STAT 525

Introduction

Dr. Qifan Song

Regression Analysis

- Regression analysis is a set of statistical processes for studying the relationships among variables.
- Given the independent (or predictor, covariate) variables X, and the dependent (or response) variable Y, the regression analysis relates Y to a function of X, in the sense $Y \approx f_\beta(X)$, where f_β is a function parametrized by some unknown parameter β. More formally, the *mean* regression model is

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

- A linear regression model considers that f_β is a linear function with coefficient β, i.e. $f_\beta(X) = X\beta$.
- A generalized linear model generalizes the linear model to $f_\beta(X) = g(X\beta)$ for some given function g.

Statistical Inference Procedure of Regression

- Design a statistical modeling, i.e., choose a proper f_β, as well as the distribution family of Y.
- Parameter estimation, including β and other nuisance parameters.
- Statistical inferences, including hypothesis testings, confidence intervals and predictions.
- Re-exam the appropriateness of statistical modeling.

Example

X_1,...,X_n are the GPA's of randomly chosen Purdue students, and we are interested in the average GPA of all Purdue students.

- Model Assumption: $X_i \sim^{iid} N(\mu_i, \sigma^2)$ where $\mu_i \equiv \mu$ or μ_i depends on sex or major
- Estimate μ and σ^2
- CI for μ and σ^2
- Model diagnose, e.g., independence and normality assumptions

Parameter Estimation

- Moment Estimation (Methods of Moment)
- Maximum Likelihood Estimation (MLE)
  - usually the optimal estimation (achieves Cramer-Rao bound bound), but biased for variance parameters
  - require optimization technique
- Restricted MLE (ReML): MLE of transformed data
  - $X \sim N(\mu 1, \sigma^2 I_n)$
  - $KX \sim N(\mu K * 1, \sigma^2 KK')$
  - choose proper K such that $K * 1 = 0$
  - Let K be $(n-1)$ by n, with $(n-1)$ orthonormal rows
  - $KX \sim N(0, \sigma^2 I_{n-1})$
  - $\hat{\sigma}^2 = X'K'KX/(n-1) = s^2$ (sample variance)

Sampling distribution

- $(n-1)s^2 = X'K'KX = [N(0, \sigma^2 I_{n-1})]'N(0, \sigma^2 I_{n-1}) \sim \sigma^2 \chi^2_{n-1}$
- $\mu = \bar{X} = (1/n)1 * X \sim N(\mu, \sigma^2/n)$
- μ and $(n-1)s^2$ are independent, since $Cov(KX, 1 * X) = \sigma^2 K * 1 = 0$.
- $$\frac{\sqrt{n}(\bar{X} - \mu)}{s} = \frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{\sqrt{s^2/\sigma^2}} = \frac{N(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}} = t_{n-1}$$
- Derive valid CIs for σ^2 and μ:

Model Diagnose

Is it reasonable to assume that X_i's are independently normal distributed?

- Normality
- Same mean across all samplers
- Independence
- Same variance across all samplers

Methods (based on hypothesis testing idea): QQ plot, histogram check and formal normality test (will be introduced later)

Other required knowledge

- F distribution
- Programming (SAS)
- Concepts and Procedure of Hypothesis Testings, Type I and II errors, Power and etc