5000 Methods Notes

Introduction

Statistics Dictionary Definitions:

- Branch of mathematics dealing with the collection, analysis, interpretation, and presentation of data
- Art and science of drawing justifiable conclusions from data

Mathematically, the simple linear regression model in matrix form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \text{where } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}).$$

The matrix formulation has

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \qquad E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}.$$

The unknown parameters are

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \text{and} \quad \sigma^2.$$

We have the following results:

- **The least squares estimator**

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix},$$

is the minimum variance linear unbiased estimator for $\boldsymbol{\beta}$.

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \mathbf{V} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}.$$

$$\mathbf{c}^\top\hat{\boldsymbol{\beta}} \sim N\big(\mathbf{c}^\top\boldsymbol{\beta},\ \mathbf{c}^\top\mathbf{V}\mathbf{c}\big).$$

- Test $H_0 : \mathbf{c}^\top\boldsymbol{\beta} = 0$ using

$$t = \frac{\mathbf{c}^\top \hat{\boldsymbol{\beta}} - 0}{\sqrt{\mathbf{c}^\top \mathbf{V} \mathbf{c}}}.$$

Statistics is the science of using information to make decisions and quantify uncertainty inherent to those decisions.

There are four basic steps in the statistical problem solving process (Deming):

1. Define the questions to be answered (Plan)
2. Gather appropriate data (Do)
3. Analyze the data (Study)
4. Interpret the results (Act)

Unit 1: Experiments

Terminology

Terminology

Experiment: an investigation in which the investigator applies (assigns) some treatments to experimental units and then observes the effect of the treatments on the experimental units by measuring one or more response variables.

Treatment: a condition or set of conditions applied to experimental units in an experiment.

Experimental Design: The assignment rule specifies which experimental units are to be observed under which treatments.

Experimental Unit: the physical entity to which a treatment is randomly assigned and independently applied. - the smallest division of material (e.g., land, plant, animal, etc.) to be studied

Response Variable: a characteristic of an experimental unit that is measured after treatment and analyzed to assess the effects of treatments on experimental units
(e.g., yield, gene expression level, etc.).

Observational Unit: the unit on which a response variable is measured. There is often a one-to-one correspondence between experimental units and observational units, but that is not always true.

Replication

- Applying a treatment independently to two or more experimental units
- Level of variability can be estimated for units that are treated alike.

Randomization

- Random assignment of treatments to experimental units
- Reduce or eliminate sources of bias (treatment groups are equivalent, *on average*, except for the assigned treatment)
- Cause and effect relationships can be demonstrated
- Create a probability distribution for a test statistic under the null hypothesis of no treatment effects

Blocking / Matching

- Group similar experimental units into blocks

- Apply each treatment to (the same number of) experimental units within each block (balance)
- Separate random assignment of units to treatments is done within each block (randomization)

Blinding

- Subjects do not know which treatment they received
- Researchers making measurements do not know the treatment assignments

Control of Extraneous Variables

- Control non-intervention factors
- Use homogeneous experimental units
- Accurate measurement of outcomes (responses)
- Tradeoff between accuracy and generalizability

Comparison to a Control Group

- Untreated (placebo) group
- Gold standard (best available treatment)

Scope

- Inferences are restricted to only those units used in the experiment
- Extending inferences beyond the units in the experiment
  - Were the units used in the experiment obtained from a **representative random sample** from some larger population?
    - Yes $\Rightarrow$ can make inferences about the population
    - No $\Rightarrow$ cannot make inferences about the population

Randomization Tests

Used for randomized experiments

Use the probability distribution imposed by the random assignment of units to treatment groups

- Under the null hypothesis
  $$H_0: \text{ treatments have the same effect}$$
  the response provided by any particular unit does not depend on the assigned treatment ($\Rightarrow \mu_1 = \mu_2$)
- Is the observed difference $\bar{y}_1 - \bar{y}_2$ inconsistent with H_0?
- Compare $\bar{y}_1 - \bar{y}_2$ with differences in sample means for all other possible random assignments of units to treatment groups
  (What if H_0 is true?)

General Comments

- The randomization test is also called the permutation test
- The randomization test (permutation test) depends on identifying units to permute, which should be the units in the experiment that are **exchangeable under the null hypothesis**, determined by the design of the experiment and the factor(s) being tested.

Observational Studies

- In some cases, the treatments cannot be assigned to experimental units by some rule.
  - For example, study of the effects of smoking on cancer with humans as the experimental units
  - Neither ethical nor possible
- We can still gather data by observing some members of the target population as they naturally exist.
  - Census: Observe all members of population
  - Haphazard (convenience) sample
  - Representative random sample
- This type of study is called an observational study and is not an experiment.

Simple Random Sampling

Without Replacement: every subset of n unique units has the same probability of being selected (more typical)

With Replacement: on each draw every member of the population has the same chance of being selected and the selected unit is put back into the population before the next unit is selected (some units may be selected more than once)

Sampling Schemes

- Only consider simple random samples, but there are many other sampling schemes that produce representative samples (Stat 521: Survey Sampling)
- The sampling procedure dictates the method of analysis
- Can make predictions and inferences about associations
- Causal inferences are not justified

Model-based Inference Overview

The Normal

A random variable Y with density function

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\}$$

is said to have a **normal (Gaussian)** distribution with

$$\text{Mean} \equiv E(Y) = \mu \qquad \text{and} \qquad \text{Variance} \equiv \text{Var}(Y) = \sigma^2.$$

The standard deviation is

$$\sigma = \sqrt{\text{Var}(Y)}.$$

We will use the notation

$$Y \sim N(\mu, \sigma^2).$$

The Standard Normal

Suppose Z is a random variable with a normal distribution where

$$E(Z) = 0 \quad \text{and} \quad \text{Var}(Z) = 1,$$

i.e.,

$$Z \sim N(0,1),$$

then Z has a **standard normal** distribution.

Linear Combinations

If Y_1 is a random variable with expectation μ_1 and variance σ_1^2 and Y_2 is a random variable with expectation μ_2 and variance σ_2^2, then

$$E(Y_1 + Y_2) = \mu_1 + \mu_2$$

$$E(aY_1 + bY_2 + c) = a\mu_1 + b\mu_2 + c$$

$$\text{Var}(Y_1 + Y_2) = \sigma_1^2 + \sigma_2^2 \quad \text{if } Y_1 \text{ and } Y_2 \text{ are independent}$$

$$\text{Var}(aY_1 + bY_2 + c) = a^2\sigma_1^2 + b^2\sigma_2^2 \quad \text{if } Y_1 \text{ and } Y_2 \text{ are independent}$$

$$\text{Var}(Y_1 + Y_2) = \sigma_1^2 + \sigma_2^2 + 2\,\text{Cov}(Y_1, Y_2)$$

$$\text{Var}(aY_1 + bY_2 + c) = a^2\sigma_1^2 + b^2\sigma_2^2 + 2ab\ \text{Cov}(Y_1, Y_2)$$

Useful Definitions

Variance:

$$\text{Var}(Y_1) = \sigma_1^2 = E\left[(Y_1 - \mu_1)^2\right].$$

Covariance:

$$\text{Cov}(Y_1, Y_2) = E[(Y_1 - \mu_1)(Y_2 - \mu_2)] = \rho_{12}\sigma_1\sigma_2,$$

where ρ_{12} is the correlation between Y_1 and Y_2.

The **correlation coefficient**

$$\rho_{12} = \frac{\text{Cov}(Y_1, Y_2)}{\sigma_1\sigma_2}$$

measures the strength of the linear relationship between Y_1 and Y_2.

Distribution of a Sample Mean

- Assuming independent observations from a population with mean μ_k, the sample mean

  $$\bar{Y}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} Y_{kj}$$

  is the best linear unbiased estimator for μ_k.

- If $Y_{k1}, Y_{k2}, \ldots, Y_{kn_k}$ are i.i.d. $N(\mu_k, \sigma_k^2)$ random variables, i.e., a simple random sample from a normal population, then

  $$\bar{Y}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} Y_{kj} \sim N\left(\mu_k,\ \frac{\sigma_k^2}{n_k}\right).$$

- $\bar{Y}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} Y_{kj}$ is a random variable (an **estimator**).
  Use

  $$\bar{y}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} y_{kj}$$

  to denote its **estimate** (observed value).

Distribution for Difference in Two Sample Means For independent simple random samples from two normal populations:

- $Y_{11}, \ldots, Y_{1n_1}$ are i.i.d. $N(\mu_1, \sigma_1^2)$,
- $Y_{21}, \ldots, Y_{2n_2}$ are i.i.d. $N(\mu_2, \sigma_2^2)$.

Then,

$$\bar{Y}_1 - \bar{Y}_2 \sim N\left(\mu_1 - \mu_2,\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

The Central Chi-Square Distribution Let Z_i, $i = 1, 2, \ldots, n$, be independent standard normal random variables.
The distribution of

$$W = \sum_{i=1}^{n} Z_i^2$$

is called the **central chi-square distribution** with n degrees of freedom.

We denote this by

$$W \sim \chi_\nu^2,$$

where ν is the number of degrees of freedom.

Estimation of Variances For

$$Y_{11}, Y_{12}, \ldots, Y_{1n_1} \overset{\text{iid}}{\sim} N(\mu_1, \sigma_1^2), \qquad Y_{21}, Y_{22}, \ldots, Y_{2n_2} \overset{\text{iid}}{\sim} N(\mu_2, \sigma_2^2),$$

- The sample variance

  $$S_1^2 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (Y_{1j} - \bar{Y}_1)^2$$

  is an unbiased estimator of σ_1^2.

- The sample variance

  $$S_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (Y_{2j} - \bar{Y}_2)^2$$

  is an unbiased estimator of σ_2^2.

- If $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (homogeneous variances), the pooled estimator is

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

Sum of Independent Chi-Squares The sum of two independent central chi-square random variables with ν_1 and ν_2 degrees of freedom has a central chi-square distribution with $\nu_1 + \nu_2$ degrees of freedom.

Consequently,

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

has a chi-square distribution with

$$(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$$

degrees of freedom.

The Student t-Distribution If

$$Z \sim N(0,1), \quad W \sim \chi^2_r,$$

and Z and W are independent random variables, then the random variable

$$T = \frac{Z}{\sqrt{W/r}}$$

has a **central Student t-distribution** with r degrees of freedom.

We denote this by

$$T \sim t_r.$$

Inference for Difference in Means with Equal Variances

Assumptions

- Two independent random samples:

$$Y_{11}, Y_{12}, \ldots, Y_{1n_1} \quad \text{and} \quad Y_{21}, Y_{22}, \ldots, Y_{2n_2}$$

- Normality:

$$Y_{1i} \sim N(\mu_1, \sigma_1^2), \quad Y_{2j} \sim N(\mu_2, \sigma_2^2)$$

- Homogeneous population variances:

$$\sigma_1^2 = \sigma_2^2$$

Distribution for Inference

Let

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

Then

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}.$$

Hypothesis Testing

Hypotheses

$$H_0 : \mu_1 = \mu_2 \quad (\mu_1 - \mu_2 = 0)$$

$$H_a : \begin{cases} \mu_1 < \mu_2 & \text{(left-tailed)} \\ \mu_1 > \mu_2 & \text{(right-tailed)} \\ \mu_1 \neq \mu_2 & \text{(two-tailed)} \end{cases}$$

Test Statistic

The observed test statistic is

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

We assess whether this value is typical under H_0 or unlikely assuming H_0 is true.

Sampling Distribution

Assuming H_0 is true,

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}.$$

If H_0 is true, we expect T to be close to zero.
Large deviations from zero are unlikely under H_0.

p-Value

Definition:
The p-value is the probability of observing a test statistic at least as extreme as the one observed, assuming H_0 is true.

Interpretation: Scale-of-Evidence Framework

p-value range	Evidence for H_a
$p > 0.10$	little to no evidence
$0.05 < p \leq 0.10$	borderline / weak evidence
$0.025 < p \leq 0.05$	moderate evidence
$0.001 < p \leq 0.025$	strong evidence
$p \leq 0.001$	overwhelming evidence

Post-hoc Assessment: Errors

- If the p-value was small:
  - H_0 is true and we unluckily/randomly made an error
  - Type I error probability:
  $$P(\text{reject } H_0 \mid H_0 \text{ true}) \leq \alpha$$
  - H_0 is false (no error committed)
- If the p-value was large:
  - H_a is true and we unluckily/randomly made an error
  - Type II error probability:
  $$P(\text{fail to reject } H_0 \mid H_0 \text{ false}) = \beta$$
  - The power of a test is $1 - \beta$
  - H_0 is true (no error committed)

Confidence Intervals

The following is for estimating *differences in means*

Assumptions

- $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ are i.i.d. $N(\mu_1, \sigma^2)$
- $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ are i.i.d. $N(\mu_2, \sigma^2)$
- Population variances are equal

- Y_{1i} and Y_{2j} are independent for all i and j

Confidence Interval

A $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{Y}_1 - \bar{Y}_2) \ \pm \ t_{n_1+n_2-2,\,1-\alpha/2}\, S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},$$

where

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

Hypothesis Test Interpretation

A $100(1-\alpha)\%$ confidence interval can be constructed by including all values of δ such that the data does not provide sufficient evidence to reject the null hypothesis

$$H_0: \ \mu_1 - \mu_2 = \delta$$

relative to the two-sided alternative

$$H_a: \ \mu_1 - \mu_2 \neq \delta$$

at the α significance level.

