Chapter 3
The Research Cycle and Literature Review
The Research Cycle

1. Formulate
2. Generate Hypothesis
3. Collect Data
4. Explore the Data
5. Model
6. Implement/Monitor
7. Sell Solution
The Research Cycle

- **Formulate**: initial understanding of the problem or situation to be modelled. The variables and emotive issues are also identified (Not too big, not too small, a few variables, one or two selling points).

- **Generate a Hypothesis**: Conventional hypothesis statements help to guide the analysis. They take the form of the null hypothesis (H_n), that there is no effect or relationship (falsification), which is compared to the alternative hypothesis (H_a), that there is an effect or a relationship.

- **Collect Data**: data collected on selected variables are usually classified into independent and dependent types.
The Research Cycle

- **Explore the Data**: using very simple tools, such as scatter plots; or to catch any oddities that exist in the data.

- **Model and Implement/Monitor**: the coefficients are computed, often using the statistical procedure of linear regression.

- **Sell Solution**: great care and time must be taken to explain to users how the model has been arrived at and how it works.

- If at any stage in the research cycle there are unsatisfactory findings, one returns to the formulation phase.
The Main Research Stages

- Specify the real problem.
  - Investigation purpose (hypotheses to be tested). Managerial vs. research (theoretical) problems

- Set up a model (from theory).
  - Break the problem (1) into parts, Select variables, Make sensible assumptions, and Determine the limits of validity.

- Formalize the model.
  - Mathematical or statistical version of the problem, Find the relationships involved, and Find a functional form.

- ‘Solve’ the problem.
  - Choose appropriate techniques, and Amend the model and/or change approach if needed

- Interpret the results.
  - Does step 4 make sense in terms of step 1? If not, why not?

- Validate the model (validation study, robustness check cross-validation).
  - How realistic were the initial assumptions? What if one or more were changed? More refining needed? At what cost? If yes, then back to step 2!

- Generalization.
  - Can the model and results be generally applied to this class of problem?

- Report the findings.
Importance of Literature Review

- The importance of a literature review can only be appreciated when we ask:
  - Has the work already been done?
  - Who are the experts in the field?
  - What are the main theoretical perspectives?
  - What are the common research methods in the topic?
  - What are the main problems in researching the topic?
  - Are there any major controversies in this topic area?
  - Is the topic open to hypothesis testing?
  - Is the topic a trivial one?

The only way you can satisfactorily answer these questions is by reading as much as you can on the directly and indirectly related research.
Importance of Literature Review

- **Has the Work Already Been Done?**
  - Very rare in business research
  - Still important to check similar research

- **Who are the Experts in the Field? Establish a ‘road map’ for your literature review**
  - Consult the Social Science or Business citation index by topic area
  - Consult the reference list of a published paper

- **What are the Main Theoretical Perspectives? (streams of research, not chronological, TCE, OLI, RBV, OL, social networks, new institutional research)**
  - Construct a conceptual framework within the topic
  - Need a theory in order to inform us what kind of data we require
  - Where to find a theory?---In the relevant literature
Importance of Literature Review

- **What are the Common Research Methods in the Topic?**
  - Is the published research in this topic mainly of a qualitative or a quantitative nature?
  - If qualitative, what is its most common basis?
  - If quantitative, what is its most common basis?
  - Is the published research usually a mix of qualitative and quantitative methods?

- **What are the Main Problems in Researching the Topic?**
  - One of the most common problems is data
  - The absence of a clear theoretical framework (contribution) in the published research
Importance of Literature Review

- Are there Any Major Controversies in this Topic Area?
  - The controversies may be of a methodological, theoretical or empirical nature.
  - Very fruitful source for deriving your own research questions
  - Research gaps: lack of theory, inconsistent findings, omitted variables?
  - (comprehensive, integrative, revisionism: main effect, mediating/moderating)

- Is the Topic Open to Hypothesis Testing?
  - Quantitative research: test clear, unambiguous, focused and testable hypotheses using an accepted statistical method
  - Qualitative research: a set of focused research questions or propositions using a logical and discursive analysis

- Is the Topic a Trivial One?
  - Little or no trace of the topic in the literature
What Should the Literature Review Do?

