Chapter 2. Looking at Data-Relationships

• Is there a relationship?
• Can we describe this relationship with a model?
• Can we use this model to predict the future values?

Response Variable: measures an outcome of a study.
Explanatory Variable: explains or causes in the response variables.

I. Relationships between Quantitative Variables

1. Scatterplots:
   • A scatterplot shows the relationship between two quantitative variables measured on the same individuals.
   • The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis.

   NOTE: Questions to ask about a scatter plot
   - What is the average pattern? Linear or Nonlinear
   - What is the direction of the pattern? Positive or Negative association
   - How much do individual points vary from the average pattern?
   - Are there any outliers?

2. (Pearson's) Correlation:
   • A number that measures the strength and direction of the linear relationship between two numerical variables.

   • Correlation is usually written as r.

   • Properties of Pearson's correlation coefficient

   1. r is between -1 and +1.
   2. r<0: negative relationship and r>0: positive relationship.
   3. r= -1 or r= +1 happens only when the points lie on an exact line.
   4. The value of r does not depend on the units of measurement. For example, the correlation between height and shoe size in inches will be the same as the correlation
between height and shoe size in feet.

3. Fitting a line to two quantitative Data (Least-Square Regression)

If we have an idea that there is some sort of relationship, we are interested in some way to summaries the relationship. We want to find a line that "best fit" the points-least squares regression line- so that we can predict values for Y.

Regression Analysis: examines the relationship between a quantitative response variable(Y) and one or more explanatory variables(X).

regression equation:
Making prediction: prediction of potential GPAs of future students, based on their SAT scores.
X: explanatory variable or independent variable
Y: response or dependent variable

(ex) regression equation: footlength=10.937+0.233*height
=> here, we can get prediction based on new data with this equation.

How to estimate the values slope and intercept
The most widely used criterion is Method of Least Squares.

II. Relationship between categorical variables

Summarizing data resulting from the measurement of two categorical variables is easy to do: Simply count the number of individuals who fall into each combination of categories, and present those counts in a table. Such displays are often called contingency tables because they cover all contingencies for combinations of the two variables. Each row and column combination in the table is called a cell. In some cases, one variable can be designated as the explanatory variable and the other as the response variable.

Age at birth of first child and breast cancer
First child at age 25 or older?

<table>
<thead>
<tr>
<th></th>
<th>Breast cancer</th>
<th>No breast cancer</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>31 (1.9%, 32.3%)</td>
<td>1597 (98.1%, 26.3%)</td>
<td>1628</td>
</tr>
<tr>
<td>No</td>
<td>65 (1.4%, 67.7%)</td>
<td>4475 (98.6%, 73.7%)</td>
<td>4540</td>
</tr>
<tr>
<td>Total</td>
<td>96</td>
<td>6072</td>
<td>6168</td>
</tr>
</tbody>
</table>

- **Conditional percentage. (Row percentage and Column percentage)**

  We can calculate the conditional percentages for the response variable by separately looking at each category of the explanatory variable.

  ex: In the above contingency table,
  
  "Yes" group : the percentage who had Breast Cancer was 31/1628=1.9%
  
  "No" group : the percentage who had Breast Cancer was 65/4540=1.4%

- **Risk, Relative risk, Odds Ratio, and Increased Risk.**

  1) To compute the relative risk of developing breast cancer based on whether the age at which a woman had her first child was 25 or older, we first find the risk of breast cancer for each group:

  Risk for women having first child at age 25 or older = 31/1628 = 0.0190
  
  Risk for women having first child before age 25 = 65/4540 = 0.0143

  2) Relative risk = 0.0190/0.0143 = 1.33

  We can also represent this by saying that the risk of developing breast cancer is 1.33 times greater for women who had their first child at age 25 or older.

  3) Increased risk = difference in risks / baseline risk

  =(0.0190-0.0143) / 0.0143

  =0.33

  4) To calculate the odds ratio, you first compute the odds of getting the disease to not getting the disease for each of the two categories of the explanatory variable. For the example concerning the risk of breast cancer,

  Odds for the women having first child at age 25 or older = 31/1597 = 0.0194

  Odds for the women having first child before age 25 = 65/4475 = 0.0145

  5) Odds ratio = 0.0194 / 0.0145 = 1.34

  You can see that the odds ratio is very similar to the relative risk of 1.33.