Lesson 2

Classes of variables, and estimating the relationships

Independent Variable

- Represents directly measurable characteristics
- Such as Year of Sales Figures, Semester of study
- Year, x is independent variable in equation $y = (a0 + a1.x + a2.x^2)$

Dependent Variable

- Represents the characteristics, for example, profit during successive years or grades awarded in successive semesters
- Values of a dependent variable depend on the value of the independent variable; May also related by a mathematical expression

Predictor Variable

- Represents an independent variable, which computes a dependent variable using some equation, function or graph, and does a prediction. For example, predicts:
- Expected sales growth of a car model after five years
- Predicts user sentiments for the model

Outcome Variable

- Represents the effect of manipulation(s) using a function, equation or experiment
- For example, an outcome is CGPA (Cumulative Grade Points Average) of the student which computes from the grades awarded in the semesters in different courses studied

Explanatory Variable

- An independent variable, which explains the behavior of the dependent variable, such as:
- Linearity coefficient
- Non-linear parameter
- Probability distribution of profit-growth as a function of additional investment in successive years

Response Variable

- A dependent variable on which a study, experiment or computation focuses
- For example, improvement in profits over the years from the investments made in successive years

Feature Variable

- A variable representing a characteristic. For example:
- Apple feature: red, pink, maroon, yellowish, yellowish green and green
- Generally represented by text characters.
- Numbers also represent features: red with 1, orange with 2, yellow with 3, …. .

Categorical variable

- A variable representing a category, a unit of observation to a particular group
- Boolean variable also a category
- Car, tractor and truck belong to the same category, i.e., a four-wheeler automobile.
- Generally represented by text characters
- Set or Group with limited, and/or fixed number of possible values, features, …

Data Analysis

- Studying relationships graphically, mathematically and statistically
- The outliers, anomalies, variances, correlations, features, categories and probability distributions
- Uses a set of variables, and other characteristics.

Relationship

- Studying relationships graphically, involves some quantifiable independent variables and the resulting dependent variable or entity

Relationship and Correlation

- Variables may exhibit a relation or correlation. The relationships:
- May be linear, nonlinear, positive, negative, direct, inverse, scattered or spread
- Shown using the Graphs, Scatter Plots and/or Charts

Outlier

- A data point for dependent variable can be an outlier
- For example, Figure 6.7 Jaguar Car Sales Percentage Increment = just 4% in 1st Year can be considered as an outlier
- Showing no relationship

Linear Relationship

- Said to exist between two quantitative variables when a curve ($y = a0 + a1.x$) can be used to fit the data points with y as a function of x
- Constant a1 positive means positive relationship, −ve means negative relationship, 0 means none
- A0 the value of y when $x = 0$

Figure 6.1 Scatter plot for linear relationship between students opting for computer courses in years between 2000 and 2017

Near linear relationship)

