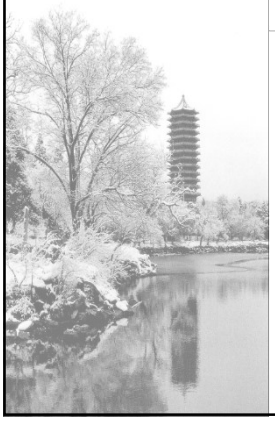


应用数理统计方法

第五章 相关与回归分析

单变量与多变量研究

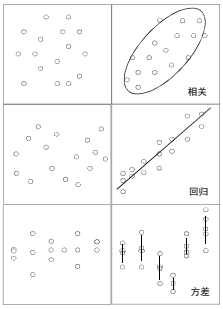
p.

- **总体与变量**
  - 总体: 研究总体中的全部个体
  - 变量: 研究者关注的总体性质
- **单变量研究**
  - 参数估值 总体大小, 离散, 分布特征表征
  - 假设检验 总体大小, 离散, 分布特征比较
- **多变量研究**
  - 方差分析 影响因素研究 假设检验
    从两总体大小比较到方差分析 – 引入影响因素概念
  - 相关分析 变量共变关系研究 参数估值与假设检验
  - 回归分析 预测与估值模型 参数估值与假设检验

多变量研究方法对比

p.229,292

- **相关分析**
  - 共变关系 两变量等价
- **回归分析**
  - 模型构建 根据自变量估计或预测因变量
- **方差分析**
  - 影响因素 判断变量是否受某些因素影响



多变量研究方法的关系

p.229,292-242

- **相关分析与回归分析**
  - 共同之处 研究两个或多个变量间关系, 数据格式相似
  - 不同之处 研究目的, 变量关系, 变量类型, 统计量, 检验方法
- **方差分析与回归分析**
  - 共同之处 一变量对另一个变量的依赖, 变量类型
  - 不同之处 研究目的, 变量关系, 结果表达, 统计量, 检验方法, 方法应用

方法	相关分析	回归分析	方差分析
研究目的	变量一起变化的程度	构建回归模型 – 估值或预测	研究影响因素
变量关系	两个独立变量	自变量与因变量	独立变量与影响因素
变量类型	均为随机变量	因变量随机, 自变量固定或随机	独立变量随机, 影响因素固定或随机
统计量	无量纲的相关系数	有单位的回归参数	无描述统计量
检验	相关是否显著	斜率是否显著, 模型选择	影响因素是否显著

多变量研究的数据

p.229,292-242

- **方差分析, 相关分析与回归分析数据比较**
  - 方差分析 研究对象为随机变量, 影响因素固定或者随机
  - 相关分析 两个随机变量
  - 回归分析 无重复数据回归: 一个自变量取值对应一个因变量取值
    有重复数据回归: 一个自变量取值对应一个以上因变量取值

相关分析
无重复数据回归
$x_i, y_i,\ i = 1...n$

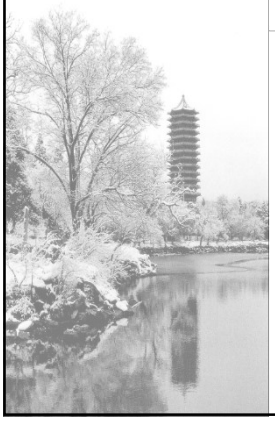
x_1	y_1
x_2	y_2
…	…
x_n	y_n

方差分析
有重复数据回归
$x_i, y_{ij},\ i=1...n, j=1...m_i$

x_1	y_{11}	y_{21}	…	y_{m1}
x_2	y_{12}	y_{22}	…	y_{m2}
…	…		…	
x_n	y_{1n}	y_{2n}	…	y_{mn}

相关与回归分析应用中的常见问题

- **混用相关分析与回归分析**
  - 研究目的不明确 没有区分共变与预测
- **忽视变量的分布特征**
  - 对非正态分布数据使用参数相关 可能导致错误结论
  - 回归分析中忽视数据分布特点 可能导致结果偏差
- **不区分因果关系与间接相关的差别**
  - 对显著相关的过度解释 错误解释相关结果, 导致对因果的误判
- **错误理解回归分析的检验结果**
  - 无重复数据的回归仅能检验斜率 不能判断模型优劣



应用数理统计方法

第五章 相关与回归分析

假设检验问题

p.63

- **特征比较**
  - 两总体或多总体大小比较 2.2 – 2.5
  - 两总体或多总体离散程度比较 3.1 – 3.2
  - 两总体分布特征比较 3.4
  - 总体分布是否服从特定理论分布 3.3 – 3.4
- **影响因素**
  - 方差分析及补充分析 4.1 – 4.4
- **变量关系**
  - 相关分析 **5.1 – 5.2**
  - 回归分析 5.3 – 5.4

变量间相关关系

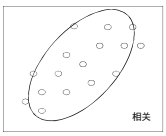
p.295

- **直接相关**
  - 变量间有因果关系 如价格与销量, 暴露剂量与毒性
  - 因果关系需非统计证据 如苯并芘致癌的流调, 动物实验, 临床, 机理 … 研究
- **间接相关**
  - 受共同因素影响的共变 如人均收入与癌症发病率, 啤酒消费量与GDP
  - 避免过度解释, 环境健康研究中常见的关联
- **假相关**
  - 互为衍生关系 如清洁能源占比与煤占比, 粘粒与沙粒占比
  - 没有意义

相关分析方法选择

p.296-297

- **研究目的**
  - 确认探讨共变关系 属于相关问题？
- **变量**
  - 确认均为随机变量 相关分析的必要前提
  - 确定变量个数 二元或多元相关？
  - 确定变量关系 是否为简单线性关系？或需要作数据变换？
- **总体特征**
  - 确认分布类型 参数或非参数方法
  - 是否需要作正态变换？



相关分析方法

p.296

- **参数与非参数方法**
  - 参数方法 正态分布数据
  - 非参数方法 非正态分布数据
- **二元与多元**
  - 二元相关 两变量共变
  - 多元相关 多变量共变
- **其它**
  - 列联表分析 类型变量关联
  - 快速方法 简便, 但不严格

方法	二元相关		多元相关
参数方法	**Pearson** 相关系数	5.1.1	偏相关系数 复相关系数 5.2 典型相关分析 5.2
非参数方法	**Spearman** 秩相关系数	5.1.2	Kendall 偏秩相关系数
	Kendall秩相关系数		Kendall 和谐系数
	双向列联表分析	5.1.2	多向列联表分析
	象限相关分析		
	Lomstead-Tukey 偶角检验		

5.1.1

5.1.1 Pearson 相关系数及其显著性检验

5.1.2 Spearman 秩相关系数

二元正态分布

p.298

- **二元正态分布**

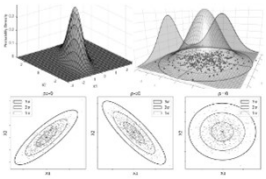
  两个正态分布总体的联合概率密度函数

  Pearson 相关分析的必要前提

- **投影**

  与 **x-z** 或 **y-z** 平面平行的剖面上的切面为正态分布

  在 **x-y** 平面投影为椭圆或圆



二元参数相关

p.296

- **Pearson 相关系数**

  二元参数相关方法

方法	二元相关		多元相关	
参数方法	**Pearson** 相关系数	5.1.1	偏相关系数	5.2
			复相关系数	5.2
			典型相关分析	
非参数方法	**Spearman** 秩相关系数	5.1.2	Kendall 偏秩相关系数	
	Kendall秩相关系数		Kendall 和谐系数	
	双向列联表分析	5.1.2	多向列联表分析	
	象限相关分析			
	Lomstead-Tukey 偶角检验			

