第五章：广义线性模型详细笔记

小狗

2024 年 10 月 14 日

目录

1 5. 广义线性模型 (Generalized Linear Models, GLMs)

1.1 5.0 引言

广义线性模型（GLMs）是一类强大而灵活的统计模型, 它扩展了普通线性回归的概念, 能够处理各种不同类型的响应变量。在本章中, 我们将深入探讨 GLMs 的核心概念、组成部分以及在 R 中的实际应用。

1.1.1 5.0.1 GLMs 的基本概念

广义线性模型由三个主要部分组成:

1. **随机成分**: 描述响应变量 Y 的概率分布。这通常是指数分布族的一个成员。
2. **系统成分**: 指定预测变量 (自变量) 如何通过线性组合形成线性预测器。
3. **连接函数**: 将线性预测器与响应变量的期望值联系起来。

1.1.2 5.0.2 GLMs 的优势

1. **灵活性**: 可以处理各种类型的响应变量 (二元、计数、连续等)。
2. **统一框架**: 为多种常见的统计模型提供了一个统一的理论框架。
3. **可解释性**: 保留了线性模型的许多有利特性, 如参数的可解释性。

1.1.3 5.0.3 本章结构

在接下来的章节中, 我们将详细探讨: - 方差函数及其在 GLMs 中的重要性 - 准似然方法及其应用 - 特定分布 (如 Gamma 分布和逆高斯分布) 在 GLMs 中的应用 - 泊松分布和二项分布的关系及其在建模中的选择

每个部分都会包含理论解释、数学推导 (适当的难度) 以及 R 代码实例, 帮助您更好地理解和应用这些概念。

1.2 5.1 方差函数 (The Variance Function)

1.2.1 5.1.1 引言

在广义线性模型中, 理解响应变量的均值和方差之间的关系是至关重要的。这种关系是通过**方差函数**来描述的, 它源于指数分布族 (Exponential Dispersion Model, EDM) 的特性。

1.2.1.1 指数分布族回顾 指数分布族是一类具有特定形式的概率分布, 其密度函数可以表示为:

$$f(y;\theta,\phi)=\exp\left(\frac{y\theta-b(\theta)}{\phi}+c(y,\phi)\right)$$

其中: - θ 是自然参数 - ϕ 是离散参数 - $b(\theta)$ 是累积函数 - $c(y,\phi)$ 是归一化常数

1.2.1.2 均值和方差的关系 对于 EDM, 我们可以通过累积函数 $b(\theta)$ 得到均值和方差:

$$E(Y|\phi,\theta)=b'(\theta)$$

$$Var(Y|\phi,\theta)=\phi b''(\theta)$$

这里,$b'(\theta)$ 和 $b''(\theta)$ 分别是 $b(\theta)$ 的一阶和二阶导数。

1.2.2 5.1.2 方差函数的定义

通过标准参数 θ 与均值参数 μ 的一一对应关系, 我们可以将方差表示为均值的函数:

$$Var(Y|\mu,\phi)=\phi V(\mu)$$

这里,$V(\mu)$ 就是我们所说的**方差函数**, 它描述了方差如何随均值变化。ϕ 是**离散参数**, 控制着总体的变异程度。

1.2.3 5.1.3 方差函数的性质

命题 5.1: EDM 可以通过方差函数 $Var(Y|\mu,\phi)=\phi V(\mu)$ 的关系唯一确定。

这个命题的重要性在于, 它告诉我们, 如果我们知道一个分布的方差函数, 我们就可以完全确定这个分布的所有特性。

证明: 为了证明这个命题, 我们需要两个关键的引理:

- **引理 5.1:** $\theta(\mu) = \int \frac{1}{V(\mu)} d\mu$
- **引理 5.2:** $b(\theta(\mu)) = \int \mu V(\mu) d\mu$