Interval Width

Confidence interval widths depend on:

- the confidence level (which is related to significance α),
- the value of σ,
- sample sizes n_1 and n_2.

Sample Size Considerations

Note: Sample size calculations refer to the experimental units to replicate, not the observational units (though they sometimes are one and the same!)

Based on Standard Error   Difference in Means

- Difference in population means $(\mu_1 - \mu_2)$:

$$\text{s.e.}(\bar{Y}_1 - \bar{Y}_2) = S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

- Assuming $n_1 = n_2 = n$, we have:

$$\text{s.e.}(\bar{Y}_1 - \bar{Y}_2) = S_p\sqrt{\frac{2}{n}}$$

- Specify an acceptable value for the standard error and solve for n:

$$\text{s.e.} = \frac{\sqrt{2}S_p}{\sqrt{n}} \quad \Rightarrow \quad n = \frac{2S_p^2}{(\text{s.e.})^2}$$

- Requires a value for S_p from:
  - a previous study
  - a pilot study
  - a guess

Based on Confidence Interval Difference in Means

- Width of the confidence interval (assuming $n_1 = n_2 = n$):

$$w = 2\, t_{2(n-1),\, 1-\alpha/2}\, S_p \sqrt{\frac{2}{n}}$$

- Find n to achieve specified width:

$$n = 8\left(\frac{t_{2(n-1),\, 1-\alpha/2} S_p}{w}\right)^2$$

- One difficulty is that n enters twice (sample size and degrees of freedom for t):
  - Compute initial value using the normal approximation:

$$n_0 = 8\left(\frac{z_{1-\alpha/2} S_p}{w}\right)^2$$

- Then improve using:

$$n = 8\left(\frac{t_{2(n_0-1),\, 1-\alpha/2} S_p}{w}\right)^2$$

Recall: Four Possible Outcomes for Hypothesis Test

Decision	H_0 is true	H_0 is false
Reject H_0	Type I Error	Good Decision
Fail to reject H_0	Good Decision	Type II Error

Based on Hypothesis Test Difference in Means

For a t-test of
$H_0 : \mu_1 = \mu_2$
against
$H_a : \mu_1 \neq \mu_2$:

- Equal sample sizes: $n_1 = n_2 = n$
- Type I error rate: α
- Power: $1 - \beta$ for detecting $\delta = \mu_1 - \mu_2$
- Pooled estimate of population variance: S_p^2

The required sample size for each group is:

$$n = \frac{\left(t_{2(n-1),\,1-\alpha/2} + t_{2(n-1),\,1-\beta}\right)^2 (2S_p^2)}{\delta^2}$$

Based on Hypothesis Test (Two-Step Approach) Difference in Means

- As before, n enters twice. Use the same two-step approach.
- First compute:

$$n_0 = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 (2S_p^2)}{\delta^2}$$

- Then update:

$$n = \frac{\left(t_{2(n_0-1),\,1-\alpha/2} + t_{2(n_0-1),\,1-\beta}\right)^2 (2S_p^2)}{\delta^2}$$

- Common to use power values of 80%, 90%, or 95%, just as arbitrary as using $\alpha = 5\%$.
- Can adapt to a one-sided alternative by replacing $\alpha/2$ with α in the formulas.

Inference Diagnostics

Assessing Equal Variances

Graphical Method

- Construct residual plots, histograms, or boxplots of values for each group/population
- Look for:
  - Outliers in each sample
  - Differences in IQR, range
  - Differences in shape of sample distributions

Summary Statistics

- Check the ratio of sample standard deviations

$$\frac{\max\{S_1, S_2\}}{\min\{S_1, S_2\}}$$

- Interpretation guidelines:
  - Between 1 and 2 — little impact
  - Between 2 and 3 — potential impact
  - Greater than 3 — likely impact

F-test

- Reject $H_0 : \sigma_1^2 = \sigma_2^2$ if

$$F_{\max} = \frac{\max\{S_1^2, S_2^2\}}{\min\{S_1^2, S_2^2\}} \geq F_{(a,b),\,1-\alpha/2}$$

- where
  - $a = n_1 - 1,\ b = n_2 - 1$ if $S_1^2 > S_2^2$
  - $a = n_2 - 1,\ b = n_1 - 1$ if $S_2^2 > S_1^2$
- Notes:
  - Very sensitive to normal distribution assumption
  - Not recommended as the only check

Brown–Forsythe Test

- Conduct a two-sample t-test on the absolute deviations from the sample medians to assess homogeneous variability

Remedies to Unequal Variance   Welch Approximation

- Very similar results to two-sample inference when sample sizes are nearly equal
- Better performance with unequal sample sizes **and** unequal variances

Transformation

- Replace Y_{ij} with $X_{ij} = h(Y_{ij})$
- Perform inference on the X_{ij}'s → e.g., compare $\bar{X}_1$ with $\bar{X}_2$
- Back-transform estimates to get conclusions on the Y scale
  - only approximate conclusions about population means on the Y scale

Transformation Cont.

- Choosing the transformation
  - Trial and error: transform and check histogram
  - Rules of thumb:
    - Data are all positive — use $\log(Y)$
    - Data are proportions — use $\arcsin(\sqrt{Y})$
    - Data are counts — use $\sqrt{Y}$
  - Use transformation based on science
    (square root of area, cube root of volume)
  - Adjust for a variance–mean relationship
    (common for variance to increase with the mean)

Assessing Normality

Graphical Methods

- Histogram of values within each group/population
  - Look for symmetric, bell shape
- Normal probability plot within each group/population
  - Compare empirical cumulative distribution function (CDF) to CDF for theoretical normal distribution
  - Most commonly done using quantiles (Q–Q plot):
    plot empirical quantiles against expected quantiles from normal distribution

Normal Q–Q Plot

- Order residuals from smallest to largest
  (say $X_{(1)}, \ldots, X_{(n)}$)
- Compute expected quantiles $(q_{(1)}, \ldots, q_{(n)})$ from a standard normal distribution
  - Expected quantiles can be calculated with tables
  - General approximation:
    $$q_i = \Phi^{-1}\left(\frac{i}{n+1}\right)$$
  - Blom approximation:
    $$q_i = \Phi^{-1}\left(\frac{i - .375}{n + .25}\right)$$
  - For $i = 5$, $n = 9$, $q_5 = \Phi^{-1}\left(\frac{5}{10}\right) = 0$
- Scatterplot of $X_{(i)}$ vs q_i should be close to a straight line with slope σ
- Curved patterns indicate non-normal distributions (or presence of outliers)

Numerical Summaries

- For any normal distribution:

  - Mean and median should be equal
  - Skewness $= E(Y - \mu)^3/\sigma^3 = 0$
    (Skewness measures the asymmetry)
  - Kurtosis $= E(Y - \mu)^4/\sigma^4 = 3$
  - Excess kurtosis = kurtosis -3
    (estimated by the *univariate* procedure in SAS)
  - The sample kurtosis measures the heaviness of the tails of the data distribution
  - Positive value: long-tail; negative value: short-tail

Tests

- Many proposed tests for normality
- Tests based on empirical CDFs: Kolmogorov–Smirnov, Anderson–Darling, etc.
- Tests based on skewness or kurtosis
- Chi-square goodness-of-fit tests
- Tests based on normal probability plots: Shapiro–Wilk, correlation tests
- Normality is almost always rejected for large sample sizes

Consequences of Non-Normality

- Large samples → few consequences (Central Limit Theorem)
- Small samples:
  - Sample distributions have same shape and
    - equal sample sizes → very little impact
    - different sample sizes → potential impact if distributions are skewed
  - Sample distributions have different shapes → impact

Remedy for Non-Normality

- Transformation (especially for skewness)
- Discussed earlier (under remedies for unequal variances)
- Detect and eliminate outliers
- Non-parametric tests

Non Parametric Tests Wilcoxon Rank–Sum Test

- Independence
- Null hypothesis: two populations have the same distribution
  - Distribution is not required to be normal
  - Implies equal medians, percentiles, means, and variances
- Can test against one- or two-sided alternative
- Can compute “exact” p-values based on the null distribution of the ranks

Wilcoxon Rank–Sum Test (Procedure)

- Order the combined $n_1 + n_2$ observations (small to large)

- Assign ranks
  - Smallest gets rank = 1, second smallest gets rank = 2, etc.
  - For tied observations, average the ranks
- Compute the sum of the ranks for one group (call it W)
- Assuming H_0 is true, compute:

$$E_0(W) = \frac{n_1(n_1 + n_2 + 1)}{2} \qquad \text{and} \qquad V_0(W) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

- Large sample Z-test:

$$z = \frac{|W - E_0(W)| - 0.5}{\sqrt{V_0(W)}}$$

- Approximate p-value:

$$2 \times P(Z > |z|)$$

Unit 2: ANOVA

Motivation

- Do the populations or treatment groups have the same mean values for the variable?
- Two sources of variation:
  - Variability among observations within each treatment group (or within each population)
  - Variability among mean responses for treatments (or between populations)
- Question:
  - Are differences among group means large relative to variation within groups?
  - Do all populations have the same mean?

Analysis of Variance (ANOVA)

- Calculate three variations based on observations Y_{ij}:
  - Variation due to group means
  - Variation due to residuals
  - Total variation
- These are called the **sums of squares (SS)**

Cell Means Model

Linear Model Form

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

- Each observation Y_{ij} can be described by two components:
  - Fixed mean value μ_i
  - Random error term ε_{ij}
- Gives an equation for each of the
  $$N = \sum_{i=1}^{r} n_i$$
  observations

Matrix Form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- The vector $\mathbf{Y}$ is length N and is the vector of observations.
- The matrix $\mathbf{X}$ is size $N \times r$ and is called the design matrix.
  It relates the observations to the parameters according to the model.
  It is fixed (non-random).
- The vector $\boldsymbol{\beta}$ is length r and is the vector of parameter values.
- The vector $\boldsymbol{\varepsilon}$ is length N and is the vector of random error terms.

Basic ANOVA

Variation due to Group Means

$$SS_{\text{among groups}} = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{r} n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$$

- Also called SS_{model}
- If the population means are the same (different), this value should be small (large)

Variation due to Residuals

$$SS_{\text{within groups}} = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2$$

$$= \sum_{i=1}^{r}(n_i - 1)S_i^2$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{n_i} e_{ij}^2$$

- Also called SS_{error} or $SS_{\text{residuals}}$

Total Variation

$$SS_{\text{total}} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 = SS_{\text{model}} + SS_{\text{error}}$$

ANOVA Table

Source of variation	Degrees of freedom	Sums of squares	Mean square	F
Model	$r-1$	SS_{model}	$MS_{\text{model}} = \dfrac{SS_{\text{model}}}{r-1}$	$\dfrac{MS_{\text{model}}}{MS_{\text{error}}}$
Error	$N-r$	SS_{error}	$MS_{\text{error}} = \dfrac{SS_{\text{error}}}{N-r}$	
Total	$N-1$	SS_{total}		

Note:

$$MS_{\text{error}} = S_p^2$$

‘

Model Assumptions

- Assumptions on random error terms:
  - ε_{ij} are i.i.d. from a normal distribution with mean 0 and variance σ^2
  - $\boldsymbol{\varepsilon}$ is multivariate normal with mean $\mathbf{0}$ and variance $\sigma^2\mathbf{I}$
- This implies that:
  - Y_{ij} are i.i.d. from a normal distribution with mean μ_i and variance σ^2
  - $\mathbf{Y}$ is multivariate normal with mean $\mathbf{X}\boldsymbol{\beta}$ and variance $\sigma^2\mathbf{I}$
- In addition, we assume groups are independent of each other

ANOVA F-test

- Null hypothesis:
$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_r$$
- Alternative hypothesis:
$$H_a : \text{at least one } \mu_i \text{ is different for } i = 1, \ldots, r$$
- Test statistic:
$$F = \frac{MS_{\text{model}}}{MS_{\text{error}}}$$
- P-value:
$$P(F_{r-1,\, N-r} > F)$$

Effects Model

Linear Effects Model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

- Each observation Y_{ij} can be described by two components:
  - **Fixed mean value:** $\mu_i = \mu + \alpha_i$
    - Overall mean value: μ
    - Treatment effects compared with overall mean: α_i
    - Goal: find which α_i's are different from 0
  - **Random error term:** ε_{ij}

Identifiability Issues

- Model has too many parameters: estimates r means with $r + 1$ parameters
- Design matrix $\mathbf{X}$ is not full column rank
- The usual inverse $(\mathbf{X}^\top \mathbf{X})^{-1}$ does not exist
- There are infinitely many least squares estimators

Solution: impose constraints on the parameters - Set $\alpha_r = 0$ (baseline constraint), or - Set

$$\sum_{i=1}^{r} \alpha_i = 0$$

(sum-to-zero constraint)