Pearson相关系数

p.298-301

- **数据**

  服从二元正态分布

- **Pearson 相关系数**

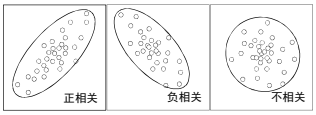
  表征两个随机变量的共变程度　一个发生变化, 另一个随之变化, 反之亦然

  取值在 -1 到 1 之间　从负相关 –1 到不相关 **0**, 再到正相关 +1

- **正相关与负相关**

  正相关: 两变量共变方向相同

  负相关: 两变量共变方向相反



偏离二元正态分布的可能影响

p.299

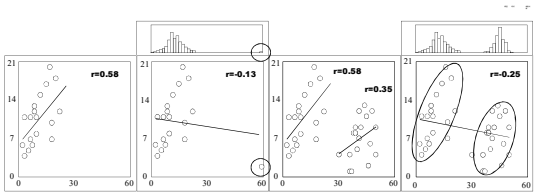
- **二元正态分布**

  Pearson 相关检验的必要前提

- **偏离二元正态分布的影响举例**

  个别异常的影响　剔除异常

  多峰分布的影响　拆分总体



相关系数计算

p.299

- **Pearson 相关系数计算**

  包括所有点 x_i, y_i 到 　、 　均值的距离

- **调整相关系数**

  修正样本量过小导致的样本相关系数对总体相关系数的偏差

- **相关系数取值**

  -1 完全负相关, **1** 完全正相关, **0** 完全不相关

  需要检验与 **0** 是否有显著差异

相关系数的显著性检验

p.299-300

- **相关系数检验**

  总体相关系数是否与 **0** 有显著差别 即是否显著相关

  与偏度-峰态系数检验类比

- **假设检验**

  双侧检验 $H_0\ \rho = 0,\ H_1\ \rho \neq 0$

  单侧检验 $H_0\ \rho = 0,\ H_1\ \rho < 0$　已知不可能为正相关

  $H_1\ \rho > 0$　已知不可能为负相关

- **判断**

  直接判断 $p < \alpha$　或 $p < 2\alpha$　单侧检验

  间接判断 $|r| > r_{\alpha[v]}$ 或 $|r| > r_{2\alpha[v]}$　单侧检验

  如果显著, 再根据计算值的正负号区分正相关/负相关

相关系数比较与公共相关系数

不要求 p.303-305

- **相关系数比较**

  比较两个或多个相关系数是否有显著差异

  类比大小比较 $H_0\ \mu_1=\mu_2,\ H_1\ \mu_1\neq\mu_2$

- **检验**

  两个相关系数比较 $H_0\ \rho_1=\rho_2,\ H_1\ \rho_1\neq\rho_2$　　可选择单侧或双侧检验

  多个相关系数比较 $H_0\ \rho_i$ 都相同, $H_1\ \rho_i$ 不都相同 无单双侧之分

- **公共相关系数**

  如果检验结果无显著差异, 可计算公共相关系数 计算略

5.1.2

5.1.1 Pearson 相关系数及其显著性检验

5.1.2 Spearman 秩相关系数

非参数相关检验方法

p.297

- **非参数相关检验**

  非二元正态分布总体　且不能做正态变换

  非线性关系　　且不能做线性变换

  仅获得秩数据　　数据变换的单向性

- **秩相关系数**

  Spearman 和 **Kendall** 检验　类似, 功效效率 **91%**

- **其它方法**

  列联系数 类型变量相关

  快速检验 简便, 不严格

方法	二元相关		多元相关	
参数方法	**Pearson** 相关系数	5.1.1	偏相关系数	5.2
			复相关系数	5.2
			典型相关分析	
非参数方法	**Spearman** 秩相关系数	5.1.2	Kendall 偏秩相关系数	
	Kendall秩相关系数		Kendall 和谐系数	
	双向列联表分析	5.1.2	多向列联表分析	
	象限相关分析			
	Lomstead-Tukey 偶角检验			

Spearman 秩相关系数

p.308-310

- **应用**

  非二元正态分布数据　可尝试正态变换后用参数方法

  非线性关系数据　　可尝试线性变换后用参数方法

- **假设**

  双侧检验 H_0: 两总体不相关, H_1: 两总体相关

  单侧检验 H_0: 两总体不相关, H_1: 两总体正相关或负相关

- **计算与判断**

  计算相伴概率 p 或 秩相关系数 r_s

  直接判断 $p<\alpha$　　或 $p<2\alpha$　单侧检验

  间接判断 $r_s\geq r_{\alpha[1,n-2]}$ 或 $r_{2\alpha\,[1,n-2]}$　单侧检验

非线性相关

- **线性相关**

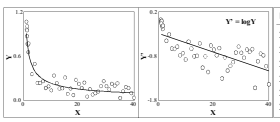
  计算 **Pearson** 相关系数

- **非线性相关**

  选择一 线性变换后, 计算 **Pearson** 相关系数

  选择二 直接用非参数方法

- **线性化举例**



曲线方程	线性化方法	线性化方法
$y=(a+bX)/X$	$y'=Xy$	$y'=a+bX$
$y=1/(a+bX)$	$y'=1/y$	$y'=a+bX$
$y=X/(1+bX)$	$y'=X/y$	$y'=a+bX$
$y=ae^{bX}$	$y'=\ln y$	$y'=\ln a+bX$
$y=aXe^{bX}$	$y'=\ln(y/X)$	$\ln(y/X)=\ln a+bX$
$y=aX^b$	$y'=\ln y,\ X'=\ln X$	$y'=\ln a+bX'$

双向列联表分析

不要求 p.313-315

- **双向列联表**

  X 与 Y 两个类型变量的频数表 f_{ij}

- **双变量列联表分析**

  检验两个类型变量的相关性, 即非独立性　简称双向表分析

- **方法**

  假设: H_0: 两变量有独立性; H_1: 两变量不具有独立性

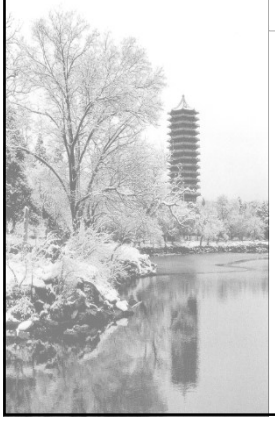
  计算: 两变量独立条件下的理论频数 $\hat{f}_{ij}$,

  根据观测频数 f_{ij} 和理论频数 $\hat{f}_{ij}$ 计算列联系数 **C**

  检验: 同拟合度卡方检验方法 参见**3.4.2**

  拒绝: $p<\alpha$ 或 $G\geq\chi^2_{\alpha[1]}$

X \ Y	1	2	…	*k*
1	f_{11}	f_{12}	…	f_{1k}
2	f_{21}	f_{22}	…	f_{2k}
…				
r	f_{r1}	f_{r2}	…	f_{rk}



应用数理统计方法

第五章 相关与回归分析

多元参数相关

p.296

- **偏相关与复相关**

  类似 **Pearson** 相关系数 涉及两个以上变量, 正态假设, 线性关系

  偏相关: 在消除其他变量影响的前提下, 考察两个变量间共变关系

  复相关: 考察一个变量与一组变量间的共变关系

- **典型相关**

  代表两组变量的两个综合变量间的相关关系 多元分析内容之一

方法	二元相关		多元相关
参数方法	**Pearson** 相关系数	5.1.1	偏相关系数 5.2 复相关系数 5.2 典型相关分析
非参数方法	**Spearman** 秩相关系数	5.1.2	Kendall 偏秩相关系数
	Kendall秩相关系数		Kendall 和谐系数
	双向列联表分析	5.1.2	多向列联表分析
	象限相关分析		
	Lomstead-Tukey 偶角检验		

