贵州统计学习

—Summer School

— Regression

居梦婷

主讲人介绍

- Eric Feigelson
- Penn State University
- 最早的天文统计学发起者
- 目前主要研究基于参数自回归建模的系外行星检测统计方法

Introduction

An astrostatistics lexicon …

Cosmology	↔	**Statistics**
Galaxy clustering	↔	Spatial point processes, clustering
Galaxy morphology	↔	Regression, mixture models
Galaxy luminosity fn	↔	Gamma distribution
Power law relationships	↔	Pareto distribution
Weak lensing morphology	↔	Geostatistics, density estimation
Strong lensing morphology	↔	Shape statistics
Strong lensing timing	↔	Time series with lag
Faint source detection	↔	False Discovery Rate
Multiepoch survey lightcurves	↔	Multivariate classification
CMB spatial analysis	↔	Markov fields, ICA, etc
ΛCDM parameters	↔	Bayesian inference & model selection
Comparing data & simulation	↔	*Uncertainty Quantification*

报告内容

1. Density estimation & Local regression
2. Fundamentals of statistical inference
3. Regression
4. Multivariate clustering & classification
5. Bayesian inference
6. Censoring & truncation
7. Time series analysis

Regression

- 两个主要的不同（与Density estimation相比）

1. 在自变量X 和应变量Y 之间有个假定的关系存在，并且这个关系要说的通（取决于X）

2. 得到一个参数化关系，根据X,Y可以得到最佳的参数（天文上的方程通常是根据天文理论得到）

Regression

Classical regression model:

$$E[Y|X] = f(X, \theta) + \epsilon$$

X： 自变量

Y： 应变量

Θ： 参数

ϵ： 随机误差

- The `error’ e is commonly assumed to be a normal (Gaussian) i.i.d. random variable with zero mean, e = N(0,s2). Note that all of the randomness is in this error term; the functional relationship is deterministic with a known mathematical form.

Warning

- 天文上经常用的那些经典参数回归方程，可能只是相比于其他的这些方程更加熟悉而已。

1. 如果没有理论方程的支持的话，density estimation 可能会更合适（经验公式）
2. 如果没有依赖关系，XY没有自变量应变量的说法了 (e.g. OLS bisector, orthogonal regression, Principal Component Analysis).

ϵ

- 数据上的误差‘structural regression model’.
- 统计方法上的误差`functional regression model’.
- 两个都有

参数估计和模型选择

Once a mathematical model is chosen, and a dataset is provided, then the `best fit' parameters are estimated by one (or more) of the techniques discussed in MSMA Chpt. 3:

- Method of moments
- Ordinary least squares (OLS, L_2)
- Least absolute deviation (L_1)
- Maximum likelihood estimation (MLE)
- Bayesian inference

方法

Seek balance between model complexity and parsimony (Occam's Razor):

- Does the ΛCDM model have a w-dot term?
- Are three or four planets orbiting the star?
- Is the star cluster an isothermal sphere or ellipsoid?

Choice of model form and complexity is called `model selection'. Methods include: χ^2_ν, BIC, AIC, …

The final model should be validated against the dataset (or other datasets) using goodness-of-fit test (e.g. Anderson-Darling test with bootstrap resamples for significance levels) and residual analysis.

Regression

- Linear指的是参数的线性，不是指X

Examples of linear regression functions:

