DTS206TC Formulas

Distributions

Exponential Dispersion Family

$$p(y;\theta,\phi)=\exp\{\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)\}$$

Normal Distribution

$$p(y;\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}\exp\{-\frac{1}{2\sigma^2}(y-\mu)^2\}$$

Binomial Distribution

$$p(y;\pi)=\binom{n}{y}\pi^y(1-\pi)^{n-y}, \text{ where } y=0,1,\cdots,n, \text{ and } \binom{n}{y}=\frac{n!}{y!(n-y)!}$$

ny has a Binomial Distribution.

$$p(y;\pi,n)=\binom{n}{ny}\pi^{ny}(1-\pi)^{n-ny}$$

where, $y=0,\frac{1}{n},\frac{2}{n},\cdots,1$.

Poisson Distribution

$$p(y;\mu)=\frac{\mu^y e^{-\mu}}{y!}$$

where,

- $y=0,1,2,\cdots$
- μ is the parameter of the distribution

Geometric distribution

$$f(y;\pi)=\pi(1-\pi)^y, y=0,1,2,\cdots$$

Estimation, Inference

$$\beta=(X'X)^{-1}X'Y$$

$$S_{xx}=\sum(x_i-\bar{x})^2$$

$$s_x^2=\frac{\sum(x_i-\bar{x})^2}{n-1}$$

$$s_R^2=MSE=\frac{RSS}{n-2}$$

$$\sigma^2\{\hat{\beta}_1\} = \frac{\sigma^2_{y|x}}{\sum (x_i - \bar{x})^2}$$
$$s^2\{\hat{\beta}_1\} = \frac{s^2_R}{\sum (x_i - \bar{x})^2}$$

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{s^2_R}{(n-1)s^2_x}}} \sim t_{n-2}$$

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2} \cdot s\{\hat{\beta}_1\}$$

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2} \sqrt{\frac{s^2_R}{(n-1)s^2_x}}$$

$$\sigma^2\{\hat{\beta}_0\} = \sigma^2_{y|x}\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right)$$

$$s^2\{\hat{\beta}_0\} = s^2_R\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right)$$

$$s^2\{\hat{\beta}_0\} = s^2_R\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s^2_x}\right)$$

$$\frac{\hat{\beta}_0 - \beta_0}{s\{\hat{\beta}_0\}} \sim t_{n-2}$$

$$\hat{\beta}_0 \pm t_{n-2,\alpha/2} \cdot s\{\hat{\beta}_0\}$$

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{s^2_R\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s^2_x}\right)}} \sim t_{n-2}$$

$$\hat{\beta}_0 \pm t_{n-2,\alpha/2} \sqrt{s^2_R\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s^2_x}\right)}$$

$$\hat{y}_0 \pm t_{n-2,\alpha/2} \sqrt{s^2_R\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s^2_x}\right)}$$

$$\hat{y}_0 \pm t_{n-2,\alpha/2} \sqrt{s^2_R\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s^2_x}\right)}$$

$p(\theta|y) \propto p(\theta)p(y)$

Bayesian topic

$$p(\boldsymbol{\beta}|Y) = \mathcal{N}(\boldsymbol{\beta}|\boldsymbol{\mu}_n, \Sigma_n)$$
$$\boldsymbol{\mu}_n = \Sigma_n(\Sigma_0^{-1}\boldsymbol{\mu}_0 + \sigma^{-2}X^TY)$$
$$\Sigma_n = (\Sigma_0^{-1} + \sigma^{-2}X^TX)^{-1}$$

Background on Regression Model

- Regression Model 回归模型.

  x: Independent or Explanatory variable. 独立或解释变量.

  y: Dependent or Response variable 依赖或响应变量.

  从n对值的样本 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ 获得不同x值的y的估计

- Regression Analysis 回归分析

  statistical methodology. 利用两个或多个变量(quantitative variable)之间的关系, 以便从一个变量中预测或响应另一个变量.

  a response or outcome variable can be predicted from other or others.

- Relations between Variables

  Deterministic: Functional Relation 确定性. 功能关系

  ↓ expressed by

  a mathematical formula

  dependent

  $y = f(x)$

  ↓ independent

  Nondeterministic: Statistical Relation 非确定性. 统计关系

  not perfect.

  $y = f(x) + \epsilon$ ← ϵ is an unknown perturbation (random variable)

- Type of Relation

Linear. $y = a + bx$ / $y = mx + b$

$\downarrow$

$y = \beta_0 + \beta_1 x$ → a straight line (linear)

$\beta_1 > 0$ → positive linear relationship

$\beta_1 < 0$ → negative linear relationship

Nonlinear $f(x) = \log(x)$

$f(x) = x^2 + x + 1$.

Absense of relation

Whenever $f(x) = c$, that is, whenever $f(x)$ doesn't depend on x.

- Measures of Linear Dependency

Covariance 协方差 $cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$ 取决于变量的测量单位

① positive linear relation → cov positive and large.

② negative linear relation → cov negative and large in absolute value

③ no relation/significantly ~~test~~ linear → cov close to zero

the Correlation coefficient 相关系数. 不依赖于测量单位

$r(x,y) = cor(x,y) = \frac{cov(x,y)}{S_x S_y}$ $S_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ $S_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$

① $-1 \le cor(x,y) \le 1$ ② $cor(x,y) = cor(y,x)$ ③ $cor(ax+b, cy+d) = sign(a)sign(c)cor(x,y)$

- Simple Linear Regression (SLR) Model with Distribution of Error Terms Unspecified.

Basic regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

↑ ↑

系数/参数

coefficients / parameters

y_i = the value of the dependent variable for the i-th observation

x_i = the value of the independent variable for the i-th observation

ε_i : the error for the i-th observation,

mean $E\{\varepsilon_i\} = 0$ 和 variance $\sigma\{\varepsilon\} = \sigma^2$ 方差 (各样本数据与平均数之差的平方和的平均数)

β_0, β_1 are parameters or ~~of~~ coefficients. β_0 : intercept 截距

β_1 : slope 斜率.

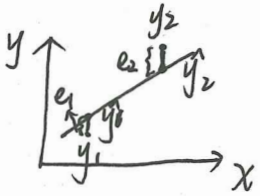
The parameters to estimate are: β_0, β_1, σ.

SLR Model

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. ← the regression line

residual 残差 $e_i = y_i - \hat{y}_i$

有正负



• Hypothesis of the SLR Model 假设.