偏相关系数

p.320-321

- **偏相关**

  研究 k 个变量间的相关关系 排除其它因素影响的两变量相关关系

  固定 $v_3 \dots v_k$ 的前提下, v_1 - v_2 间共变 相当于固定其它条件的实验设计

  记为 $r_{\mathbf{12}.3..k}$ 如 $r_{12.3}$, $r_{12.34}$ 共3个或4个变量

  取值在 -1 到 +1 之间 与二元相关相同

- **检验**

  假设 $H_0\ \rho_{12.3..k} = 0$, $H_1\ \rho_{12.3..k} \neq 0$ 单侧 $\rho_{12.3..k} > 0$, $\rho_{12.3..k} < 0$

  拒绝条件 $p < \alpha$ 或 $|r| \geq r_{\alpha[k,v]}$ 单侧 $p < 2\alpha$ 或 $|r| \geq r_{2\alpha[k,v]}$

  如果显著, 再根据计算值的正负号确定正相关或负相关

复相关系数

p.322-323

- **复相关**

  研究一个变量与一组变量间共变 可依次研究每个变量与其它所有变量间关系

  研究 v_1 与 $v_2 \dots v_k$ 的共变 局限性

  记为 $R_{\mathbf{1}.2..k}$ 如 $R_{1.2..3}$, $R_{1.2..4}$

  取值在 0 到 1 之间 无正负相关之分

  局限性 不能区分变量 2 .. k 的独立贡献

  无单侧检验

- **检验**

  假设 H_0 复相关系数与 0 无显著差别, H_1 与 0 有显著差别

  拒绝条件 $p < \alpha$ 或 $R \geq r_{\alpha[k,v]}$

Kendall 偏秩相关系数

p.325-326

- **偏秩相关**

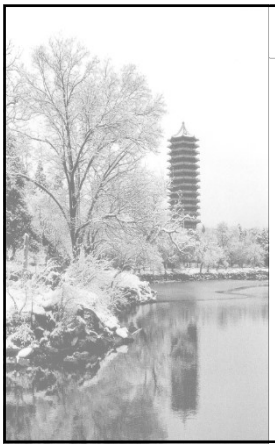
  从 Kendall 秩相关系数演变而来

  不要求多元正态分布, 不要求线性关系

  表达类似偏相关系数, 如 $\tau_{12.3}$

- **方法局限**

  无显著性检验手段, 因此用途有限



应用数理统计方法

第五章 相关与回归分析

5.1 二元相关分析
5.2 多元相关分析
应用举例
5.3 一元线性回归
5.4 其他回归方法
应用举例

应用实例 巧克力与诺贝尔奖的关系

Messerli, New Eng. J. Med. 2015

- **问题与方法**

  背景: 可可中的黄烷醇可有效减缓年龄增长导致的认知功能下降

  人均诺贝尔奖总数可在一定程度上衡量国家的整体认知功能？

  研究: 研究巧克力消费与人均诺奖获奖人数的相关关系

- **检验与结果**

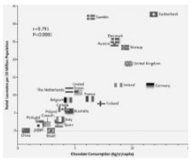
  假设: 人均巧克力消费量和人均诺奖获奖人数无关

  结果: 检验结果显著

  结论: 巧克力消费与整体认知水平正相关

  讨论: 典型的间接相关 无意义　隐含因果关系

  经济, 社会, 教育, 投入, 积累 …



应用实例 影像分辨率对景观格局与地表温度关系的影响

Li et al., LUP, 2013

- **问题与方法**

  问题:地表温度与景观格局指标间关系　不同分辨率遥感影像 30m, 10m, 2.44m

  绿地覆盖率, 斑块面积, 斑块密度, 景观形状, 边界密度, 形状指数, 邻域距离

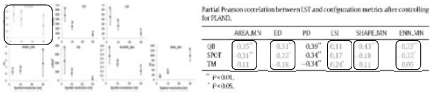
  研究: 数据分辨率对地表温度与相关指标相关关系的影响

- **检验与结果**

  计算: 偏相关系数 排除指标间相互影响, 比较不同分辨率结果

  发现: 分辨率影响大部分指标, 影响方向各异

  如: 分辨率越高, 绿地覆盖率与地表温度相关越显著, 高分辨图像可以识别较小的斑块



Partial Pearson correlation between LST and configuration metrics after controlling for PLAND.

	AREA_MN	ED	PD	LSI	SHAPE_MN	ENN_MN
QB	−0.35**	0.32*	0.39**	0.41*	0.43*	0.37*
SPOT	−0.31**	0.33*	0.34**	0.41*	0.38*	0.36*
TM	−0.11	0.18	−0.14*	0.15	0.16	0.09

** P<0.01.
* P<0.05.

应用实例 牙周炎指标与口腔指标的关系

Kim et al., KAOMS, 2021

- **问题与方法**

  问题: 影响牙周炎的主要口腔指标 **BOP**是否与牙周炎有关

  研究: 募集 **84** 例牙周炎患者测定 **sRAGE** 等牙周炎指标及 **BOP** 等口腔指标

  按 5 mm牙周探测深度分两组

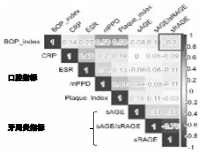
  研究牙周炎指标与口腔指标的关系 **检验与结果**

  结果: 口腔指标与测定参数无显著相关关系 如 探诊出血

  ≥5 mm 组的 BOP/sRAGE 偏相关系数 $p < 0.05$

  调整参数包括: mPPD, PI, sAGE, ESR

  讨论: 细分总体可揭示某些规律



应用实例 黄酒酸度与酿酒红曲米特征关系

林等, 中国食品学报, 2019.

- **问题与方法**

  问题: 红曲米特性对黄酒酸度的影响 pH,干失重,容重,淀粉,蛋白,氨基酸,液化力

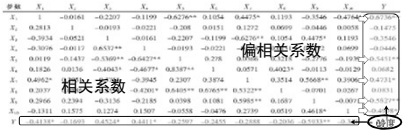
  研究: 相关与偏相关分析　构建相关系数与偏相关系数矩阵, 含各因素间关系

- **检验与结果**

  结果 影响酸度的主要影响因子包括　pH, 蛋白质, 发酵力 等

  讨论: 特性参数间相关

  偏相关系数能更好反映主要影响因素



应用实例 构建评价刊物引用的综合指标

Bradshaw and Brook, Plos One 2016

- **研究问题与方法**

  背景: 不同引用指标各有特点 IF, IM, SNIP, SJR, h5/log10

  目标: 构建综合评价指标　基于上述指标

  数据: 25个生态学及交叉刊物 κ-再抽样法 vs. 对 188 个科学家调查结果

  统计: 相关分析　排序数据的 **Spearman** 秩相关分析

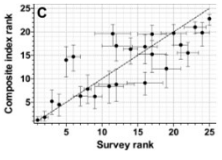
- **检验与结果**

  结果: 显著相关

  纵标: 综合指数平均秩±95% 不确定性;

