Week2

Table of Contents

Note 2

Summary of Key Concepts (English)

1. Goodness of Fit for Poisson Regression: Deviance

- **Deviance in Poisson Regression**: The deviance is a measure of goodness of fit for a Poisson regression model. It is calculated as negative two times the log-likelihood of the model evaluated at the maximum likelihood estimates (MLEs). A smaller deviance indicates a better fit.
- **Null Deviance**: This is the deviance of a model that only includes an intercept term (no predictors). It reflects how well the simplest model fits the data.
- **Saturated Deviance**: This is the deviance of a model that includes a unique parameter for each data point, essentially overfitting the data.
- **Residual Deviance**: The difference between the deviance of the model of interest and the saturated model. It is used to assess the goodness of fit by comparing it to a chi-squared

distribution.

2. Pearson's Chi-Squared Test

- **Pearson's Chi-Squared Statistic**: This test provides an alternative measure of goodness of fit for Poisson regression. The statistic compares observed counts to expected counts under the model, with the sum of squared differences standardized by the expected counts.
- **Pearson Residuals**: These residuals are calculated as the difference between observed and expected counts, standardized by the square root of the expected counts. They are used for diagnostic checks, similar to residuals in linear regression.

3. Overdispersion in Poisson Regression

- **Concept**: Overdispersion occurs when the variance of the data exceeds the mean, which violates the assumption of the Poisson distribution (where variance equals mean).
- **Causes**: Overdispersion can arise from real phenomena like zero inflation or spatial/temporal correlation, or apparent issues like missing predictors or inappropriate link functions.
- **Detection**: Overdispersion can be detected using the Pearson chi-squared statistic divided by its degrees of freedom, with a value significantly greater than one indicating overdispersion.
- **Solutions**:
  - **Quasilikelihood Methods**: Adjust the standard errors of the estimates by introducing a dispersion parameter, ϕ, to account for overdispersion.
  - **Negative Binomial Regression**: A more flexible alternative to Poisson regression that directly models overdispersion by introducing an additional parameter.

总结（中文）

1. 泊松回归的拟合优度：偏差

- **泊松回归中的偏差**: 偏差是衡量泊松回归模型拟合优度的指标。计算方法是模型的对数似然函数在最大似然估计（MLE）处的值的负二倍。偏差值越小，模型拟合效果越好。
- **零偏差**: 这是仅包含截距项（无预测变量）的模型的偏差，反映最简单模型对数据的拟合程度。
- **饱和偏差**: 这是一个包含每个数据点独特参数的模型的偏差，本质上是对数据的过拟合。
- **残差偏差**: 所关心的模型的偏差与饱和模型的偏差之间的差值。通常通过将其与卡方分布进行比较来评估拟合优度。

2. 皮尔逊卡方检验

- **皮尔逊卡方统计量**: 这是评估泊松回归拟合优度的另一种方法。该统计量通过标准化观察值与模型期望值的平方差来计算。

- **皮尔逊残差**: 这些残差计算为观察值与期望值之间的差异，并以期望值的平方根进行标准化。它们用于诊断检查，类似于线性回归中的残差。

3. 泊松回归中的过度离散

- **概念**: 过度离散发生在数据的方差超过均值时，这违反了泊松分布的假设（方差等于均值）。
- **原因**: 过度离散可能来自真实现象，如零膨胀或空间/时间相关性，也可能是缺失预测变量或不适当的链接函数等表面问题引起的。
- **检测**: 通过皮尔逊卡方统计量除以自由度，可以检测过度离散，若值显著大于1则表明存在过度离散。
- **解决方法**:
  - **准似然方法**: 通过引入离散参数ϕ来调整标准误差，以考虑过度离散的影响。
  - **负二项回归**: 这是泊松回归的更灵活的替代方法，通过引入一个额外参数直接建模过度离散。

在Poisson回归中，**偏差（Deviance）** 是用来衡量模型拟合优度的一个重要指标。具体来说，偏差衡量了当前模型与完全饱和模型（即能够完美拟合数据的模型）之间的差异。计算偏差的公式与模型的对数似然函数（Log-Likelihood）有关。

偏差的计算

对于Poisson回归模型，偏差的计算公式如下：

$$\text{Deviance} = -2 \times (\log L(\text{current model}) - \log L(\text{saturated model}))$$

其中：

- $\log L(\text{current model})$ 是当前模型的对数似然函数值。
- $\log L(\text{saturated model})$ 是饱和模型的对数似然函数值。

然而，通常我们并不直接计算饱和模型的对数似然值，而是通过以下公式来计算偏差：

$$\text{Deviance} = -2 \times \sum_{i=1}^{n} \left[y_i \log \left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i) \right]$$

这里：

- y_i 是第 i 个观测的实际值。
- $\hat{\mu}_i$ 是第 i 个观测的预测值。

这个公式基于Poisson分布的性质，直接计算了当前模型与实际数据之间的偏差。

具体步骤

1. **计算对数似然函数值**：
   - 对数似然函数值 $\log L$ 可以通过将模型的预测值 $\hat{\mu}_i$ 和实际观测值 y_i 代入Poisson分布的似然函数来计算。
2. **计算偏差**：
   - 使用上面的公式，计算模型的偏差。一般来说，偏差越小，表示模型与数据的拟合越好。

在R中如何计算

在R中，可以通过拟合Poisson回归模型后直接获得偏差值。例如：

```r
# 拟合Poisson回归模型
model <- glm(num_accidents ~ speed_limit, family = poisson, data = data)

# 获取偏差值
deviance_value <- model$deviance

# 输出偏差值
print(deviance_value)
```

在这个例子中，`model$deviance` 会返回拟合模型的偏差值。R会自动计算模型的对数似然函数值并根据公式计算偏差。

总结

- **偏差** 是衡量模型拟合优度的一个关键指标。
- 偏差通过模型的对数似然函数值计算，并反映了当前模型与完全拟合模型之间的差异。
- 在R中，通过 `glm` 函数拟合Poisson回归模型后，可以直接获得偏差值。

这个指标在模型选择和评估过程中非常重要。希望这些解释能帮助你更好地理解Poisson回归中偏差的计算过程。如果有进一步的问题，随时提问！