Introduction to Binomial Regression

中文总结

题目：二项回归简介

1. **二项回归模型的构建**：
   - **广义线性模型的三大组成部分**：随机成分（响应变量）、系统成分（预测变量及其参数）、连接函数（Link Function）。二项回归中，响应变量来自二项分布，每个观测值可能来自不同的二项分布。
   - **伯努利分布的特殊情况**：当样本大小$n_i = 1$时，响应变量为0或1，属于伯努利分布，可用于二元分类问题，如逻辑回归（Logistic Regression）。
   - **线性预测器与概率的关系**：线性组合预测器无法直接用于概率预测，需要通过连接函数将其转化为概率值。
2. **连接函数的选择**：
   - **Logit连接函数**：常见的连接函数，定义为成功概率p_i与失败概率$1 - p_i$的对数比。这个函数与指数分布族中的自然参数θ相联系。
   - **Probit连接函数**：另一种连接函数，通过标准正态分布的累积分布函数的反函数来连接概率和线性预测器。
3. **二项回归参数估计**：
   - **最大似然估计（Maximum Likelihood Estimation, MLE）**：用于估计模型参数，通过对似然函数的对数进行最大化，通常使用迭代算法进行参数估计。
4. **二项回归的参数解释**：
   - **截距（Intercept）**：表示所有预测变量为零时的成功概率的对数几率。
   - **斜率参数（Slope Parameters）**：解释某一预测变量增加一个单位时，在其他变量不变的情况下，成功的对数几率增加量。通过指数化可以解释为成功几率的乘法增量。
5. **R中的二项回归应用**：
   - **数据集分析**：使用来自UCI机器学习库的数据集，预测办公室是否被占用。通过GLM函数拟合二项回归模型，输出系数估计并解释其含义。

English Summary

Title: Introduction to Binomial Regression

1. **Construction of the Binomial Regression Model**:

   - **Three Components of Generalized Linear Models**: The random component (response variable), systematic component (predictors and parameters), and the link function. In binomial regression, the response variable is from a binomial distribution, and each observation may come from a different binomial distribution.
   - **Special Case of Bernoulli Distribution**: When the sample size $n_i = 1$, the response variable is either 0 or 1, representing a Bernoulli distribution, useful for binary classification problems like logistic regression.
   - **Relationship Between Linear Predictor and Probability**: The linear combination of predictors cannot directly predict probability and must be linked to the probability value using a link function.
2. **Choice of Link Function**:
   - **Logit Link Function**: A common link function defined as the log odds of success p_i over failure $1 - p_i$. It is connected to the canonical parameter θ in the exponential family of distributions.
   - **Probit Link Function**: Another link function that uses the inverse of the cumulative distribution function of the standard normal distribution to link probability and linear predictor.
3. **Parameter Estimation in Binomial Regression**:
   - **Maximum Likelihood Estimation (MLE)**: Used to estimate model parameters by maximizing the log-likelihood function, often done using iterative algorithms.
4. **Interpretation of Binomial Regression Parameters**:
   - **Intercept**: Represents the log odds of success when all predictors are zero.
   - **Slope Parameters**: Explain the increase in log odds of success for a one-unit increase in a predictor while holding others constant. The exponentiated values represent multiplicative changes in the odds of success.
5. **Application of Binomial Regression in R**:
   - **Data Analysis**: Analyzed a dataset from the UCI Machine Learning Repository to predict office occupancy using binomial regression. The GLM function is used to fit the model, and the coefficient estimates are interpreted accordingly.