Least Squares Estimator of $\boldsymbol{\beta}$

When

$$\sum_{i=1}^{r} \alpha_i = 0,$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = \begin{pmatrix} \frac{1}{r}\sum_{i=1}^{r} \bar{Y}_{i\cdot} \\ \bar{Y}_{1\cdot} - \frac{1}{r}\sum_{i=1}^{r} \bar{Y}_{i\cdot} \\ \bar{Y}_{2\cdot} - \frac{1}{r}\sum_{i=1}^{r} \bar{Y}_{i\cdot} \\ \vdots \\ \bar{Y}_{(r-1)\cdot} - \frac{1}{r}\sum_{i=1}^{r} \bar{Y}_{i\cdot} \end{pmatrix} = \begin{pmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_{r-1} \end{pmatrix}.$$

Cautions

- The above two types of constraints are not the only ways to model the means
- The choice of constraint affects the least squares estimator $\hat{\boldsymbol{\beta}}$
- You must determine which constraint was applied before interpreting parameter estimates
- The interpretation of parameters (elements of $\boldsymbol{\beta}$) depends on the parametrization

Fixed vs. Random Effects

Fixed Effects

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

- The r treatments (or groups) examined in the study are the only ones under consideration
- Research questions concern treatment means or differences in means
  - e.g., two drugs, four pesticides

Random Effects

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

- The r treatments (or groups) are a random sample from a larger population of possible treatments (or groups)
- Research questions concern variability among sets of treatments (or groups) that could be selected for different studies
- Additional assumptions:
  $$\alpha_i \sim N(0, \sigma_\alpha^2),$$
  and α_i is independent of ε_{ij}

ANOVA Diagnostics and Remedies

ANOVA Assumptions

- ε_{ij} are i.i.d. $N(0, \sigma^2)$
- Independence of groups and observations
- Homogeneous (equal) variance:
  $$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2 = \sigma^2$$
- Normal distribution:
  - Random error terms are normally distributed

Model Diagnostics

- Many results from two-sample model diagnostics apply:
  - Independence: critical aspect
  - Equal variances: important
  - Normality: only a concern for small sample sizes or very skewed distributions
  - Outliers: results not robust
- Use residuals to assess model assumptions:
  $$e_{ij} = Y_{ij} - \bar{Y}_{i\cdot}$$

Independence Assumption

Data collection: - Random sample(s) from multiple populations - Observations from multiple independent groups

- Study designed to produce independent responses

Equal Variance Assumption (Graphical Checks)

- Construct histograms of residuals for each group
- Construct boxplots of residuals for each group
- Plot residuals versus predicted values (there should be no trend)
  - Beware of interpretation if n_i's are very unequal
  - Expect larger range of e_{ij} if n_i is larger
- Study ratio of sample standard deviations:

$$\frac{\max\{S_i\}}{\min\{S_i\}}$$

Equal Variance Assumption (Formal Tests)

- Tests for equality of variances:
  - Brown–Forsythe test
  - Levene's test
  - etc.
- Consequences of unequal variances on the F-test:
  - Minor if sample sizes are the same
  - Large distortion of α level if sample sizes are very unequal
  - Decreased power

Normality Assumption

- Histogram of residuals
- Normal probability plot of residuals
- Numerical summaries:
  - Skewness
  - Kurtosis
- Tests for normality:
  - Shapiro–Wilk
  - Kolmogorov–Smirnov
  - Cramér–von Mises
  - Anderson–Darling

Non-Parametric

Kruskal–Wallis Test

- Combine the data into a single data set
- Order the N observations from smallest to largest
- Assign ranks R_{ij}:
  - Smallest observation gets rank 1, second smallest gets rank 2, etc.

  - For tied observations, average the ranks

- Calculate $\bar{R}_{i\cdot}$ = mean rank of observations in group i
- Test statistic:

$$H = (N-1)\frac{\sum_{i=1}^{r} n_i(\bar{R}_{i\cdot} - \bar{R})^2}{\sum_{i=1}^{r}\sum_{j=1}^{n_i}(R_{ij} - \bar{R})^2}$$

  where

$$\bar{R} = \frac{N+1}{2}$$

- If H_0 is true, H has an approximate χ^2 distribution with $r-1$ degrees of freedom
- Approximation is best when $n_i \geq 5$ for all i
- p-value:

$$P(\chi^2_{r-1} > H)$$

ANOVA Contrasts

Motivation

- Inference for a single population mean
- Linear combinations of means, including contrasts
- Pairwise comparisons

Inference for Single Population Mean

- $100(1-\alpha)\%$ confidence interval for a single group mean:

$$\bar{Y}_{i\cdot} \ \pm \ t_{N-r,\,1-\alpha/2}\sqrt{\frac{MS_{\text{error}}}{n_i}}$$

- Notes:
  - MS_{error} is the estimate of the population variance σ^2
  - Degrees of freedom for the t distribution: $N-r$
  - Valid for inference on a *single* population mean (not used for comparison between means)

Contrast

- A **contrast** is a linear combination of the population means with $\sum_{i=1}^{r} c_i = 0$:

$$\gamma = \sum_{i=1}^{r} c_i \mu_i$$

Orthogonal Contrasts

- Two contrasts

  $$\gamma_1 = \sum_i c_i\mu_i, \qquad \gamma_2 = \sum_i b_i\mu_i$$

  are **orthogonal** if

  $$\sum_i \frac{b_i c_i}{n_i} = 0$$

- If γ_1 and γ_2 are orthogonal:
  - They represent statistically unrelated pieces of information
  - One contrast conveys no information about the other
  - Estimates $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are uncorrelated
  - Hypothesis tests for γ_1 and γ_2 are independent (i.e., results of one test do not affect results of the other)
  - Confidence intervals for γ_1 and γ_2 are independent

Why Are Orthogonal Contrasts Useful?

- The F-test from the ANOVA table:
  - Tests whether all groups have the same mean
  - We do not always care about the omnibus F-test
- Contrasts:
  - Focus attention on specific scientific questions
  - Require the researcher to explicitly specify those questions
- Orthogonality implies:
  - Independence of test results
  - Tests for contrasts can be interpreted individually
  - A natural partitioning of sums of squares into *"interesting"* components and *"everything else"*

Multiple Comparisons

Pairwise Comparisons

Each pairwise comparison has Type I error level α, or confidence level $100(1-\alpha)\%$.

When there are r groups, we perform

$$\binom{r}{2}$$

pairwise comparisons.

If r is large, some significant differences are expected by chance even if all of the population means are the same.

Comparison-wise Type I Error Rate

The comparison-wise Type I error rate is defined as

$$P\big(\text{reject } H_0 \text{ for one test} \mid H_0 \text{ is true for that test}\big).$$

Experiment-wise Type I Error Rate

The experiment-wise Type I error rate is defined as

$$P\big(\text{reject at least one of the } H_0\text{'s} \mid \text{all } H_0\text{'s are true}\big).$$

Multiple Comparisons Adjustment

Multiple comparisons adjustments are used to avoid too many false significant findings. The goal is to make the experiment-wise Type I error rate reasonably small.

These adjustments are equivalent to constructing simultaneous confidence intervals; that is, all confidence intervals in a set include their individual targets with a specified probability.

Basic Approach

Adjust the critical value $t_{N-r,1-\alpha/2}$ used in individual $100(1-\alpha)\%$ confidence intervals or individual α-level t-tests.

The cost of this approach is lower power, meaning it is less likely to detect a non-zero effect.

The benefit is that the experiment-wise Type I error rate is no larger than the specified α.

Least Significant Difference (LSD)

(Note: This is a Comparison-wise adjustment)

First conduct the overall F-test of

$$H_0: \ \mu_1 = \mu_2 = \cdots = \mu_r$$

at the α level.

If H_0 is not rejected, then declare all means the same. In this case, the chance of any false declaration of a significant difference is less than α.

If H_0 is rejected, then calculate confidence intervals or conduct hypothesis tests for pairwise comparisons.

This method is commonly used, but there can be substantial loss of power when only a few groups have different means.

(Note: What follows are examples of Experiment-wise adjustments)

Scheffé's Method

Scheffé's method works for any number of linear contrasts, including all possible linear contrasts.

It is the most conservative multiple comparison procedure, but it is relatively easy to apply.

In place of $t_{N-r,1-\alpha/2}$, use

$$\sqrt{(r-1)F_{r-1,N-r,1-\alpha}}.$$

Declare a significant difference between groups i and j if

$$\left|\bar{Y}_i - \bar{Y}_j\right| \geq \sqrt{(r-1)F_{r-1,N-r,1-\alpha}}\sqrt{MS_{\text{error}}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}.$$

Tukey–Kramer Honest Significant Difference (HSD)

The Tukey–Kramer procedure is based on the distribution of the studentized range.

The studentized range statistic is

$$q_{(r,N-r)} = \frac{\max_i \bar{Y}_i - \min_i \bar{Y}_i}{S_p/\sqrt{n}}.$$

For confidence intervals, use the critical value

$$\frac{1}{\sqrt{2}}q_{(r,N-r,1-\alpha)}.$$

For hypothesis tests, declare a significant difference if

$$\left|\bar{Y}_i - \bar{Y}_j\right| \geq \frac{1}{\sqrt{2}} q_{(r,N-r,1-\alpha)} \sqrt{MS_{\text{error}}\left(\frac{1}{n} + \frac{1}{n}\right)}.$$

Bonferroni Method

If we conduct m tests (or confidence intervals), replace α with α/m for each test (or confidence interval).

This method is easy to implement.

Declare a significant difference if

$$\left|\bar{Y}_i - \bar{Y}_j\right| \geq t_{N-r,1-\alpha/(2m)} \sqrt{MS_{\text{error}}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}.$$

The Bonferroni method is conservative, especially when m is large and the tests are not independent, resulting in an experiment-wise Type I error rate less than α.

The number of comparisons m must be specified in advance.

False Discovery Rate (FDR)

- FDR (or pFDR = positive FDR) is an alternative error rate that can be useful for RNA-seq experiments or other genomic studies.
- **Table of Outcomes for m Tests**

Hypothesis	Accept Null	Reject Null	Total
Null true	U	V	m_0
Alternative true	T	S	m_1
Total	W	R	m

- FDR (Benjamini and Hochberg, 1995)

$$\text{FDR} \;=\; E\left(\frac{V}{R}\,\middle|\, R > 0\right)\Pr(R > 0).$$

Conceptually

- Suppose a scientist conducts 100 independent RNA-seq experiments.
- For each experiment, the scientist produces a list of genes declared to be differentially expressed by testing a null hypothesis for each gene.
- For each list consider the ratio of the number of false positive results to the total number of genes on the list (set this ratio to 0 if the list contains no genes).
- The FDR is approximated by the average of the ratios described above.

Blocking

Variation within Groups: **Problem**

- When σ^2 is large compared to differences between means
  - Fail to reject H_0 of equal means even when differences between means exist.
- Why would σ^2 be large?
  - Response variable has large amount of variation.
  - Experimental units are not homogeneous with respect to response variable.

Variation within Groups: **Solution**

- Choose more homogeneous experimental units.
  - Reduces variation in response variable — more likely to produce significant result.
  - Reduces generalizability of experimental results.
- Use more heterogeneous experimental units.
  - Increases variation in response variable — less likely to produce significant result.
  - Increases generalizability of experimental results.

Block

A group of experimental units that, prior to treatment, are expected to be more like one another (with respect to response variables) than experimental units in general.

In simple words, blocks are groups of similar experimental units.

Types of Blocking

Sorting

- You are interested in the effect of two different instructional methods on achievement in mathematics of 8th graders.
- Sort students by their Iowa Test math scores from 7th grade.
- Students within each block will have similar Iowa Test math scores.

Subdividing

- You are interested in the yield of three varieties of soy beans.
- You have 12 fields across Iowa that you can use.
- Divide each field into 3 sections and plant one variety on each section.

Reusing

- You are interested in determining which of two brands of golf balls travels the furthest when hit with a five iron.
- Have each person hit both types of golf ball (reuse each person).

Matching

- You are interested in determining which of two brands of golf balls travels the furthest when hit with a five iron.
- Pair two golfers with the same skill level.
- Have one person hit one brand of golf ball and the other person hit the other brand of golf ball.

Matched Pairs

Experiments with Two Treatments

- Experiments with two treatments
- Blocks have one or two experimental units
  - **One unit (reuse)**
    - Receives both treatments
    - Order of treatments is random
  - **Two units (match)**
    - Two treatments randomly assigned to pair
    - One unit receives one treatment
    - Other unit receives other treatment

Hypothesis Test

- $H_0 : \mu_d = 0 \quad \text{vs.} \quad H_a : \mu_d \neq 0$
- **Test Statistic:**

$$t \ = \ \frac{\bar{D}}{s_d/\sqrt{n}}$$

- **p-value:**

$$2 \times P(t_{n-1} > |t|)$$

Confidence Interval

$100(1-\alpha)\%$ Confidence Interval for μ_d:

$$\bar{D} \ \pm \ t_{n-1,\,1-\alpha/2}\,\frac{s_d}{\sqrt{n}}$$

Diagnosing Assumptions **Model assumptions are:**

- Blocks are independent $\Rightarrow$ differences are independent
- D_i are i.i.d. $N(\mu_d = \mu_1 - \mu_2,\ \sigma_d^2)$

where

$$\sigma_d^2 \ = \ \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$$

Independence of Differences

- Examine study to determine if responses from one block could affect responses from any other block.
- Critical problem if this fails.
- Observations from the same block will usually be positively correlated.

