Lesson 5

Correlation

Correlation

- An analysis to find the association or the absence of the relationship between two variables, x and y
- A measure of strength of the relationship between the model and the dependent variable on a convenient 0-100% scale
- .

Correlation

- A statistical technique
- Measures and describes the ‘strength’ and ‘direction’ of the relationship between two variables
- Does y increase or decrease with x? Does expenditure increases with income?

Correlation giving Answers of Questions

- Does the number of patients decrease with proper medication?(Direction)
- Suppose y does increase with x; then, how fast? Is the relation between x and y strong? Can reliable predictions be made about y from the x?
- Can one tell the income, can the expenditure be predicted?

R-squared (R^2)

- R— a measure of correlation between the predicted values y and the observed values of x.
- R^2 — a goodness-of-fit measure in linear-regression model
- — A coefficient of determination.

R-squared (R^2)

- — the coefficient of multiple correlations
- Includes additional independent (explanatory) variables in regression equation

R-squared (R^2) Interpretation

- Larger the better fits the observations the regression model, better correlation
- Theoretically, if a model shows 100% variance, then the fitted values are always equal to the observed values, and therefore, all the data points would fall on the fitted regression line

Pearson product-moment correlation coefficient *r*

- $r = \{1/(n-1)\}\sum\{(x_i - \bar{x})/\sigma_x \times (y_i - \bar{y})/\sigma_y\}$ (equation 6.8a)
- Sum over all values of n,
- $n = 1, 2, 3, \ldots n$; $\bar{x}$ = sample mean of x; $\bar{y}$ = sample mean of 6
- σ_x = standard deviation of x, σ_y = standard deviation of y

r^2

- The square of sample correlation coefficient between the observed outcomes and the observed predictor values
- Includes intercept on y-axis in case of linear regression

Sample Pearson correlation metric c_r

- Measures how well two sample datasets fit on a straight line
- (equation 6.8b)

Similarities based on correlation

(i) Constrained Pearson correlation– a variation of Pearson correlation that uses midpoint instead of mean rate.

Similarities based on correlation

(ii) Spearman rank correlation –similar to Pearson correlation, except that the ratings are ranks.

Similarities based on correlation

(iii) Kendall's Γ correlation – Similar to the Spearman rank correlation, but instead of using ranks themselves, only the relative ranks are used to calculate the correlation

Strength of the relationship

- A function of *r*
- Refer Table 6.1
- When r > 0; positive correlation
- When r< 0; negative correlation
- When r = 0; no correlation

