Welcome to Stat 425!

- Personnel
  - Instructor: Liang, Feng (OH: Tuesday, noon-1pm)
  - TA: Huang, Xichen
- Websites: Piazza and Compass and my page
- Homework
  - When, where and how to submit your homework
  - No late submissions will be accepted
  - You get 100% for homework, if finish no less than 85% of all the assignments
  - Grading policy
- Two Exams, One Project, No Final Exam

Communication

- For questions related to homework/lectures, please post your question on Piazza. By default, you are anonymous to your classmates, but not to the instructor.

- If you want to send email to my Illinois account, please

  1. Write from your Illinois email account (so I would know who you are)

  2. Start your subject line with the course number, e.g., "[stat425] cannot attend exam I" (since I'm teaching two courses this semester)

  3. Sign with your full name

  4. Don't send unexpected attachments.

What You'll Learn

Let's first take a look of the final project in the past years.

- Fall 2015: Bike rental forcasting
- Fall 2014: Walmart store sales forecasting
- Fall 2013: Champaign-Urbana housing data
- Fall 2012: Titanic disaster data

- Regression analysis is used to explain the dependence between a response variable Y and one or more explanatory variables $X_1, X_2, \ldots X_p$.

- In regression analysis, we assume

$$\mathbb{E}[Y \mid X_1, \ldots, X_p] = g(X_1, \ldots, X_p),$$

  where g could be a linear or non-linear function of the p covariates. The task is to estimate g based on data: $\{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^n$.

- Two major goals of regression:
  - Prediction
  - Exploration

- Take the Zillow data as an example. Here are the questions we would hope to answer by analyzing the data
  - What would be the fair market price of a house?
  - Which of the variables Size (sqft), # Bathrooms, Age, Location has the largest estimated effect on Price?
  - Is it worth adding an additional bathroom?
  - Identify, given this data set, the best deal for the buyer and the best deal for the seller.

Type of Regression Models

- We begin with linear regression, which models the mean function as a linear combination of the X_j's:

$$\mathbb{E}[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$

  - Linear models (Y, X_k's: numerical);
  - Analysis of variance models (Y: numerical; X_k's: categorical);
  - Analysis of covariance models (Y: numerical; X_k's: categorical/numerical);
- Generalized linear models (Y: categorical);
- Mixed effects models (data are correlated);
- Nonparametric regression models (g is a smoothed curve).

Course Overview

1. Linear regression
    - Simple linear regression
    - Multiple linear regression
    - Regression diagnostics
    - Transformation and variable selection
    - Experimental design and ANOVA
2. Generalized linear regression
3. Nonparametric regression
4. Linear models with mixed effects

General Expectations

- Finish the reading assignments
- Review the notes
- Get familiar with `R`
- Feedback and questions
- Finish homework independently

  You can discuss homework problems with other students but should write your answers independently using your own words.

Prerequisites

You should be comfortable with the following jargons/concepts (Check the posted Prerequisite_Stat425.pdf)

- CDF, pdf, density functions, expectations, variance, independence, conditional distributions;
- likelihood functions, random samples, estimator, mean-squared error, hypothesis testing, p-value, confidence interval,
- vector, matrix, matrix multiplication, matrix transpose, inverse of a matrix, full rank.

Useful Distributions

Refresh your memory on the following distributions: Normal, Student t, and F distribution.

The Univariate Normal Distribution

- $Y \sim \mathsf{N}(\mu, \sigma^2)$, $\mathbb{E}(Y) = \mu, \mathsf{Var}(Y) = \sigma^2$, with pdf

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

- $Z \sim \mathsf{N}(0, 1)$: the standard normal rv. $\Phi(z)$ denotes its CDF.

$$\Phi(-z) = 1 - \Phi(z), \quad z > 0.$$

- Linear transformations of normals are still normal. $Y \sim \mathsf{N}(\mu, \sigma^2)$, then

$$aY + b \sim \mathsf{N}(a\mu + b, a^2\sigma^2), \quad \frac{1}{\sigma}(Y - \mu) \sim N(0, 1).$$

- Linear combinations of normal rv's are normal? Not true in general, but true for almost all cases we'll encounter in 425.

The Multivariate Normal Distribution

- Let $\mathbf{Z} = (Z_1, \ldots, Z_n)$ where Z_i's are iid $\sim \mathsf{N}(0,1)$ rv's. Then $\mathbf{Z}$ follows a multivariate normal distribution, denoted by $\mathsf{N}_n(\mathbf{0}, \mathbf{I}_n)$, with pdf

$$f(\mathbf{z}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-z_i^2} = \frac{1}{(2\pi)^{n/2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} z_i^2 \right\}$$

$$= \frac{1}{(2\pi)^{n/2}} \exp\left\{ -\frac{1}{2} \mathbf{z}^t \mathbf{z} \right\},$$

and moment generating function

$$M(\mathbf{t}) = \mathbb{E}[\exp\{\mathbf{t}^t \mathbf{Z}\}] = \exp\left\{ \frac{1}{2} \mathbf{t}^t \mathbf{t} \right\},$$

and mean and covariance

$$\mathbb{E}(\mathbf{Z}) = \mathbf{0}, \qquad \mathsf{Cov}(\mathbf{Z}) = \mathbf{I}_n.$$

