AP® Research Handbook

2019-07-09

Contents

Research Philosophy & Ethics

- No Plagiarism
- Institutional Review Board
- Data privacy standards

This AP® Research Handbook[1] is still a work in progress. Please excuse any blank sections and filler text.

[1]*AP is a registered tradebook of the College Board, which does not endorse and is not involved in the ongoing production of this handbook.

Chapter 1

Research Question

You can label chapter and section titles using `{#label}` after them, e.g., we can reference Chapter 1. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter 8.

Figures and tables with captions will be placed in `figure` and `table` environments, respectively.

Chapter 2

Literature Review

List resources for conducting literature review. Show example of literature review with inline citations. Show ways to keep track of sources for bibliography.

- How to Write a Literature Review
  - contains example literature reviews from political science, philosophy, and chemistry.

Consider using a reference management system like Mendeley to organize your sources as you conduct your literature review. In fact, Mendeley has a Literature Search function, so you can manage sources and conduct literature reviews at the same time. See the Bibliography Management Section for more information on managing sources.

- Databases for Literature Reviews
  - Directory of Open Access Journals
    * Browse by subjects in the humanities and sciences. This can be your starting point if you have not developed a research topic.
  - arXiv
    * Open-access journal articles in fields such as mathematics, statistics, economics, physics, quantitative biology, quantiative finance, and electrical engineering
    * arXiv to BibTex: Outputs automated citations in BibTeX and other formats by typing the arXiv number of the article. For instance, just type in 1905.03758 into the search engine if the article is labeled arXiv: 1905.03758.
    * Alternatively, use Mendeley Web Importer to import article into Mendeley Desktop for automated citation outputs.
  - Mendeley Literature Search
    * Download Mendeley Desktop and register for a free account. Mendeley Desktop syncs with your online Mendeley account, but the literature search is currently only available in the desktop version.
    * Mendeley is primarly a reference managements software, so you can organize your citations as you conduct your literature review.
  - CORE
    * Search engine with the world's largest collectin of open-access research papers.
    * For batch searches of metadata and full texts, you may consider requesting a free API key to use the Core API.
  - ScienceOpen
    * Search for content, authors, collections, and journals in the advanced search, where you have the option to search by discipline or key word.
  - Dimensions
    * Search for articles in clincial sciences, biochemistry, public health, physical chemistry, and materials engineering.

  - EBSCO Open Access
    * Search open-access journals and dissertations. Note that dissertations can vary in quality, since they have not gone through peer review.
    * AP Research students should have access to a free EBSCO account from the AP Capstone program.
  - SSRN
    * Many of the social science articles are free access.
  - ERIC: Institute of Educadtion Sciences
    * Search for articles related to to education research.
    * The search engine includes the open to search for full-text articles.
  - dblp: Computer Science Bibliography
    * Index of major computer science publications.
    * Option to search for open-access articles.
  - EconBiz
    * Search for journal articles, working papers, and conference papers in economics and business.
    * Option to search for open-access articles.
  - MyJSTOR
    * You can sign up for a free MyJSTOR account to access up to six articles a month for free.
    * This may be helpful for accessing articles that are not open access.

- Tips for Accessing Paywalled Articles
  - Search for the author's website. Many researchers have draft manuscripts on their websites or research profiles on sites such as ResearchGate.
  - Consult your school's research librarian for other ways to access the article.
  - Send the author an e-mail to request for a digital copy of the article. You should provide context in the e-mail request by including a brief description of your AP Research project and its relevance and connection to the author's article.

Chapter 3

Bibliography Management

While the bibliography is placed at the end of research papers, reference management begins as soon as you begin your literature review.

- Managing Citations in LaTeX
- Mendeley Desktop Download
  - The download will prompt you to create a free account. Mendeley Desktop is synced with your online account.
- Mendeley Web Importer
  - As you search for research articles online, you can use the Mendeley Web Importer to import citations into your Mendeley Desktop. If the importer can recognize the online article's metadata, it will automatically populate the citation entries. If not, you can still enter the citation entries manually and import into the Mendeley Desktop to keep track of your sources.
- Mendeley Tutorials
- Exporting .bib files from Mendeley Desktop
- Install MS Word plugin
- Import Mendelay sources into LyX
- Import .RIS Files into Mendeley

Chapter 4

Paper Guidelines

- AP Research Proposal Guidelines
- AP Research Paper Guidelines (LaTeX)
  - draft in progress

Chapter 5

File Organization

Chapter 6

Research Concepts

6.1 Reliability & Validity

6.2 Accuracy vs. Precision

6.3 Bias vs. Variance Tradeoff

- Understanding the Bias-Variance Tradeoff
  - As you introduce more variables to a model, bias tends to decrease. This means that on average the model will make predictions closer to the true value.
  - However, at the same time, variance tends to increase with the added variables. This means that each time you input the model with another random set of data, the outputted predictions will vary a lot. Why? More variables in the model means you might be in danger of overfitting your specific data. This means your model is so specifically tailored to your data at hand that it probably won't generalize well if you applied the model to other randomly drawn datasets.