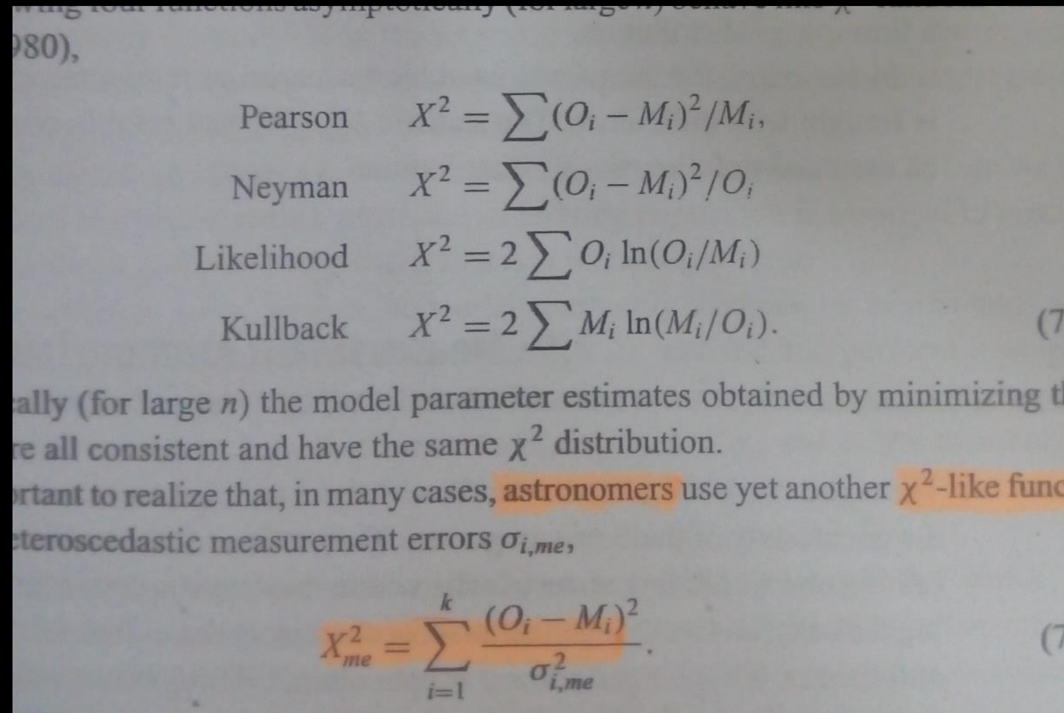
$Y = \beta_0 + \beta_1 X + \epsilon$	1st order polynomial	Linear
$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$	high order polynomial	
$Y = \beta_0 e^{-X} + \epsilon$	exponential decay	
$Y = \beta_0 + \beta_1 \cos X + \beta_2 \sin X + \epsilon$	periodic sinusoid with fixed phase	

Examples of non-linear regression functions:

$Y = \left(\frac{X}{\beta_0}\right)^{-\beta_1} + \epsilon$	power law (Pareto)	Non-linear
$Y = \frac{\beta_0}{1 + (X/\beta_1)^2} + \epsilon$	isothermal sphere	
$Y = \beta_0 + \beta_1 \cos(X + \beta_2) + \beta_3 \sin(X + \beta_2) + \epsilon$	sinusoid with arbitrary phase	
$Y = \begin{cases} \beta_0 + \beta_1 X & \text{for } X < x_o \\ \beta_2 + \beta_3 X & \text{for } X > x_o \end{cases}$	segmented linear	

Regression

- 最小二乘
- 卡方最小
- 最大似然法



980),

$$\text{Pearson} \quad X^2 = \sum (O_i - M_i)^2 / M_i,$$
$$\text{Neyman} \quad X^2 = \sum (O_i - M_i)^2 / O_i$$
$$\text{Likelihood} \quad X^2 = 2 \sum O_i \ln(O_i / M_i)$$
$$\text{Kullback} \quad X^2 = 2 \sum M_i \ln(M_i / O_i). \quad (7.$$

ally (for large n) the model parameter estimates obtained by minimizing th
e all consistent and have the same χ^2 distribution.
rtant to realize that, in many cases, astronomers use yet another χ^2-like funct
eteroscedastic measurement errors $\sigma_{i,me}$,

$$X^2_{me} = \sum_{i=1}^{k} \frac{(O_i - M_i)^2}{\sigma^2_{i,me}}. \quad (7.$$

A better approach uses a more complicated likelihood that includes the measurement errors & model error, and proceeds with MLE or Bayesian inference. See important article by Brandon C. Kelly, ApJ 2007

还有的问题

But poor practice does occur:

- Overuse of heuristic models
- Ill-defined response variable
- Improper used of `minimum chi-squared’ method
- Inadequate model selection
- Inadequate residual analysis
- Overuse of Bayesian inference with uninformative priors