Figure 6.4 Part a : Perfect and linear positive and negative relationships

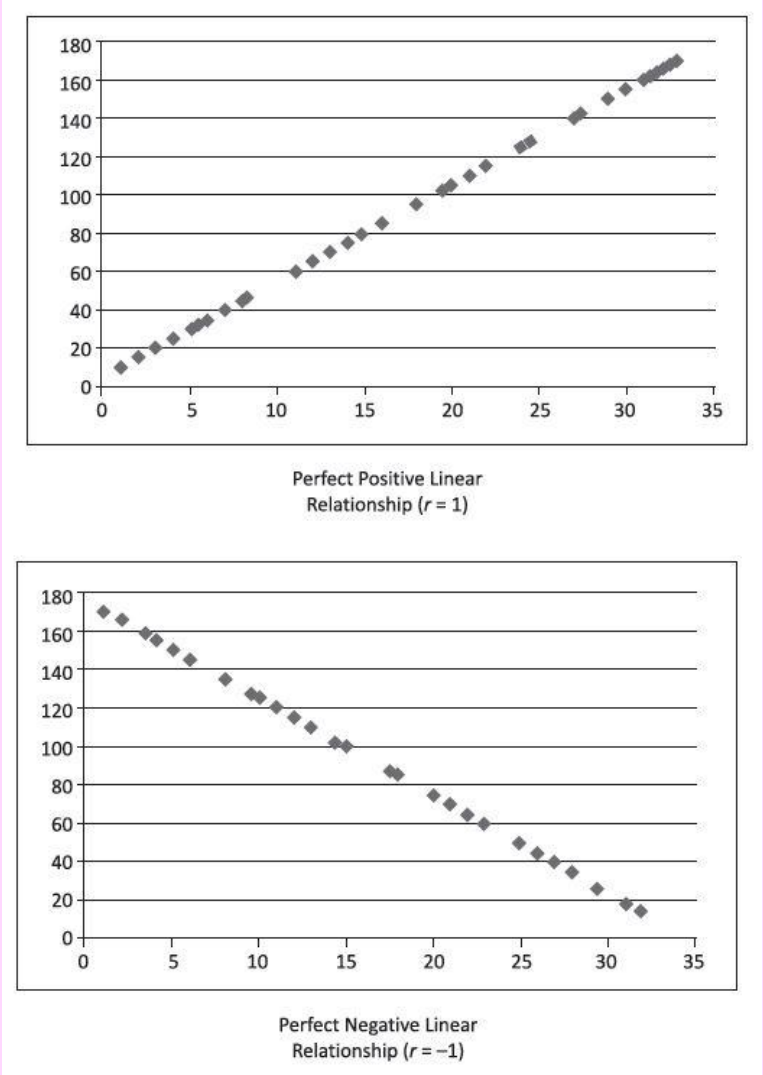
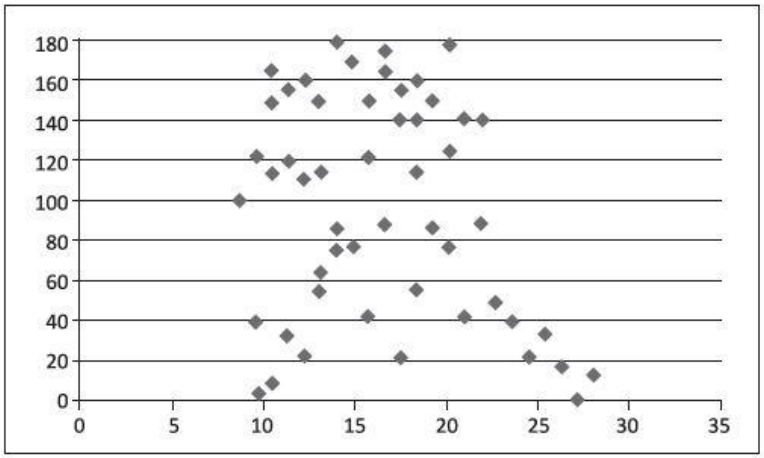
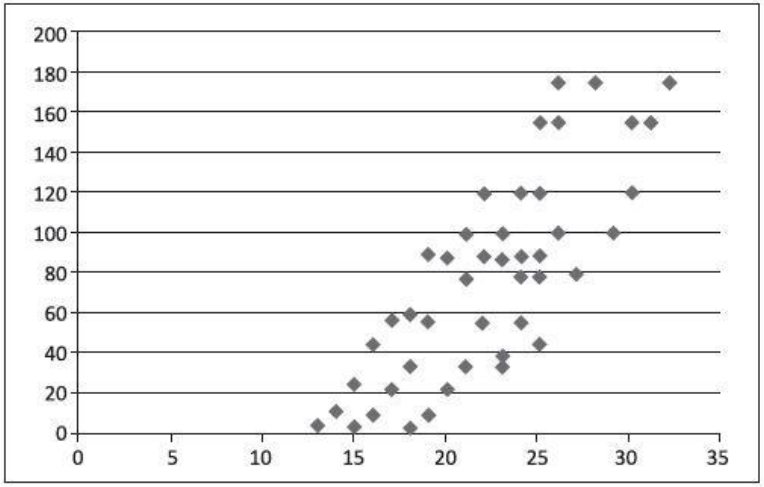


Figure 6.4 Part b : No relationship and imperfect linear positive relationships



No Relationship ($r \sim 0$)



Positive Linear Relationship ($r = 0.9$)

Summary

We learnt:

- Correlation function
- Pearson Correlation Coefficient
- Strength of relationship

End of Lesson 5 on Correlation