  横标: 专家调查结果平均秩±标准差

  讨论: 结论？ 单一指标呢？



应用实例 影响伊春河河水溶解态有机碳的因素

Tao et al., Water Res., 1998

- **问题**

  背景: 伊春河河水富含溶解态有机碳 DOC 10-15 mgC/L　沿河多沼泽湿地, 冻土

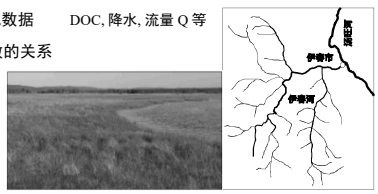
  伊春市主要水源, DOC是消毒副产物的前体物　$CHCl_3$等致癌物

  研究: 影响河水 DOC 含量的因素　预测消毒副产物水平

- **方法**

  数据: 非封冻期日变化数据　DOC, 降水, 流量 Q 等

  分析: DOC和相关参数的关系



应用实例 影响伊春河河水溶解态有机碳的因素

Tao et al., Water Res., 1998

- **发现**

  现象: DOC与流域降水量密切相关 表现为与流量 **Q** 的同步变化

  应用: 用 Q 预测 DOC 很容易通过水位观测获得即时流量 **Q**

- **结果与讨论**

  结果: DOC 与 Q 日数据显著相关 $r = \mathbf{0.76}$, $p < 0.001$

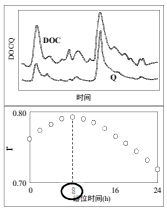
  分析: DOC 变化似乎滞后, 向后错位计算相关系数

  错位 **8** h, 相关系数达峰值 $r = \mathbf{0.79}$

  结论: DOC 变化比 Q 滞后 **8 h** 湿地滞水与输入

  讨论: 潜在应用 – 预测消毒副产物生成量

  然后…



应用实例 室内外颗粒物浓度关系

WHO, 2020

- **研究背景**

  背景: $PM_{2.5}$ 呼吸暴露是最重要的环境风险 2019 导致 **120** 万例过早死亡

  室内暴露主导 成年人平均 **86%** 时间在室内

  室外向室内渗透是室内 $PM_{2.5}$ 主要源 没有强内源的条件下

  错觉: 室外: **280-290** μg/m³, 室内 **190-240** μg/m³ **2015.11.8**, 北大逸夫二楼

- **研究方法**

  观测: 同步观测室内外 $PM_{2.5}$ 浓度 2013/14 取暖季, 低成本在线传感器

  统计: 相关分析

应用实例 室内外颗粒物浓度关系

Han et al., Environ. Pollut, 2015

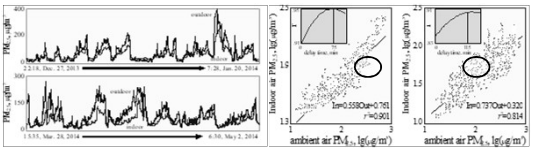
- **结果与检验**

  室内外浓度显著相关, 室内滞后 高/低浓度范围的差别

  Pearson 相关系数, 时间错位 5 min 步长

  冬季平均滞后 **75** min 开窗率低

  春季平均滞后 **115** min



要点

- **多变量方法**

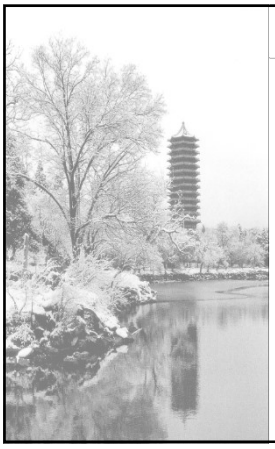
  方差, 相关, 回归 目的, 数据, 变量, 联系 …

- **相关**

  正态与非正态

  线性与非线性

  Pearson, 偏复相关, Spearman 等



应用数理统计方法

第五章 相关与回归分析

5.1 二元相关分析
5.2 多元相关分析
应用举例
5.3 一元线性回归
5.4 其他回归方法
应用举例

多变量研究方法对比 复习

p.229,292

- **相关分析**

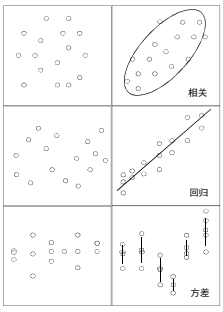
  共变关系 两变量等价

- **回归分析**

  模型构建 根据自变量估计或预测因变量

- **方差分析**

  影响因素 判断变量是否受某些因素影响



多变量研究方法的关系 复习

p.229,292-242

- **相关分析与回归分析**

  共同之处 研究两个或多个变量间关系, 数据格式相似

  不同之处 研究目的, 变量关系, 变量类型, 统计量, 检验方法

- **方差分析与回归分析**

  共同之处 一变量对另一个变量的依赖, 变量类型

  不同之处 研究目的, 变量关系, 结果表达, 统计量, 检验方法, 方法应用

方法	相关分析	回归分析	方差分析
研究目的	变量一起变化的程度	构建回归模型 – 估值或预测	研究影响因素
变量关系	两个独立变量	自变量与因变量	独立变量与影响因素
变量类型	均为随机变量	因变量随机, 自变量固定或随机	独立变量随机, 影响因素固定或随机
统计量	无量纲的相关系数	有单位的回归参数	无描述统计量
检验	相关是否显著	斜率是否显著, 模型选择	影响因素是否显著

假设检验问题

p.63

- **特征比较**

  两总体或多总体大小比较 2.2 – 2.5

  两总体或多总体离散程度比较 3.1 – 3.2

  两总体分布特征比较 3.4

  总体分布是否服从特定理论分布 3.3 – 3.4

- **影响因素**

  方差分析及补充分析 4.1 – 4.4

- **变量关系**

  相关分析 5.1 – 5.2

  回归分析 **5.3 – 5.4**

5.3.1

5.3.1 无重复因变量数据的一元线性回归

5.3.2 有重复因变量数据的一元线性回归

回归分析

p.330

- **回归分析**

  建立根据自变量预测或估算因变量的统计模型

- **应用**

  预测: 根据新的自变量值计算对应的因变量值

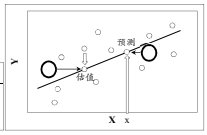
  估值: 对观测数据中已有因变量值均值-置信区间的估计

- **回归分析的模型**

  模型 I / II, 自变量为固定变量 / 随机效应

  类比方差分析模型 – 自变量与影响因素 参见**4.3.1**

模型	自变量类型	以估值为目的	以预测为目的
I	固定变量	最小二乘法	最小二乘法
II	随机变量	主轴法，约化主轴法	最小二乘法



一元线性回归

p.331-333

- **一元线性回归**

  据一个自变量预测/估计一个因变量

  因变量为随机变量, 自变量是随机或固定变量 二元相关两变量均为随机变量

- **数据**

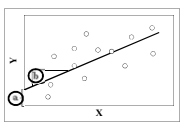
  无重复因变量数据, 一个 **x** 对应一个 **y** $x_i\ y_i\ \ i = 1 .. n$