Normal Distribution for Differences

- Normal probability plot for differences
- Effects of non-normality:
  - t-test sensitive to outliers
  - t-test sensitive to skewness of the distribution of possible differences
  - If sample size is large, t-test is fairly robust to these problems

If Smaller Sample Sizes with Non-Normal Differences

- Wilcoxon signed rank test
- Sign test

RCBD

Block

A group of experimental units that, prior to treatment, are expected to be more like one another (with respect to one or more response variables) than experimental units in general.

(In simple words, groups of similar experimental units.)

Randomized Complete Block Design (RCBD)

Experimental design in which separate and completely randomized treatment assignments are made for each of multiple blocks in such a way that all treatments have at least one experimental unit in each block.

Typical RCBD Set-up

- J treatments
- n blocks with J units in each block
  - Units within each block are similar
  - Within each block, randomly assign J treatments to the units so that one experimental unit receives each treatment
  - Each block is essentially a repetition of the experiment

Model (experiments with one unit per treatment per block)

$$Y_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij}$$

- $i = 1, \ldots, n$ indexes blocks
- $j = 1, \ldots, J$ indexes treatments
- τ_j are fixed treatment effects (with $\sum_{j=1}^{J} \tau_j = 0$)
- β_i are block effects
  - Could be fixed effects with $\sum_{i=1}^{n} \beta_i = 0$
  - Could be random effects with $\beta_i \sim N(0, \sigma_\beta^2)$
- Additive model (same treatment effects in each block)
- $\varepsilon_{ij} \sim N(0, \sigma_e^2)$

ANOVA Table

Source of variation	Degrees of freedom	Sums of squares
Blocks	$n-1$	$J \sum_{i=1}^{n} (\bar{Y}_{i.} - \bar{Y}_{..})^2$
Treatments	$J-1$	$n \sum_{j=1}^{J} (\bar{Y}_{.j} - \bar{Y}_{..})^2$
Error	$(n-1)(J-1)$	$\sum_{i=1}^{n} \sum_{j=1}^{J} (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2$
Total	$nJ-1$	$\sum_{i=1}^{n} \sum_{j=1}^{J} (Y_{ij} - \bar{Y}_{..})^2$

Expectations for Mean Squares

- Residual (error) Mean Square:

$$E(MS_{\text{error}}) = \sigma_e^2$$

- Fixed Treatment Effects (with $\bar{\tau} = \sum_{j=1}^{J} \tau_j$):

$$E(MS_{\text{treatments}}) = \sigma_e^2 + \frac{n}{J-1} \sum_{j=1}^{J} (\tau_j - \bar{\tau})^2$$

- Fixed Blocks (with $\bar{\beta} = \sum_{i=1}^{n} \beta_i$):

$$E(MS_{\text{blocks}}) = \sigma_e^2 + \frac{J}{n-1} \sum_{i=1}^{n} (\beta_i - \bar{\beta})^2$$

- Random Blocks:

$$E(MS_{\text{blocks}}) = \sigma_e^2 + J\sigma_\beta^2$$

Tests for Treatment Effects

- Test the null hypothesis of no treatment effects:

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_J$$

against the alternative that at least one mean is different.

- Reject H_0 if

$$F = \frac{MS_{\text{treatments}}}{MS_{\text{error}}} \;\geq\; F_{(J-1,\,(n-1)(J-1)),\,1-\alpha}.$$

Efficiency and Diagnostics

Efficiency

Is RCBD Better than CRD?

- If the experiment was repeated on similar experimental units (e.u.'s), should you block?
- Not a question about how to analyze the observed data.
  Analysis should match the design.

How to Measure "Better"?

- Consider the **error variance** for each design:

$$\sigma^2_{\text{CRD}} \quad \text{versus} \quad \sigma^2_{\text{RCBD}}$$

- **Efficiency of RCBD relative to CRD** is

$$\text{Efficiency} \;=\; \frac{\sigma^2_{\text{CRD}}}{\sigma^2_{\text{RCBD}}}.$$

- Efficiency $> 1 \;\Rightarrow\;$ RCBD provides more precise estimates of treatment mean contrasts.

Efficiency in Terms of Sample Sizes

- The variance of a difference in treatment means is

$$\text{Var}\big(\bar{Y}_{.j} - \bar{Y}_{.k}\big) \;=\; \sigma_e^2 \left(\frac{2}{n}\right).$$

- To have $\mathrm{Var}(\bar{Y}_{.j} - \bar{Y}_{.k})$ the same for both designs, we need

$$\sigma^2_{\text{CRD}} \left(\frac{2}{n_{\text{CRD}}} \right) = \sigma^2_{\text{RCBD}} \left(\frac{2}{n_{\text{RCBD}}} \right).$$

- Therefore,

$$\text{Efficiency} = \frac{\sigma^2_{\text{CRD}}}{\sigma^2_{\text{RCBD}}} = \frac{n_{\text{CRD}}}{n_{\text{RCBD}}}.$$

- For example, **Efficiency = 1.5** implies that the CRD requires **50% more units per treatment** than the RCBD.

Fisher's Adjustment for Degrees of Freedom

- Fisher used the *relative amount of information*, an adjusted efficiency:

$$\frac{(df_{\text{RCBD}} + 1)(df_{\text{CRD}} + 3)\, \hat{\sigma}^2_{\text{CRD}}}{(df_{\text{RCBD}} + 3)(df_{\text{CRD}} + 1)\, \hat{\sigma}^2_{\text{RCBD}}},$$

to account for differing degrees of freedom.

Practical Notes

- Typical values of efficiency depend on the subject matter.
- Values between **1.10 and 1.30** are common (i.e., blocking often reduces the number of units needed by **10–30%**).

Diagnostics

Assumptions *(treatments used equally often in each block, i.e., balanced)*

- Independence of errors
- Homogeneous error variance
- Normality of errors
- Block and treatment effects are additive (no interaction)

Diagnose Assumption: **Additive Model**

- **Additivity**: treatment effect is the same within each block.

Additive Model:

$$Y_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij}$$

- **Non-additivity**: treatment effect varies depending on block.

Non-Additive Model:

$$Y_{ij} = \mu + \beta_i + \tau_j + (\beta\tau)_{ij} + \varepsilon_{ij}$$

- Unless there are replicates of treatments within blocks, we cannot test for significance of the interaction $(\beta\tau)_{ij}$.

Diagnose Assumption: **Tukey's Test for Non-Additivity**

- Used when there are no replicates of treatments within blocks.
- Detects one specific pattern of non-additivity: **multiplicative interaction** between block and treatment effects.

Tukey Model:

$$Y_{ij} = \mu + \beta_i + \tau_j + \kappa\,\beta_i\tau_j + \varepsilon_{ij}$$

- Tukey constructed an F-test for

$$H_0: \ \kappa = 0 \quad \text{vs.} \quad H_a: \ \kappa \neq 0.$$

Latin Squares

Latin Squares Design

- Two blocking variables
- Number of levels for each blocking factor = number of treatments (or its multiple)
  - 3 treatments: each block has three levels (or 6, 9, 12, etc.)
  - 4 treatments: each block has four levels (or 8, 12, 16, etc.)
- Each block contains only one unit for each treatment
- Each level of each blocking variable gets all treatments

Advantages

- Can estimate treatment effects in a small study
- Can use two blocking factors to reduce variability

Limitations

- Levels of each blocking variable must equal (or be a multiple of) the number of treatments
- Analysis assumes no interactions between blocking factors and treatments
  - Critical, because each block contains only one unit for each treatment
- Few degrees of freedom for error
  - Can increase by using multiple Latin squares

Model

$$Y_{ijk} = \mu + \beta_i + \gamma_j + \tau_k + \varepsilon_{ijk}$$

where $i, j, k = 1, 2, \ldots, r$,

- β_i: first blocking factor effect
- γ_j: second blocking factor effect
- τ_k: fixed treatment effect
- k is the treatment and is determined by (i, j)
- $\varepsilon_{ijk} \sim N(0, \sigma^2)$

ANOVA

Source	d.f.	SS
Block 1	$r-1$	$r\sum_i(\bar{Y}_{i..} - \bar{Y}_{...})^2$
Block 2	$r-1$	$r\sum_j(\bar{Y}_{.j.} - \bar{Y}_{...})^2$
Treatment	$r-1$	$r\sum_k(\bar{Y}_{..k} - \bar{Y}_{...})^2$
Error	$(r-1)(r-2)$	SS_{error}
Total	r^2-1	$\sum_i\sum_j(Y_{ij.} - \bar{Y}_{...})^2$

Multi-factor Designs

Factor & Levels

- A **factor** is an explanatory variable studied in an investigation.
- The different values of a factor are called **levels**.
- Often correspond to treatments in an experiment.

Factorial Experimental Design

- **Factorial designs** use combinations of levels of two or more factors as treatments.

Example

- Factor A: 3 levels (a_1, a_2, a_3)
- Factor B: 2 levels (b_1, b_2)
- Combinations of A and B ⇒ 6 treatments:

$$(a_1b_1,\ a_1b_2,\ a_2b_1,\ a_2b_2,\ a_3b_1,\ a_3b_2)$$

Terminology

- **Complete (full) factorial**: all possible combinations of factor levels are used.
- **Fractional factorial**: only a subset of the possible combinations is used.

Notation: Cell Means Model

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \qquad \varepsilon_{ijk} \text{ are i.i.d. } N(0, \sigma^2)$$

- μ_{ij} = mean response to level i of factor A and level j of factor B
- $\bar{\mu}_{i\cdot} = \frac{1}{b}\sum_j \mu_{ij}$
  = mean response of factor A at level i, averaging across the levels of factor B
- $\bar{\mu}_{\cdot j} = \frac{1}{a}\sum_i \mu_{ij}$
  = mean response of factor B at level j, averaging across the levels of factor A
- $\bar{\mu}_{\cdot\cdot} = \frac{1}{ab}\sum_i \sum_j \mu_{ij}$
  = overall mean response, averaging across the levels of both factors
- σ^2 = variance of responses in level i of factor A and level j of factor B

Research Questions

- Are the 6 response means (μ_{ij}) the same?
- Are mean responses to copper levels the same, averaging over zinc levels?

$$\bar{\mu}_{1\cdot} = \bar{\mu}_{2\cdot} \ ?$$

- Are mean responses to zinc levels the same, averaging over copper levels?

$$\bar{\mu}_{\cdot 1} = \bar{\mu}_{\cdot 2} = \bar{\mu}_{\cdot 3} \ ?$$

- Are differences in mean responses between copper levels the same across zinc levels?

$$(\mu_{11} - \mu_{21}) = (\mu_{12} - \mu_{22}) = (\mu_{13} - \mu_{23}) \ ?$$

Factors and Levels

Factor & Levels

- A **factor** is an explanatory variable studied in an investigation.

- The different values of a factor are called **levels**.

- Often correspond to **treatments** in an experiment.

Factorial Experimental Design

- **Factorial designs** use combinations of levels of two or more factors as treatments.

Example

- Factor A: 3 levels (a_1, a_2, a_3)

- Factor B: 2 levels (b_1, b_2)

Combinations of A and B $\Rightarrow$ 6 treatments:

$$(a_1b_1,\ a_1b_2,\ a_2b_1,\ a_2b_2,\ a_3b_1,\ a_3b_2)$$

Terminology

- **Complete (full) factorial**: all possible combinations of factor levels are used.
- **Fractional factorial**: only a subset of combinations is used.

Notation: Cell Means Model

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \qquad \varepsilon_{ijk} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

- μ_{ij}: mean response at level i of factor A and level j of factor B

Marginal means:

$$\bar{\mu}_{i\cdot} = \frac{1}{b}\sum_j \mu_{ij}$$

$$\bar{\mu}_{\cdot j} = \frac{1}{a}\sum_i \mu_{ij}$$

$$\bar{\mu}_{\cdot\cdot} = \frac{1}{ab}\sum_i\sum_j \mu_{ij}$$

- σ^2: variance of responses within cell (i, j)

Research Questions

- Are the ab response means μ_{ij} the same?
- Are mean responses to factor A the same, averaging over factor B?

$$\bar{\mu}_{1\cdot} = \bar{\mu}_{2\cdot} = \cdots$$

- Are mean responses to factor B the same, averaging over factor A?

$$\bar{\mu}_{\cdot 1} = \bar{\mu}_{\cdot 2} = \cdots$$

- Are differences between levels of factor A the same across levels of factor B?

$$(\mu_{11} - \mu_{21}) = (\mu_{12} - \mu_{22}) = (\mu_{13} - \mu_{23})\ ?$$

Factor Effects

- **Main effect**: difference (contrast) between levels of one factor averaged over all levels of the other factor(s).
- **Simple effect**: difference (contrast) between levels of one factor at a specific level of the other factor.
- **Interaction** exists when simple effects are not the same:

  - Equivalent to non-parallel lines in a plot of means.
  - Can differ in magnitude or direction.