线性 Linearity: The existing relation between x and y is linear.

data look reasonably straight $f(x) = \beta_0 + \beta_1 x$

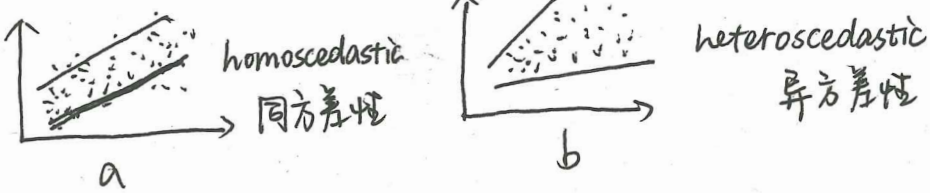
齐性 Homogeneity: The mean value of the error is zero.

$$E[\epsilon_i] = 0$$

方差齐性 Homoscedasticity: The variance of the errors is constant

$$Var(\epsilon_i) = \sigma^2$$

数据分散,必须恒定. dispersion of the data must be constant.



独立性 Independence: The observations are independent.

数据必须独立.

$$E[\epsilon_i \epsilon_j] = 0$$

↓

Any two response y_i and y_i are uncorrelated

① y_i a random variable: (i) constant term: $\beta_0 + \beta_1 x$ (ii) random term ϵ

② $E\{y\} = E\{\beta_0 + \beta_1 x + \epsilon\} = \beta_0 + \beta_1 x + E\{\epsilon\} = \beta_0 + \beta_1 x$

③ y_i 或者大于或者小于 (y) $= \beta_0 + \beta_1 x + \epsilon_i$.

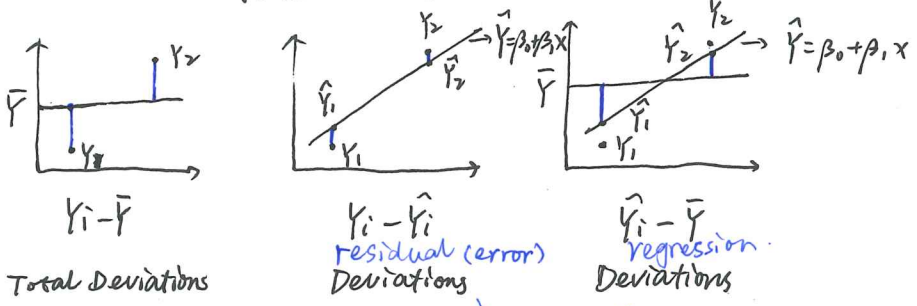
④ The error term ϵ_i are assumed to have constant variance σ^2.
ϵ_i and ϵ_j are uncorrelated.

⑤ cause ϵ_i and ϵ_j not correlated → y_i and y_j. so are the responses.

Total Deviation Partition.

↓

检查回归线的拟合度.



$Y_i-\bar{Y}$ Total Deviations

$Y_i-\hat{Y}_i$ residual (error) Deviations

$\hat{Y}_i-\bar{Y}$ regression Deviations

y_i 最简单的回归估计为 $\bar{y}$ (仅有截距的模型)

total error: 观测值 y_i 与 y_i 的平均值的差异 $y_i-\bar{y}$

↓be broken down as two component

$$y_i-\bar{y}=(\hat{y}_i-\bar{y})+(y_i-\hat{y}_i)$$

regression residual

↓ 这种变化导致的回归分析的基本方程.

$$\sum(y_i-\bar{y})^2=\sum(\hat{y}_i-\bar{y})^2+\sum(y_i-\hat{y}_i)^2$$

$$SSTO = SSR + SSE$$

↓ Total Sum of Squares

所有 y 相同, SSTO = 0. y 的变化越大, SSTO 越大 (离散程度个)

SSE Measure of after predictor effect

↓ x.

error sum of squares 考虑 x 后 SSE 仍是 y 的变化度量.

$$SSE=\sum(y_i-\hat{y}_i)^2$$

SSR Regression sum of square (回归平方和) SSTO - SSE = SSR

↓ regression sum of squares SSTO 和 SSE 之间的差为 SSR.

$$SSR=\sum(\hat{y}_i-\bar{y})^2$$

小写 y_i observation (x不考虑)

表示样本点与样本均值之间的差值的平方和

决定系数 Coefficient of Determination R^2

$$R^2=\frac{SSR}{SSTO}=1-\frac{SSE}{SSTO} \qquad 0\le SSE\le SSTO \Rightarrow 0\le R^2\le 1$$

① $R^2=0$ 时: regression line explains none of the variability in y and the regression line is no better than using $\bar{y}$ as our prediction of y.

回归线不如 $\bar{y}$ 用来预测 y.

② $R^2=1$ 时. a perfect fit and regression line explains all ~~the~~ of the variability. 这种情况下所有的数据点都落在回归线上且没有残差 residual variation

↓

不测量斜率的大小或测量直线模型的适当性.

not measure magnitude of the slope

or measure the appropriateness of the straight-line model.

R^2 值很大时并不意味着适当的模型.

① high R^2 ~~doesn't do that~~ means that there is a good linear fit between predictor and output. 不一定有很好的线性拟合 (实际曲线关系可近似以很好的线性模型拟合可能会导致 high R^2)

② low R^2 doesn't mean that there is ~~no~~ relationship between input and output 可能有关系, 但线性函数无法很好解释这种关系. (样本量大时) n-1 degrees of freedom

y 的方差 variance of y. $\hat{\sigma}^2_y$

$$\hat{\sigma}^2_y=\frac{\sum(y_i-\bar{y})^2}{n-1}=\frac{SSTO}{n-1}$$

由于计算和已经均值而产生的 1 个线性约束.

linear constraints

预测变量 x 纳入方差. variance of y given x $\hat{\sigma}^2_{y|x}$ → (Residual standard error)^2

$$\hat{\sigma}^2_{y|x}=\frac{\sum(y_i-\hat{y}_i)^2}{n-2}=\frac{SSE}{n-2}$$

n-2 degrees of freedom

β_0 和 β_1 估计产生的线性约束

$(n-1)-(n-2)=-1+2=1$

1 degree of freedom: 只依赖一个参数, 即斜率 β_1, 因为截距与均响应 mean response 相关联消失了, 仅保留 slope 作为单一自由度.

ANOVA Table 方差分析表. 用于汇总回归结果.

↓

the analysis of variance

n 为样本大小.

p 为模型中所包含的预测变量的数量. 自变量个数.