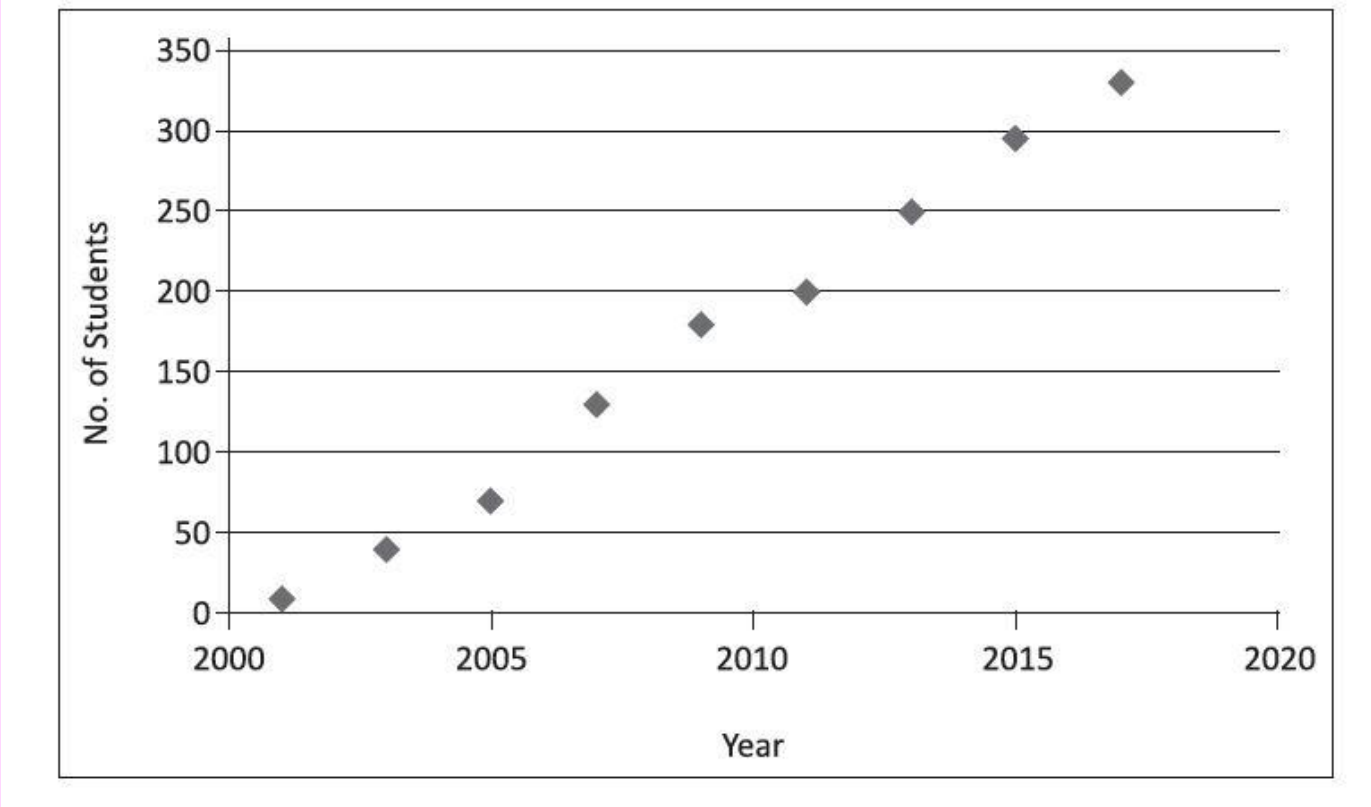
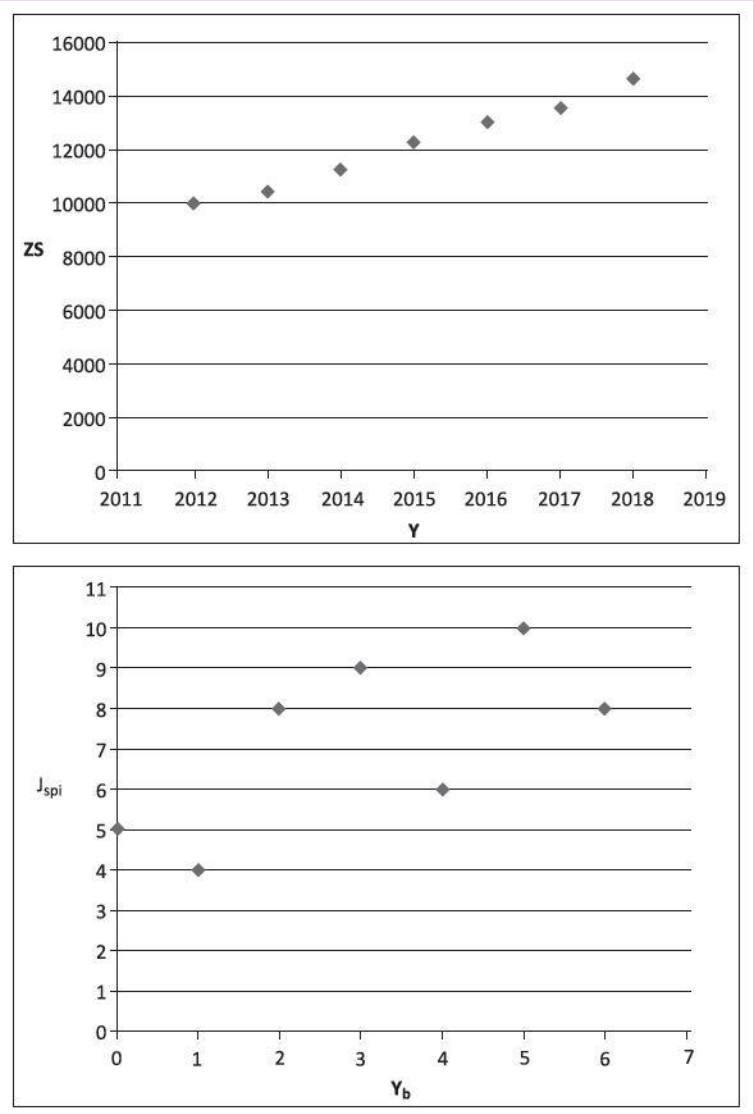


Figure 6.7 Scatter plots for two set of data points

Zest Car Sales as a function of Year (Near linear relationship)