Two-Way ANOVA Table

Source	df	Sum of Squares
Factor A	$a-1$	$nb\sum_i(\bar{Y}_{i\cdot\cdot}-\bar{Y}_{\cdot\cdot\cdot})^2$
Factor B	$b-1$	$na\sum_j(\bar{Y}_{\cdot j\cdot}-\bar{Y}_{\cdot\cdot\cdot})^2$
Interaction AB	$(a-1)(b-1)$	$n\sum_i\sum_j(\bar{Y}_{ij\cdot}-\bar{Y}_{i\cdot\cdot}-\bar{Y}_{\cdot j\cdot}+\bar{Y}_{\cdot\cdot\cdot})^2$
Error	$ab(n-1)$	$\sum_i\sum_j\sum_k(Y_{ijk}-\bar{Y}_{ij\cdot})^2$
Total	$abn-1$	$\sum_i\sum_j\sum_k(Y_{ijk}-\bar{Y}_{\cdot\cdot\cdot})^2$

Expected Mean Squares

$$E(MS_{\text{error}}) = \sigma^2$$

$$E(MS_A) = \sigma^2 + \frac{nb}{a-1}\sum_i(\bar{\mu}_{i\cdot}-\bar{\mu}_{\cdot\cdot})^2$$

$$E(MS_B) = \sigma^2 + \frac{na}{b-1}\sum_j(\bar{\mu}_{\cdot j}-\bar{\mu}_{\cdot\cdot})^2$$

$$E(MS_{AB}) = \sigma^2 + \frac{n}{(a-1)(b-1)}\sum_i\sum_j(\mu_{ij}-\bar{\mu}_{i\cdot}-\bar{\mu}_{\cdot j}+\bar{\mu}_{\cdot\cdot})^2$$

F-tests

Overall Treatment Effects

$$H_0:\ \mu_{ij} \text{ equal for all } i,j$$

$$F = \frac{MS_{\text{model}}}{MS_{\text{error}}}$$

Factor A Main Effect

$$H_0:\ \bar{\mu}_{1\cdot} = \bar{\mu}_{2\cdot} = \cdots = \bar{\mu}_{a\cdot}$$

$$F = \frac{MS_A}{MS_{\text{error}}}$$

Factor B Main Effect

$$H_0:\ \bar{\mu}_{\cdot 1} = \bar{\mu}_{\cdot 2} = \cdots = \bar{\mu}_{\cdot b}$$

$$F = \frac{MS_B}{MS_{\text{error}}}$$

Interaction Effect

$$H_0:\ (\mu_{ij} - \mu_{kj}) = (\mu_{ir} - \mu_{kr}) \quad \forall\, i \neq k,\ j \neq r$$

$$F = \frac{MS_{AB}}{MS_{\text{error}}}$$

Interpretation Considerations

No Interaction Effect

- Marginal means are straightforward to interpret.
- Main effects summarize average differences across the other factor.

Interaction Effect Present

- Main effects may be misleading.
- Simple effects are conditional on the other factor.
- Consider:
  - Whether effects are additive on another scale
  - Practical significance of the interaction
  - Reporting simple effects rather than marginal means

Two-way ANOVA

Effects

- Estimate $a \times b$ treatment means with
  - μ
  - a effects for Factor A
  - b effects for Factor B
  - $a \times b$ interaction effects for Factors A and B
- Impose constraints on main effects and interaction effects to reduce number of parameters to $a \times b$

Baseline Constraints Effects Model

$$Y_{ijk} = \mu + \alpha_i + \tau_j + (\alpha\tau)_{ij} + \varepsilon_{ijk}$$

- Set $\alpha_a = 0$ (level a of Factor A)
- Set $\tau_b = 0$ (level b of Factor B)
- Set $(\alpha\tau)_{aj} = 0$ for all $j = 1, \ldots, b$
  (All interaction effects with level a of Factor A)
- Set $(\alpha\tau)_{ib} = 0$ for all $i = 1, \ldots, a$
  (All interaction effects with level b of Factor B)

Interpretation (Baseline Constraints)

- μ is the mean of the baseline cell $(i = a, j = b)$.
- α_i is the difference between level i of Factor A and the baseline level a, evaluated at $j = b$.
- τ_j is the difference between level j of Factor B and the baseline level b, evaluated at $i = a$.
- $(\alpha\tau)_{ij}$ is the additional deviation for cell (i, j) beyond what is explained by the main effects, relative to the baseline cell.

Sum-to-Zero Constraints Effects Model

$$Y_{ijk} = \mu + \alpha_i + \tau_j + (\alpha\tau)_{ij} + \varepsilon_{ijk}$$

- Set $\sum_{i=1}^{a} \alpha_i = 0$
- Set $\sum_{j=1}^{b} \tau_j = 0$
- Set $\sum_{i=1}^{a} (\alpha\tau)_{ij} = 0$ for all j
- Set $\sum_{j=1}^{b} (\alpha\tau)_{ij} = 0$ for all i

Interpretation (Sum-to-Zero Constraints)

- μ represents the grand mean across all $a \times b$ treatment combinations.
- α_i represents the deviation of level i of Factor A from the grand mean, averaged over levels of Factor B.
- τ_j represents the deviation of level j of Factor B from the grand mean, averaged over levels of Factor A.
- $(\alpha\tau)_{ij}$ represents the remaining cell-specific deviation after accounting for both main effects, subject to averaging to zero across rows and columns.

Diagnostics

Model Assumptions

$$\varepsilon_{ijk} \text{ are i.i.d. } N(0, \sigma^2)$$

- Independence
- Homogeneous (Equal) Variance
- Normality

Model Diagnostics

- Independence
  - Check details of data collection
- Homogeneous variance
  - Plot residuals vs. estimated means
  - Plot residuals vs. levels of each factor
- Normality
  - Normal probability plot for residuals
  - Histogram of residuals
  - Tests for normality

Additional Factorial Designs

2 Factors Plus Block Setup

- RCBD with full factorial treatment design ($r = ab$ treatments)
- Different random assignment of units to treatments in each of the n blocks

- One experimental unit for each block–treatment combination
- Assume no interaction between block and treatment effects
- Model:

$$Y_{ijk} = \mu + \beta_i + \alpha_j + \tau_k + (\alpha\tau)_{jk} + \varepsilon_{ijk}$$

ANOVA Table

Variation	d.f.	Sums of Squares
Block	$n-1$	$ab\sum_{i=1}^{n}(\bar{Y}_{i..} - \bar{Y}_{...})^2$
Factor A	$a-1$	$nb\sum_{j=1}^{a}(\bar{Y}_{.j.} - \bar{Y}_{...})^2$
Factor B	$b-1$	$na\sum_{k=1}^{b}(\bar{Y}_{..k} - \bar{Y}_{...})^2$
$A \times B$ interaction	$(a-1)(b-1)$	$n\sum_j\sum_k(\bar{Y}_{.jk} - \bar{Y}_{.j.} - \bar{Y}_{..k} + \bar{Y}_{...})^2$
Error	$(ab-1)(n-1)$	SS_{error}
Corrected total	$abn-1$	$\sum_i\sum_j\sum_k(Y_{ijk} - \bar{Y}_{...})^2$

Three Factors Notation

- Factor A with a levels: $i = 1, \ldots, a$
- Factor B with b levels: $j = 1, \ldots, b$
- Factor C with c levels: $k = 1, \ldots, c$
- n replications for each treatment: $l = 1, \ldots, n$

$$Y_{ijkl} = \mu + \alpha_i + \tau_j + \delta_k + (\alpha\tau)_{ij} + (\alpha\delta)_{ik} + (\tau\delta)_{jk} + (\alpha\tau\delta)_{ijk} + \varepsilon_{ijkl}$$

ANOVA Table

Source of variation	Degrees of freedom	Sums of squares
Factor A	$a-1$	$nbc\sum_i(\bar{Y}_{i...} - \bar{Y}_{....})^2$
Factor B	$b-1$	$nac\sum_j(\bar{Y}_{.j..} - \bar{Y}_{....})^2$
Factor C	$c-1$	$nab\sum_k(\bar{Y}_{..k.} - \bar{Y}_{....})^2$
Interaction AB	$(a-1)(b-1)$	$nc\sum_i\sum_j(\bar{Y}_{ij..} - \bar{Y}_{i...} - \bar{Y}_{.j..} + \bar{Y}_{....})^2$
Interaction AC	$(a-1)(c-1)$	$nb\sum_i\sum_k(\bar{Y}_{i.k.} - \bar{Y}_{i...} - \bar{Y}_{..k.} + \bar{Y}_{....})^2$
Interaction BC	$(b-1)(c-1)$	$na\sum_j\sum_k(\bar{Y}_{.jk.} - \bar{Y}_{.j..} - \bar{Y}_{..k.} + \bar{Y}_{....})^2$
Interaction ABC	$(a-1)(b-1)(c-1)$	SS_{ABC}
Error	$abc(n-1)$	$\sum_i\sum_j\sum_k\sum_l(Y_{ijkl} - \bar{Y}_{ijk.})^2$
Total	$abcn-1$	$\sum_i\sum_j\sum_k\sum_l(Y_{ijkl} - \bar{Y}_{....})^2$

2^K **Studies Set-up**

- K factors with $j = 1, 2, \ldots, r_k$ levels for the kth factor
- Known as an $r_1 \times r_2 \times \cdots \times r_K$ factorial design
- There are $r_1 \times r_2 \times \cdots \times r_K$ experimental units, and exactly one unit is assigned to each treatment

- If all possible interactions are included in the model, there are no degrees of freedom left for computing MS_{error}
- We will only consider the special situation of K factors with exactly two levels for each factor — 2^K factorial designs

Special Features

- All main effect and interaction contrasts have 1 d.f.
- Set up the model with
  - One column in the model matrix X for each main effect using $+1/-1$ coding
  - Interaction columns obtained by multiplication of appropriate main effect columns
  - All columns of X are orthogonal

Unit 3: Simple Linear Regression

Introduction

Research Question

- Study the relationship of two or more quantitative variables
  - quantitative: numbers, usually continuous

  - qualitative: classes, identify groups

- Is there a significant linear relationship between the response variable and the explanatory variable?

- What mean value of response would we predict for a given value of the explanatory variable?

- What value of response would we predict for a given value of the explanatory variable?

SLR Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad i = 1, \dots, n$$

- $i = 1, \dots, n$ is the number of observations
- Y_i is the response or dependent variable
- X_i is the predictor, explanatory variable, or independent variable, treated as known and fixed
- ε_i is the random error term representing individual variation and measurement error

Model Assumptions

- x's are fixed (or conditioned upon)
- The expected response is a linear function of the explanatory variable:

$$E(Y_i \mid X_i = x_i) = \beta_0 + \beta_1 x_i$$

- additive random errors:

$$Y_i = E(Y_i \mid X_i = x_i) + \varepsilon_i$$

- independent (uncorrelated) random errors
- homogeneous error variance:

$$\text{Var}(\varepsilon_i) = \sigma^2$$

- normally distributed random errors:

$$\varepsilon_i \sim N(0, \sigma^2)$$

Least Squares Estimates

$$b_0 = \bar{Y} - b_1\bar{X}$$

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{X})^2} = \frac{\sum_{i=1}^n (x_i - \bar{X})Y_i}{\sum_{i=1}^n (x_i - \bar{X})^2}$$

Multivariate Normal Distribution

Suppose

$$Z = \begin{bmatrix} Z_1 \\ \vdots \\ Z_m \end{bmatrix}$$

is a random vector whose elements are independently distributed standard normal random variables.

For any $n \times m$ matrix A, we say that

$$Y = \mu + AZ$$

has a multivariate normal distribution with mean vector

$$E(Y) = E(\mu + AZ) = \mu + AE(Z) = \mu + A0 = \mu$$

and variance–covariance matrix

$$\text{Var}(Y) = A[\text{Var}(Z)]A^T = AA^T \equiv \Sigma.$$

Multivariate Normal Linear Combinations

If $Y \sim N(\mu, \Sigma)$, then

$$W = c + BY \sim N(c + B\mu, B\Sigma B^T)$$

for any non-random c and B.

Inference

Regression Analysis: ANOVA

- Write the deviation from the overall sample mean as

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}) \quad \text{where} \quad \hat{Y}_i = b_0 + b_1 X_i$$

- Partition the corrected total sums of squares

$$SS_{\text{total}} = \sum_i (Y_i - \bar{Y})^2 = \sum_i (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2$$

$$= \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{Y})^2 + 2\sum_i (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$$

$$= \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{Y})^2$$

$$= SS_{\text{residuals}} + SS_{\text{model}}$$

ANOVA Table

Source	df	Sums of Squares
Model	1	$SS_{\text{model}} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$
Error	$n-2$	$SS_{\text{error}} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
Total	$n-1$	$SS_{\text{total}} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$

F-test for Significance of Model

- $H_0 : \beta_1 = 0 \;\rightarrow\; Y_i = \beta_0 + \varepsilon_i$
- $H_a : \beta_1 \neq 0 \;\rightarrow\; Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- Test Statistic:
$$F = \frac{MS_{\text{model}}}{MS_{\text{error}}}$$
- Reject H_0 if
$$F = \frac{MS_{\text{model}}}{MS_{\text{error}}} > F_{1,n-2,1-\alpha}$$

Coefficient of Determination (R^2)

$$R^2 = \frac{SS_{\text{model}}}{SS_{\text{total}}}$$

- Fraction of variation in the response variable that can be explained by the linear regression model with the explanatory variable x.
- Expressed as percentage: $0\% \leq R^2 \leq 100\%$
- Large values of R^2 indicate better model fit.

