Week 2 Note 1 - Poisson Regression

Poisson回归是一种广泛用于建模**计数数据**的回归模型。它用于描述响应变量是非负整数（通常是计数值）的情况，尤其适用于建模事件发生次数的数据。Poisson回归假设响应变量（即因变量）服从Poisson分布。

Poisson回归的基本概念

1. **Poisson分布**：
   - Poisson分布是用于描述在固定时间段或空间区域内某个事件发生次数的概率分布。Poisson分布的一个重要性质是其均值等于方差。
   - 如果一个随机变量 Y 服从参数为 λ 的Poisson分布，则 Y 的概率质量函数为：

$$P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}, \quad y = 0, 1, 2, \ldots$$

     其中，λ 是Poisson分布的均值（也是方差）。
2. **Poisson回归模型**：
   - 在Poisson回归中，我们假设响应变量 Y_i（事件发生的计数）对于给定的解释变量 X_i 服从Poisson分布，其均值 λ_i 由解释变量通过指数函数连接起来：

$$\lambda_i = \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip})$$

   - 这里，$\beta_0, \beta_1, \ldots, \beta_p$ 是模型的参数，而 $X_{i1}, X_{i2}, \ldots, X_{ip}$ 是解释变量。
3. **对数连接函数**：
   - Poisson回归模型通常使用对数连接函数将线性预测器与均值参数 λ_i 相关联：

$$\log(\lambda_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

   - 通过对数转换，Poisson回归可以处理响应变量的非负性（即计数值总是非负的）。

应用场景

Poisson回归模型通常用于以下情况：

- **事件计数**：在固定时间或空间内事件发生的次数，例如一小时内接到的电话次数。
- **稀有事件的发生率**：如罕见疾病的发病率或某产品的故障率。

- **交通事故数据**：如某特定路段在一定时间内的交通事故次数。

Poisson回归的R代码示例

假设我们有一个数据集，其中我们记录了某特定时间段内的交通事故次数，并且我们有不同的预测变量如天气、路段类型等，我们可以使用Poisson回归来建模：

```
# 假设数据已经加载到dataframe 'df' 中，并且响应变量为 'accidents'
# 预测变量包括 'weather' 和 'road_type'

# 使用glm()函数拟合Poisson回归模型
poisson_model <- glm(accidents ~ weather + road_type, family = poisson(link = "log"), data = df)

# 查看模型摘要
summary(poisson_model)
```

在这个例子中， `glm()` 函数用于拟合广义线性模型，而 `family = poisson(link = "log")` 指定了Poisson回归模型及其对数连接函数。

Poisson回归的参数估计和解释

在 Poisson 回归模型中，参数的估计和解读是理解模型的关键部分。我们将详细讨论如何估计这些参数、 `beta_j` 和 `exp(beta_j)` 的含义，以及如何解读你给出的 `glm` 结果。

Poisson 回归中的参数估计

在 Poisson 回归中，参数 β_j 是通过最大似然估计（Maximum Likelihood Estimation, MLE）来估计的。具体来说，模型试图找到一组 β_j 值，使得在给定自变量 X_j 的条件下，观测到的因变量 Y 的概率最大化。R 中的 `glm` 函数使用迭代算法（如 Fisher Scoring）来找到这些参数的最优值。

参数 β_j 和 e^{β_j} 的含义

1. β_j **的含义**：
   - β_j 是自变量 X_j 对响应变量 Y 的对数期望值的影响。具体来说，β_j 表示当 X_j 增加一个单位时，Y 的对数期望值会改变多少。
   - 在对数链接函数的作用下，Poisson 回归模型的形式为：

$$\log(\lambda_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

这里，λ_i 是响应变量 Y_i 的期望值。

2. e^{β_j} **的含义**：
   - e^{β_j} 是自变量 X_j 对响应变量 Y 的相对风险比（Relative Risk, RR）。它表示当 X_j 增加一个单位时，响应变量 Y 的期望值 λ 按比例增长多少。
   - 例如，如果 $e^{\beta_j} = 1.5$，则表示 X_j 每增加一个单位，事件发生的概率（或期望值）增加 50%。

解读 `glm` 结果

假设 `p` 数据集中包含以下变量：

- `num_awards`：学生获得的奖项数量（计数变量，响应变量）。
- `prog`：学生所参加的项目类别（因子变量，有三个水平：General、Standard、Honors）。
- `math`：学生的数学成绩（数值变量）。

运行的模型为：

```
glm(formula = num_awards ~ prog + math, family = "poisson", data = p)
```

结果解读

```
Coefficients:
                Estimate   Std. Error    z value    Pr(>|z|)
(Intercept)     -5.2471    0.6585        -7.97      1.6e-15 ***
progStandard    1.0839     0.3583        3.03       0.0025 **
progHonors      0.3698     0.4411        0.84       0.4018
math            0.0702     0.0106        6.62       3.6e-11 ***
```

1. **(Intercept)**：
   - **估计值**: -5.2471
   - 解释：在 `prog` 为基线类别（General）、 `math` 为 0 的情况下，奖项数量的对数期望值为 -5.2471。
   - $e^{(\text{Intercept})}$ 约等于 $e^{-5.2471}$，表示的是在这个基础条件下，奖项数量的期望值，非常接近于 0。
2. **progStandard**：
   - **估计值**: 1.0839
   - 解释：与基线类别 `progGeneral` 相比， `progStandard` 学生的奖项数量对数期望值增加 1.0839。

- $e^{1.0839}$ 约为 2.956，表示 `progStandard` 学生的奖项数量的期望值是 `progGeneral` 学生的 2.956 倍。

3. **progHonors**：
   - **估计值**: 0.3698
   - 解释：与基线类别 `progGeneral` 相比， `progHonors` 学生的奖项数量对数期望值增加 0.3698。
   - $e^{0.3698}$ 约为 1.447，表示 `progHonors` 学生的奖项数量的期望值是 `progGeneral` 学生的 1.447 倍。
   - 注意：p值（0.4018）表明这个系数在常见的显著性水平下并不显著。
4. **math**：
   - **估计值**: 0.0702
   - 解释：数学成绩每增加 1 分，奖项数量的对数期望值增加 0.0702。
   - $e^{0.0702}$ 约为 1.073，表示数学成绩每增加 1 分，奖项数量的期望值增加约 7.3%。

小结

在 Poisson 回归中，β_j 和 e^{β_j} 对模型的解读至关重要。通过估计值 β_j，你可以了解自变量对响应变量的对数期望值的影响，而 e^{β_j} 则直接告诉你这些自变量对事件发生率的影响比例。

总结

Poisson回归是处理计数数据的有效工具，适合用于分析那些响应变量为非负整数且分布不对称的场景。通过对数连接函数，将线性模型与Poisson分布的均值关联起来，提供了对计数数据建模的强大方法。