Source of Variation	Sum of Squares	df	MS	Variance Ratio (F)	p-value. F统计量的显著性水平.
Regression	SSR	p	MSR	$F=\frac{MSR}{MSE}$ ~ $F_{p,n-p-1}$	$Pr(F_{p,n-p-1}>F)$
Error	SSE	n-p-1	MSE		
Total	SSTO	n-1			

F-statistic.

MS 平均平方和 每个来源的平方和除以其自由度.

$$MSR=\frac{SSR}{p} \quad \text{Regression Mean Square}$$

$$MSE=\frac{SSE}{n-p-1} \quad \text{Error Mean Square}$$

- A T-test will tell if a single variable is statistically significant.
- An F-test will tell if a Group of variables are jointly significant.

F Test for SLR (p=1 for SLR)

对于 MSR 的期望是 $E(MSR)=\sigma^2_{y|x}+\beta_1^2\sum(x_i-\bar{x})^2$

对于 MSE 的期望是 $E(MSE)=E(s^2_{y|x})=\sigma^2_{y|x}$

两个方差相等 ($\sigma_1^2=\sigma_2^2$) 的情况下, 两个方差的比值服从 F 分布 $\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}\sim F_{n_1-1,n_2-1}$

在回归线的真实斜率为零 (原假设下 $H_0:\beta_1=0$) MSR 和 MSE 是 $\sigma^2_{y|x}$ 的检验估计.

因此 $F^*=\frac{MSR}{MSE}\sim F_{p,n-p-1}$ 回归的方与残差的方的比率服从具有 p 和 n-p-1 个自由度的 F 分布

F 检验用于检验包含协变量的模型是否与仅包含截距项的模型相比. 导致残差平方和的显著减少.

(covariate(s))

带有自变量 x 的回归模型 $y=\beta_0+\beta_1 x+\varepsilon$

仅有常数项的模型 $y=\beta_0+\varepsilon$

如果原假设为真 ($H_0:\beta_1=0$) F 比的期望值应该为 1. 如果 H_0 错误, F 比的期望值应该大于 1.

H_0 H_1.

F test of $\beta_1=0$ vs $\beta_1\neq 0$.

Two-sided test $H_0:\beta_1=0$. $H_1:\beta_1\neq 0$.

T-test statistic from before $t^*=\frac{\hat{\beta}_1-0}{s\{\hat{\beta}_1\}}$.

ANOVA statistic $F^*=\frac{MSR}{MSE}$

先进行 F 检验再进行 t 检验.

判断 x 是否显著 判断 β_1 是否显著

t-Test:

$t=\frac{\hat{\beta}_1}{\hat{SE}(\hat{\beta}_1)}$ where $t\sim t_{n-p-1}$ under H_0

F-Test.

If $x\sim t_n$, then $x^2\sim F_{1,n}$

Motivation.

The partial F-test is used to in model building and variable selection to help decided if a variable or term can be removed from a model without making the model significantly wrose.

在使得模型不明显变差的情况下，使用F检验用于模型构建和变量选择.

以帮助确定是否可以用于从模型中删除变量或项.

Patial F-test is used to compare Nested model. 用于比较嵌套模型.

Full (or complete) model 包括所有预测因子. predictors of interest.

Reduced model. 完整模型的某个子集 subset

在MLR Model的背景下. 总体F检验: Overall F-test

整组自变量是否对y的预测有显著贡献

$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$

H_1: at least one of the $\beta_k \neq 0$.

Full model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$

Reduced model $y = \beta_0$

偏F检验: Partial F test 比较完整模型和简化模型的SSE.

测试一组变量，添加一些自变量是否显著增加了y的预测，

超过了模型中已存在的其他自变量所实现的预测.

$H_0: \beta_1^* = \beta_2^* = \cdots = \beta_k^* = 0$

H_1: at least one of the $\beta_k^* \neq 0$

Full model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \beta_1^* x_{p+1} + \cdots + \beta_k^* x_k$

Reduced mode $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

对于Reduced Model: 或 简化模型是否会导致模型误差增加，从而降低模型的预测能力. 误差平方和 SSE

对于Full Model: 模型中包含(新的)自变量是否会降低误差并显著提高模型的预测能力 或 是否显著改善了模型.

要更简单的模型

某种程度上，SSE可以用来总结模型的拟合程度.

H_0: No significant difference in SSE of Full and Reduced Models

H_1: Full Model has a significantly lower SSE than the Reduced Model.

F Test Statistic

$$F^* = \frac{(SSE_{Reduced} - SSE_{Full})/k}{MSE_{Full}} \sim F_{k, n-p-k-1}$$

k: change in the numbe of parameters.

越大 ⇒ 两个模型之间的SSE越大

k: 增加了k个自变量x

p: 原来有p个自变量x.

Covariance 协方差 对于线性相关性的度量. Linear dependency.

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

正线性相关. cov 为正且大.

负线性相关 cov 为负且绝对值大.

没有关系 变量之间. cov 为0或趋近于0.

The Correlation Coefficient 相关系数. → 协方差的标准化.

$$r(x,y) = cor(x,y) = \frac{cov(x,y)}{S_x S_y} \quad S_x^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} \quad S_y^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}$$

$-1 \le cor(x,y) \le 1 \quad cor(x,y) = cor(y,x)$

$cor(ax+b, cy+d) = \underbrace{sign(a)sign(c)}_{\text{取 a 和 c 的符号 (±)}} cor(x,y)$

Standardization VS Normalization ?

标准化 归一化

$X \sim N(\mu,\sigma^2)$ → 均值 方差 → 公式 $\frac{X-\mu}{\sigma}$ → 标准正态

保留了原始数据的形状和分布.

将特征的值映射到一个固定区间. 例如 [0,1] 或 [-1,1]

常用方法:

Min-Max: $\frac{(x-min)}{(max-min)}$

Z-score $\frac{(x-\mu)}{max-min}$

Residuals VS Error ?

观测值与模型中 $\hat{y}_i$ 的差值. 观测值与真实值的差值.

Residual $e_i = y_i - \hat{y}_i$

Hypothesis → SLRModel

$f(x) = \beta_0 + \beta_1 x$

$E[\varepsilon_i] = 0$ mean value of the error is 0

$Var(\varepsilon_i) = \sigma^2$ variance error is constant

$E[\varepsilon_i \varepsilon_j] = 0$ observations are independent.

