Applied Regression Analysis and Other Multivariable Methods

Chapter 3-4 outlines

Yi Zhou

2015-10-02

Outline

Chapter 3 Basic Statistics: A Review

3-1 Preview

3-2 Descriptive Statistics

3-3 Random variables and Distributions

3-4 Sampling Distributions of t, χ^2, and F

3-5 Statistical Inference: Estimation

3-6 Statistical Inference: Hypothesis Testing

3-7 Error Rates, Power, and Sample Size

Chapter 4 Introduction to Regression Analysis

4-1 Preview

4-2 Association vs. Causality

4-3 Statistical vs. Deterministic Model

3-1 Preview

- Statistics: methods and procedures for collecting, classifying, **summarizing and analyzing** data.
- The primary goal of most statistical analysis:
  making statistical inferences.
- Population, sample, parameter, statistic,
  descriptive statistics.
- Statistical inferences: estimation and hypothesis testing.

3-2 Descriptive Statistics

- Central tendency: average value (e.g. sample mean).
- Central Limit Theorem.
- Variability/dispersion
  sample variance: S^2
  sample standard deviation: S
  attention: $\dfrac{1}{n-1}$
- Succinct picture of data: $\bar{X} \pm S_X$, $\bar{Y} \pm S_Y$.

3-3 Random variables and Distributions I

- Random variable: X, Y, Z.
- Possibility distribution of the random variable:
  relative frequencies associated with the possible values (presented by table, graph, mathematical expression ...).
- **Discrete random variables**: $Pr(X = a)$
  countable values, gappy distributions graphed as a series of lines (heights of these lines).
  e.g. the number of death/arrivals.
- **Continuous random variables**: $Pr(a < X < b)$
  uncountable values, nongappy distributions graphed as smooth curves (an area under the curve).
  e.g. the blood pressure/weight.
  note: $Pr(X = a) = 0$

3-3 Random variables and Distributions II

- **Binomial distribution**: discrete, $X \sim B(n, p)$.
- **Normal distribution**: continuous, $X \sim N(\mu_X, \sigma_X^2)$.
  - **Standard normal distribution**: $Z \sim N(0, 1)$.
  - The conversion formula.
  - The normal approximation of the binomial distribution, when n is *moderately* large.

3-4 Sampling Distributions of t, χ^2, and F I

- **The (student's) t distribution**: $t(n-1)$.
  - Symmetric about 0.
  - X is normal distribution.
  - σ^2 is unknown and is estimated by S^2.
  - Estimated standard error of $\bar{X}$: $\dfrac{S}{\sqrt{n}}$.
  - Degree of freedom: $n-1$. (tabulated percentile)
  - Pooled sample variance: estimate the common variance. (e.g. $N(\mu_1, \sigma), N(\mu_2, \sigma)$)

$$S_p^2 = \frac{\sum_{i=1}^{k}(n_i - 1)S_i^2}{\sum_{i=1}^{k}(n_i - 1)}$$

3-4 Sampling Distributions of t, χ^2, and F II

- **The chi-square (χ^2) distribution**: $\chi^2(n-1)$.
  - Nonsymmetric distribution.
  - Nonnegative.
  - Sample variance: S^2
  - Sample size is n and from a normal distribution.
  - Widespread application in categorical data (e.g. contingency table).

3-4 Sampling Distributions of t, χ^2, and F III

- **The F distribution**: $F(n_1 - 1, n_2 - 1)$.
  - Skewed to the right.
  - The ratio of independent estimators of two populaton variances. (e.g. for $N(\mu_1, \sigma_1), N(\mu_2, \sigma_2)$, we have $S_1, \sigma_1, S_2, \sigma_2$
  - $T^2 \sim F(1, v)$ if and only if $T \sim t(v)$).

    Note: Hotelling's T^2 distribution.
    Multivariate testing vs. multiple testing.

3-5 Statistical Inference: Estimation

- Specific value of a parameter.
- $\hat{\theta}$ is the **point estimator** of θ. $(\mu, p, \mu_1 - \mu_2, \theta_1/\theta_2)$
  - Method of maximum likelihood estimation
  - Method of moment estimation
- **Procedure**. (sample→point estimator→variablity→CI)
- Confidence interval (CI).
  - Two bundary points
  - θ is fixed.
  - 95%CI for $\mu_1 - \mu_2$ contains 0 or not.

3-6 Statistical Inference: Hypothesis Testing

- Making a decision (Reject/accept H_0).
- The null hypothesis (H_0) is unlikely to be true.
  (that is, the estimated value is different enough from the hypothesized value)
- Procedure (seven steps).
- Type I error, significant level α, type II error, β, power.
- Rejection/critical region, acceptance region, critical point
- **P-value**.

3-7 Error Rates, Power, and Sample Size

- H_0, H_A, α, β.
- Power of the test.
- Sample size formula.
- Multiple-testing problem.

4-1 Preview

- Regression analysis: one of the methods in multivariate techniques.
  - Wide applicability.
  - Simplest to implement.
  - R package:
- Independent variables/predictors ($X : X_1, X_2, ...$), dependent variable/response (Y).
- Examples. (Variables, relationships and equations)

4-2 Association vs. Causality

- Bias.
- Statistically significant association does not mean a causal relationship.
- Causality.
- Path analysis.
- The criteria by Bradfor Hill (1971)

4-2 Association vs. Causality

- Association.
- Statistical model:
  - **Regression analysis**.
  - Discrimination analysis.
  - Factor analysis.
  - Analysis of variance & covariance.
  - ...
- Deterministic model.

Appendices I

Multivariate distribution

- Multivariate normal distribution
  The multivariate normal distribution of a k-dimensional random vector $\mathbf{X} = X_1, X_1, \ldots, X_k$ can be written in the following notation:

$$\mathbf{X} \ \sim \ \mathcal{N}_k(\boldsymbol{\mu}, \, \boldsymbol{\Sigma})$$

  $\boldsymbol{\mu} = [\mathrm{E}[X_1], \mathrm{E}[X_2], \ldots, \mathrm{E}[X_k]],$
  $\boldsymbol{\Sigma} = [\mathrm{Cov}[X_i, X_j]], i = 1, 2, \ldots, k; j = 1, 2, \ldots, k.$
- Hotelling's T^2 distribution
  Hotelling's T-squared statistic is a generalization of Student's t statistic that is used in multivariate hypothesis testing.

Appendices II

Sample size determination

- R: *samplesize, pwr (for power analysis)*
- PASS
- Factors increasing the sample size: $\alpha \downarrow, (1-\beta) \uparrow, \sigma^2 \uparrow, \Delta = |\mu_0 - \mu_1| \downarrow$

Multiple testing

- Bonferroni correction, fixed-sequence, fallback, and gatekeeping procedures.

Appendices III

Bias

- Bias in randomized controlled trials (RCTs).
  - Selection bias
  - Performance bias
  - Detection bias
  - Attrition bias
  - Reporting bias
  - Others

Path analysis

- PLS-PM model: package *plspm* in R