- **It should**
  - Enable you to sharpen and focus your initial research questions or even suggest new research questions.
  - Provide you with a wide and deep knowledge of the theoretical, empirical and methodological issues within your chosen research topic.
  - Provide a ‘bridge’ between your research questions and your research findings (relate to the literature).
  - Enable you to speak with authority on your research topic and the wider subject area.
  - Enable you to compare your research methods, theoretical framework and findings with work already done.
  - Enable you to set the scope and range of your research topic. (beyond the scope of our study, competitive manipulation)
What Should the Literature Review Do?

- Linking to your Research Questions
  - Discard ideas which are considered trivial
  - Discard any questions extremely difficult
  - Frame your research questions in the context of the main theories
  - Identify research questions that will fill in the gap

- Linking to Your Research Methods
  - Identify the ‘Limits of Validity’ of any findings

- Linking to Your Theoretical Framework
  - The most important!

- Linking to Your Research Findings
  - Evaluate your findings and sharpen in the light of your understanding of the literature
Types of Literature Review

- An Evaluative Review
  - Provide a discussion of the literature in terms of its coverage and contribution to knowledge in a particular area. \textit{meta-analysis}

- An Exploratory Review --- Most Common
  - Seek to find out what actually exists in the academic literature in terms of theory, empirical evidence and research methods
  - Create a ‘path’ between previous and current research

- An Instrumental Review
  - Used exclusively as a source of information on how to conduct some research on a highly specific research problem
Some General Points in Literature Reviewing

- Reviewing academic literature is not the same as just reading it!
- For any given piece of work:
  - Is there a theoretical framework?
  - If so, what is it and how does it fit into this topic?
  - Does the work provide links to other work in the topic?
  - Is there an empirical aspect to the work?
  - If so, what is its basis?
  - Does the work relate to a specific social group?
  - Does it relate to a particular place?
  - How applicable might it be outside the latter two?
  - How old is the work?
  - Is it still valid?
Some General Points in Literature Reviewing

- Summarize the literature review (**naturally leads to your research!**)
  - What does the previous research tell us about this topic?
  - What does it not tell us?
  - What are the key weaknesses in terms of theory, methods and data?

*Now you have a platform from which to launch your own research, interpret the findings and evaluate what you have achieved in comparison with the literature.*

- Should be written in a critical and reflective style
Obtaining and Searching the Literature

- Sources: journals, books, reports, abstracts, and electronic websites
  Top-tier and good journals (B) in your fields and related fields

- The credibility of sources
  - The safest sources are academic journals and the websites of academic departments

- Make notes and record accessing details

- Searching the literature
  - A systematic and methodical search of published sources of information to identify items relevant to a particular requirement

- Why do a search?
  - Help in topic selection
  - Discover information and provide knowledge of the subject area
Obtaining and Searching the Literature

○ Break your topic down into several separate information searches
  How many: 50-80 articles on the topic, 20-30 closely related articles, 2-5 benchmarks

○ Plan the Search
  ➢ Scope: the information you require
  ➢ Timescale: how far back is the information of relevance?
  ➢ Range: e.g., National or international?
    “working papers, unpublished manuscripts”

○ Select the key concepts and generate the keywords

○ The type of information
  ➢ Is it general theory (textbooks), current analysis and comment (newspapers and journals), names of contacts (directories), statistical, governmental, legal, technical or bibliographical?
Assess the Quality of Literature

- Tell ‘good’ and ‘not-so-good’ for a piece of work
  - Which organization is responsible and its legitimacy information?
  - Is there a statement as official approval of the organization?
  - Is there a statement giving the organization’s name as copyright holder?
  - Who wrote the article and his or her qualifications for writing on this topic?
  - Who is ultimately responsible for the content of the material?
  - Can the sources for any factual information be verified?
  - Are there editors monitoring the accuracy of the information being published?
  - If an academic journal article, when was it submitted and accepted to the journal?
Examples of literature review
Components of an Approach

- Objectives/Theoretical Foundations (theories, e.g., price vs. demand previous works)
- Analytical Model (your model or flow chart)
- Research Questions
- Hypotheses (relation between 2 or 3 vars)
- Specification of the Information Needed
Development of Research Questions and Hypotheses