Important features → SLRM odel.

$$E\{y\} = E\{\beta_0 + \beta_1 x + \varepsilon\} = E\{\varepsilon\} + \beta_0 + \beta_1 x = \beta_0 + \beta_1 x$$

最小二乘法. Least Squares

① 设 $\hat{y}_i = \beta_0 + \beta_1 \hat{x}_i$

② 计算 $\bar{x}$. $\bar{y}$

③ 计算 $\beta_1 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$ S_{xx}

④ 计算 $\beta_0 = \bar{y} - \beta_1 \bar{x}$

⑤ 得出 SLRM

Arithmetic Mean 算术平均值.

$Mean = \frac{\sum x_i}{n}$ ← x_i ran data value; ← n sample size

Discrete Random Variable 离散随机变量.

Cumulative probability distribution function → c.d.f.

累积概率分布函数. 取值范围 [0,1] 单调递增

↓

随机变量小于或等于某个特定值的概率.

离散

期望 $E(Y) = \sum_{i=1}^{n} y_i p_i$ μ_Y ← 可用此表示. 加权平均值.

Expected value.

总体方差 Variance.

$Var(Y) = \sigma_Y^2 = E[(Y-\mu_Y)^2] = \sum_{i=1}^{n}(y_i-\mu_Y)^2 p_i$

样本方差 Sample Variance

$S^2_Y = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$ unbiased estimat. ($\bar{y}$: sample mean)

标准差. standard deviation $\sigma_Y = \sqrt{\sigma_Y^2}$ square root of variance.

$Var(Y) = \frac{1}{N}\sum_{i=1}^{N}(y_i-\mu_Y)^2$ 此时. p_i 都一样.

↑ Bessel's correction

除以 n-1 的行为可以提供对总方差的无偏估计.

连续.

$E(Y) = \mu_Y = \int y f(y)\,dy$ 积分(求面积)

$Var(Y) = \sigma_Y^2 = \int (y-\mu_Y)^2 f_Y(y)\,dy$

$Var(X) = E(X^2) - E(X)^2$

Continuous Random Variables 连续随机变量.

Probability density function (pdf) 概率密度函数.

↓

概括连续随机变量 cdf 的导数为 pdf.

各个取值发生的相对概率.

$P(a \le Y \le b) = \int_a^b f_Y(y)\,dy$

$P(-\infty \le Y \le \infty) = 1 \Rightarrow \int_{-\infty}^{\infty} f_Y(y)\,dy = 1$

Distribution 分布.

Normal Distribution 正态分布 $N(\mu, \sigma^2)$ 均值 方差

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

For standard normal distribution 标准正态分布 $N(0,1)$ $\mu=0$ $\sigma=1$

the standard normal PDF 用 φ 表示. $\varphi(c) = \Phi'(c)$

the standard normal CDF 用 Φ 表示 $\Rightarrow \Phi(c) = P(Z \le c), Z \sim N(0,1)$

The Chi-Squared Distribution 卡方分布 (涉到样本类型的假设时需要)

M 平方和主标准正态分布随机变量的总和服从具有 M 自由度的卡方分布.

The sum of M squared independent standard normal distributed random variables follows a chi-squared distribution with M degrees of freedom:

M 个标准正态分布的平方和

$Z_1^2 + \cdots + Z_M^2 = \sum_{m=1}^{M} Z_m^2 \sim \chi_M^2$ with $Z_m \overset{i.i.d}{\sim} N(0,1)$

自由度为 M. 众数为 M-2. 期望为 M 方差 Var 为 2M

for M ≥ 2

$$pdf\ f(x) = \begin{cases} \left[\frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}\right] \cdot x^{(\frac{n}{2}-1)} \cdot e^{-\frac{x}{2}}, & x>0 \\ 0, & \text{其他} \end{cases}$$

未完(分布)

The F Distribution F分布

两个独立的 χ^2 分布式随机变量除以自由度 M 和 N 的比率为F分布

$$\frac{W/M}{V/N} \sim F_{M,N} \text{ with } W \sim \chi^2_M \text{ (卡方分布)}, V \sim \chi^2_N$$

分子自由度 M，表示为 $F_{M,N}$

分母自由度 N

当分母自由度很大，以至于 $F_{M,N}$ 分布可以近似于 $F_{M,\infty}$ 分布

⇓

对于正态总体来说，$W/M \sim F_{M,\infty}$, $W \sim \chi^2_M$

两个总体的方差比较可以用F-分布来检验

$$F = \frac{S_1^2}{S_2^2} = \frac{\frac{\sum(X_1-\bar{X})^2}{n_1-1}}{\frac{\sum(X_2-\bar{X})^2}{n_2-1}}$$

Lec2/Exercise

Bernoulli Distributed random variable 伯努利分布随机变量.

binomial distribution. 伯努利实验中的成功次数服从二项分布

二项分布. $k \sim B(n,p)$

↑ the number of success

$f(k) = P(k) = \binom{n}{k} \cdot p^k (1-p)^{n-k}$.

The Student t Distribution t分布

Z 为 standard normal variate 标准正态变量.

Z 与 W 独立. M 为自由度.

$$\frac{Z}{\sqrt{W/M}} =: X \sim t_M$$ X follows a Student t distribution

t_M 分布取决于 M. M ↑ t分布越像正态分布

当 $M \geq 30$. t分布可近似为正态分布. 有证.

⇓

t_∞ 分布为标准正态分布.

当 $M > 1$: t_M 分布随机变量 X 有期望 $E(X) = 0$, $M > 1$

当 $M > 2$ 有方差 $Var(X) = \frac{M}{M-2}$, $M > 2$

Lecture 6: 线性回归.

CI. 置信区间

对真实总体均值具有一定置信水平的范围估计.

平均响应：通过回归方程来估计.

$$\hat{y}_0 = \beta_0 + \beta_1 x_0$$

在估计平均响应($\hat{y}_0$)的情况下，置信区间可以计算为：

$$\hat{y}_0 \pm t_{n-2,1-\frac{\alpha}{2}} * SE(\hat{y}_0) \Rightarrow \beta_0 + \beta_1 x_0 \pm t_{n-2,1-\frac{\alpha}{2}} * SE(\hat{y}_0)$$

- $\hat{y}_0$：表示被估计的平均响应
- $t_{n-2,1-\frac{\alpha}{2}}$：表示t分布的临界值.
- $SE(\hat{y}_0)$：表示标准误差估计的平均响应.

$$\hat{y}_0 \pm t_{n-2,\alpha/2} \sqrt{S^2 R \left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{(n-1)S_x^2}\right)}$$

S^2R：

$$S^2R = \frac{\sum e_i}{n-2}.$$ (无偏估计).

