**地学统计分析课程复习**

复习时间：Oct 15, 2024

**零假设 (null hypothesis)** 通常是**指一个没有显著变化、没有效应、没有相关性或差异的假设**。它是“默认的”假设，假设两组数据之间没有关系，或者某个参数没有发生变化。分布是**正态的**

参数检验和非参数检验：

A screenshot of a test

Description automatically generated

F-test is parametric test

Chi-square test is non-parametric test

Powerful: 统计功效

Robst: 稳健型

**游程检验**（**Runs test** 或 **Runs test for randomness**）是一种非参数统计检验，用来检测一组数据中的**观测值是否是随机的**，特别是用于判断数据序列中的模式是否具有随机性或是否存在某种系统性的趋势。

 **原假设（H0H\_0H0​**）\*\*：数据是随机的，游程的数量**符合随机分布**。

 **备择假设（H1H\_1H1​**）\*\*：数据不是随机的，存在某种系统性的模式。

* **Skewness** tells you about the **symmetry** of the distribution:
  + Positive skew = right tail longer, negative skew = left tail longer.
* **Kurtosis** tells you about the **peakedness** of the distribution:
  + Higher kurtosis = more extreme values (heavy tails), lower kurtosis = fewer extreme values (light tails).

A screenshot of a screen

Description automatically generated

让数据变成正态分布的方法：

A table with text and words

Description automatically generated with medium confidence

**自由度：**

A screenshot of a computer

Description automatically generated

章文波老师老师的outline：

1. 一个或两个均值的参数检验
   1. 单样本检验
   2. 双样本t检验
   3. 双样本配对t检验
   4. 突变检验：滑动t检验
2. 两独立总体大小的非参数检验

A close-up of a chart

Description automatically generated

Wilcoxon Rank-Sum Test

1. 两相关总体大小的非参数检验

A close-up of a sign

Description automatically generated

Wilcoxon Signed-Rank Test

1. 离散、分布和综合特征的假设检验
   1. 卡方检验：比较一个总体方差与特定值
   2. 比较两个总体方差的F检验
   3. 比较两个总体变异方差的t检验
   4. 多总体方差的对数方差分析
   5. 多总体方差的Bartlett检验

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A white text with black text

Description automatically generated

A screenshot of a test

Description automatically generated

A table with text and symbols

Description automatically generated

**多重共线性 (Multicollinearity)**

检测方法：

A screenshot of a computer

Description automatically generated

**解决办法：**

A screenshot of a computer

Description automatically generated

**自变量的选择准则：**

在回归分析中，选择合适的\*\*自变量\*\*（或称为预测变量）对于构建合理、有效的回归模型至关重要。合理选择自变量可以提高模型的解释力、避免过拟合、提高预测能力，并减少噪声的干扰。以下是几种常见的\*\*自变量选择原则\*\*：