6.4 Curse of Dimensionality

6.5 Correlation vs. Causation

Chapter 7

Research Design

Before you decide how you will conduct your research, read through the list of research designs in the USC library guide.

- What is Research Design?

7.1 Action

7.2 Case Study

7.3 Case-Control Study

- Case-Control Studies
  - The diagram on the first page provides a clear overview of case-control study design.
- Overview of Case-Control Design

7.4 Causal

7.5 Cohort

1. Sample a cohort of people from a population of interest
2. Follow this cohort for many years (often decades)
3. Record their attributes and exposures.
4. Record who develops outcome of interest (ongoing process).

- This method can be costly, both monetarily and time-wise.
- Compare this method with case-control studies, which are a type of observational study often used in epidemiology.

7.6 Cross Sectional

7.7 Descriptive

7.8 Experimental

7.9 Exploratory

7.10 Historical

7.11 Longitudinal

7.12 Meta-Analysis

7.13 Mixed Methods

7.14 Observational

7.15 Philosophical

7.16 Sequential

7.17 Systematic Review

7.18 Quasi-experimental

Chapter 8

Qualitative Research Methods

- Qualitative Research Methods Field Guide

8.1 Case Study

8.2 Narrative

8.3 Phenomenological

8.4 Ethnography

8.5 Grounded Theory

- Grounded Theory as Scientific Method

Chapter 9

Quantitative Methods

Causal Inference

These notes are based on Professor Masten's online course on Causal Inference at the Social Science Research Institute at Duke.

- Causal effect is often easy to detect with simple actions for which the effect immediately follows (e.g., you caused the alarm clock to stop ringing by pressing the snooze button)
- With multiple causes and delayed effects, causality is much harder to detect.
- Measurement:
  - Unit of analysis: countries, city blocks, people, firms, etc.
  - Outcome variable: the characteristic of the unit of analysis that we want to affect
  - Policy/treatment variable: the characteristic that we use to change the outcome varialbe
- A lot of characteristics cannot readily be quantified, so we often use proxy variables. For example, GDP could be a proxy for economic development.
- Causality: how an intervention in the policy variable affects the outcome variable
- Data:
  - The value of the policy variable has to vary in the dataset. Without this variation, you can't analyze how changes in the policy variable might affect the outcome variable.
  - Larger standard deviation = larger variation
- Correlation vs. Causation
  - If the policy and outcome variables are correlated, this does not necessarily imply a casual relationship.
  - Selection Problem: when units get to choose their policy variable, correlations between policy and outcome variables are unlikely to be causal.
    * Example: Neighborhoods with a lot of trees tend to have less crime.
    * If this were a casual relationship, then we could plant more trees in a neighborhood and expect crime to go down. However, this is unlikely. More likely, people who tend to commit less crimes chose to live in neighborhoods with tree-lined streets.
- Average Treatment Effect
  - Causal effects vary among people, so there is a distribution of causal effects in the population.

  - Theoretical ideal: you would know the unit level of causal effect for each person and thus make individualized treatment decisions. This is impossible in practice. You can't know the effect of receiving and not receiving treatment for an individual.
  - Unit-level causal effect: difference in outcome between treatment & control, holding all other variables fixed
  - Avg. treatment effect (ATE): avg. of all values for unit-level causal effects in a population
  - Avg. outcome under the policy: avg. outcome when everyone is affected by the policy (i.e., receives treatment)
  - Avg. outcome without the policy: avg. outcome when everyone is not affected by policy (i.e., does not receive treatment)
  - ATE = Avg. outcome under policy - Avg. outcome w/o policy

9.0.1 Experiments

- Controlled Experiments
  - Control group does not receive treatment
  - Experimental group receives treatment
  - All possible factors that could affect the outcome are identical for both groups, except for the treatment
  - Difference in the outcome between the two groups is the treatment effect
  - Typically used in hard sciences, but difficult to achieve in social sciences given the myriad of factors, many of which are difficult to measure and control
- Randomized Experiments
  - Split units randomly into two large groups: treatment or control
  - Right after randomization and before the experiment, both groups should be similar (i.e., avg. values of factors should be about the same), because the split was done randomly and the groups are very large
  - Since the two groups are similar in all factors except treatment, changes in the *average* outcomes are due to treatment
  - Complications:
    * Noncompliance: Even when you randomly assign treatment, people in the treatment group may not all decide to take the treatment. Also, some people in the control group, who should not receive the treatment, might decide to get the treatment.
      · Solution 1: Intent to Treat Analysis
        · The *intent* to provide treatment is by design random regardless of treatment non-compliance.
        · Thus, we can examine the causal effect of the option of providing treatment.
        · Downside: cannot analyze causal effect of treatment itself
        · For example, in the Oregon Health Experiment, while a lottery randomly selected people to receive free Medicaid, there was noncompliance in both the treatment/control groups. Original interpretation (effect of Medicaid on health outcomes) can be revised to effect of Medicaid lottery assignment on health outcomes.
      · Solution 2: Instrumental Variables
        · Advantage: We can analyze the causal effect of the treatment (not just the option of treatment) for a subset of the population.
        · Downside: cannot analyze average treatment effect over the entire population
      · Solution 3: Assume random compliance
        · Assume people comply with their treatment assignment.
        · Just drop the entries of the non-compliers.
        · Advantage: We can analyze the causal effect of the treatment over the entire population.
        · Downside: Decision to not comply is probably not random. We don't observe the reasons for non-compliance.
      · Solution 4: Bounds analysis