Hypothesis Test for β_1

- Null and Alternative Hypotheses
$$H_0 : \beta_1 = 0 \qquad H_a : \beta_1 \neq 0$$
- Test Statistic
$$T = \frac{b_1 - 0}{S_{b_1}}$$
- Reject H_0 if
$$|T| > t_{n-2,1-\alpha/2}$$
- Note that $T^2 = F$, this t-test for β_1 is the same as the F-test for significance of model from ANOVA Table.
- One-sided alternative hypothesis is possible for the t-test:
$$H_a : \beta_1 > 0 \quad \text{or} \quad H_a : \beta_1 < 0$$

Confidence Interval for β_1

- $100(1-\alpha)\%$ confidence interval for β_1:
$$b_1 \pm t_{n-2,1-\alpha/2} S_{b_1}$$

Prediction

Predict the value for Y at given x:

$$Y_{\text{new}} = \beta_0 + \beta_1 x + \varepsilon$$

- Estimate is still $\hat{Y} = b_0 + b_1 x$
- Standard error is

$$S_{\text{pred}} = \sqrt{MS_{\text{error}} \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)}$$

- $100(1 - \alpha)\%$ prediction interval:

$$(b_0 + b_1 x) \pm t_{n-2,1-\alpha/2}\, S_{\text{pred}}$$

Model Diagnostics

SLR Model and Assumptions

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{where} \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

- Values of Y_i are independent (independent random errors)
- Values of x_i are fixed
- $\mu_{Y|x_i}$ is a linear function of x_i
- Homogeneous error variance: $\text{Var}(\varepsilon_i) = \sigma^2$
- Normally distributed errors: $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

Interpretation

- The regression line describes the *conditional mean* of Y given x.
- All randomness is attributed to the error term ε_i, not to x_i.
- These assumptions justify least squares estimation, standard errors, and inference.

Independence

- Check independence of observations through details of data collection
- Beware of:
  - Observations over time
  - Clustering of observations
  - Spatial elements to observations
- Crucial assumption — must use other methods if violated

Interpretation

- Lack of independence leads to underestimated standard errors and invalid tests.
- Time series, spatial data, or grouped data often violate this assumption.
- Remedies include correlation structures, random effects, or generalized least squares.

Fixed Values of x

- Assume x is measured without error
- Check through variable definition and through details of data collection
- If violated, model the error in x using a random effect

Interpretation

- Treating x as fixed simplifies inference and variance calculations.
- Measurement error in x typically biases slope estimates toward zero.
- Errors-in-variables models are needed when this assumption fails.

Linearity

- Scatterplot: Plot of Y_i versus x_i → linear pattern
- Residual plot: Plot of residuals e_i versus x_i → no pattern
- Violations of linearity:
  - Transform Y_i values so that relationship with x_i is linear
  - Common transformations: log and power (Y^2, Y^3, $\sqrt{Y}$, etc.)
  - Conduct analysis with transformed Y values
  - Undo transformation in drawing conclusions

Interpretation

- Linearity concerns the *mean structure*, not the raw data cloud.
- Residual patterns indicate missing nonlinear structure.
- Transformations allow linear regression tools to be used legitimately.

Regression Analysis – Residuals

- Residuals do not have homogeneous variances

$$e_i \sim N\left(0,\ \sigma^2\left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right)\right)$$

- Sometimes use

$$r_i = \frac{e_i}{\sqrt{MSE\left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right)}}$$

(known as studentized residuals)

Interpretation

- Raw residuals have different variances depending on leverage.
- Studentized residuals standardize variability, enabling fair comparison.
- Large studentized residuals suggest potential outliers.

Residual Plots

- Plot residuals versus X
  - In simple linear regression, this is the same as previous
  - In multiple regression, it will be useful
- Plot residuals versus other possible predictors (e.g., time)
  - Detect important lurking variable
- Plot residuals vs. lagged residuals
  - Detect correlated errors
- Normal probability plot of residuals
  - Detect non-normality

Interpretation

- Residual plots diagnose unmet assumptions rather than model fit.
- Patterns imply omitted variables, dependence, or heteroskedasticity.
- Normal Q–Q plots assess whether inference based on normality is reasonable.

Diagnostics

Leverage

- Extreme values of x are called high leverage cases because they exert a large "pull" on SLR
- Measure of "potential" influence of observation on SLR
- Leverage of the ith observation is:

$$h_i = \frac{1}{n-1}\left(\frac{x_i - \bar{x}}{s_x}\right)^2 + \frac{1}{n}$$

- Properties of h_i:
  - $\frac{1}{n} \leq h_i \leq 1$
  - $\sum_{i=1}^{n} h_i = 2 \;\Rightarrow\; \bar{h} = \frac{2}{n}$

Interpretation

- High leverage points have unusual x-values, not necessarily unusual Y.
- They can strongly affect slope estimates even with small residuals.
- Leverage alone does not imply influence—residual size also matters.

Outliers

- Extreme Y_i value for a given x_i
- Three assessment methods:
  - Residuals
  - Internally studentized residuals

  - Externally studentized residuals

Interpretation

- Outliers are unusual *conditional responses*, not unusual predictors.
- Studentized residuals allow formal cutoff rules (e.g., $|r_i| > 2$).
- Influential points combine high leverage *and* large residuals.

Influence

- Concerned about unusual cases that have a big influence on both:
  - $\hat{Y}_i$ for some x_i
  - estimated slope $\hat{\beta}_1$
- Could delete the case, refit model, and examine the change

Interpretation

- Influence combines information about *outliers* (large residuals) and *leverage* (extreme x values).
- An influential case can substantially alter fitted values and slope estimates.
- Sensitivity analysis (refitting with/without the case) helps assess robustness of conclusions.

Influence (Cook's Distance)

- Cook's D — effect of deleting the ith case on the least squares regression model

$$D_i = \left(\frac{r_i^2}{2}\right)\left(\frac{h_i}{1-h_i}\right)$$

- D_i is large when r_i is large and h_i is large
- $D_i > 2\sqrt{2/n}$ indicates substantial influence

Interpretation

- Cook's distance summarizes influence in a *single diagnostic*.
- Large D_i values indicate cases that strongly affect fitted values and parameter estimates.
- Influence does not imply data error—such cases may be scientifically meaningful.
- Follow-up typically includes plotting Cook's D, checking data quality, and refitting models with and without influential observations.

Lack of Fit

Lack of Fit Test

- One method for model checking.
- Suppose we have multiple observations at one or more of the x_i values.
- Notation: Y_{ij} is the jth observation at x_i.
- Three models:
  1. $Y_{ij} = \mu + \varepsilon_i$ (common mean)

2. $Y_{ij} = \beta_0 + \beta_1 x_i + \varepsilon_i$ (regression)
3. $Y_{ij} = \mu_i + \varepsilon_i$ (separate means)

Interpretation

- Model (1) ignores x entirely.
- Model (2) imposes a linear structure on the mean response.
- Model (3) makes no functional-form assumption and serves as a flexible benchmark.
- The lack-of-fit test compares the linear model to the saturated "separate means" model.

Lack of Fit Test (Sum of Squares Decomposition)

- SSE from regression model (Model 2):

$$SS_{\text{error}} = \sum_i \sum_j (Y_{ij} - \hat{Y}_i)^2$$

$$= \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2 + \sum_i \sum_j (\bar{Y}_{i\cdot} - \hat{Y}_i)^2$$

$$= SS_{\text{pure error}} + SS_{\text{lack-of-fit}}$$

- $SS_{\text{Pure Error}}$ is the error sum of squares for Model 3.
  - Measures variability of observations about the mean response for each x.
  - Does **not** assume the regression model is correct.
- $SS_{\text{Lack-of-Fit}}$ measures lack of fit of the regression model.
- Let r = number of distinct x values.

Interpretation

- Pure error reflects natural variability at fixed x.
- Lack-of-fit captures systematic deviation from linearity.
- Replication at the same x values is essential to separate these two components.

New and Improved ANOVA Table

Source of variation	Degrees of freedom	Sums of squares
Regression	1	$SS_{\text{regression}}$
Lack-of-Fit	$r-2$	$SS_{\text{lack-of-fit}}$
Pure Error	$n-r$	$SS_{\text{pure error}}$
Total	$n-1$	SS_{total}

Interpretation

- Total variation is partitioned into explained variation, lack-of-fit, and pure error.
- The lack-of-fit row isolates model misspecification.
- When $r = n$ (no replication), lack-of-fit cannot be tested.

Example: Lack-of-Fit Test

- F-test for lack of fit

Null and alternative hypotheses:

$$H_0 : \; E(Y_{ij} \mid x_i) = \beta_0 + \beta_1 x_i$$

$$H_a : \; E(Y_{ij} \mid x_i) = \mu_i = \beta_0 + \beta_1 x_i + g(x_i)$$

- Expected mean squares:

$$E(MS_{\text{Pure Error}}) = \sigma^2$$

$$E(MS_{\text{Lack-of-Fit}}) = \sigma^2 + \frac{\sum_{i=1}^{r} n_i [g(x_i)]^2}{r-2}$$

- Reject H_0 if

$$F = \frac{MS_{\text{Lack-of-Fit}}}{MS_{\text{Pure Error}}} > F_{(df_{\text{LoF}},\, df_{\text{PE}}),\, 1-\alpha}$$

Interpretation

- The test assesses whether departures from linearity are larger than expected from random error.
- A significant result indicates the linear model is inadequate.
- Failure to reject does **not** prove linearity—only that deviations are not detectable given the data.
- Power depends strongly on replication at each x value.

Correlation

Population Correlation Coefficient

- Measure of linear relationship between two quantitative variables (X and Y) in the population
- Denoted as ρ
- Defined as

$$\rho = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - E(X))(Y - E(Y))]}{\sigma_X \sigma_Y}$$

Interpretation

- ρ measures **linear association**, not causation.
- $-1 \le \rho \le 1$:
  - $\rho > 0$: positive linear relationship
  - $\rho < 0$: negative linear relationship

  - $\rho = 0$: no linear relationship (may still be nonlinear)

- Scale-free: unaffected by changes in units of X or Y.
- Sensitive to outliers and extreme values.

Sample Correlation Coefficient

- Estimate ρ by taking a sample from the population and calculating

$$r = \frac{1}{n-1}\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{S_X S_Y}\right)$$

- r has the same properties as ρ

Interpretation

- r is a **sample-based estimate** of the population correlation ρ.
- Takes values between -1 and 1 with the same directional interpretation as ρ.
- Strongly influenced by outliers.
- Describes the *strength and direction* of linear association in the observed data.

r and R^2

- r is a function of R^2

$$r = \pm\sqrt{R^2}, \qquad r^2 = R^2$$

- r is a numerical summary of the direction and strength of the linear relationship between X and Y
- R^2 is a numerical summary of the percentage of variability in Y that can be explained by the linear regression with X

Interpretation

- The **sign of** r is determined by the sign of the slope $\hat{\beta}_1$.
- R^2 measures *explanatory power*, not strength of association alone.
- High R^2 does not imply causation or a correct model.
- In simple linear regression, R^2 and r^2 are equivalent summaries of linear fit.

Multiple Linear Regression

Introduction

Research Questions

- Does the MLR model significantly explain the response variable Y_i and how well does it explain the variation in the response variable Y_i?

- Which explanatory variables are significant in the MLR model?

- Which set of explanatory variables are significant in the MLR model?
- What value of the conditional mean of Y_i would we predict for given values of $x_{i1}, x_{i2}, \dots, x_{ik}$?
- What value of Y_i would we predict for given values of $x_{i1}, x_{i2}, \dots, x_{ik}$?

MLR Model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & x_{33} & \cdots & x_{3k} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

MLR Assumptions

- Fixed values of the explanatory variables, $x_{i1}, x_{i2}, \cdots, x_{ik}$
- Conditional mean of Y given the values of $x_{i1}, x_{i2}, \cdots, x_{ik}$ is linear:

$$\mu_Y|_{x_{i1}, x_{i2}, \cdots, x_{ik}} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

- Additive random errors:

$$Y_i = \mu_Y|_{x_{i1}, x_{i2}, \cdots, x_{ik}} + \epsilon_i$$

- Independent (uncorrelated) random errors - Homogeneous error variance:

$$\text{Var}(\epsilon_i) = \sigma^2$$

- Normally distributed random errors:

$$\epsilon_i \sim N(0, \sigma^2)$$

Interpretation and Estimation

Parameters (Coefficients)

Interpretation of parameters $\beta_0, \beta_1, \dots, \beta_k$ depends on the presence or absence of other explanatory variables in the model.

Example:

- Model 1: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \epsilon_i$
- Model 2: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$

Interpretation of parameters β_0, β_1, and β_2 are **NOT** the same in the two models.

Least Squares Estimation

Find **b**, the least squares estimator for β, that minimizes:

$$q(b) = \sum_{i=1}^{n}(Y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik})^2$$

This can be written in matrix form as:

$$q(b) = (Y - Xb)^T(Y - Xb) = e^T e$$

where $e = Y - Xb$ is the vector of residuals.