### 1. \*\*统计显著性原则\*\*：

- \*\*原则\*\*：选择那些与因变量显著相关的自变量，通常通过显著性检验（如 t 检验或 F 检验）来判断自变量的贡献。

- \*\*应用\*\*：对每个自变量进行显著性检验，保留那些 p 值小于显著性水平（如 0.05 或 0.01）的自变量。

- \*\*优点\*\*：能确保选择的自变量与因变量有统计上的显著关系。

- \*\*局限\*\*：可能遗漏那些虽然单独不显著但与其他自变量一起有重要作用的变量。

### 2. \*\*理论依据原则\*\*：

- \*\*原则\*\*：根据已有的理论或先验知识，选择那些在研究领域中被认为与因变量有逻辑关系的自变量。

- \*\*应用\*\*：在构建模型之前，依据理论假设确定哪些自变量有助于解释因变量，从而提高模型的合理性。

- \*\*优点\*\*：确保选择的自变量符合研究的理论背景和逻辑关系。

- \*\*局限\*\*：如果理论不够完善或理论假设错误，可能导致误选自变量。

### 3. \*\*多重共线性原则\*\*：

- \*\*原则\*\*：避免选择那些彼此之间有高度线性相关性的自变量，以减少多重共线性对模型解释力的影响。

- \*\*应用\*\*：通过计算自变量之间的相关系数矩阵、方差膨胀因子（VIF）等方法检测共线性问题，去除高度共线的自变量。

- \*\*优点\*\*：减少共线性问题，使回归系数更加稳定、易于解释。

- \*\*局限\*\*：去除相关性强的自变量可能会丢失部分信息，因此需要谨慎选择。

### 4. \*\*逐步回归（Stepwise Regression）原则\*\*：

- \*\*原则\*\*：通过算法自动选择最优的自变量子集，基于变量对模型拟合度的贡献逐步添加或删除自变量。

- \*\*常见方法\*\*：

- \*\*前向选择法（Forward Selection）\*\*：从无变量模型开始，逐步添加显著性最高的自变量，直到没有新的变量显著。

- \*\*后向消除法（Backward Elimination）\*\*：从包含所有自变量的模型开始，逐步删除显著性最低的自变量，直到剩下的自变量都是显著的。

- \*\*逐步回归法（Stepwise Selection）\*\*：结合前向选择和后向消除，动态添加和删除变量，直到得到最优的自变量组合。

- \*\*优点\*\*：通过自动化过程选择最重要的变量，减少人工干预的误差。

- \*\*局限\*\*：可能依赖于样本数据的特定情况，容易受到异常值或噪声的影响。

### 5. \*\*AIC、BIC 准则\*\*：

- \*\*原则\*\*：使用信息准则（如 AIC、BIC）来选择自变量组合。AIC（Akaike Information Criterion）和 BIC（Bayesian Information Criterion）都是平衡模型拟合优度和模型复杂性的准则。

- \*\*应用\*\*：比较不同自变量组合的模型，选择 AIC 或 BIC 值最小的模型。AIC 和 BIC 都考虑了模型的残差和自变量数量，但 BIC 对模型复杂性有更强的惩罚。

- \*\*优点\*\*：自动平衡模型的拟合度和复杂性，减少过拟合的风险。

- \*\*局限\*\*：AIC、BIC 主要用于模型比较，而非直接用于单个变量的显著性检验。

### 6. \*\*Mallows' \(C\_p\) 准则\*\*：

- \*\*原则\*\*：选择最优的自变量组合，使得 Mallows' \(C\_p\) 值接近自变量的数量 \( p \)。

- \*\*应用\*\*：常用于回归模型中的变量选择。通过全子集回归法或逐步回归法，计算不同自变量组合的 \(C\_p\) 值，选择 \(C\_p\) 值最接近 \(p\) 的模型。

- \*\*优点\*\*：在选择自变量时考虑了模型的复杂度，避免过拟合和欠拟合。

- \*\*局限\*\*：需要计算不同子集的模型，较为耗时。

### 7. \*\*主成分回归（Principal Component Regression, PCR）\*\*：

- \*\*原则\*\*：通过主成分分析将原始自变量转换为新的互相正交的主成分，选择那些解释大部分方差的主成分进行回归分析。

- \*\*应用\*\*：将多个相关自变量转换为一组不相关的主成分，减少多重共线性问题。

- \*\*优点\*\*：减少多重共线性问题，提高回归模型的稳定性。

- \*\*局限\*\*：主成分回归不直接保留原始自变量，因此解释性较差。

### 8. \*\*Lasso 回归和岭回归\*\*：

- \*\*原则\*\*：通过正则化方法选择自变量。\*\*Lasso 回归\*\*通过引入 L1 正则化将不重要的系数缩小为零，从而自动选择自变量；\*\*岭回归\*\*通过 L2 正则化减小多重共线性问题，但不会完全消除变量。