Fig. 2.4

Components of the Marketing Research Problem

Objective/ Theoretical Framework

Analytical Model

Research Questions

Hypotheses
# The Role of Theory in Research

<table>
<thead>
<tr>
<th>Research Task</th>
<th>Role of Theory</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Conceptualizing and identifying key variables</td>
<td>Provides a conceptual foundation and understanding of the basic processes underlying the problem situation. These processes will suggest key dependent and independent variables.</td>
</tr>
<tr>
<td>2. Operationalizing key variables</td>
<td>Theoretical constructs (variables) can suggest independent and dependent variables naturally occurring in the real world.</td>
</tr>
<tr>
<td>3. Selecting a research design</td>
<td>Causal or associative relationships suggested by the theory may indicate whether a causal or descriptive design should be adopted.</td>
</tr>
<tr>
<td>4. Selecting a sample</td>
<td>The theoretical framework may be useful in defining the population and suggesting variables for qualifying respondents, imposing quotas, or stratifying the population (see Chapter 11).</td>
</tr>
<tr>
<td>5. Analyzing and interpreting data</td>
<td>The theoretical framework (and the models, research questions and hypotheses based on it) guide the selection of a data analysis strategy and the interpretation of results (see Chapter 14).</td>
</tr>
<tr>
<td>6. Integrating findings</td>
<td>The findings obtained in the research project can be interpreted in the light of previous research and integrated with the existing body of knowledge.</td>
</tr>
</tbody>
</table>
Research Questions and Hypotheses

- **Research questions** (RQs) are refined statements of the specific components of the problem.

- A **hypothesis** (H) is an unproven statement or proposition about a factor or phenomenon that is of interest to the researcher. Often, a hypothesis is a possible answer to the research question.
Department Store Project

- RQ: Do the customers of Sears exhibit store loyalty?

- H1: Customers who are store-loyal are less knowledgeable about the shopping environment.

- H2: Store-loyal customers are more risk-averse than are non-loyal customers.
Models

An **analytical model** is a set of variables and their interrelationships designed to represent, in whole or in part, some real system or process.

\[ Y = f(x), \quad Y = a_1X_1 + a_2X_2 + \cdots \] (linear, nonlinear)

In **verbal models**, the variables and their relationships are stated in prose form. Such models may be mere restatements of the main tenets of a theory. (chart & statements)
Graphical Models

Graphical models are visual. They are used to isolate variables and to suggest directions of relationships but are not designed to provide numerical results (e.g., AIDA model)

- Awareness
- Understanding: Evaluation
- Preference
- Patronage
Mathematical models explicitly specify the relationships among variables, usually in equation form.

\[ y = a_0 + \sum_{i=1}^{n} a_i x_i \]

Where

\[ y \] = degree of preference

\[ a_0, a_i \] = model parameters to be estimated statistically
Example of Literature, Theoretical Framework and Hypotheses development
Chapter 4
Research Design and Sampling
Research Design

- Research design is a master plan specifying the methods and procedures for collecting and analyzing the needed information.

- Research Designs in Terms of the Controlling Method
  - Experimental Design
    Researchers plan to measure the response variable depending (dependent variable) on the explanatory variable (independent variable)---Randomization

  - Quasi-experimental Design
    A mixed design where random and non-random experiments are employed together---Lack of randomization

  - Observational Design
    Prospective or Retrospective
Research Design

- Research Designs in Terms of Time Sequences
  - Prospective Design
    Researcher follows the participants and measures or observes the behavior of the participants: clinical trials or cohort design

  - Retrospective Design
    Researcher gathers data at once and classifies the participants simultaneously into the group categories: case-control studies vs. cross-sectional studies
Research Design

- Research Designs in Terms of Sampling Methods
  - Clinical Trial
    *Randomly allocates participants to the various groups of interest and measures differences in the future*
  - Cohort Study
    *Participants have a right to choose the group they want to join. The researcher measures differences between groups without randomization*
  - Case-control Study
    *Researcher gathers the data at once and then looks into the past of the participants to classify them*
  - Cross-sectional Study
    *Researcher gathers the data at once like case-control studies and then classifies them simultaneously on the classification (more than two categories) and their current responses.*
Cross-Sectional (slice of time) Designs

- Involve the collection of information from any given sample of population elements only once.