Solution:

- Solve the set of normal equations: $(X^T X)b = X^T Y$
- Solution (assuming X is of full column rank):

$$b = (X^T X)^{-1} X^T Y$$

is the unique solution to the normal equations.

Properties of Least Squares Estimators Variance of b:

$$\text{Var}(b) = \sigma^2 (X^T X)^{-1}$$

Derivation requires: - Uncorrelated errors - Homogeneous error variances - *Note:* Normality is **not** required for this derivation (normality is needed for inference procedures)

Estimating σ^2:

An unbiased estimator for σ^2 is:

$$s_e^2 = MS_{\text{error}} = \frac{(Y - Xb)^T(Y - Xb)}{n - (k+1)} = \frac{e^T e}{df_{\text{error}}} = \frac{\sum e_i^2}{df_{\text{error}}}$$

where $df_{\text{error}} = n - (k+1)$.

Variance Decomposition

Total variability in response variable

$$SS_{\text{Total}} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Total variability explained by the model

$$SS_{\text{model}} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

Total variability not explained by the model

$$SS_{\text{error}} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

ANOVA Table

Source of Variation	Degrees of Freedom	Sums of Squares
model	k	$SS_{\text{model}} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$
error	$n-(k+1)$	$SS_{\text{error}} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$
Total	$n-1$	$SS_{\textbf{total}} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$

F-test for Significance of Model Hypotheses

- $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$
- H_a : at least one $\beta_j \neq 0, \quad j = 1, \ldots, k$

Test Statistic

$$F = \frac{MS_{\text{model}}}{MS_{\text{error}}}$$

Decision Rule

Reject H_0 if $F > F_{k,\, n-(k+1),\, 1-\alpha}$

Model Comparison Interpretation

The F-test from the ANOVA Table is comparing two models:

- Model under H_0: $Y_i = \beta_0 + \epsilon_i$
- Model under H_a: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$

  Note: We almost always reject H_0 in this test (if the model explains any meaningful variation in the data).

More on Estimation

Coefficient of Determination

$$R^2 = \frac{SS_{\text{model}}}{SS_{\text{Total}}}$$

- Fraction of variation in the response variable that can be explained by the multiple linear regression model
- Expressed as percentage: $0\% \leq R^2 \leq 100\%$
- Adding explanatory variables to the model will always increase the value of R^2

Adjusted R^2

$$adj\ R^2 = 1 - \frac{MS_{\text{error}}}{SS_{\text{total}}/(n-1)}$$

- Expressed as percentage: $0\% \leq adj\ R^2 \leq 100\%$
- Adjusts for the number of explanatory variables in model through degrees of freedom of $MS_{\text{error}} = n - (k+1)$
- Used primarily for model comparisons

Hypothesis Tests For Population Coefficient

Null and Alternative Hypotheses

$H_0 : \beta_j = 0$ vs. $H_a : \beta_j \neq 0$

Test Statistic

$$T = \frac{b_j - 0}{S_e \sqrt{(X^T X)^{-1}_{[j+1,j+1]}}} = \frac{b_j - 0}{S_{b_j}}$$

Decision Rule

Reject H_0 if $|T| > t_{n-(k+1),\,1-\alpha/2}$

Confidence Interval for Population Coefficient

- $100(1-\alpha)\%$ CI for β_j is:

$$b_j \pm t_{n-(k+1),\,1-\alpha/2} \cdot S_{b_j}$$

Partial F-Test

Understanding the Difference in Sum of Squares

$$SSE_{\text{reduced}} - SSE_{\text{full}}$$

- Amount of error explained by adding the m explanatory variables to the model
- The only difference in these two models is the m explanatory variables
- Difference has m degrees of freedom
- Compare amount of error explained to MSE_{full}

Test Statistic Formula

$$F = \frac{(SSE_{\text{reduced}} - SSE_{\text{full}})/m}{MSE_{\text{full}}}$$

- Large values of F indicate group of m explanatory variables should be included in the model

Hypothesis Testing Framework

- $H_0 : \beta_j = 0$ for the m explanatory variables
- H_a : at least one $\beta_j \neq 0$ for the m explanatory variables
- Test Statistic:

$$F = \frac{(SSE_{\text{reduced}} - SSE_{\text{full}})/m}{MSE_{\text{full}}}$$

- Decision: Reject H_0 if $F > F_{m,\, n-(k+1),\, 1-\alpha}$
- **Important**: Conclusion about the significance of the m explanatory variables depends on the presence of the other $k - m$ explanatory variables in the model.

Inference for Conditional Means

Estimate the conditional mean response $\mu_{Y|X}$ under specific values for vector $x = (1, x_1, x_2, \ldots, x_k)^T$

Point Estimate

$$\hat{\mu}_{Y|X} = x^T \hat{\beta}$$

Standard Error

$$S_{\hat{\mu}_{Y|X}} = \sqrt{MS_{\text{error}}}\, x^T (X^T X)^{-1} x$$

Confidence Interval

A $(1 - \alpha) \times 100\%$ confidence interval for $\mu_{Y|X}$ is:

$$\hat{\mu}_{Y|X} \pm t_{n-(k+1),\, 1-\alpha/2}\, S_{\hat{\mu}_{Y|X}}$$

Simultaneous Confidence Region (Scheffé's Method)

For an entire line segment:

$$\hat{Y} \pm \sqrt{(k+1) F_{k+1,\, n-k-1,\, 1-\alpha}}\, S_{\hat{\mu}_{Y|X}}$$

Prediction Intervals

Predict value of $Y_i = \mathbf{x}^T \beta + \epsilon_i$ that will be observed under specific values for vector $\mathbf{x} = (1, x_1, x_2, \ldots, x_k)^T$

Predictor

$$\hat{Y}_i = \mathbf{x}^T \hat{\beta}$$

Standard Error for Predictor

$$S_{\hat{Y}} = \sqrt{MS_{\text{error}} + S^2_{\hat{\mu}_{Y|X}}}$$

Prediction Interval

A $(1-\alpha)\times 100\%$ prediction interval is:

$$\hat{Y}_i \pm t_{n-(k+1),\,1-\alpha/2}\, S_{\hat{Y}}$$

Categorical Predictors

We also have different possible parametrizations when our predictors (random variables) are categorical. An easier example where the predictor takes only 2 dinstict values, e.g., Male/Female, Yes/No, etc.

Baseline Coding of Categories

Baseline coding of categories:

$$x_{2i} = \begin{cases} 1, & \text{student is female} \\ 0, & \text{student is male} \end{cases}$$

Sum-to-Zero Coding of Categories

Sum-to-zero coding of categories:

$$x_{2i} = \begin{cases} 1, & \text{student is female} \\ -1, & \text{student is male} \end{cases}$$

Motivating Example: Sum-to-Zero with Interaction

We can consider a model with an interaction term between x_1 and x_2 — height and gender.

For Females

$$Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_{1i} + \varepsilon_i$$

For Males

$$Y_i = (\beta_0 - \beta_2) + (\beta_1 - \beta_3)x_{1i} + \varepsilon_i$$

This model allows for a different relationship between students' height and the height of their ideal romantic partner.

- **Different slopes:** $(\beta_1 + \beta_3)$ vs. $(\beta_1 - \beta_3)$
- **Different intercepts:** $(\beta_0 + \beta_2)$ vs. $(\beta_0 - \beta_2)$

Model Selection

Importance of Model Selection

- Including too few variables in the model leads to inaccurate estimates of coefficients and response means.
- Including too many variables leads to unnecessary excess variability in estimates of the coefficients and mean response.

Theory

- Adding a predictor to Model A (adding a column to X that is not a linear combination of the columns already in X):
  - decreases bias (or may leave it the same) of $\hat{Y}$
  - increases the total variance of the estimates of the response means $\hat{Y}$, because the column rank of X, which is also the rank of the new P_X, increases by 1
- If we fit the "true" model, Model B, then:
  - bias = 0
  - variance $= \sum_i \text{Var}(\hat{Y}_i) = \sigma^2(k + 1 + \dim(Z))$

Criterion: R^2

$$R^2 = \frac{SS_{\text{model}}}{SS_{\text{Total}}}$$

- Larger values indicate better model
- Maximizing R^2 is equivalent to minimizing SS_{error}
- R^2 never decreases when adding an explanatory variable to the model
- Most useful for comparing two models with the same number of explanatory variables

Criterion: adjusted R^2

$$\text{adj } R^2 = 1 - \frac{MS_{\text{error}}}{SS_{\text{Total}}/(n-1)}$$

- Larger values indicate better model
- Maximizing adjusted R^2 is equivalent to minimizing $MS_{\text{error}} = \hat{\sigma}^2$
- Does not necessarily increase when adding an explanatory variable to the model
- Most useful in comparing models with different numbers of explanatory variables

Criterion: C_p

$$C_p = \frac{SS_{\text{error}}}{\hat{\sigma}^2} - [n - 2(k+1)]$$

- SS_{error} from fitted model
- $\hat{\sigma}^2$ is MS_{error} for the model containing all explanatory variables
- $p = k + 1$ is the number of coefficients in the fitted model

Also:

- Full name: Mallow's C_p
- Good models have C_p around $p = k + 1$
  - Why?
- $C_p < p$ is no problem (sampling error)
- Large C_p indicates poor model
- Let m denote the size of the biggest possible model with $m - 1$ explanatory variables and m regression coefficients
- For the model containing all explanatory variables, $C_p = m$
- Limited to MLR models

Criterion: AIC

$$\text{AIC} = n \log\left(\frac{SS_{\text{error}}}{n}\right) + 2(k+1)$$

- Full name: Akaike Information Criterion
- Smaller values indicate better models
- Favors models with a slightly larger number of explanatory variables, i.e., may include a few non-significant explanatory variables
- Not limited to MLR models

Criterion: BIC

$$\text{BIC} = n \log\left(\frac{SS_{\text{error}}}{n}\right) + (k+1)\log(n)$$

- Full name: Bayesian Information Criterion
- Smaller values indicate better models
- Leads to smaller models than AIC (larger penalty for explanatory variables)
- Not limited to MLR models

Selection Techniques

- Many different approaches
- Measures that focus on fit:
  - R^2 (fit using SS_{error}): bad
  - adjusted R^2 (fit using MS_{error})
- Measures that combine fit and complexity:
  - general idea: fit + penalty for model complexity
  - Mallows C_p: least penalty
  - AIC: larger penalty
  - BIC: largest penalty (usually)
  - Often C_p, AIC, and BIC lead to the same model
    - When they differ, smaller penalty $\Rightarrow$ more variables
    - C_p selects most variables
    - BIC selects fewest

All Possible Subsets

- Set of k explanatory variables
- Fit all $2^k - 1$ possible models
- Compare models using some criterion (adj-R^2, C_p, AIC, BIC)
- Works up to about $k = 20$ (i.e., takes a reasonable amount of time to process $2^k - 1$ possible models)
- Review the best models of each size: $1, 2, \ldots, k$

Stepwise Methods

- Enter or delete one variable at a time from model according to algorithm
- Less time to compute than all possible subsets
- Possible algorithms:
  - Forward selection
  - Backward elimination (selection)
  - Stepwise selection

Difficulties with Model Selection

- Multicollinearity: high correlation between some explanatory variables
- Example: Suppose x_j and x_ℓ have a high correlation (near -1 or 1) and are both in the model
  - Significance test for either β_j or β_ℓ: does x_j or x_ℓ significantly add to the model that includes all other explanatory variables?
  - Once one of the variables is in the model, the other is not likely to significantly add to the model due to their close association

Variance Inflation Factor (VIF)

- Measures the degree to which the standard error of an estimated coefficient $\hat{\beta}_j$ is inflated by the correlations with the other explanatory variables:

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the R^2 value from the MLR with response variable x_j on the remaining explanatory variables.