- \*\*应用\*\*：在高维度数据中，Lasso 回归可以自动选择重要的变量，而岭回归更适合解决多重共线性问题。

- \*\*优点\*\*：自动选择变量，特别适用于高维数据集。

- \*\*局限\*\*：正则化模型中引入了惩罚项，可能会导致模型的解释力下降。

### 9. \*\*交叉验证（Cross-Validation）\*\*：

- \*\*原则\*\*：通过交叉验证（如 k 折交叉验证）来选择自变量组合，以模型在训练集和验证集上的表现为依据选择模型。

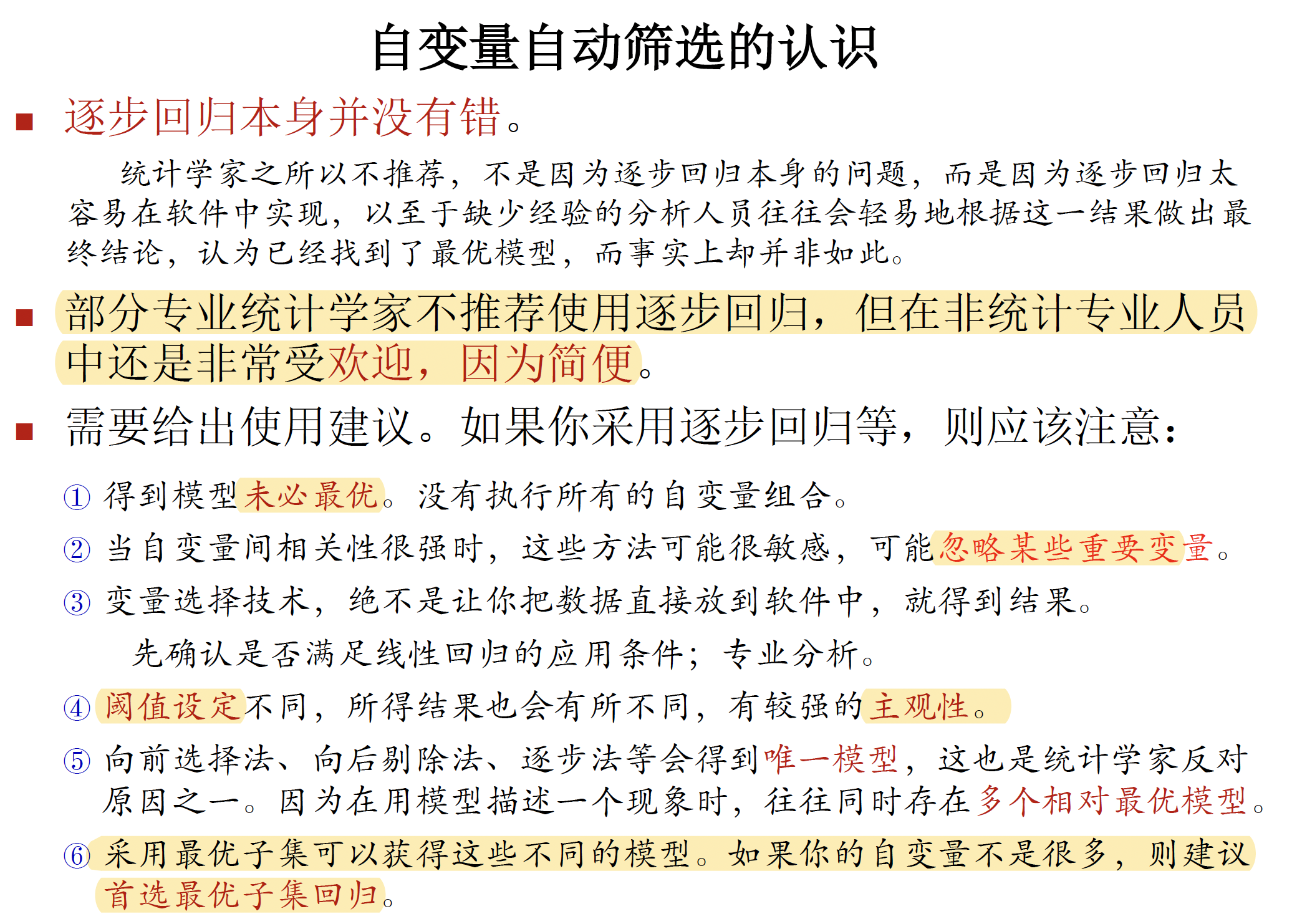
- \*\*应用\*\*：将数据分为训练集和验证集，反复测试不同的自变量组合，选择能够在新数据上表现最好的模型。

- \*\*优点\*\*：能有效避免过拟合，确保模型具有良好的泛化能力。

- \*\*局限\*\*：交叉验证需要较大的计算量，尤其是当变量较多时。

### 总结：

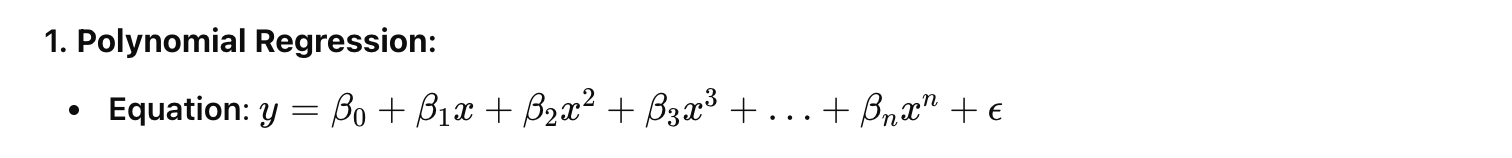
自变量的选择对构建有效的回归模型至关重要。常见的自变量选择原则包括统计显著性原则、理论依据原则、多重共线性原则、逐步回归法、AIC/BIC 准则、Mallows' \(C\_p\) 准则、主成分回归、Lasso 和岭回归、以及交叉验证等。选择哪种方法取决于数据的特性、模型的复杂度、以及研究者的具体需求。