        - Get lower/upper bounds of average treatment effects using extreme scenarios.
        - Upper bounds: assume maximum value for outcome variable for noncompliers
        - Lower bounds: assume minimum value for outcome variable for noncompliers
    * Survey nonresponse
        - If nonresponse is not random, you cannot interpret the treatment effect as causal.
        - Example: People in the treatment group with negative outcomes responded to surveys at higher rates than those with positive experiences. Data becomes biased toward negative outcomes for the treatment group.
    * Sample Size: Even with a great research design, small sample size limits statistical inference.
    * Control: You may not be able to control the assignment of treatment.
  - Issues:
    * Ethics: Random assignment of treatment may have difficult ethical considerations (e.g., withholding a potentially life-saving drug to a terminally ill patient assigned to a control group in a randomized trial).
    * Extrapolation: It may be hard to extrapolate the results of a randomized experiment to another study if the treatment conditions and features are different.

- Natural Experiments
  - Researchers not involved in the research design and data collection in natural experiments, unlike in randomized experiments.
  - observational data used instead
  - Example: charter school lotteries

1. True Natural Experiments
   - treatment was randomly assigned, just not by researcher
2. As-If Natural Experiments
   - treatment not actually randomly assigned, but the treatment/control groups appear randomized as though treatment assignment were random)
   - treatment assignment not related to any variables that could affect outcome
   - balance check: characteristics of all observed variables (other than outcome variable) need to be similar between the treatment/control groups
     * There could still be differences between groups in unobserved variables.
     * Thus, we cannot prove treatment assignment is truly random, but balanced observed variables between groups would be part of a convincingly arugment that the observational data represents an as-if natural experiment.

9.0.2 Regression

In general, regression analysis can show only correlations between variables but not causation. The following are some factors that would prevent you from making the leap from correlation to causation in regression analysis:

1. Confounding Variables

- A confounding variable is a hidden variable that has an influence on both the dependent variable and independent (or outcome) variable.
- This distorts the asssocation between the dependent and independent variable.
- Example: Suppose we want to test the effect of the number of books at home (dependent variable) on a child's reading proficiency (independent variable). A confounding factor could be the mother's level of education, which could effect both the number of books at home and the child's reading proficiency. Without accounting for the mother's level of education, we could be overstating the effect of books on reading proficiency.

2. Reverse Causality

- In your study of the effect of A on B, you spot an association between the two variables. However, the direction of causality acturally could be reversed (i.e., B has an effect on A).

- Example: Suppose you notice that as the debt of a country increases, its economic growth decreases. The cause-and-effect relationship could be the other way around. Maybe a faltering economy is the reason that countries have to accumulate more debt to meet their budgets.

3. Simultaneity

- Similar to reverse causality, except that that direction of causality goes both ways at the same time. X causes Y, and Y causes X.

4. Mediating Variable

- Example: Does consuming caffeine stunt children's growth? We might question this causal link if we examine the mediating variable of sleep duration. Perhaps drinking coffee disrupts sleep patterns, which in turn affects growth.

Unconfoundedness Assumption

- Under the unconfounded assumption (a.k.a. **selection on observables assumption**), we have controlled for enough variables that we assume no confounding variables exists. With this assumption, we can consider the association between the treatment and outcome variable to be causal, since conditioning on all the other variables makes the treatment assignment essentially random.
- In practice, this assumption is difficult to make, because we cannot truly account for every possible variable relevant to the outcome variable. You would have to assume that all unobserved variables are not related to the treatment and do not contribute any effects to the outcome variable.

Ordinary Least Squares (OLS) Regression

- In OLS regressions, we want to draw a line through a set of points that represents the data.
- We calculate the distance between each data point and the line. We then square that distance. Do this for all data points and find the sum of all thes squared distances.
- The OLS regression line is the line that minimizes the sum of all these squared distances.
- Assumptions:
  - sample is random
  - outcome variable is continuous

Interaction Effects

- If an independent variable z changes the effect of another independent variable x on the outcome variable y, we have an **interaction effect** between x and z. The original regression model $y = \beta_0 + \beta_1 x + e$ turns into $y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3\ x * z + e$.
- This concept is also called **statistical moderation**, since z moderates the effect of x on y.