  有重复因变量数据, 一个 **x** 对应若干 **y** $x_i\ y_{ij}\ \ i = 1 .. n, j = 1 .. m$

- **回归方程**

  $\hat{Y} = a + bX,$ a 截距 b 斜率

  模型取决于自变量类型 类似方差分析



最小二乘法

p.331-332

- **最小二乘法**

  用途 构建回归模型 – 模型 I 预测与估值, 模型 II 预测

  目标 观察数据到回归方程的垂直距离平方和最小

  限定 仅考虑 y 方向上的随机波动

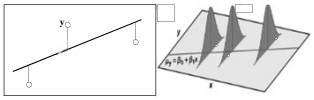
- **数据要求**

  独立性 对任何给定自变量, 因变量独立 即不受其它因变量取值影响

  正态性 因变量服从正态分布 最小二乘的合理性

  同质性 因变量有相同的方差 加和权重的一致性

  理论要求与实际应用的差距？

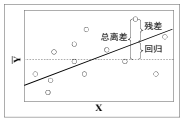


可决系数 *p.331*

- **定义**
  - 回归平方和与总平方和之比, 记为 r^2　　多元回归的可决系数记为 R^2
  - 反映数据与回归方程的吻合程度, 取值在 0 到 1 之间
- **与相关系数的差别**
  - 可决系数: 没有对应的总体参数, 不能检验, 不可比较
  - 相关系数: 有对应的总体参数, 可检验, 可比较

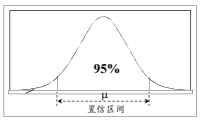
$$r=\frac{\sum((x_i-\bar{x})(y_i-\bar{y}))}{\sqrt{\sum(xi-\bar{x})^2\sum(yi-\bar{y})^2}}$$ 所有点到 x 和 y 均值的距离

$$r^2=\frac{(\sum(xi-\bar{x})(yi-\bar{y}))^2}{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}$$ 仅 y 方向距离



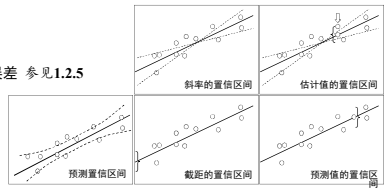
置信区间 复习 *p.43*

- **置信区间的一般表达**
  - 一个区间及该区间包括待估总体参数的概率
  - $P\{L_1 \le$ 总体参数 $\le L_2\} = p$
  - L_2, L_1 为该区间的上下界, p 为该区间覆盖总体参数的概率
- **例 算术均值的置信区间**
  - $P\{L_1 \le \mu \le L_2\} = p$　　$L_1 = \bar{x} - t_{\alpha[n-1]} s / n^{0.5}$, $L_2 = x + t_{\alpha[n-1]} s / n^{0.5}$
  - 如 $P\{5.5 \le \mu \le 6.7\} = 95\%$ 范围 5.5-6.7 包含总体算术均值的概率为 95%
- **应用**
  - 点估计的可靠性表达　　即区间估计
  - 据预设可靠性估算样本量　见上式, 见下节举例



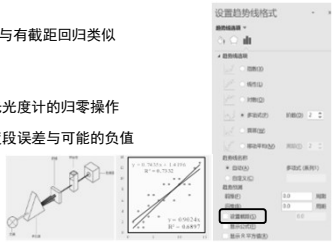
回归参数的置信区间 *p.334-335*

- **回归参数的置信区间**
  - 截距的置信区间　$P\{L_1 \le \alpha \le L_2\} = 1 - \alpha$
  - 斜率的置信区间　$P\{L_1 \le \beta \le L_2\} = 1 - \alpha$
  - 估计值的置信区间 $P\{L_1 \le$ $\le L_2\} = 1 - \alpha$,
  - 预测值的置信区间 $P\{L_1 \le$ $\le L_2\} = 1 - \alpha$　预测结果的置信范围, 图中虚线
- **计算**
  - 基于标准误差
  - 类比算术均值的标准误差 参见**1.2.5**



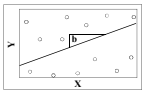
过原点回归 *p.338*

- **固定截距的回归**
  - 增加固定截距的限定条件　过原点回归是 a = 0 的特例, Excel 中的选项
  - $\hat{Y} = bX$　　a = 0, b 为斜率, 拟合方法不变, 拟合效果变差
- **拟合效果描述**
  - 可决系数, 参数的置信区间 与有截距回归类似
- **应用举例**
  - 理论上截距应为零　如分光光度计的归零操作
  - 拟合效果与预测误差　低浓度段误差与可能的负值



回归的显著性检验 *p.333*

- **可决系数**
  - 不可检验　　拟合效果的定量描述, 不是对总体参数的估计
  - 仅体现拟合好坏 0 – 1　不反映模型优劣
- **针对斜率的假设检验**
  - 检验回归模型的斜率　因变量是否有随自变量改变而发生显著变化的趋势
  - 假设 $H_0\ \beta = 0$, $H_1\ \beta \ne 0$　β 为总体斜率
  - 用 t-检验 或 方差分析　检验结果显著仅意味着回归直线不是“水平的”
  - 局限性　　不代表模型优劣, 不说明模型是否有预测能力
- **错误解释**
  - 如“检验结果显著, 模型可用于预测”



斜率比较与公共斜率 *不要求 p.348-350*

- **回归方程比较**
  - 比较多个回归方程的斜率是否有显著差异　同类问题
- **检验方法**
  - 两个斜率比较 $H_0\ \beta_1 = \beta_2$, $H_1\ \beta_1 \ne \beta_2$　　t-检验, 单侧或双侧
  - 多个斜率比较 $H_0\ \beta_i$ 都相同, $H_1\ \beta_i$ 不都相同　方差分析
- **公共斜率**
  - 若检验结果无显著差异, 可计算公共斜率　计算略 参见**5.1.1**
- **类比相关分析**
  - 相关系数比较与公共相关系数

5.3.2

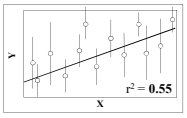
5.3.1 无重复因变量数据一元线性回归

5.3.2 有重复因变量数据一元线性回归

无重复因变量数据一元线性回归的局限性

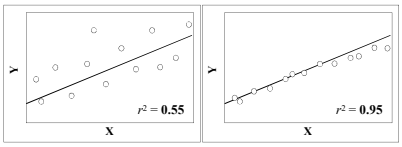
p.341

- **影响拟合效果的因素**
  - 选择模型的合理性　是否反映总体自变量与因变量的真实关系？
  - 因变量随机波动大小　因变量变异越大, 拟合结果越差
- **拟合效果的描述**
  - 可决系数 r^2　受模型合理性与随机波动大小两者的共同影响
- **模型显著性检验**
  - 不能区分模型选择与随机波动的影响　没有随机波动信息



课堂讨论

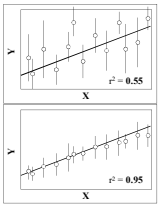
- **实例讨论**
  - 图中哪个模型拟合更好？
  - 前者受随机波动影响大, 后者呈现非线性关系　两者均影响拟合效果
- **如何区分**
  - 没有随机波动信息的情况下, 无从判断　不能判断模型的优劣
  - 只有获得重复因变量数据才能区分两者贡献　类比大小比较的重复数据



无重复因变量数据一元线性回归的局限性

p.341

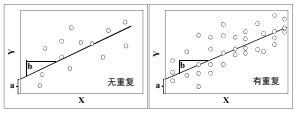
- **无重复因变量数据简单回归的局限性**
  - 不能区分随机波动和模型合理性两个因素的影响
  - 类比大小比较和方差分析中的随机波动 参见**2.2.4**, **4.2.1**
- **无重复因变量数据简单回归的结果与应用**
  - 结果: 获得既有条件下的最佳拟合
  - 应用: 提供既有条件下的最优估值与预测手段
  - 缺陷: 不能证明模型选取是否合理　无从判断模型优劣
  - 改进: 获得随机波动信息　关注因变量 – 预测需要