PI. 预测区间

提供了一个范围的估计值，在该范围内，总体的单个观测值预计会以一定的置信度水平下降.

$$\hat{y}_0 \pm t_{n-2,1-\frac{\alpha}{2}} \cdot SE(y_0).$$

- $\hat{y}_0$：表示给定x_0的y预测值.
- $t_{n-2,1-\frac{\alpha}{2}}$：表示t分布的临界值.
- $SE(y_0)$：表示预测值的标准误差.

区别CI和PI (PPT, 问原话)：

We need to distinguish between PI and CI;

PI is prediction intervals, where we predict outcome of one single experiment, and confidence intervals. (预测单个实验结果，以及置信区间).

CI is where we predict the mean value of future outcomes. (预测未来结果的平均值).

of a Mean Response;

平均响应

均响应?

极大似然.

$$Pr(x|M,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-M)^2/2\sigma^2}.$$

M：均值.

σ：方差.

σ小，曲线更瘦高

σ大，曲线更矮胖.

$N(M, \sigma^2)$.

M, σ^2(确定).

$$L = \left(\frac{1}{\sqrt{2\pi}\sigma^2}\right)^n e^{-\sum(y-\mu_y)^2/2\sigma^2}$$

$$\Rightarrow \frac{\Delta}{\mu} = 0$$

σ, M(确定).

$y = e^{ax^2+bx+c}$

$\ln e^{y} = \ln e^{ax^2+bx+c}$

$\ln e^{y} = b \ln e$;

Hypothesis Test

线性!!!

① on the slope / linearly / SLM ($y = \beta_0 + \beta_1 x$) etc.

没有 β_1，就是常数的了. 所以：Assume

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

H0的reject区域

The rejection region for the null hypothesis is:

要背 ☆

$$\left| \frac{\hat{\beta}_1}{\sqrt{S_R^2 / (n-1) S_X^2}} \right| > t_{n-2,\alpha/2}$$

two-tails

题目中应用 like: 先算 $\left| \frac{\hat{\beta}_1}{\sqrt{\frac{S_R^2}{(n-1)S_X^2}}} \right| = \cdots = \cdots > ?< t_{n-2,\alpha/2}$

② 方法二：

同样，如果 the value zero is outside of the CI for β_1

at $1-\alpha$ level (one-tail), we reject the null hypothesis at this level.

The p-value of the test is:

③ $$\text{p-value} = 2\Pr\left(t_{n-2} > \left| \frac{\hat{\beta}_1}{\sqrt{S_R^2/(n-1)S_X^2}} \right| \right)$$

t表越远越好!

if p-value $< \alpha$, we can reject H_0.

~~Hence,~~

↳ Probability value

④ 假设检验 F^*

F^* 分布在 $F(1, n-2)$ when H_0 holds,

将该错误风险控制在 I error 时，F statistic 符合

$F^* \backsim F(1, n-2)$ 分布 ⇒

① if $F^* \leq F(1-\alpha; 1, n-2)$ conclude H_0

② if $F^* > F(1-\alpha; 1, n-2)$ conclude H_1

R语言解读：

Multiple R-squared: 多重决定系数，因变量的比例 e.g. 0.57/22.

57.22% of the variability in y is explained by the linear relationship with x.

Adjusted R-squared: 调整后决定系数 考虑了自变量，我们可以用 Adjusted R 校正.

F-statistic: F统值，如果F-statistic的值 < 0.05，我们认为模型显著的，至少有一个自变量对因变量有影响.

$df = n-(1+p)$ 只要它对应的 p-value < 0.05，模型显著

F-sta...: xx on p and n-(1+p) DF.

*表示小于多少，只要***数 < α level

(P小于0.001 --)

↳ " Significance test "

④

Lecture 13 和 14

GLM 广义线性 (Generalized !)

→ 不止解释一个变量

general linear model

$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$ 误差.

⇒ 正态分布

响应变量 response variable y_i

解释 explanatory x_j, $j = 1 \ldots p$.

我们 assume ϵ_i 是独立且同分布的

$E(\epsilon_i) = 0$

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

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

指数族有 正态、泊松利、泊松等.

⇓

为进行模型拟合和推断,

应该使用最大似然. (MLE) (Maxmum likelihood Estimation)

区别:

General 是 linear in the parameters, assume a random compent that follows a normal distribution

- the range of y is restricted
- identify link function

Generalized allowing for different random distributions, such as Normal, Poisson 等.

non-linear systematic components.

- flexible choice of link functions based on the distribution type.

GLM 的三个组成部分: $c(y, \phi)$ y 和 ϵ

① Random Component ⇒ response varaible (y) and assume a probability distribution for it

这里 y is assumed to follow a distribution from the exponential family

② Systematic ⇒ specify 解释 ϕ what the predictor $(X_1, X_2 \ldots)$ ⇒ 线性 manner ⇒ predictor

or $\beta_1 \quad X_i$

③ Link ⇒ the relation between the mean and expected value of random component and systematic

$\eta = X\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$.

component

linear / non-linear

Link function $g(\cdot)$:

connect x and μ

$g(\mu_i) = \eta_i$

$g^{-1}(\eta_i) = \mu_i$

⇒ 对于正态分布的 general

$\eta_i = \beta_0 + \beta_1 x_1 + \cdots$

link: $g(\mu_i) = \mu_i$

variance $V(\mu_i) = 1$

满足 $g(E(y_i|x_i)) = \sum_{j=1}^{p} \beta_j x_{ij}$ 就是. (or $X\beta$) linear predictor

$g(E[y|X]) = X\beta$.

$\phi \; V(\mu)$ 离散系数 discrete disperson parameter

$E[y_i|x_i] = g^{-1}(X\beta)$

二项 Binomial

n_i 试验次数 p_i 成功概率.

y_i/n_i ⇒ 成功的相对比例

$E[\frac{y_i}{n_i}] = p_i$ ⇒ 期望

$\text{Var}(\frac{y_i}{n_i}) = \frac{1}{n_i} p_i(1-p)$

$V(\mu_i) = \mu_i(1-\mu_i)$

$g(\mu_i) = \log(\frac{\mu_i}{1-\mu_i})$

↓ bθ.