Calculating Average Treatment Effects

Table 9.1: Potential Outcomes of Treatment by Gender

Person	Gender	Treated	Outcome with Policy	Outcome without Policy	Unit-Level Causal Effect	Observed Outcome
1	M	Y	80	60	20	80
2	M	Y	75	70	5	75
3	M	Y	85	80	5	85
4	M	N	70	60	10	60
5	F	Y	75	70	5	75
6	F	N	80	80	0	80
7	F	N	90	100	-10	100
8	F	N	85	80	5	80

- In the example table above, we have two potential outcomes for each person: `Outcome with Policy`

and `Outcome without Policy`. In practice, you would never be able to observe both potential outcomes for each person. You either observe `Outcome with Policy` if the person received treatment or `Outcome without Policy` if the person did not receive treatment. In actual data, you would only obtain `Observed Outcome`.

- Unit-Level Causal Effect = Outcome with Policy - Outcome without Policy
- Average Treatment Effect (ATE) = average of all the unit-level causal effects
  - Using the example data above, ATE = $\frac{20+5+5+10+5+0-10+5}{8} = 5$.
- Conditional Avg. Treatment Effect (CATE) = average treatment effect for a subset of the units
  - Example: CATE for men is $\frac{20+5+5+10}{4} = 10$.
  - Example: CATE for women is $\frac{5+0-10+5}{4} = 0$.
- Average Treatment Effect on the Treated (ATT) = average treatment effect for only the people who received treatment
  - In the table above, Persons 1, 2, 3, and 5 received treatment.
  - To calculate ATT, we take the average of the unit-level causal effects of these treated people.
  - ATT = $\frac{20+5+5+5}{4} = 8.75$
- Difference in Mean Outcomes b/w Treated and Untreated (Control) Groups = ATT + Bias
  - left-hand side (LHS) = $\frac{80+75+85+75}{4} - \frac{60+80+100+80}{4}$ = -1.25
  - ATT = 8.75
  - Bias = difference in the averages in `Outcome without Policy` between treated and untreated groups
    - If this bias term is nonzero, we can say that the data has **selection bias**, because the selection into treatment is associated with potential outcomes.
    - For `Treated = Y`, average `Outcome without Policy` = $\frac{60+70+80+70}{4} = 70$.
    - For `Treated = N`, average `Outcome without Policy` = $\frac{60+80+100+80}{4} = 80$.
    - Bias = 70 - 80 = -10.
    - In this example, people selected for treatment have a much lower potential outcome without treatment than that of people not selected for treatment.
    - This negative selection bias will understate the actual impact of the treatment.
    - Similarly, positive selection bias will overstate the actual impact of the treatment.
  - right-hand side (RHS) = ATT + Bias = $8.75 - 10 = -1.25$
  - Thus, LHS = RHS.
  - In randomized experiments, the assignment of treatment is randomized, making the selection bias zero. Thus, the LHS (difference in the mean observed outcomes between the treated/untreated groups) equals ATT. Note that ATT is the average treatment effect (ATE) conditioned on being in the treatment group. Under randomization, we can assume that the potential outcomes are independent of the treatment assignment, so the average treatment effect is the same regardless of conditioning on treatment assignment. Thus, LHS = ATT + zero bias = ATE.
  - In this example data, we know that treatment is not randomly assigned, since more men received treatment than women. This means that treatment is correlated with gender. Thus, the calculations done above *do not* represents estimates of the average treatment effect due to the selection bias.
  - Instead, we assume that treatment assignment is randomized within each gender group. We calculate the estimated conditional average treatment effects $\widehat{CATE}(m)$ and $\widehat{CATE}(f)$ for males and females, respectively, by taking the difference between the average treatment group outcome and the average control (untreated) group outcome *within* each gender.
    - $\widehat{CATE}(m) = \frac{80+75+85}{3} - 60 = 20$
    - $\widehat{CATE}(f) = 75 - \frac{80+100+80}{3} = -11.67$
  - We add a hat on top of CATE to denote that the calculations are estimates. These estimates $\widehat{CATE}(m) = 20$ and $\widehat{CATE}(f) = -11.67$ are different from $CATE(m) = 10$ and $CATE(f) = 0$ that we calculated earlier. The CATEs without the hats are the true conditional average effects based on the difference in the two potential outcomes (with/without treatment) for each person. In practice, we will never be able to calculate the true CATEs, because we can never know both potential outcomes for each person. We only have one observed outcome for each person that we

use to estimate CATE with $\widehat{CATE}$.
- To estimate ATE, we take the weighted average of the estimated CATEs. In this case, both gender groups are equally weighted, so $\widehat{ATE} = \frac{1}{2}\widehat{CATE}(m) + \frac{1}{2}\widehat{CATE}(f) = 20 - 11.67 = 4.16$.
- Note that $\widehat{ATE} = 4.16$ is positive, whereas earlier calculation of LHS (difference in the mean observed outcomes in the treated/untreated groups) is negative, which understates the average treatment effect due to negative selection bias arising from non-random assignment of treatment between gender.