- $\mathbf{Y}$ has a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance Σ, denoted by $\mathsf{N}_n(\boldsymbol{\mu}, \Sigma)$ if its moment generating function is

$$M_Y(\mathbf{t}) = \exp\left\{\mathbf{t}^t\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^t\Sigma\mathbf{t}\right\}.$$

Why don't we define $\mathbf{Y}$ via its pdf? (It may not exist.)

- Recall the definition of the covariance matrix for a random vector $\mathbf{Y}$. Any covariance matrix Σ should be symmetric and *positive semi-definite* (psd), where psd means

$$\mathbf{a}^t\Sigma\mathbf{a} \geq 0.$$

This is because

$$0 \leq \text{Var}(\mathbf{a}^t\mathbf{Y}) = \mathbf{a}^t\Sigma\mathbf{a}.$$

Any symmetric psd matrix has a *spectral decomposition*

$$\Sigma = \Gamma^t\Lambda\Gamma, \qquad \Lambda = \text{diag}(\lambda_i)_{i=1}^n,$$

and $\Gamma_{n\times n}$ is a orthonormal matrix, i.e., $\Gamma\Gamma^t = \mathbf{I}_n$.

- If $\lambda_n > 0$, i.e., Σ is of full rank, then $|\Sigma| > 0$ and Σ^{-1} exists.

  Then the pdf of $\mathsf{N}_n(\boldsymbol{\mu}, \Sigma)$ is given by

$$f(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{y} - \boldsymbol{\mu})\right\}.$$

Properties of Multivariate Normals

- Affine transformations of a normal vector are still normal:

$$\mathbf{Y} \sim \mathsf{N}_n(\boldsymbol{\mu}, \Sigma) \Longrightarrow A_{m\times n}\mathbf{Y} + b_{m\times 1} \sim \mathsf{N}_m(A\boldsymbol{\mu} + b, A\Sigma A^t).$$

- Marginals of a normal are still normal.

- Conditionals of a normal are still normal.

$$\mathbf{Y}_1|\mathbf{Y}_2 \sim \mathsf{N}_m\Big(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{Y}_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Big)$$

- For multivariate normals, uncorrelated $=$ independent.

Distributions Related to Normals

- Z_i iid $\sim N(0,1)$, then $Z_1^2 + \cdots + Z_n^2 \sim \chi_n^2$.

$$W \sim \chi_n^2, \qquad \mathbb{E}(W) = n, \qquad \mathsf{Var}(W) = 2n.$$

- $Z \sim N(0,1)$ and $W \sim \chi_n^2$ are independent, then

$$\frac{Z}{\sqrt{W}} \sim t_n \text{ (student } t\text{-dist).}$$

- $W_1 \sim \chi_n^2$, $W_2 \sim \chi_m^2$ and they are independent, then

$$\frac{W_1}{W_2} \sim F_{n,m}.$$

- Chi-square and Student t-dist have one df (degree of freedom) and F-dist has two dfs.

Basic Statistical Inference

Refresh your memory on basic statistical inference, such as

- point estimation: bias, unbiased, MSE;
- interval estimation: 95% CI (confidence interval);
- hypothesis testing: significance level, type I error, type II error, p-value.

Consider the following example: $Z_1, \ldots, Z_n$ iid $\sim \mathsf{N}(\theta, \sigma^2)$, where θ and σ^2 are unknown.

- What's the MLE of θ? Is it unbiased? What's the MSE (mean-squared error) of the MLE?
- What's the MLE of σ^2? Is it unbiased? If yes, find an unbiased one.
- How to test $\theta = 1$ against a two-sided alternative $\theta \neq 1$? How to calculate the p-value?
- How to test $\theta = 1$ against a one-sided alternative $H_a : \theta > 1$?
- How to construct a 95% confidence interval (CI) for θ?

MLE

Suppose we collect n iid samples $Z_1, \ldots, Z_n$ from $\mathsf{N}(\theta, \sigma^2)$ where θ is unknown. First, write the likelihood function

$$\mathsf{Lik}(\theta; Z_1, \ldots, Z_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(Z_i - \theta)^2}{2\sigma^2} \Big).$$

The MLE of θ is the one that maximizes the likelihood function (given data $Z_{1:n}$)

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta} \mathsf{Lik}(\theta) = \arg\max_{\theta} \log \mathsf{Lik}(\theta) \\ &= \arg\max_{\theta} -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Z_i - \theta)^2 = \arg\min_{\theta} \sum_{i=1}^{n} (Z_i - \theta)^2 \\ &= \frac{1}{n}(Z_1 + \ldots Z_n) = \bar{Z}. \end{aligned}$$

- Note that $\hat{\theta}$, as a function of the data $(Z_1, \ldots, Z_n)$, is a random variable

$$\mathbb{E}\hat{\theta} = \mathbb{E}\frac{1}{n}(Z_1 + \cdots + Z_n) = \theta, \quad \mathsf{Var}(\hat{\theta}) = \frac{\sigma^2}{n}.$$

Under the iid normal assumption, we have $\hat{\theta} \sim \mathsf{N}(\theta, \sigma^2/n)$.