泊松 Poisson

$E[y_i] = \lambda_i$

$\text{Var}[y_i] = \lambda_i$

$g(\mu_i) = \log(\mu_i)$

↓ bθ.

$\|L\|_p^p = \sqrt[p]{\|\beta_1\|^p + \cdots}$

一般来说 对方差的处理是

$$\log(y_i) = \beta_0 + \beta_1 x_1 + \epsilon_i$$

$$\Rightarrow E(\log y_i) = \beta_0 + \beta_1 x_1$$

$$\log E(y_i) = \beta_0 + \beta_1 x_1 \qquad V(\mu) = \mu$$

Transformed-data

$$\log(y_i) = \beta_0 + \beta_1 x_1 + \epsilon_i$$

$$E(\log(y_i)) = \beta_0 + \beta_1 x_1$$

GLMs

$$\log E(y_i) = \beta_0 + \beta_1 x_1$$

$$E(y_i) = e^{\beta_0 + \beta_1 x_1}$$

指数family

$$f(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\}$$

θ: parameter of ~~insts~~ interest

ϕ: ⇒ 和 variance 相关

$$\begin{cases} E[y] = b'(\theta) \\ \text{var}(y) = b''(\theta)\phi \end{cases}$$

($g(\cdot)$ canonical if $\theta_i = \eta_i$)

PDF CLT MLE NT (正态检验)

Gaussian ⇒ linear regression

Binary ⇒ odds ratio interpretation

Poisson ⇒ rate ratio interpretation

Normal Distribution:

$\theta = \mu$ $\quad b(\theta) = \frac{1}{2}\mu^2$

$\phi = \sigma^2$ $\quad c(y,\phi) = -\frac{y^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)$

$E^{\theta}(y) = b'(\theta) = \frac{\partial}{\partial\mu}\frac{1}{2}\mu^2 = \mu$ → 偏微分

$\text{Var}(y) = b''(\theta)\alpha(\phi) = \left(\frac{\partial}{\partial\mu}b'(\theta)\right)\sigma^2 = \sigma^2$

$\alpha(\phi)$

这些可以用链式法则求导 因为是指数

X^T ⇒ 矩阵转置

$\mu = \mu(X) = E(y|X)$ $\quad \mu(X) = X^T\beta$

Link: $g(\mu(X)) = X^T\beta$ $\quad \mu = \mu|X = E(y|X)$

Normal Distribution

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

Binomial Distribution

$$P(y;x) = \binom{n}{y}x^y(1-x)^{n-y}$$

$$\binom{n}{y} = \frac{n!}{y!(n-y)!}$$

n ⇒ trial $\quad y$ ⇒ success

推导: $p(y;x,n) = \binom{n}{y}x^{ny}(1-x)^{n-ny}$

$$= \exp\left(\frac{y\log\left(\frac{x}{1-x}\right) + \log(1-x)}{\frac{1}{n}} + \log\binom{n}{ny}\right)$$

$\theta = \log\left(\frac{x}{1-x}\right)$ $\quad b(\theta) = \log(1+\exp(\theta))$ $\quad \alpha(\phi) = 1/n$

$c(y,\phi) = \log\binom{n}{ny}$

过程: $= \exp\left\{\log\binom{n}{y}x^{ny}(1-x)^{n-ny}\right\}$

$= \exp\{\log\binom{n}{y} + ny\log x + (n-ny)\log(1-x)\}$

$= \exp\{\log\binom{n}{y} + ny\log x + n\log(1-x) - ny\log(1-x)\}$

$= \exp\{\log\binom{n}{y} + n[\log xy - \log(1-x)y + \log(1-x)]\}$

$= \exp\{\log\binom{n}{y} + \log\frac{x}{1-x}y + \log(1-x)\}$

$\because \theta = \log\frac{x}{1-x}$

$e^{\theta} = \frac{x}{1-x}$ $\quad e^{\theta} - e^{\theta}x = x$ $\quad x = \frac{e^{\theta}}{1+e^{\theta}}$ $\quad \therefore 1-x = \frac{1}{1+e^{\theta}}$

$\log(1+e^{\theta})^{-1}$

$= \log(1+\exp$

Poisson

$$P(y;\mu) = \frac{\mu^y e^{-\mu}}{y!}$$

$$= \exp\left(\log\left(\frac{\mu^y e^{-\mu}}{y!}\right)\right)$$

$$= \exp(y\log(\mu) - \mu - \log(y!))$$

$\theta = \log\mu$ $\quad b(\theta) = \exp(\theta) = \mu$

$\alpha(\theta) = 1$ $\quad c(y,\theta) = -\log(y$

$E(y) = b'(\theta) = \exp(\theta) = \mu$

$\text{var}(y) = b''(\theta) = \exp(\theta) = \mu$

Link function: geometric distribution 几何分布

为什么推导:

对数 $-r\log(1-e^{\theta})$

$g(h(\theta))$

$h(\theta) = 1-e^{\theta}$ $\quad g(\mu) = -r\log(\mu)$

① 先求 $g(\mu) = g'(\mu) = -r\left(\frac{1}{\mu}\right)\times u' = -\frac{r}{\mu}$

② 求 $h(\theta)$ 导 $= -e^{\theta}$

③ 链式法则 $f(\theta)' = g'(h(\theta)) \times h'(\theta)$

$= \frac{+r}{1-e^{\theta}} \times (+e^{\theta})$

$= \frac{re^{\theta}}{1-e^{\theta}}$

对log求导

$y = \log_a x$ $\quad (\log a^x)' = \frac{1}{x\ln a}$

$y' = \frac{1}{x\ln a}$ $\quad (\ln x)' = \frac{1}{x}$

$(e^x)' = e^x$ $\quad (re^x)' = re^x$ $\quad a^x = a^x\ln a$

对上下求导 $\left(\frac{v(x)}{u(x)}\right)'$

$= \frac{u(x)v'(x) - u'(x)v(x)}{(u(x))^2}$

概念 12.15

~~[redacted]~~ Partial F-test

Lecture 12 :

感兴趣的预测因子

Full (or complete) model : include all predictors of interest

Reduced model : Some subset of the full model

Partial F-test is used to compare Nested models

~~F~~ statistic 是 _ :

Lecture 15

simplest test 3种似然理论的推导

inferential approach 极限分配系数

Wald Tests ⇒ depend on the estimated coefficients and standard errors

ϕ is known !!!

score ⇒

likelihood ratio ⇒ Compare two nested GLMs, are nonlinear functions of β that must be solved iteratively

优化方法 Optimization Methods

Fisher-scoring Method ⇒ solve Score equations (对于Hessian matrix)

~~Netwon~~

Newton-Raphson Method ⇒ For models with the canonical link

Iteratively ⇒

① start with inital estimate

② calculate working Responses and working weights

③ ~~Calcula~~ Calculate $\beta^{(r+1)}$ by weighted least squares

④ Repeat 2 and 3 till convergence 收敛.