有重复因变量数据的一元线性回归

p.341

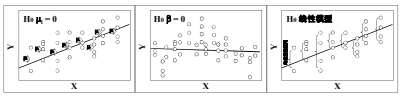
- **因变量重复数据**
  - 因变量重复数据可提供随机波动信息　模型优劣验证的数据基础
  - 每个自变量取值有≥2 个因变量取值　允许个别数据缺失 – 方差同质性
  - 无重复 $X_i, Y_i,\ i = 1 \dots n$
  - 有重复 $X_i, Y_{ij},\ i = 1 \dots n, j = 1 \dots m_i$　类比方差分析 参见**4.2.1**
- **计算**
  - 拟合及参数计算与无重复回归相同　公式, 回归参数, 可决系数, 置信区间等
  - 可决系数同样不能反映模型优劣

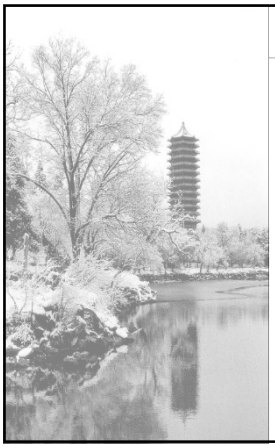


有重复因变量数据回归的假设检验

p.342

- **检验假设**
  - $H_0\ \mu_i$ 都相同　$H_1\ \mu_i$ 不都相同　不同自变量取值, 因变量有无差异, 即方差分析
  - $H_0\ \beta = 0$　$H_1\ \beta \neq 0$　斜率与 0 无显著差异, 与无重复数据回归相同
  - H_0 线性关系好 H_1 线性关系不好 是否选择了合适的模型
- **检验与判断**
  - 方差分析方法, 检验顺次进行　体现系统逻辑思路, 第一步可省略
  - 任何一步检验不显著, 即终止　均值有差异不等于斜率为 0
  - 模型优劣通过第三步检验判断　排除因变量随机波动影响的拟合效果





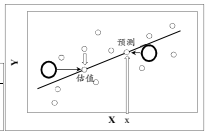
应用数理统计方法

第五章 相关与回归分析

回归分析 复习

p.330

- **回归分析**
  建立根据自变量预测或估算因变量的统计模型
- **应用**
  预测: 根据新的自变量值计算对应的因变量值
  估值: 对观测数据中已有因变量值均值-置信区间的估计
- **回归分析的模型**
  模型 I / II, 自变量为固定变量 / 随机效应
  类比方差分析模型 – 自变量与影响因素参见**4.3.1**

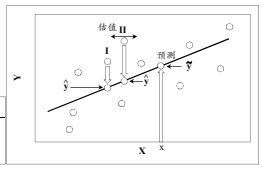


模型	自变量类型	以估值为目的	以预测为目的
I	固定变量	最小二乘法	最小二乘法
II	随机变量	主轴法，约化主轴法	最小二乘法

最小二乘法的局限性

p.355

- **最小二乘法**
  仅考察垂直方向的随机波动　数据点到回归线的垂直距离
- **最小二乘法的应用**
  用于预测与模型 **I** 估值　自变量为固定处理, 或待预测自变量固定
  不适用于模型 **II** 估值　自变量也是随机变量
- **模型 II 估值方法**
  同时考虑两个方向的随机波动

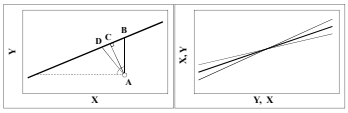


模型	自变量类型	以估值为目的	以预测为目的
I	固定变量	最小二乘法	最小二乘法
II	随机变量	主轴法, 约化主轴法	最小二乘法

以估值为目的的模型II回归

p.355

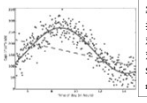
- **最小二乘**
  最小二乘法不适用　**A-B** 到拟合线距离最短, 垂直于 X 轴
- **模型 II 回归估值**
  主轴法, 最小正交平方法　**A-C** 到拟合线距离最短, 垂直于回归线, 两变量权重？
  约化主轴法　考虑变量尺度差异, 基于标准化的主轴法
  等分线　**A-D** 到 **X** 和 **Y** 方向两条垂线的夹角平分线
  夹角平分线　X, Y 互换做两次最小二乘, 取两个方程的夹角平分线



非线性回归

p.361-363

- **基于特定理论的非线性回归**
  已知理论方程, 根据观测结果获得拟合参数　可检验假设理论
  如一级反应动力学　C - 浓度, C_0 - 初始浓度, k - 速率常数, t - 时间
  应用: 获取相关参数, 构建预测方程, 验证特定机理
- **基于经验的非线性回归**
  根据观测数据寻找合适的方程形式　试算 = 凑
  常用曲线方程举例
  应用: 构建预测方程

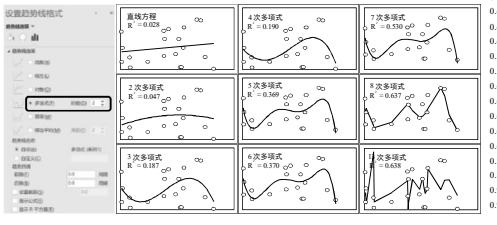


曲线	方程	
幂函数	$\hat{Y} = aX^b$	
指数函数	$\hat{Y} = ae^{bX}$	
对数函数	$\hat{Y} = a + b\log X$	
双曲线函数	$\hat{Y} = a + b/(X+c)$	
S 函数	$\hat{Y} = 1/(a+bc^X)$	
n 次多项式	$\hat{Y} = a_0 + a_1X + a_2X^2 + \ldots + a_nX^n$	

多项式回归

p.362

- **多项式拟合**
  一元 m 次多项式公式　$= b_0 + b_1x + b_2x^2 + \ldots + b_mx^m$
  常用非线性方法, 使用灵活
  不能为追求拟合效果任意增加阶数　例: 随机生成的一组数据

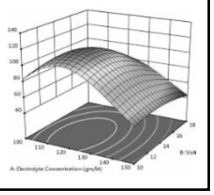


X	Y
0.01	0.73
0.01	0.14
0.02	0.42
0.10	0.24
0.10	0.04
0.29	0.15
0.46	0.75
0.46	0.26
0.47	0.59
0.51	0.79
0.58	0.30
0.61	0.60
0.63	0.04
0.64	0.04
0.66	0.41
0.67	0.81
0.81	0.97
0.85	0.95
0.90	0.24
0.99	0.02

多元回归

不要求

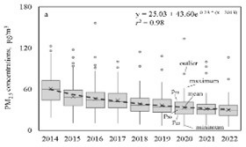
- **多元线性回归**
  - 一个以上自变量的线性回归 $\hat{Y} = a + b_1X_1 + \ldots + b_aX_a$
  - 计算与检验　类似一元线性回归
  - 可决系数与调整可决系数　R^2 与 R^2_{adj}, 后者可消除自变量增加造成的偏离
  - 标准化多元回归　自变量标准化以消除尺度差异的影响
- **多元非线性回归**
  - 类似一元非线性回归　如曲面拟合



非参数回归 – 顺序检验

p.366

- **非参数回归**
  - 顺序检验, 粗略判断动态变化的简单趋势　不是本来意义上的回归
  - 自变量为时间的 **Spearman** 秩相关分析　本质上是非参数相关
- **检验**
  - 假设: 某变量是否有显著的动态变化趋势　随时间升降的一般趋势
- **应用举例**
  - 判断空气质量多年变化的一般趋势