- Is $\hat{\theta}$ unbiased?

$$\mathsf{Bias}(\hat{\theta}) = \mathbb{E}\hat{\theta} - \theta.$$

- What's its MSE?

$$\mathsf{MSE}(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta)^2 = \mathsf{Bias}^2 + \mathsf{Var}(\hat{\theta}).$$

Here, we have 0 bias, and therefore $\mathsf{MSE}(\hat{\theta}) = \sigma^2/n$.

If θ and σ^2 are unknown, you'll find that the MLE of θ is the same as what we derived before. Also we can find the MLE of σ^2:

$$\hat{\sigma}^2_{\text{mle}} = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})^2$$

Using the following equality

$$\sum_{i=1}^{n}(Z_i - \bar{Z})^2 = \sum_{i=1}^{n} Z_i^2 - n\bar{Z}^2,$$

we can show that

$$\mathbb{E}\hat{\sigma}^2_{\text{mle}} = \frac{n-1}{n}\sigma^2,$$

that is $\hat{\sigma}^2_{\text{mle}}$ is biased. It is easy to obtain an unbiased one

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z})^2.$$

What's the distribution of $\hat{\sigma}^2$? $\hat{\sigma}^2 \sim \sigma^2\chi^2_{n-1}/(n-1)$

Hypothesis Testing

- Suppose we want to test

$$H_0 : \theta = a, \quad \text{versus} \quad H_a : \theta \neq a.$$

- Construct a test statistic (which tend to take extreme values under H_a)

$$\frac{\hat{\theta} - a}{\mathsf{se}(\hat{\theta})}.$$ [a]

Under H_0, the statistic follows T_{n-1}, i.e., student T dist with $(n-1)$ degree-of-freedom. [b]

[a] For this normal example, $\mathrm{se}(\theta) = \hat{\sigma}^2/n$.

[b] When the sample size n is large, the test statistic follows $\mathsf{N}(0,1)$ approximately, even if Z_i's are not normally distributed.

- Given the data, we can calculate the test statistic – suppose it's t_0. Then the p-value is defined to be 2 $\times$ the area under the T_{n-1} dist more extreme than the observed statistic t_0.

  That is, p-value $= 2 \times F(|t_0|)$, where F is the CDF for T_{n-1}.

- If p-value $<$ the pre-specified significant level, say 5%, then we reject H_0 (small p-values are evidence against H_0).

Confidence Intervals

- The $(1-\alpha)$ confidence interval (CI) for θ is given by

$$\left(\hat{\theta} - t_{n-1}^{(\alpha/2)}\mathsf{se}(\hat{\theta}),\ \hat{\theta} + t_{n-1}^{(\alpha/2)}\mathsf{se}(\hat{\theta})\right),$$

or we sometimes write it as

$$\hat{\theta}\ \pm\ t_{n-1}^{(\alpha/2)}\mathsf{se}(\hat{\theta})$$

where $t_{n-1}^{(\alpha/2)}$ is the $(1-\alpha/2)$ percentile of T_{n-1}.

- Suppose $\alpha = 5\%$. The 95% CI (constructed) above is random (since it depends on the data). We CAN say that this random interval covers θ with probability 95%.

- Suppose given a data set, we calculate the CI, which is $(2.1, 3.5)$. Then for this particular interval, 95% is confidence, not chance.

  We CANNOT say that this particular interval $(2.1, 3.5)$ covers θ with probability 95%.

  This is because $(2.1, 3.5)$ is a fixed interval and θ is a fixed number (although it's unknown), so $(2.1, 3.5)$ either covers θ or not, and there is no probability attached to $(2.1, 3.5)$.

So how should we interpret the 95% CI $(2.1, 3.5)$?

- Based on the data, we are 95% confident that θ is between 2.1 and 3.5.

- We do not know whether $(2.1, 3.5)$ covers θ or not, but we know: if we were to repeat this process—collect samples from the same population and calculate 95% CI—many times, then about 95% of the resulting CIs will cover the true θ.

- The interpretation I like is based on a nice duality between testing and CI. The interval $(2.1, 3.5)$ contains a set of plausible values for θ, in the sense that for any value $\theta_0 \in (2.1, 3.5)$, based on the data, we cannot reject the null hypothesis $H_0 : \theta = \theta_0$ at the 5% significant level.