W1 - 1 Introduction to Generalized Linear Models

中文总结

题目：广义线性模型导论

1. **线性回归模型的假设**：
   - **线性关系（Linearity）**：假设参数在模型中是线性进入的，即响应变量的期望值是设计矩阵与参数向量的乘积。
   - **误差独立性（Independence of Errors）**：假设误差是相互独立的，未满足该条件时误差可能会相关。
   - **误差方差恒定性（Homoscedasticity）**：假设误差的方差在所有测量中是恒定的。
   - **误差正态分布（Normality of Errors）**：假设误差是独立同分布的正态分布。
2. **违反线性回归假设的案例**：
   - **二项分布（Binomial Distribution）**：如在预测选举结果时，响应变量是离散的，违反了正态性和恒定方差假设。
   - **残差分析（Residual Analysis）**：对于满足和不满足线性回归假设的数据进行残差分析，显示了模型适配度的差异。
3. **广义线性模型（Generalized Linear Models, GLM）**：
   - **随机成分（Random Component）**：响应变量来自指数分布族（Exponential Family of Distributions）。
   - **系统成分（Systematic Component）**：预测变量和参数的线性组合。
   - **连接函数（Link Function）**：将系统成分与随机成分联系起来。
4. **指数分布族的定义与性质**：
   - **指数分布族（Exponential Family）**：定义了广义线性模型中的响应变量需要符合的分布形式，包括二项分布、泊松分布等。
   - **指数分布族的形式**：概率质量函数或概率密度函数可以写为指数形式，涉及位置参数（Canonical Parameter）和尺度参数（Dispersion Parameter）。
   - **均值和方差**：指数分布族的均值为函数$b(\theta)$的一阶导数，方差为其二阶导数乘以尺度参数。

拟合优度（Goodness of Fit）是统计学中的一个概念，用来评估一个统计模型对观测数据的拟合效果。换句话说，它衡量的是模型所预测的值与实际观测值之间的差异程度。

具体来说，拟合优度可以帮助回答以下问题：

- **模型是否能够准确描述数据的特征?**
- **模型预测的值与实际观测值的偏差有多大?**

在不同的统计模型中，拟合优度的衡量标准有所不同。以下是一些常见的拟合优度评估方法：

1. **残差平方和 (Residual Sum of Squares, RSS)**: 常用于线性回归模型，是观测值与预测值之间的差异的平方和。RSS越小，表示模型拟合得越好。
2. **R平方 (R-squared)**: 同样常用于线性回归，表示模型解释观测数据总变异的比例。R平方值越接近1，表示模型的解释力越强，拟合优度越好。
3. **偏差 (Deviance)**: 用于广义线性模型（如泊松回归），衡量模型与最大似然估计的对数似然函数的差异。偏差越小，拟合优度越高。
4. **卡方检验 (Chi-square Test)**: 用于分类数据的拟合优度检验，通过比较实际观测频数与期望频数来判断模型的拟合效果。如果卡方统计量较小，表明模型拟合得较好。
5. **AIC/BIC (Akaike Information Criterion/Bayesian Information Criterion)**: 这两个信息准则用于模型选择，AIC/BIC值越小，表示模型拟合效果更好，且考虑了模型的复杂度。

总的来说，拟合优度是一个衡量模型预测能力的重要指标，能够帮助我们判断一个模型是否适合用于描述数据或进行预测。

English Summary

Title: Introduction to Generalized Linear Models

1. **Assumptions of Linear Regression Models**:
   - **Linearity**: The parameters enter the model linearly, meaning the expected value of the response is a product of the design matrix and parameter vector.
   - **Independence of Errors**: Errors are assumed to be independent, meaning uncorrelated.
   - **Homoscedasticity**: The variance of the errors is constant across all measurements.
   - **Normality of Errors**: Errors are normally distributed, i.i.d., with a mean of zero and constant variance.
2. **Examples of Violating Linear Regression Assumptions**:
   - **Binomial Distribution**: For example, when predicting election outcomes, the binary response violates the normality and homoscedasticity assumptions.
   - **Residual Analysis**: Residual plots for data that meet and violate the linear regression assumptions show differences in model fit.
3. **Generalized Linear Models (GLM)**:
   - **Random Component**: The response variable comes from the Exponential Family of Distributions.
   - **Systematic Component**: A linear combination of predictor variables and parameters.

- **Link Function**: Connects the systematic component to the random component.

4. **Exponential Family Definition and Properties**:
   - **Exponential Family**: Defines the form of the response variable's distribution in GLMs, including binomial, Poisson distributions, etc.
   - **Form of Exponential Family**: Probability mass or density functions can be expressed in exponential form, involving the canonical parameter and dispersion parameter.
   - **Mean and Variance**: The mean is the first derivative of the function $b(\theta)$, and the variance is the second derivative multiplied by the dispersion parameter.