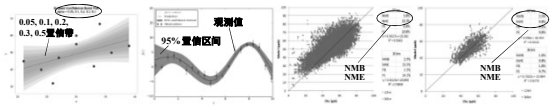


回归的不确定性表征

p.332

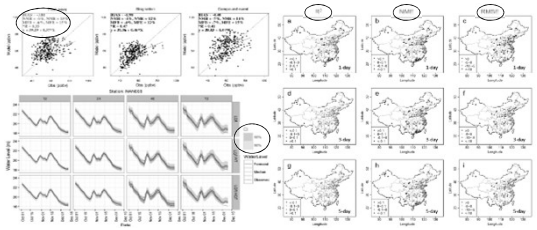
- **主要指标**
  - 可决系数　Coeff. of determination
  - 回归系数的置信区间　斜率, 估计值, 预测值
  - 归一化平均偏差　Normalized mean bias
  - 归一化平均误差　Normalized mean error
  - 均方根误差　Root mean square error

y 观测因变量, 计算因变量, 预测因变量　观测因变量均值　观测自变量均值



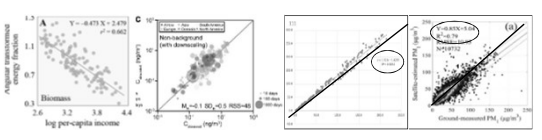
更多图示举例

- **举例**
  - 多种参数并用　观测 vs 模型估计
  - 多重置信区间　观测, 预测, 置信区间
  - 参数空间分布　模型的 R^2, *NME*, *RMSE*



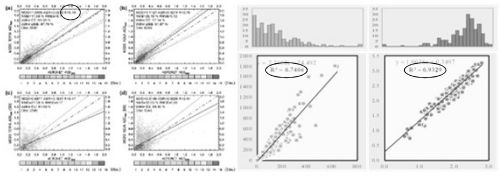
回归结果图示

- **区分两类散点图**
  - 回归: 因变量 / 自变量　可包含回归方程, 和各类误差表达
  - 验证: 预测值 / 观测值　**1:1**直线, 虚线误差范围, 颜色分类, 大小样本量 等
- **常见错误举例**
  - 预测值 / 观测值 散点图　使用没有意义的拟合直线或可决系数
  - 非正态分布数据　常见非正态分布, 特别是对数正态分布例子 参见下页



对数正态分布数据举例及模拟

- **对数正态分布**
  - 常见　微量污染物浓度, 风速, …
  - 基于原始数据回归的偏斜　高值点拟合权重贡献过高
- **基于对数变换数据的回归**
  - 结果更合理　有时也能改善拟合效果



多变量研究方法对比 复习

p.229,292

- **相关分析**

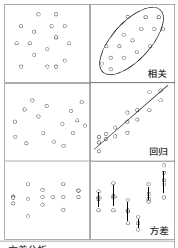
  共变关系 两变量等价

- **回归分析**

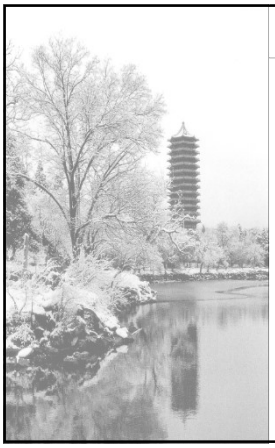
  模型构建 根据自变量估计或预测因变量

- **方差分析**

  影响因素 判断变量是否受某些因素影响



方法	相关分析	回归分析	方差分析
研究目的	变量一起变化的程度	构建回归模型 - 估值或预测	研究影响因素
变量关系	两个独立变量	自变量与因变量	独立变量与影响因素
变量类型	均为随机变量	因变量随机, 自变量固定或随机	独立变量随机, 影响因素固定或随机
统计量	无量纲的相关系数	有单位的回归参数	检验统计量 F



应用数理统计方法

第五章 相关与回归分析

应用实例 据教育水平等参数预测死亡率

Muller, BMJ, 2002

- **问题与方法**

  背景: 人群死亡率与代表性社会经济因素有关 基尼指数, 人均收入, 教育水平

  目的: 构建用相关参数预测死亡率的模型 基于美国州数据

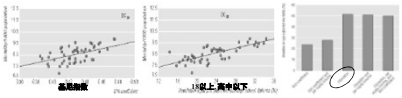
- **结果和讨论**

  假设: H_0 $\beta = 0$ 自变量不会影响人群死亡率

  结果: 拒绝原假设 其中单一教育因素的预测能力最强 $r^2_{adj} = 0.54$

  增加另两个因素不能提高预测能力

  讨论: 自变量相关



应用实例 据显示器销售量预测回收量

许等, 厦大学报, 2017

- **问题与方法**

  目的: 根据某品牌显示器累积销售量预测累积回收量

  数据: 收集典型数据 显示器 E1913c

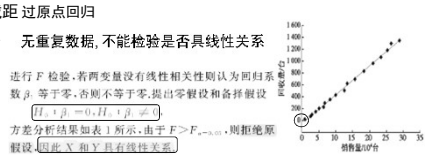
  方法: 建立回归模型: 回收量 = 0.00487 × 销售量 – 60.57

- **结果和讨论**

  检验: H_0 $\beta = 0$, H_1 $\beta \neq 0$, $F > F_{0.05}$ 拒绝H_0, 可用销售量预测回收量

  讨论: 不合理截距 过原点回归

  检验结论 无重复数据, 不能检验是否具线性关系



应用实例 据人均GDP预测机动车黑炭排放因子

Wang et al., Environ. Atom., 2012

- **问题与方法**

  背景: 机动车黑炭排放因子 $\mathbf{EF}_{BC}$ 单位质量耗油排放的黑炭量

  目的: 构建 $\mathbf{EF}_{BC}$ 数据库和模型, 预测各国汽车排放量

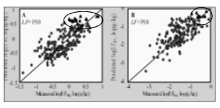
- **结果和讨论**

  发现: 人均 GDPc 是很好的预测变量 $\log EF_{BC} = k\,GDPc + C$

  注意到对发展中国家的高估 左图中黑色三角符号

  改进: 引入变量 Y_{3000} 达到 **$3000** 年份 $\log EF_{BC} = kGDPc + aY_{3000} + C$

  解释: 证实了发展中国家的后发优势 学习了发达国家的管理和技术



应用实例 据价格等参数预测土焦产量

Xu et al., PNAS 2018

- **问题与方法**

  背景: 土法炼焦是最重要的污染物排放源之一 **1998** 年被禁, **2011** 年消失

  目的: 评估禁止土法炼焦的健康效益 $PM_{2.5}$和苯并芘两类污染物

  方法: 利用遥感数据, 估算 1982-2015 土焦产量

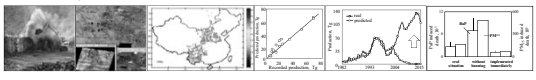
- **结果和讨论**

  建模: 不禁焦, 基于历史数据的多元线性回归 焦炭, 煤, 钢产量和焦炭价格

  土焦 = $0.96P_{coke}$-$0.15P_{coal}$-$0.39P_{steel}$-0.78 $Price_{coke}$+170 R^2_{adj}=0.98

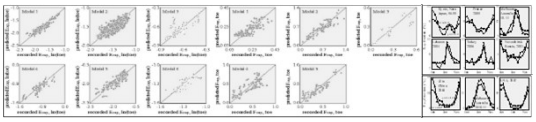
  估算: 三种情景的产量, 排放和健康危害 真实, 不禁, 1998立即终止

  结果: $PM_{2.5}$导致的过早死亡 实际 **36.6** 万, 无禁焦 **106** 万, 98年禁 **18** 万



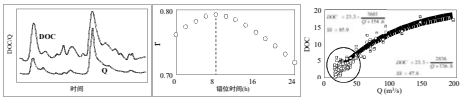
应用实例 据社会经济参数预测生活能耗季节变化 *Chen et al., Appl. Energy 2016*