- In **single cross-sectional designs**, there is only one sample of respondents and information is obtained from this sample only once (one time T1).

- In **multiple cross-sectional designs**, there are two or more samples of respondents, and information from each sample is obtained only once. Often, information from different samples is obtained at different times (annual surveys).

- **Cohort analysis** consists of a series of surveys conducted at appropriate time intervals, where the cohort serves as the basic unit of analysis. A cohort is a group of respondents who experience the same event within the same time interval (class of students, 2015, freshmen, sophomores, , ,).
Longitudinal Designs

- A fixed sample (or samples) of population elements is measured repeatedly (over time, T1, T2, T3,...) on the same variables
  Stock prices, interest rates, CPIs
  Other time series

- A longitudinal design differs from a cross-sectional design in that the sample or samples (people or cases) remain the same over time (where multiple cross-sectional designs use different samples every time)
Cross-Sectional vs. Longitudinal

Cross-Sectional Design
Sample Surveyed at $T_1$

Longitudinal Design
Sample Surveyed at $T_1$
Same Sample also Surveyed at $T_2$
Sources of Data

- Two sources of information normally used for research purposes
  - Primary source of data
    - Obtain from a survey or experiment
  - Secondary source of data
    - Already available or have been collected for other purposes

- Types of data
  - Qualitative and quantitative

- Measurement of data
  - Nominal Scale: The numbers or letters assigned to objects serve as labels for identification or classification
  - Ordinal Scale: Arranges objects or categorical variables according to an ordered relationship---When a nominal scale follows an order
  - Interval Scale: Not only indicate order, they also measure the order or distance in units of equal intervals---When an ordinal scale measure the differences BW events
  - Ratio Scale: Have absolute rather than relative quantities (e.g., age, money)---When an interval scale has an absolute zero
Method of Data Collection

- Two phases in data collection
  - Pre-test and main study
    *Pre-test checks the data collection form to minimize errors due to improper design elements using a small-scale data*
    *Pilot study*

- The unit of analysis
  - Organizations, departments, work groups, individuals, or objects
  - *(multi-level: national, organizational/firm, team/group, individual)*
Sampling techniques

- Sampling is the process or technique of selecting a suitable sample for the purpose of determining parameters or characteristics of the whole population.

- Decisions to be made:
  - Size of the sample
  - What method to sample
  - The cost for sampling
  - How ‘representative’ is the sample

- Two basic sampling techniques
  - Probability sample: every element of the population has an equal chance of being selected.
  - Non-probability sampling: sample units are selected on the basis of personal judgment

- Sampling frame: the list of elements from which the sample may be drawn

- Sampling unit is a single element or group of elements subject to selection
Representative Sampling Plans

- Simple Random Sample---applied to homogeneous populations
  - Selections are made from a specified and defined population (i.e., the frame is known).
  - Every unit in the population has an equal (known) chance of selection.
  - The method of selection is specified, objective and replicable.

- Stratified Random Sampling---applied to heterogeneous populations
  - Samples are drawn equally or proportionately from each stratum (i.e., each homogeneous subpopulation).

- Systematic (Quasi-random) Sampling
  - Range the population from which selections are to be made in a list or series, choose a random starting point and then count through the list selecting every n-th unit
Representative Sampling Plans

- **Cluster (Multistage) Sampling**
  - Have a number of clusters which are characterized by heterogeneity in between and homogeneity within.

- **Sequential (Multiphase) Sampling**
  - A sampling scheme where the researcher is allowed to draw sample on more than one occasions.

- **Non-probability Sampling Methods**
  - The probability of selecting population elements is unknown

- **Convenience Sampling**
  - Unrestricted non-probability sampling---the cheapest and easiest
Representative Sampling Plans

- **Purposive Sampling**: A non-probability sample that conforms to certain criteria
  - Judgment sampling: A cross-section of the sample selected by the researcher conforms to some criteria.
  - Quota Sampling: control characteristics such as gender or social status to draw a representative sample of the population.