9.0.3 Matching Methods

In this method, we want to find units with similar explanatory variables but assigned to different treatments. One way to do this is propensity score matching.

9.0.4 Instrumental Variables

9.1 Statistical Tests

- Choosing a Statistical Test
- Hypothesis Testing Roadmap
- Chosing the Correct Statistical Test in SAS, Stata, SPSS, and R
- Uses & Misuses of Statistics
- Statistical Tests in R
- 1 group
  - interval variables
    * 1-sample t test for the mean
    * chi-squared test for variance
  - categorical variables
    * z test for proportions (2 categories)
    * chi-squared goodness-of-fit
  - ordinal or interval
    * one-sample median test
- 2 groups (independent groups)
  - interval variables
    * 2 independent sample t-test (equal variances)
    * 2 independent sample t-test (unequal variances)
    * F test for difference between 2 variances
  - categorical variables
    * z test for difference between 2 proportions
    * chi-squared test for difference between 2 proportions
    * Fisher's exact test
- 2 groups (dependent or paired groups)
  - paired t-test (interval variables)
  - McNemar's test (categorical variables)
  - Wilcoxon signed ranks test (oridinal or interval variables)
- more than 2 groups (independent groups)
  - one-way ANOVA (for interval variables)
  - Kruskal Wallis (for ordinal or interval variables)

  - chi-squared test (for categorical variables)

- more than 2 groups (dependent groups)
  - one-way repeated measures ANOVA (for interval variables)
  - repeated measures logistic regression (for categorical variables)
  - Friedman test (for ordinal or interval)

9.1.1 1-sample t-test

- Assumptions:
  - data is a simple random sample from population
  - sample mean follows a normal distribution
  - by the Central Limit Theorem, with sample size $n >= 30$, the sample mean is normally distributed regardless of the population distribution

- Two-tailed Hypothesis:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

- Test Statistic:

$$T = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}} \sim t_{(n-1)}$$

  - $\overline{X}$ = sample mean
  - μ_0 = hypothesized population mean
  - S = sample standard deviation
  - $t_{(n-1)}$ = t distribution with $n-1$ degrees of freedom

9.1.2 chi-squared test for variance

9.1.3 z test for proportions

- Assumptions:
  - sample proportion $p = \frac{X}{n}$ comes from random sample in population, where X is number of events of interest in sample size n.
  - p follows a binomial distribution, but we can assume normality when X and $n - X$ are each at least 5 (old standards) or at least 15 (current standards)

- Two-tailed Hypothesis:

$$H_0 : \pi = \pi_0$$

$$H_1 : \pi \neq \pi_0$$

- Test Statistic:

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim \mathcal{N}(0, 1)$$

  - π_0 = hypothesized proportion
  - p = sample proportion

9.1.4 t-test for 2 independent samples

- Assumptions:
  - two independent samples are randomly selected from two populations with the same variance
  - if you cannot use the assumption of same variance, use the Welch two-sample t-test
    * test statistic is the same as below, but degrees of freedom are adjusted
  - if populations are not normally distributed, the sample sizes n_1 and n_2 from the two populations needs to be at least 30 to ensure that the distribution of the sample means are normal by the Central Limit Theorem
- Two-tailed Hypothesis:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

  - μ_1 = population mean of 1st sample
  - μ_2 = population mean of 2nd sample
- Test Statistic:

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{(n_1+n_2-2)}$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$$

  - S_p = pooled variance
  - $\overline{X}_1$ = mean of 1st sample
  - $\overline{X}_2$ = mean of 2nd sample
  - S_1^2 = variance of 1st sample
  - S_2^2 = variance of 2nd sample
- More Info
- R Example
  - includes examples under both assumptions of equal and unequal variances
  - Andrew Heiss provides a brief tutorial with frequentist, simulation-based, and Bayesian approaches to comparing means between two groups. Also see Matti Vuorre's tutorial for more details.

9.1.5 paired t-test

- Assumptions:
- More Info with R Example

9.1.6 chi-squared test for proportions

- The chi-squared test for 2 x 2 frequency tables is equivalent to the square of the z-test for two proportions. See this link for detailed explanation.

9.1.7 chi-squared test for independence

- Explain connection between chi-squared test for independence and log-linear models, which are Poisson models for categorical data.