- **问题与方法**
  - 目的: 构建全球生活能耗模型, 预测季节变化　全球国家级年数据, 极少季节数据
- **结果和讨论**
  - 假设: 能耗时-空变化受控于同样因素 HDD, CDD, 人均 GDP, 住房面积等
  - 方法: 时空置换: 用空间数据建模, 预测动态变化 多元线性回归
  - 模型: 区分国家类别　**6** 个燃料模型和 **5** 个电能模型
  - 验证: 空间与动态验证　**9** 组季节模拟验证



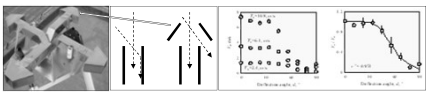
应用实例 据流量预测河水溶解态有机碳含量 *Tao et al., Water Res., 1998*

- **问题与方法**
  - 背景: 发现伊春河河水 DOC 含量与河水流量显著相关, 滞后 8 小时
  - 目的: 根据上述关系预测 DOC 含量 继而预测供水中消毒副产物前体物浓度
- **结果和讨论**
  - 结果: 比较两个回归模型
    - 直接回归　$DOC_t = \frac{a}{Q_t+154.6}$
    - 滞后 8 小时回归 $DOC_{t-8} = \frac{a}{Q_t+136.8}$
  - 讨论: 对数变换



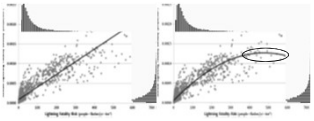
应用实例 模拟大气定向被动采样器的定向效果 *Tao et al., Environ. Pollut. 2008*

- **问题与方法**
  - 背景: 来自不同方向气团的污染物浓度可能不同　源区方向
  - 目的: 研制大气定向被动采样器　　采集四个方向气团样品, 测定多环芳烃
  - 设计: 屏蔽侧向气流的定向口设计　　防侧风
  - 研究: 风向 ***d*** 对流速 ***V*** 影响实验　　测定入风 ±45° 范围内的衰减率 V_d/V_0
- **结果和讨论**
  - 拟合: 非线性 $\frac{V_d}{V_0} = \frac{1}{1+\left(\frac{d}{5.72}\right)^{5.72}}$　$r^2 = 0.974$　d 超过 30° 明显衰减
  - 讨论: 缺少对比



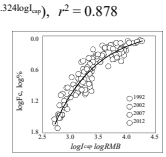
应用实例 据闪电密度预测闪电致死风险 *Roeder et al., Nat. Hazards, 2015*

- **问题与方法**
  - 目的: 根据人口加权闪电密度预测闪电致死密度　云地闪电
  - 数据: **1997-2010** 全美数据
- **结果和讨论**
  - 拟合: 线性回归　$y = 3.21\times10^{-6}x + 8.95\times10^{-5}$　　$r^2 = 0.820$
    - 二项回归　$y = 5\times10^{-9}x^2 + 5\times10^{-6}x + 6\times10^{-5}$　$r^2 = 0.864$
  - 讨论: 数据分布 – 尖峰右偏, 基于对数变换的回归
    - 二项式的不合理非单调变化



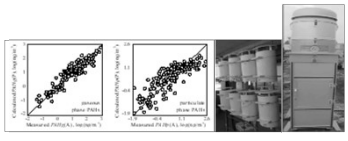
应用实例 据人均收入预测农村居民清洁炊事占比 *Tao et al., Nature Energy, 2018*

- **问题与方法**
  - 背景: 中国农村居民直接能源结构发生显著变化 炊事清洁能源占比上升
  - 数据: 农村居民 **1992-2012** 年能源结构　　**34,000** 多户入户调查
  - 目的: 分析影响变化的主要驱动因素　　清洁能源占比 **Fc**
- **结果和讨论**
  - 因素: 可负担性人均收入 $\mathbf{I_{cap}}$ 主导, 可获取性秸秆和煤炭产量有较小影响
  - 模型: 人均收入为预测变量 $\log\mathbf{Fc} = -62.4 + 62.6\,(1 - e^{-1.324\log I_{cap}})$,　$r^2 = 0.878$
  - 结果: 能源结构转变的优势驱动因素　**87.8%**
  - 用途: 可预测时空变化趋势



应用实例 据分子量预测多环芳烃被动采样效率 *Tao et al., ES&T 2007*

- **问题与方法采样**
  - 背景: 被动采样器采集大气多环芳烃　16 种多环芳烃, 引入分子量参数 MW_t
  - 数据: 主动和被动采样器同步采样　　A 主动; P 被动
  - 方法: 构建校验模型　　气态与颗粒态分别拟合
- **结果和讨论**
  - 气态: $\log PAH_A = 0.77\log PAH_P - 2.2\times10^{-9}\,MW_t^{3.8} + 1.62$　$r^2 = 0.880$
  - 颗粒: $PAH_A = PAH_P/e^{3.70-0.0314MWt}$　　$r^2 = 0.877$
    - PAH_A 主动采样浓度
    - PAH_P 被动采样浓度
    - MW_t 分子量



应用实例 据服务型领导和自我效能预测服务质量 *Qiu et al., Tour. Manag. 2020*

- **问题与方法**

  目的: 研究服务型领导 X 与自我效能 Y 对服务质量 Z 的影响

  数据: **673** 个餐馆和 **317** 个酒店雇员问卷调查打分数据　餐馆与酒店

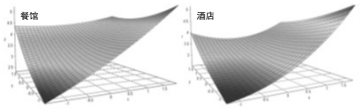
- **结果和讨论**

  假设: 服务型领导与自我效能对服务质量无显著影响　二元/多项式回归

  餐馆: $Z = b_0 + b_1X^{**} + b_2Y^{**} + b_3X^2 + b_4XY^{**} + b_5Y^{2**}$　r^2=0.333/0.342

  酒店: $Z = b_0 + b_1X^{**} + b_2Y + b_3X^{2**} + b_4XY^{**} + b_5Y^2$　r^2=0.333/0.342

  讨论: 影响 ?



应用实例 模拟土壤水溶性有机碳的淋出过程 *Tao et al., WASP. 2000*

- **问题与方法**

  背景: 土壤结构 非均相系统, 有机矿物复合体-土壤团粒

  土壤水分 重力水和毛管水分别指土壤团粒间和团粒内的水

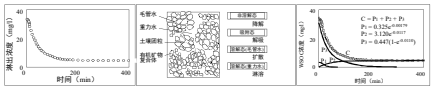
  目的: 定量分析土壤剖面中水溶性有机碳的淋出过程 土柱模拟实验

- **结果和讨论**

  拟合: 一次动力学负指数曲线拟合, 有明显偏离 肩, 折点, 常数项

  机理: 多步过程 三项加和: 对流–高斯, 扩散–负指数, 解吸–负指数

  讨论: 对假设过程的定量证明



要点

- **多变量方法**

  方差, 相关, 回归　目的, 数据, 变量, 联系 …

- **相关**

  正态与非正态, 线性与非线性 Pearson, 偏复相关, Spearman 等

- **回归**

  预测与估值

  一元线性回归, 无重复因变量 可决系数, 置信区间, 不确定性, 局限性

  一元线性回归, 有重复因变量 可检验拟合优劣

  其它 过原点, 多元, 多项式, 非线性 …

谢谢