 \text{ (weekly hours of PA)} \]

\[ R^2 = 0.22; \text{ p-value for the slope} = 0.04 \]

What is the null hypothesis? Alternative?

\[ H_0: \beta_{\text{weekly hours of PA}} = 0 \]
\[ H_1: \beta_{\text{weekly hours of PA}} \neq 0 \]

Question 1:

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

\[ \text{Weight (Kg)} = 3.5 - 1.32 \text{ (weekly hours of PA)} \]

\[ R^2 = 0.22; \text{ p-value for the slope} = 0.04 \]

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

Negative

Question 1:

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

\[ \text{Weight (Kg)} = 3.5 - 1.32 \text{ (weekly hours of PA)} \]

\[ R^2 = 0.22; \text{ p-value for the slope} = 0.04 \]

What is the magnitude of this association?

\[ -1.32 \Rightarrow \text{One hour increase of PA in a week decreases weight by 1.32 Kg.} \]
**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)

\[ R^2 = 0.22; \ p\text{-value for the slope} = 0.04 \]

Is the association significant at a level of 0.05?

Because the p-value of the B1 is < 0.05; then reject H0 and conclude that weekly hours of PA provide significant information for predicting weight.


**Question 2:**

What is the dependent/independent variable?

Dependent variable: Length of hospital stay

Independent Variable: ISS - Injury severity score

Interpret the value of \( R^2 \)

ISS explains 40.7% of the variation observed in length of hospital stay.
**In-class questions**

**Question 2:**

What is the null hypothesis? Alternative?

- H₀: $B_{ISS} = 0$
- $H_1$: $B_{ISS} \neq 0$

**Question 2:**

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

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

What is the magnitude of this association?

$0.661 \Rightarrow$ Increasing ISS by 1 unit increases length of hospital stay by 0.661 days.

---

**In-class questions**

**Question 2:**

What is the null hypothesis? Alternative?

- H₀: $B_{ISS} = 0$
- $H_1$: $B_{ISS} \neq 0$

**Question 2:**

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

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

What is the magnitude of this association?

$0.661 \Rightarrow$ 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

- Selection bias
- Information bias
- Confounding bias

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

<table>
<thead>
<tr>
<th>Coffee</th>
<th>LC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>80</td>
</tr>
<tr>
<td>No</td>
<td>20</td>
</tr>
</tbody>
</table>

\[ OR = \frac{(80 \times 85)}{(15 \times 20)} = 22 \]

What would you conclude???

Confounding Bias: Minimize bias

- **Research Design:**
  - Use of randomized clinical trial
  - Restriction

- **Data Analysis:**
  - Multivariate statistical techniques
**Bivariate Analysis**

<table>
<thead>
<tr>
<th>Variable 1</th>
<th>2 LEVELS</th>
<th>&gt;2 LEVELS</th>
<th>CONTINUOUS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$X^2$</td>
<td>$X^2$</td>
<td>$t$-test</td>
</tr>
<tr>
<td>2 LEVELS</td>
<td>chi square test</td>
<td>chi square test</td>
<td></td>
</tr>
<tr>
<td>&gt;2 LEVELS</td>
<td>$X^2$</td>
<td>$X^2$</td>
<td>ANOVA (F-test)</td>
</tr>
<tr>
<td>CONTINUOUS</td>
<td>$t$-test</td>
<td>ANOVA (F-test)</td>
<td>-Correlation -Simple linear Regression</td>
</tr>
</tbody>
</table>

**Multivariate analyses**

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

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