9.1.8 ANOVA

- F Distribution and Basic Principle Behind ANOVAs

9.2 Regression

9.2.1 Simple Linear Regression

A simple linear regression has the form $y = \beta_0 + \beta_1 x + e$, where

- y is the dependent (or outcome) variable
- x is the indepndent (or explanatory) variable
- β_0, the regression intercept, represents the average value of the outcome variable y when $x = 0$
- β_1 is the regression coefficient of x
  - for each 1-unit increase in x, y changes by β_1
  - β_1 is the slope of this regression line
- e is the error term. Some textbooks will use the variable u instead of e to emphasize that this term includes all unobserved variables.

9.2.2 Multiple Linear Regression

Linear regressions with multiple explanatory variables are called **multiple linear regressions**. If we have n explanatory variables, the multiple linear regression will have the form $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + e$, where

- β_0, the regression intercept, represents the average value of the outcome variable y when all the explanatory variables $x_1, x_2, ..., x_n$ are zero.
- β_1 represents how much y changes when x_1 increases by 1 unit, holding all other explanatory variables constant. The Latin phrase *ceteris paribus*, meaning "other things equal," is often used to describe this concept. This does not mean the other explanatory variables do not change. Rather, we are isolating the effect of x_1 on y under the theoretical condition that the other explanatory variables are held constant.
- This interpretation of β_1 extends to all the other regresssion coeffients.

9.2.3 Logistic Regression

- outcome variable is binary (e.g., 0 or 1; yes or no)

9.2.4 Mixed Effects Models

- Linear Models and Linear Mixed Effects Models in R: Tutorial 1
- A Very Basic Tutorial for Performing Linear Mixed Effects Analyses: Tutorial 2

9.3 Numerical Methods

In AP Calculus, you mostly encountered problems that can be solved analytically. However, in research, many differential equation models do not have analytical forms and must be solved numerically. Matlab is often used in applied math, engineering, and physical sciences for such cases as well as other modeling applications. Octave is an open-source alternative to Matlab. While R not the first language that comes to mind for numerical methods, many numerical R packages have been developed as well as integration with Matlab, Octave, and Julia.

- Numerical Computing with Matlab
  - This site has PDF versions of Cleve Moler's textbook on numerical computing alongside a video series with lectures on differential equations and linear algebra by Prof. Gilbert Strang and computational video tutorials by Moler.

- Numerically Solving Differential Euqations with R

9.3.1 Root-Finding Algorithms

- Newton-Raphson Method Using R
- Bisection Method Using R
- Secant Method Using R

9.3.2 Numerical Solutions to Differential Equations

- Integrating ODEs in R
- Euler Method Using Matlab
- Euler Method with Python
- Runge-Kutta Methods

9.4 Text Analysis

- Text Mining with R
- Mosteller and Wallace (1963) and the Federalist Papers
  - How Statistics Solved a 175-Year-Old Mystery About Alexander Hamilton
  - Replication of Analysis in Mosteller and Wallace (1963)

9.5 Network Analysis

- Introduction to Network Analysis with R

9.6 Geospatial Analysis

- Geocomputation with R

Resources by Discipline

Biology & Biostatistics

- Handbook of Biological Statistics
- An R Companion for the Handbook of Biological Statistics

Economics & Econometrics

- Introduction to Econometrics with R
- Principles of Econometrics with R
- Introduction to Data Science
- Using R for Introductory Econometrics
- Examples:
  - Annotated Sample Econometrics Paper
  - Microeconomic example of utility maximization constrained by budget lines

Psychology

- Psychology Research Methods

Public Health & Epidemiology

- Examples:
  - SIR Model Using R

Social Sciences

- Social Science Methods Modules
- Applied Causal Analysis

Chapter 10

Data

10.1 Data Sources by Discipline

10.1.1 Demography and Official Statistics

- U.S. Census Data
  - American Fact Finder
  - IPUMS
    * U.S. census microdata with social, economic, and health variables.
    * Create custom data sets or use online tool.
- UK Office for National Statistics
- Statistics Canada

10.1.2 Economics

- Panel Study of Income Dynamics
- University of Michigan Surveys of Consumers

10.1.3 Education

- Institute of Education Sciences: Data Files
- National Assessment of Educational Progress Data Explorer

10.1.4 Law

- Caselaw Access Project
  - Digital access to U.S. state and federal cases from the 1600s to present.

10.1.5 Social Sciences

- ICPSR

10.2 Data Documentation

Cite the source of your data. Provide links to the original data source and accompanying codebook, if any. Your data documentation will document your data analysis from the download of the raw data to the final steps of data analysis.