IRLS (Iteratively Re-weighted Least Squares (IRLS))

极大似然：

概念：在所有候选模型中，选择使观察到的数据出现概率最大的模型作为最佳模型。

公式：$\theta = \arg\max_\theta P(D|\theta)$；

所有可能的θ值中选择使似然函数$P(D|\theta)$最大的θ值作为参数θ的估计值。

如果我们想要估计高斯分布的两个参数：

首先在给定的$\{x_1, x_2, \cdots, x_n\}$服从$N(\mu, \sigma^2)$的高斯分布；

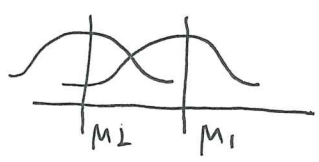
那么在μ和σ^2的所有可能值中，极大似然法会选择使$\{x_1, x_2 \cdots x_n\}$在高斯分布$N(\mu, \sigma^2)$下出现概率最大μ和σ^2作为参数估计值。

~~$f(x_1, x_2, \cdots x_n | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\{(-\frac{\sum(x_i-\mu)^2}{2\sigma^2})$~~

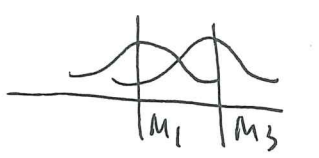
$$Pr(x_i | \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

two parameters

① μ：确定正态分布均值的位置。



$\mu_2 < \mu_1$；较小的μ值将分布的均值向左移动。



$\mu_1 < \mu_3$；较大的μ值将分布的均值至左边。

② σ：标准差。

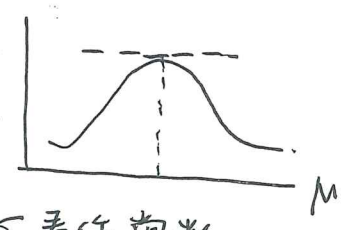
σ较小：正态曲线更高更窄。

σ较大：正态曲线更低更宽。

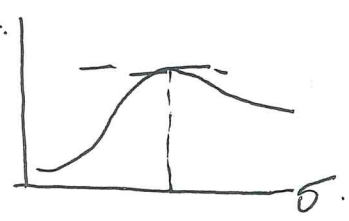
现在来计算正态分布的似然：

$$L(\mu, \sigma | x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \rightarrow \text{求出到最佳参数，} \mu \text{和} \sigma.$$

μ和L的关系：



σ和L的关系：

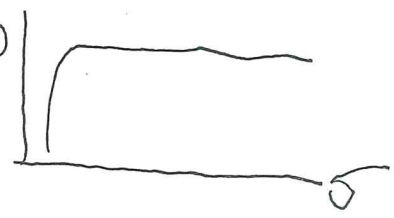


当求解μ的时候，把σ看作常数：

$$L(\mu, \sigma | x_1, x_2, \cdots, x_n) = L(\mu, \sigma | x_1) \times \cdots L(\mu, \sigma | x_n).$$

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

σ和log(L)的关系：



$L' = 0$时，找到L的最大值时μ是多少。

求导前① $\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\right)$.

化简：step 1) $\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} \times e^{-(x_1-\mu)^2/2\sigma^2}\right)$

2) $\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \ln\left(e^{-(x_1-\mu)^2/2\sigma^2}\right)$.

3) $\ln\left[(2\pi\sigma^2)^{-\frac{1}{2}}\right] - \frac{(x_1-\mu)^2}{2\sigma^2}\ln(e)$.

4) $-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x_1-\mu)^2}{2\sigma^2}$

5) $= -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^2 - \frac{(x_1-\mu)^2}{2\sigma^2}$

6) $-\frac{1}{2}\ln(2\pi) - \ln\sigma - \frac{(x_1-\mu)^2}{2\sigma^2}$

对总体 ⟹

$\ln[L(\mu,\sigma|x_1 \cdots x_n)]$

$= -\frac{1}{2}\ln(2\pi) - \ln(\sigma) - \frac{(x_1-\mu)^2}{2\sigma^2} - \cdots - \frac{1}{2}\ln(2\pi) - \ln(\sigma) - \frac{(x_n-\mu)^2}{2\sigma^2}$

有n个$(-\frac{1}{2}) \Rightarrow (-\frac{n}{2})$.

有n个$(-\ln(\sigma))$.

$= -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{(x_1-\mu)^2}{2\sigma^2} \cdots - \frac{(x_n-\mu)^2}{2\sigma^2}$

求导（σ为常数）.

这个是化简后我们要求导的表式：

$\frac{\partial}{\partial\mu}\ln[L(\mu,\sigma|x_1 \cdots x_n)]$

$= 0 - 0 + \frac{2(x_1-\mu)}{2\sigma^2} + \cdots + \frac{x_n-\mu}{\sigma^2}$.

$= \frac{(x_1+\cdots x_n) - \mu n}{\sigma^2}$.

$= \frac{1}{\sigma^2}\left[(x_1+\cdots x_n) - \mu n\right]$.

求导（μ为常数）.

$\frac{\partial}{\partial\sigma}\ln[L(\mu,\sigma|x_1 \cdots x_n)]$.

$= 0 - \frac{n}{\sigma} - \underbrace{\frac{(x_1-\mu)^2}{2}}_{\text{常数}}\sigma^{-2} - \cdots \frac{(x_n-\mu)^2}{2}\sigma^{-2}$.

$= -\frac{n}{\sigma} + \frac{(x_1-\mu)^2}{2}2\sigma^{-3} + \cdots + \frac{(x_n-\mu)^2}{2}2\sigma^{-3}$.

$= -\frac{n}{\sigma} + (x_1-\mu)^2\sigma^{-3} + \cdots + (x_n-\mu)^2\sigma^{-3}$.

$= -\frac{n}{\sigma} + \sigma^{-3}\left[(x_1-\mu^2) + \cdots + (x_n-\mu)^2\right]$.

求导公式：

$(\log_a x)' = \frac{1}{x\ln a}$

$(a^x)' = a^x\ln a$.