- **Snowball (Network or Chain) Sampling**
  - A small number of the samples initially selected by the researcher are then asked to nominate a group who would be prepared to be interviewed for the research; these in turn nominate others, and so forth.

- **Summary: Random sampling**
  - provide protection against selection bias
  - enables the precision of estimates to be estimated
  - enables the scope for non-response bias to be addressed
Purpose of Research

- For explanatory or theoretical research
  - Homogeneous samples are acceptable
  - Such as student samples in experiments
  - Randomization and exclude extraneous factor
  - Findings due to the variables of interest

For descriptive or predictive research

- Random or heterogenous samples may be preferred
- To enhance robustness, external validity
Sample Size Determination

- Various questions may arise:
  - How big should the sample be?
  - How small can we allow it to be?
  - Whether the sample size is adequate in relation to the goals of the study?

- Time and cost

- One of the most popular approaches
  - The power of a test of hypothesis

- Specify a hypothesis test on a parameter $\theta$ (the population mean).
- Specify the level of significance of the test $\alpha$.
- Specify an effect size $\hat{\theta}$ that reflects an alternative of scientific interest. (This effect size is unknown and hence hypothetical.)
- Specify a target value $\tilde{\pi}$ of the power of the test when $\theta = \tilde{\theta}$. 
Test of Significance for Population Mean

- The formula for determine sample size in the case of testing hypothesis of population means is:

\[
n_0 = \frac{Z_{\alpha/2}^2 \cdot (SD)^2}{d^2}
\]

where \( n_0 \) = sample size,

\( Z \) = Standardised normal value, usually taken as 1.96 for a 95 per cent confidence interval,

\( \alpha \) = Level of significance,

\( SD \) = Standard deviation (assumed to be known from prior survey or can be guessed or other published studies can inform on this),

\( d \) = Precision range (the required confidence interval).

- Sample size, effect size, and significant levels
- Large samples, small effect, but still significant!
Test of Significance for Population Proportion

- The formula for determine sample size in the case of testing hypothesis of population proportion (e.g., percentage of voters, or prevalence rates of a disease) is:

\[ n_0 = Z_{\alpha/2}^2 \frac{p(1-p)}{d^2} \]

where
- \( n_0 \) = sample size,
- \( Z \) = Standardised normal value,
- \( \alpha \) = Level of significance,
- \( p \) = Estimated rate,
- \( d \) = Precision range.
Key Statistical Concepts

- Population
  - Can be enumerated
  - **Population parameters**: population mean (μ) and population standard deviation (σ)
  - Population Totals
  - Proportions of the Population Having an Attribute

- Sub-population
  - **Sample statistics**: sample mean and sample variance
  - Ratios for sub-population
  - Statistical estimates: correlation coefficient, multiple regressions coefficients, and factor scores.
  - Depends on model: number of variables, parameters to be estimated and cells
    - No. >20 in each cell
    - (e.g., SEM, identification of the model, No. ≤ 200)
Some Problems With Random Sample Surveys

- Non-response Bias: distort selection probabilities
- Limited Sample Size and the ‘Inverse Square Law’: useful even if not very precise
- Sampling Distribution, Sampling Bias and Sampling Variance
  - Sampling distribution: the complete set of estimates that could be obtained from all the samples that could be selected under the sample design
  - Sampling bias: a measure of the location of the sampling distribution, relative to the population value.
  - Sampling variance: the variance of sampling distribution around the true population values.
- Sampling error: the tendency for estimates to differ from the population value
  - Sampling error = sampling bias + sampling variance
- Confidence Intervals
  - In the form of n standard errors (commonly n = 2)
The Normal Distribution

- The normal distribution is a statistical distribution that has a particular shape and known properties.
  - Symmetrical
  - A known and fixed relationship between the standard deviation of the distribution (the standard error) and the percentiles of the distribution
Inferential Statistics (vs. Bayesian)

- Based on the assumption of normal distribution (e.g., correlation and regression)
- The extent to which sample represents the population
- Such as assumptions may be violated
- Results in unreliable parameter estimates
- Problems in sample selection bias and self-selection
- Solutions: recollect data or transformation of existing data by taking a log or squared terms or quadratic terms
- May results in multicollinearity in interactions of variables
  Cui&Lui 2005: firm size and entry order, variance inflation factors (VIF<10), extremely large or reversed estimates
  Ridge regression
Sample selection bias