- Create a Codebook
  - List of codebook creation tools with guides and download links.
- Guide to Writing a Codebook
- How to Use R Codebook Package
  - Codebook Documentation
- Creating R Package for Research Documentation

Chapter 11

Analysis

11.1 Logical Fallacies

```
Read about the [common fallacies](https://medium.com/@pnhoward/12-common-fallacies-used-in-social-resea

- fallacies of authority
- fallacies of logic
- fallacies of emotion
```

11.2 Biases

```
- Sampling bias
    + e.g., 1948 U.S. presidential election (see this [case study](https://www.math.upenn.edu/~deturck/
    + even very large samples could have sampling biases if sampling methods are poor and unrepresentat
- Omitted variable bias
- Nonresponse bias
- Selection bias
- Survivorship bias
    + e.g., when bankrupt companies are removed from a stock index and replaced with profitable compani
- Recall bias
```

11.3 Model Selection

- Model Selection in R
- Linear Model Selection
- Forward/Backward Selection
- AIC/BIC
- Nested F-tests
  - model comparisons for linear regressions
- Likelihood ratio tests
  - model comparisons for generalized linear models
- Parsimony

Data Programming

11.4 R

- R
  - To download R, choose a CRAN mirror closest to your geographic location.
  - In order to build R packages, you should also download the latest recommended version of Rtools. Currently, the latest recommended version is `Rtoools35.exe`.
  - During the installation of Rtools, you may need to add in `"C:\Rtools\mingw_64\bin;"` to the path.
- R Studio
  - R Studio is an integrated development environment (IDE) for R. After downloading R Studio, you should be able to type the following command at the console to download some common R packages for data analysis and visualization.

```
install.packages(c("dplyr", "tidyr", "ggplot2", "esquisse", "stats", "xtable"))
```

- R Search Engine

11.5 Cleaning and Reshaping Data

```
library(reshape2)
library(tidyr)
library(xtable)
library(stringr)
library(knitr)
options(kableExtra.latex.load_packages = FALSE)
library(kableExtra)
library(pander)


#original data is organized by id/trial (two locations per entry)
game <- data.frame(id = c(rep("X",3), rep("Y",3), rep("Z",3)),
           trial = rep(c(1,2,3), 3),
           location_A = round(rnorm(9, mean = 0, sd = 1), 1),
           location_B = round(rnorm(9, mean = 0, sd = 1), 1))

# reshape data from wide to long (each entry is unique by id/trial/location)
game_long <- melt(game, id = c("id","trial"), value.name = "score")
game_long$variable <- str_sub(game_long$variable,-1,-1)
colnames(game_long)[3] <- "location"
```

```
# reshape data back to wide (same as original data)
game_wide <- dcast(game_long, id + trial ~ location, value.var = "score")
# reshape data into even wider form (one entry per id with 6 value columns: 2 locations X 3 trials)
game_wider <- dcast(game_long, id ~ location + trial, value.var = "score")

# using tidyr and dplyr to reshape data
game_long2 <- game %>% gather(label, score, location_A, location_B) %>%
    separate(label, c("label_p1","location"), sep = "_") %>%
    dplyr::select(-label_p1)

game_wide2 <- game_long2 %>% spread(location, value = score)

#unite() function creates the location X trial combinations first in long format # then apply the sprea
#just like in game_wide, each entry in game_wide2 is unique by id
game_wider2 <- game_long2 %>% unite(location_trial, location, trial) %>%
    spread(location_trial, value = score)

#xtable method
#print(xtable(game, caption = "Wide Data Listed by Person/Trial (Scores by Location)"), type="html")

#kable method
#kable(game, caption = "Wide Data Listed by Person/Trial (Scores by Location)", booktabs = TRUE) %>%
#    kable_styling(latex_options = c("hold_position"))

#pander method (most flexible)
pandoc.table(game, caption = "(\\#tab:wide) Wide Data Listed by Person/Trial (Scores by Location)")
```

Table 11.1: Wide Data Listed by Person/Trial (Scores by Location)

id	trial	location_A	location_B
X	1	-0.2	-0.2
X	2	1.3	1
X	3	1	-0.4
Y	1	-2.2	0.1
Y	2	1.2	-0.2
Y	3	0.8	1.1
Z	1	0.3	0
Z	2	-1.9	0
Z	3	-1.4	-0.1

```
pandoc.table(game_wider, caption = "(\\#tab:wider) Wider Data Listed by ID (Scores by Location/Trial)")
```

Table 11.2: Wider Data Listed by ID (Scores by Location/Trial)

id	A_1	A_2	A_3	B_1	B_2	B_3
X	-0.2	1.3	1	-0.2	1	-0.4
Y	-2.2	1.2	0.8	0.1	-0.2	1.1
Z	0.3	-1.9	-1.4	0	0	-0.1

```
pandoc.table(game_long, caption = "(\\#tab:long) Long Data")
```