- Explanatory variables with $\text{VIF}_j > 4$ should be investigated further
- Explanatory variables with $\text{VIF}_j > 10$ indicate severe multicollinearity

MLR Diagnostics

MLR Model and Assumptions

$$Y_i = \mu_{Y|x} + \varepsilon_i \quad \text{where } \varepsilon_i \text{ i.i.d. } N(0, \sigma^2)$$

$$= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

- Observations Y_i are independent
- Values of $\mathbf{x}$ are fixed
- $\mu_{Y|x}$ is a linear function of $\mathbf{x}$
- Homogeneous error variance: $\text{Var}(\varepsilon_i) = \sigma^2$
- Normally distributed errors: ε_i i.i.d. $N(0, \sigma^2)$

Weighted Least Squares

- Assume $\text{Var}(\varepsilon_i) = \sigma_i^2$ for $i = 1, \ldots, n$
- Define diagonal matrix W to have elements $w_{ii} = 1/\sigma_i^2$
- Weighted least squares estimate of $\boldsymbol{\beta}$ is

$$(X^T W X)^{-1} X^T W Y$$

- Observations with smaller σ_i^2 get a larger weight in the weighted least squares estimate than observations with larger σ_i^2
- Must know or be able to estimate values of w_{ii}:
  - If the ith observation is an average of n_i equally variable observations, then

$$\text{Var}(Y_i) = \sigma^2/n_i \quad \text{and} \quad w_{ii} = n_i$$

- If the ith observation is a total of n_i observations, then

$$\text{Var}(Y_i) = n_i\sigma^2 \quad \text{and} \quad w_{ii} = 1/n_i$$

- If variance is proportional to some predictor x_j, then

$$\text{Var}(Y_i) = x_{ij}\sigma^2 \quad \text{and} \quad w_{ii} = 1/x_{ij}$$

- In some cases, the values of the weights may be based on theory or prior research

Model Selection Assessment

- Stepwise procedures tend to overfit the sample data. Would the model perform as well in making predictions for new cases randomly selected from the population?
- Model validation: Split data into two parts:
  - Training sample (perhaps 2/3 of the data)
  - Validation sample (the remainder of the data)
  - Use training sample to select model
  - Use validation sample to assess model performance and fit
- Compute

$$\text{MSE}_{\text{Validation}} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left(Y_i - \hat{Y}_i\right)^2$$

- Should be approximately equal to $\text{MSE}_{\text{Training}}$ from selected model
- $\text{MSE}_{\text{Validation}}$ will be substantially larger if model is overfit to the training sample
- Use as model selection technique:
  - Fit many models to the training sample
  - Compute $\text{MSE}_{\text{Validation}}$ for the validation sample for each selected model
- PRESS is doing this over-and-over with the size of the test sample equal to one case

Linear Model Theory

Introduction

Linear Model Setup

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

is a random vector.

1. $E(\mathbf{Y}) = X\boldsymbol{\beta}$ is a vector of expected responses for some known matrix X of constants and unknown parameter vector $\boldsymbol{\beta}$.
2. $\text{Var}(\mathbf{Y}) = \Sigma$.
3. Complete the model by specifying a probability distribution for the possible values of $\mathbf{Y}$ or $\boldsymbol{\varepsilon}$.

Gauss–Markov Model

The linear model

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

is called a **Gauss–Markov model** if

$$\text{Var}(\mathbf{Y}) = \text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I$$

for some unknown constant σ^2.

Normal Theory Gauss–Markov Model

A normal theory Gauss–Markov model is a Gauss–Markov model where $\mathbf{Y}$ (and $\boldsymbol{\varepsilon}$) has a multivariate normal distribution.

$$\mathbf{Y} \sim N(X\boldsymbol{\beta}, \sigma^2 I) \quad \text{implying} \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I)$$

ANOVA Table

Variation	d.f.	Sums of Squares	Mean Square
Model	2	$\sum_{i=1}^{5}(\hat{Y}_i - \bar{Y})^2 = \mathbf{Y}^T(P_X - P_1)\mathbf{Y}$	$\frac{1}{2}SS_{\text{model}}$
Error	2	$\sum_{i=1}^{5}(Y_i - \hat{Y}_i)^2 = \mathbf{Y}^T(I - P_X)\mathbf{Y}$	$\frac{1}{2}SS_{\text{error}}$
Total	4	$\sum_{i=1}^{5}(Y_i - \bar{Y})^2 = \mathbf{Y}^T(I - P_1)\mathbf{Y}$	

Estimation

Ordinary Least Squares (OLS) Estimator For a linear model with

$$E(\mathbf{Y}) = X\boldsymbol{\beta},$$

any vector $\mathbf{b}$ that minimizes the sum of squared residuals

$$Q(\mathbf{b}) = \sum_{i=1}^{n} \left(Y_i - \mathbf{x}_i^T \mathbf{b}\right)^2 = (\mathbf{Y} - X\mathbf{b})^T (\mathbf{Y} - X\mathbf{b})$$

is an ordinary least squares (OLS) estimator for $\boldsymbol{\beta}$.

Normal Equations

For $j = 1, 2, \ldots, k$, solve the set of equations

$$0 = \frac{\partial Q(\mathbf{b})}{\partial b_j} = 2 \sum_{i=1}^{n} \left(Y_i - \mathbf{x}_i^T \mathbf{b}\right) x_{ij}.$$

These equations are expressed in matrix form as

$$\mathbf{0} = X^T (\mathbf{Y} - X\mathbf{b}) = X^T \mathbf{Y} - X^T X \mathbf{b},$$

or equivalently,

$$X^T X \mathbf{b} = X^T \mathbf{Y}.$$

These are called the **normal equations**.

Uniqueness of the OLS Estimator

If $X_{n \times k}$ has full column rank, $\text{rank}(X) = k$, then:

- $X^T X$ is non-singular,
- $(X^T X)^{-1}$ exists and is unique.

This means we can solve the normal equations for $\mathbf{b}$ as

$$X^T X \mathbf{b} = X^T \mathbf{Y},$$

$$(X^T X)^{-1} (X^T X) \mathbf{b} = (X^T X)^{-1} X^T \mathbf{Y},$$

$$\mathbf{b} = (X^T X)^{-1} X^T \mathbf{Y},$$

and $\mathbf{b}$ is unique.

Generalized Inverse

For a given $m \times n$ matrix A, any $n \times m$ matrix G that satisfies

$$AGA = A$$

is a **generalized inverse** of A.

Projection Matrix Define the projection matrix P_X to be

$$P_X = X(X^T X)^- X^T,$$

where $(X^T X)^-$ is a generalized inverse of $X^T X$.

Uniqueness of Mean Estimation

The estimation of the mean vector (predicted response vector)

$$\hat{\mathbf{Y}} = X\mathbf{b} = X(X^T X)^- X^T \mathbf{Y} = P_X \mathbf{Y}$$

is unique.

Residuals The vector of residuals is

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - X\mathbf{b} = \mathbf{Y} - P_X \mathbf{Y} = (I - P_X)\mathbf{Y}.$$

Uniqueness of Residuals

Because the projection operator

$$P_X = X(X^T X)^- X^T$$

is invariant with respect to the choice of $(X^T X)^-$, the residuals are invariant with respect to the choice of $(X^T X)^-$. That is,

$$\mathbf{e} = \mathbf{Y} - X\mathbf{b} = (I - P_X)\mathbf{Y}$$

is the same for any solution

$$\mathbf{b} = (X^T X)^- X^T \mathbf{Y}$$

to the normal equations.

Estimability

Identifiable For a linear model $E(Y) = X\beta$, the parameter vector β is *identifiable* if

$$X\beta_1 = X\beta_2 \implies \beta_1 = \beta_2.$$

Identifiability and Estimability

- Only *identifiable* parameters can be estimated.
- Linear functions of identifiable parameters are called *estimable*.
- Unbiased estimators can be found for estimable functions of model parameters.

Example: One-Way ANOVA Cell Means Model Write:

$\beta_1 = (\beta_1, \beta_2, \beta_3)^\top$

$\beta_2 = (\beta_1^*, \beta_2^*, \beta_3^*)^\top$

- For $X\beta_1 = X\beta_2$, we must have $\beta_1 = \beta_2$.
- For this model, β is identifiable.
- The vector of response means uniquely determines the values of the parameter vector β.

Estimable Functions An *estimable function* is a linear function of identifiable parameters.

- Estimable functions are reasonable quantities to estimate.
- Estimable functions have the same interpretation regardless of the constraints placed on the parameters to obtain a solution to the normal equations.
- Least squares estimates of estimable functions are not affected by the choice of constraints placed on the parameters to obtain a solution to the normal equations.

Formal Definition

For a linear model

$$Y = X\beta + \varepsilon,$$

we say that a linear function of β, $C\beta$, is estimable if

$$C\beta = AX\beta = AE(Y)$$

for some matrix A.

Rules for Estimable Functions

For a linear model

$$Y = X\beta + \varepsilon :$$

- The expectation of any observation is estimable.
- A linear combination of estimable functions is estimable.
- Each element of β is estimable if and only if $\text{rank}(X) = k$, where k is the number of columns in X.
- Every $c^T\beta$ is estimable if and only if $\text{rank}(X) = k$, the number of columns in X.
- Let X_j be the jth column of X. The parameter β_j is *not* estimable if and only if

$$X_j = \sum_{j \neq \ell} c_\ell X_\ell$$

for some set of scalars $\{c_\ell : j \neq \ell\}$.

General Remarks

For a linear model

$$Y = X\beta + \varepsilon :$$

- Definitions of estimable functions of the elements of the parameter vector β depend on the linear model for the expected responses,

$$E(Y) = X\beta.$$

- No assumption is made about $\mathrm{Var}(Y)$ or $\mathrm{Var}(\varepsilon)$ or the shape of the distribution of Y or ε.

The Gauss–Markov Theorem For the Gauss–Markov model,

$$Y = X\beta + \varepsilon,$$

with

$$E(Y) = X\beta \quad \text{and} \quad \mathrm{Var}(Y) = \sigma^2 I,$$

the OLS estimator $C\mathbf{b}$ of an estimable function $C\beta$ is the **unique best linear unbiased estimator (BLUE)** of this estimable function.

Hypothesis Testing

Quadratic Form Let $\mathbf{Y}$ be an n-dimensional random vector and let $\mathbf{A}$ be a non-random $n \times n$ matrix. A *quadratic form* is a random variable defined by

$$\mathbf{Y}^\top \mathbf{A} \mathbf{Y}.$$

Theorem

If $E(\mathbf{Y}) = \boldsymbol{\mu}$ and $\mathrm{Var}(\mathbf{Y}) = \boldsymbol{\Sigma}$, then

$$E\big(\mathbf{Y}^\top \mathbf{A} \mathbf{Y}\big) = \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top \mathbf{A} \boldsymbol{\mu}.$$

Central Chi-Squared Distribution

Let

$$\mathbf{Z} = \begin{bmatrix} Z_1 \\ \vdots \\ Z_n \end{bmatrix} \sim N(\mathbf{0}, \mathbf{I}),$$

i.e., the elements of $\mathbf{Z}$ are n independent standard normal random variables.

The distribution of

$$W = \mathbf{Z}^\top \mathbf{Z} = \sum_{i=1}^{n} Z_i^2$$

is called the **Central Chi-Squared distribution** with n degrees of freedom.

Non-central Chi-Squared Distribution

Let

$$\mathbf{Y}^\top = \begin{bmatrix} Y_1 & \cdots & Y_n \end{bmatrix} \sim N(\boldsymbol{\mu}, \mathbf{I}),$$

i.e., the elements of $\mathbf{Y}$ are independent normal random variables with

$$Y_i \sim N(\mu_i, 1).$$

The distribution of the random variable

$$W = \mathbf{Y}^\top \mathbf{Y} = \sum_{i=1}^{n} Y_i^2$$

is called a **Non-central Chi-Squared distribution** with n degrees of freedom and non-centrality parameter

$$\delta = \frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\mu} = \frac{1}{2}\sum_{i=1}^{n} \mu_i^2.$$

Distribution of Quadratic Forms Let $\mathbf{A}$ be an $n \times n$ symmetric matrix with rank($\mathbf{A}$), and let

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ is an $n \times n$ symmetric positive definite matrix.
If

$$\mathbf{A}\boldsymbol{\Sigma} \text{ is idempotent,}$$

then

$$\mathbf{Y}^\top \mathbf{A} \mathbf{Y} \sim \chi^2_{\text{rank}(\mathbf{A})}\left(\frac{1}{2}\boldsymbol{\mu}^\top \mathbf{A} \boldsymbol{\mu}\right).$$

In addition, if

$$\mathbf{A}\boldsymbol{\mu} = \mathbf{0},$$

then

$$\mathbf{Y}^\top \mathbf{A} \mathbf{Y} \sim \chi^2_{\text{rank}(\mathbf{A})}.$$

Independence of Quadratic Forms
Let

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

and let $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_p$ be $n \times n$ symmetric matrices.
If

$$\mathbf{A}_i \boldsymbol{\Sigma} \mathbf{A}_j = \mathbf{0} \quad \text{for all } i \neq j,$$

then

$$\mathbf{Y}^\top \mathbf{A}_1 \mathbf{Y},\ \mathbf{Y}^\top \mathbf{A}_2 \mathbf{Y},\ \ldots,\ \mathbf{Y}^\top \mathbf{A}_p \mathbf{Y}$$

are independent random variables.

Central F Distribution

If

$$W_1 \sim \chi^2_{n_1}, \quad W_2 \sim \chi^2_{n_2},$$

and W_1 and W_2 are independent, then the distribution of

$$F = \frac{W_1/n_1}{W_2/n_2}$$

is called the **Central F distribution** with n_1 and n_2 degrees of freedom.

Non-central F Distribution

If

$$W_1 \sim \chi^2_{n_1}(\delta_1), \quad W_2 \sim \chi^2_{n_2},$$

and W_1 and W_2 are independent, then the distribution of

$$F = \frac{W_1/n_1}{W_2/n_2}$$

is called a **Non-central F distribution** with n_1 and n_2 degrees of freedom and non-centrality parameter δ_1.

Testable Hypotheses For the Gauss–Markov model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

with

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \quad \text{and} \quad \text{Var}(\mathbf{Y}) = \sigma^2 \mathbf{I},$$

we say that the hypothesis

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$$

is **testable** if:

- $\mathbf{C}\boldsymbol{\beta}$ is estimable;
- $\text{rank}(\mathbf{C}) = m$, where m is the number of rows in $\mathbf{C}$.