- Due to deliberate sampling or missing data (homemakers with jobs, people who did not respond to direct marketing)
- Selecting targets based on their previous responses also leads to endogeneity (previous purchase leads to being selected as target)
- Results in biased estimates
- Corrections
- The two-stage Heckman procedure (CuiWongWan2015)
- Other related problems
- Survival bias (subject mortality)
- Cui&Lui2005: pioneers (that survived) vs. latecomers
- Volkswagen vs. Peugeot
Self-selection Bias

- Built in the system (Expedia requires purchase while TripAdvisor does not, different reviewer profiles)
- Self-directed behavior (choice of schools)
- Solutions: control as many variables as possible, especially those suspected of self selection (national, industry, and firm level factors)
- Repeated resampling:
  - Bootstrap good for estimating confidence intervals
  - Permutation for testing hypotheses
Second, we adopt the Delta method to estimate the confidence interval of the extreme point. The extreme point through first order derivative is 1.214. We then estimate the confidence interval of 1.214 and compare it with the range of institutional distance (Lind & Mehlum, 2010). The 95% confidence interval ranges from 0.88 to 1.54. Thus, the confidence interval of the extreme point falls within the range of the data [−1.72, 1.64]. Lastly, we performed 2,000 bootstrap resampling to estimate the parameters of regulatory distance (RD $\beta=0.74$ and RD2 $\beta=-.30$, both sig. at the .001 level). These results suggest that the logit model with the quadratic function of regulatory distance is reliable to great extent and that the relationship between regulatory distance and the probability of location choice exhibits an inverted-U shape, thus supporting hypothesis one. In addition, the bootstrap procedure also helps to account for any potential self-selection by firms.
Sample selection: Permutation

- We have addressed the potential sample selection bias using nonparametric permutation tests, which have been widely applied to test partial regression coefficients in a linear model for samples with selection bias. The rationale behind the permutation method is to generate a reference distribution by recalculating a statistic for many permutations of the data. The method uses repeated permutations of the outcome vector to estimate the distribution of measured importance for each variable, in a non-informative setting. The P-value of the observed importance provides a corrected measure. A major advantage of using permutation testing is that it does not require the normality assumption. The use of permutation methods for exact inference dates back to Fisher in 1935. Since then, such methods have gained popularity given the increased computing power. Please refer to Anderson and Legendre (1999) and Ernst (2004) for details.
Figure A-1. The Permutation Test Results for Study 2a

P-Corrected (empirical) = 0.013; P-estimated (theoretical) = 0.009

Figure A-1 shows the permutation test results for Study 2a with 2,000 permutations. The histogram in this figure shows the computed empirical distribution. The blue line is the theoretical Student’s t distribution. The green line indicates the location of the test statistic. The figure indicates that the difference between the empirical distribution and the theoretical distribution is noticeable but not very significant. The P-value for the test statistic under the theoretical null distribution is 0.009 where its P-value is 0.013 under the empirical null distribution. Therefore, though the significance of the manipulation factor was slightly inflated by sample selection, the result was not notably different after correction (Figure A-2). Meanwhile, the permutation test results show that for Study 2b, the difference in the P-values of the theoretical distribution and empirical distribution (0.0325 vs. 0.0305) is not significant, suggesting that sample selection does not affect the results.
Self-selection

- We must consider the existence of potential factor selection bias due to the varying reviewer populations of the respective websites and their self-selection, which may lead to different reviews. Our approach cannot determine the difference in reviewer propensity to post reviews across the sites due to lack of reviewer identity data. Meanwhile, for any given website, reviewer propensity typically manifests in a consistent manner. Thus, we can infer that - for a set of randomly selected products - the systematic difference in aggregated review measures between two websites is a constantly varying factor, independent of review manipulations. Let ( ) denote this factor. In the linear model, (where K is an intercept, is control variables) can be reduced to , which can be further reduced to . It is hence plausible that the constant factor K does not affect the estimation of . Thus, the potential factor selection bias, if any, may not counterbalance or enhance the impact of manipulations on sales.