**非线性回归**

内线性回归模型（曲线估计）

类型：Polynomial Regression, **Exponential Regression**, **Logarithmic Regression**, **Power Regression**.









A screenshot of a white background

Description automatically generated A white text on a white background

Description automatically generated

A white paper with black text

Description automatically generated

**非线性回归模型的求解**通常依赖于**迭代优化算法**，如**梯度下降法**、牛顿法、拟牛顿法和 Levenberg-Marquardt 算法。这些算法通过最小化损失函数（通常为残差平方和）来逼近最优参数。不同的非线性模型有不同的应用场景，如指数模型、对数模型和幂律模型等。根据问题的复杂性和数据特性，选择合适的求解算法至关重要。

**主成分分析**

**因子分析**

主成分法

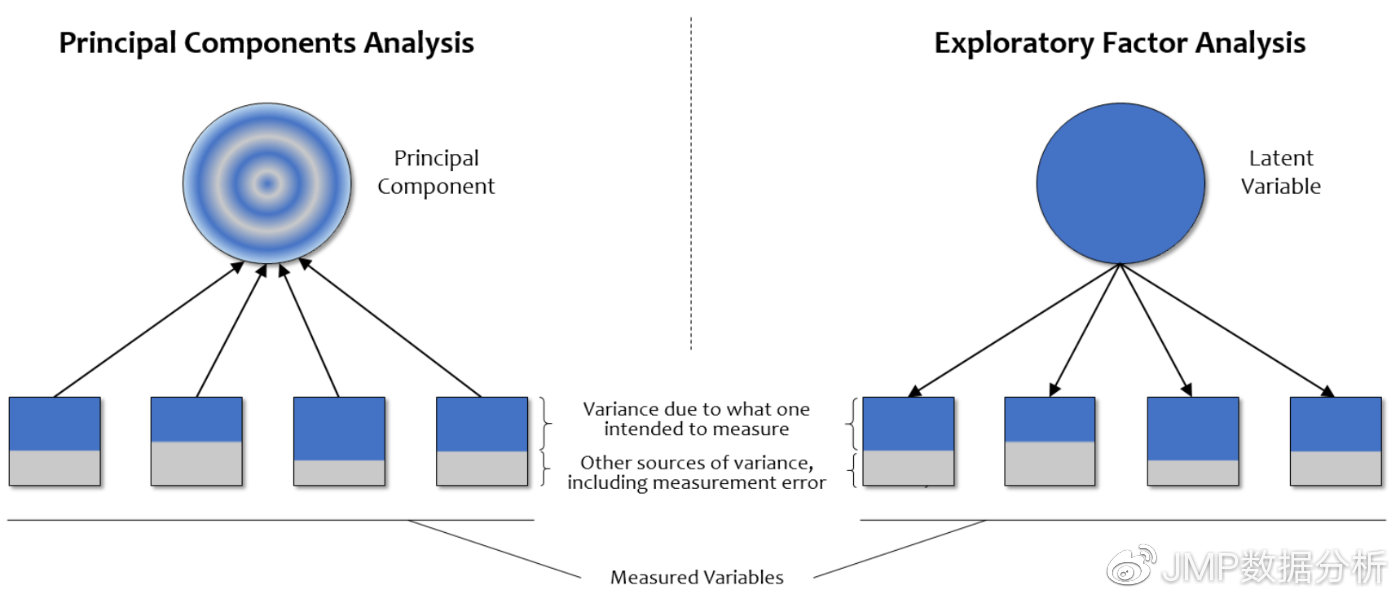
极大似然法 (Maximum Likelihood Factor)

主因子法 (Principal Factor)

迭代主因子法 (Iterated Principal Factor)

A diagram of a computer algorithm

Description automatically generated



A screenshot of a computer

Description automatically generated