Table 11.3: Long Data

id	trial	location	score
X	1	A	-0.8
X	2	A	0.4
X	3	A	1.1
Y	1	A	-0.4
Y	2	A	-0.2
Y	3	A	-0.3
Z	1	A	-0.1
Z	2	A	0.9
Z	3	A	0.2
X	1	B	-0.3
X	2	B	1
X	3	B	-0.3
Y	1	B	0.4
Y	2	B	2.6
Y	3	B	0.8
Z	1	B	0.3
Z	2	B	-0.4
Z	3	B	0.5

- Data Wrangling with dplyr and tidyr

11.6 Regular Expressions

- Regular Expressions in R
- UC Business Analytics R Programming Guide: Dealing with Regular Expressions
- Basic Regular Expressions in R Cheat Sheet

Literate Programming

11.7 LaTeX

- MiKTeX
  - First, download MiKTeX. Choose the version corresponding to your operating system (Windows, Mac, or Linux). Skip this step if you decide to use ShareLaTeX, which is an online LaTeX editor and does not require your computer to have underlying LaTeX packages via MiKTeX.
  - Recommended, download the basic installer, which will download other uninstalled packages on the fly on an as-needed basis. If you want to download all packages, you can choose the Net Installer, but this may take up a lot of space.
- Review of LaTeX Editors
  - Overleaf/ShareLaTeX
  - TeXstudio
  - LyX
- LaTeX Guides
- LaTeX Cheat Sheet
- Q and A:
  - Reference File in Parent Folder

11.8 Beamer

Beamer is a LaTeX class for presentations.

11.9 knitr (R + LaTeX)

- Using knitr in LyX
- Configure Texstudio to use knitr
- Create LaTeX Tables with kable
  - To avoid a incompatibility warning about the LaTeX `xcolor` package, place `options(kableExtra.latex.load_pac = FALSE)` in your R chunk before `library(kableExtra)`. See Hao Zhu's explanation in page 4 of the link above.
- kableExtra Vignettes
  - vignettes for using outputting tables from R into HTML, LaTeX, and Word
- xtable and stargazer Examples

- pander Tutorial

11.10 R Markdown

- Markdown Reference
- R Markdown Cheat Sheet
- Writing a Reproducible Paper in R Markdown

11.11 R Bookdown

- Authoring Books with R Bookdown
- R Markdown: The Definitive Guide
- Writing Thesis with Bookdown
  - Section on outputting into Microsoft Word using `bookdown::preview_chapter()`
- Writing Academic Papers with R Markdown

11.12 Rmd to MS Word

- Rmd to docx
- Discussion on Using knitr for Word output

Version Control

- Git

11.13 Github

3 Parts of Version Control

1. Working Directory: This is the folder where you store your files (e.g., `C:\Home\hw1`)
2. Staging Area: Every time you make changes to your files, you need to stage the files. This tells git that these files are ready
3. Repository: Every working directory that uses git for version control will have a `.git` folder, which records info about each version of the working directory.

Basic git commands:

```
git --version
```

- Check version of git

```
git config --global user.name "John Doe"
git config --global user.email "johndoe@example.com"
```

- Configure git to recognize you

```
git config --global user.name
git config --global user.email
```

- Confirm that user name/e-mail are configured

```
git init
touch readme.md
git add readme.md
git commit -m "added readme file"
```

- Initialize new repository
  - If you are at the file path `C:\Home\hw1`, `git init` will initialize the repository in the `hw1` folder
  - An invisible `.git` folder will be created in the `hw1` folder to track files.
  - If you are at the file path `C:\Home` and have not created the `hw1` folder, then you can type `git init hw1`, which will create the `hw1` folder and initialize the repo.
  - The code above did the following:
    1. `git init`: initialize repository
    2. `touch readme.md`: added a blank readme Markdown file
    3. `git add readme.md`: staged readme.md (this tells git that you are ready to commit this file to the repo)
    4. `git commit -m "added readme file"`: committed changed to the repo (`-m` allows us to add a message)

```
git diff
```

- Check differences between in the staging area and working directory.
- Output: difference between each file that has changed since you last staged files (`git add`)
- If you have staged all files, this command will output nothing, since nothing is different between your working directory and staging area.

```
git diff --staged
```

- Check the differences between staging area and repository.

```
git log
```

- See list of commits.

```
git reset file1.md
```

- Suppose `file1.md` was staged, but I want to remove it from the staging area.
  - Now, `file1.md` will be untracked.

```
git remote add origin https://github.com/<username>/<repo-name>.git
git push -u origin master
```

- Create a new repository in Github. Link your local working directory with the Github repo with these commands.
  - `git remote add`: sends commits to the location `origin`
  - We associated `origin` as a shortcut for the Github repo `https://github.com/<username>/<repo-name>.git`.
  - `git push -u origin master`
    * Send `master` branch to remote repo called `origin`
    * Option `-u` allows us to use `git push` as a shortcut for `git push origin master` in the future
- Best Practices Using Github in RStudio
- Tutorial on Git for Behavioral Sciences
- Github and R