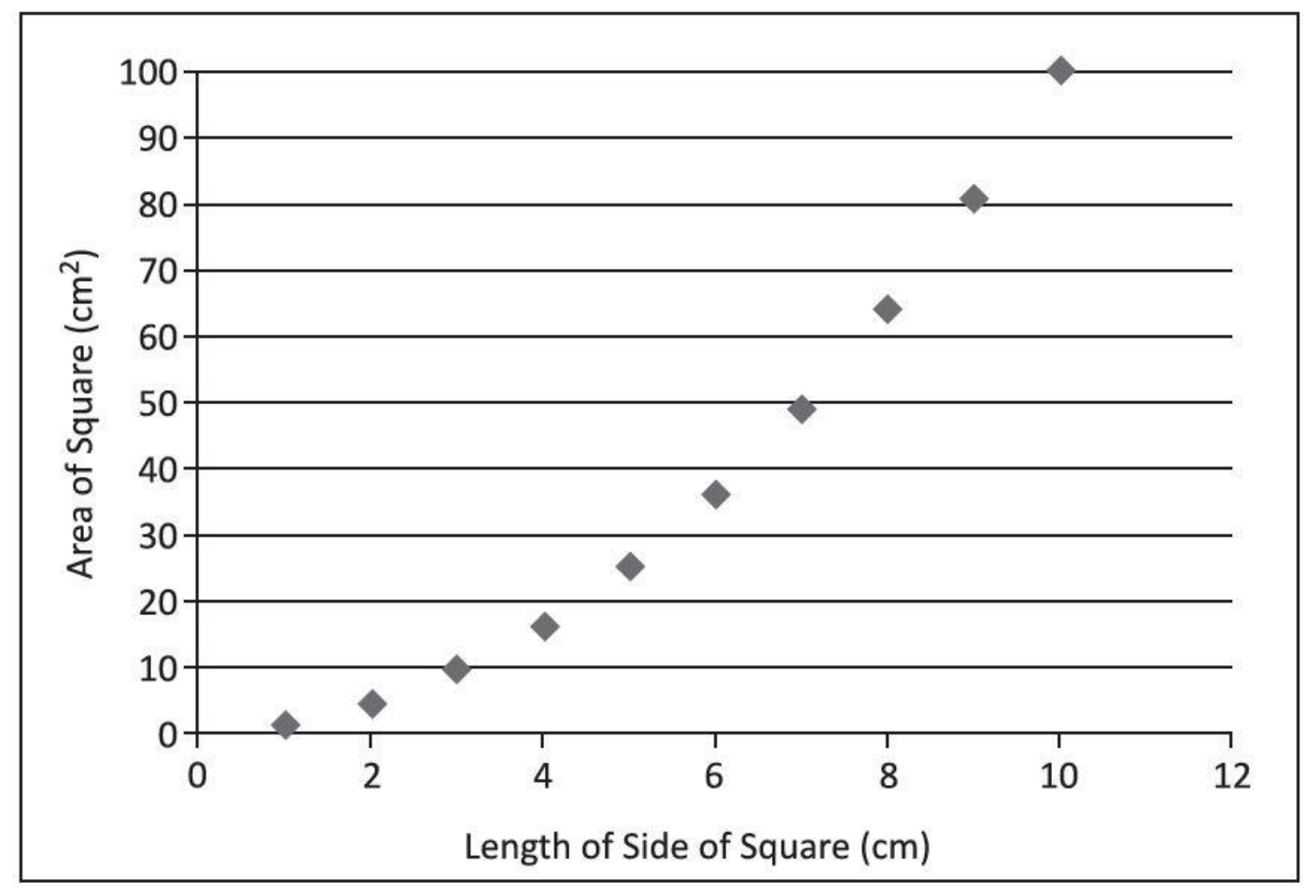
Jaguar Car Sales Percentage Increments as a function of Years after a base year



Nonlinear Relationship

- Example: relationship of data points (x, y) such as
- $y = (a0 + a1.x + a2.x^2 + \ldots)$ can be used to fit the data points. The fit should be with at least some reasonable degree of accuracy for the fitted parameters, a0, a1, a2 … Expression

Figure 6.2 Scatter plot in case of a non-linear relationship between side of square and its area



Estimating the Relationships

- Means finding a mathematical expression
- Giving the value of the response (dependent) variable according to its relationship with other variables

Examples

- Sales of a car model m $y_m = (a0 + a1.x + a2.x^2)$ in x^{th} year of the start of manufacturing that model (Quadratic relationship)
- Popularity index, $y_p = a0.\exp(a1.t)$ [Exponential Growth with time, t)

Estimating no relationship data points: Outliers

- Anomalous situation
- Presence of a previously unknown fact
- Human error (errors due to data entry or data collection)

Estimating no relationship data points: Outliers

- Participants intentionally reporting incorrect data (This is common in self-reported measures and measures that involve sensitive data which participant doesn't want to disclose)
- · Sampling error (when an unfitted sample is collected from population).

Estimating no relationship data points: Outliers

- Sampling error (when an unfitted sample is collected from population).

Summary

We learnt meanings of:

- Classes of Variables:
- Features
- Categorical
- Relationships
- Outliers

End of Lesson 2 on Classes of variables, and estimating the relationships