$(f\cdot g)'(x) = f(x)g'(x) + g(x)'f(x)$.

$\frac{f(x)'}{g(x)'} = \frac{f(x)'g(x) + f(x)g(x)'}{g(x)^2}$.

$\frac{\partial}{\partial\mu}\ln[L(\mu,\sigma|x_1,\cdots x_n)] = \frac{1}{\sigma^2}[(x_1+\cdots+x_n) - n\mu]$

$\frac{\partial}{\partial\sigma}\ln[L(\mu,\sigma|x_1,\cdots x_n)] = -\frac{n}{\sigma} + \frac{1}{\sigma^3}[(x_1-\mu)^2+\cdots+(x_n-\mu)^2]$

⇓

为了找最大L；令$\frac{\partial}{\partial\mu}\ln = 0$.

$\frac{1}{\sigma^2}[(x_1+\cdots+x_n) - n\mu] = 0$.

$\mu = \frac{x_1+\cdots+x_n}{n}$

因此μ的最大似然估计是测量值的平均值.

令$\frac{\partial}{\partial\sigma}\ln = 0$

$-\frac{n}{\sigma} + \frac{1}{\sigma^3}[(x_1-\mu)^2+\cdots+(x_n-\mu)^2] = 0$.

$-n + \frac{1}{\sigma^2}[(x_1-\mu)^2+\cdots+(x_n-\mu)^2] = 0$.

$\frac{1}{\sigma^2}[(x_1-\mu)^2+\cdots+(x_n-\mu)^2] = n$.

$[(x_1-\mu)^2+\cdots+(x_n-\mu)^2] = n\sigma^2$

$\sigma^2 = \frac{\sum(x_i-\mu)^2}{n}$.

σ的最大似然估计是测量值的标准差.

19

Bayesian Linear regression model

Data: $D = (x_1, y_1) \dots (x_n, y_n)$

Predict: y_* for a new unseen X_*

① Likelihood: ~~Defind~~ Define the likelihood

$$p(Y \mid \theta, X)$$

② Prior: Define the prior $p(\theta)$

③ Learning: Do inference by applying Bayes's therom

$$p(\theta \mid Y, X) \propto p(Y \mid \theta, X) p(\theta)$$

④ Prediction: compute predictive distribution by marginalizing 边缘计算.

$$p(y_* \mid x_*, Y, X) = \int p(y_* \mid \theta, x_*) p(\theta \mid Y, X) d\theta$$

后验 posterior

$y_i = x_i^T \beta + \epsilon_i$ ⇒ ~~BLM~~ BLR

$\epsilon_i \sim N(0, r^{-1})$ $r = \sigma^{-2}$ $\beta \sim p(\beta)$

会给怎么算成？

$$\begin{cases} p(Y \mid \beta) = N(Y; X\beta, \sigma^2 I) & \text{likelihood} \\ p(\beta) = N(\beta; \mu_0, \Sigma_0) & \text{prior} \end{cases}$$

⇒ $p(\beta \mid Y)$ ⇒ μ_n, Σ_n 都在公式纸上.

VS

predictive 预测

求的对 (y_*, X_*)

$p(y_* \mid \beta) = N(y_*; X\beta, \sigma^2 I)$

$p(\beta \mid Y) = N(\beta; \mu_n, \Sigma_n)$ 也是一样的，除了 →

$p(y_* \mid Y) = N(y_*; \mu_*, s_*)$

$\mu_* = X_*^T \mu_n$

$s_* = \sigma^2 + X_*^T \Sigma_n X_*$

Regularization least squares fitting procedure

$$RSS = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} \beta_j x_{ij} \right)^2$$

过拟合 ⇒ 在训练模型上很好

在其它测试集上表现很差.

Regularization: allows complex models to be trained by data sets of limited size without severe overfitting, essentially by limiting the effective model complexity

在有限的数据集上训练复杂模型 without overfitting

↓ by 限制 effective mode complexity

~~Ridge regularization~~

Lasso: ⇒ 线性回归方法 L1 正则加入.

Bayesian Linear Regression

学习参数模型 就是找到 最适合模型 训练的值

OLS
MLE

⇒ unknown and fixed

The data were generated randomly

Frequentist approach

$\alpha > 0$ $p \sim Beta(\alpha, \alpha)$ ⇒ prior

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

D ↓ 观测数据

$$P(y) = \int p(y, \theta) d\theta = \int p(y|\theta)\, p(\theta)\, d\theta$$

$P(\theta)$ prior ⇒ before collected data.

$P(\theta|D)$ posterior ⇒ after inferring data.

$P(D|\theta)$ likelihood of data.

$P(D)$ the marginal likelihood

$P(y^*|y)$ posterior predictive

$p(\theta|y)$ posterior $p(y)$ marginal likelihood
$p(\theta)$ prior

example 得病
90% 的人吃汉堡.
100,000 之一的人有病
假设一半的人吃汉堡.

⇒ 求一个吃汉堡的人患病可能性.

① 设 KJ = 有kj病 H = 吃汉堡.

$P(H = yes \mid kj = yes) = 90\%$

$P(ky = yes) = 0.001\%$

$P(H = yes) = 50\%$

$P(kj = yes \mid H = yes)$

	H=yes	H=No
KJ=Yes	9	1
KJ=No	499991	499999

$$= \frac{P(kj = yes, H = yes)}{P(H = yes)}$$

$$= \frac{P(kj = yes, H = yes)}{P(kj = yes, H = yes) + P(kj = no, H = yes)}$$

吃汉堡人用点数

MAP maximum a posteriori (MAP)

Bayesian approach 和 regularisation 正则化

MAP 是 θ 后验 posterior reaches its max.

$$\hat{\theta}_{MAP} = \arg_\theta \max p(\theta|y)$$

$$= \arg_\theta \max p(y|\theta)\, p(\theta)$$

$$= \arg_\theta \max [\ln p(y|\theta) + \ln p(\theta)]$$

$$p(\theta|y) = \frac{p(y|\theta)\, p(\theta)}{p(y)}$$

L^2 ⇒ 正则 regularised linear regression

$$\hat{\theta}_{L_2} = \arg_\theta \max [\ln p(y|\theta) + \ln p(\theta)]$$

当 $p(\theta) \propto \|\theta\|_2^2$ 时，θ的 MAP 和 L^2 的正则化 regularised 估计相同

$p(\theta)$ in $N(\beta; \mu_0, \Sigma_0)$