这两个引理告诉我们如何从方差函数推导出自然参数 θ 和累积函数 $b(\theta)$。

通过这两个引理, 我们可以: 1. 从 $V(\mu)$ 得到 $\theta(\mu)$ 2. 从 $\theta(\mu)$ 得到 $b(\theta)$ 3. 有了 $b(\theta)$, 我们就可以完全确定 EDM 的分布

这个过程说明, 知道方差函数就足以确定整个分布。

1.2.4 5.1.4 例子 - 正态分布

让我们以正态分布为例, 来看看方差函数是如何工作的。

对于**正态分布**: - 方差函数 $V(\mu) = 1$ (方差与均值无关) - 累积函数 $b(\theta) = \frac{\theta^2}{2}$

我们可以验证: 1. $\theta(\mu) = \int \frac{1}{V(\mu)} d\mu = \int 1 d\mu = \mu$ 2. $b(\theta) = \int \mu V(\mu) d\mu = \int \mu d\mu = \frac{\mu^2}{2} = \frac{\theta^2}{2}$

这与我们知道的正态分布特性完全一致。

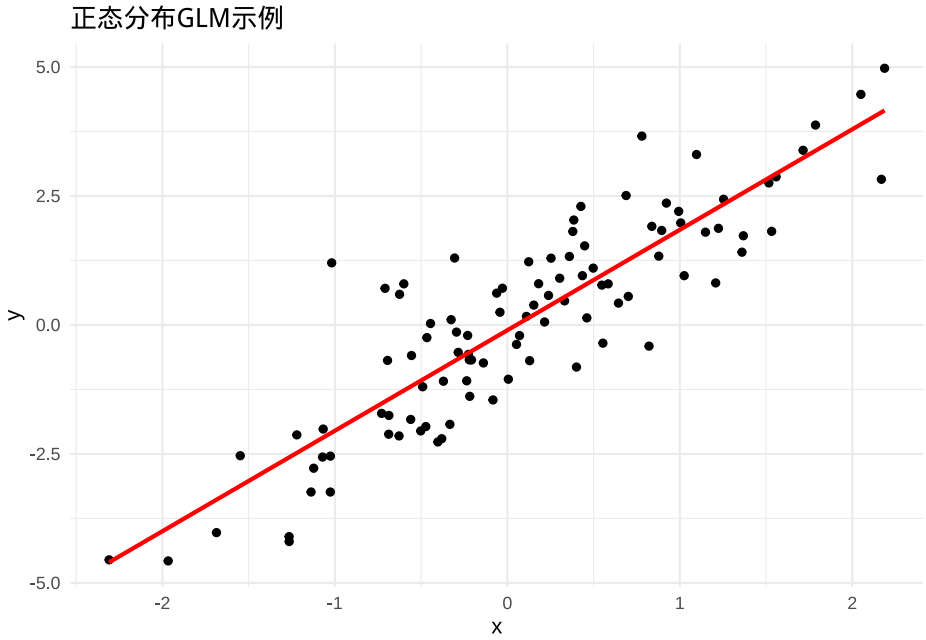
让我们用 R 来拟合一个正态分布的 GLM:

```r
# 生成模拟数据
set.seed(123)
n <- 100
x <- rnorm(n)
y <- 2 * x + rnorm(n)

# 拟合正态分布的广义线性模型
glm_normal <- glm(y ~ x, family = gaussian())
summary(glm_normal)
```

```
##
## Call:
## glm(formula = y ~ x, family = gaussian())
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.10280    0.09755  -1.054    0.295
## x            1.94753    0.10688  18.222   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.942287)
##
##     Null deviance: 405.218  on 99  degrees of freedom
## Residual deviance:  92.344  on 98  degrees of freedom
## AIC: 281.82
##
## Number of Fisher Scoring iterations: 2
```

```
# 可视化
ggplot(data.frame(x = x, y = y), aes(x = x, y = y)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = " 正态分布 GLM 示例", x = "x", y = "y") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



在这个例子中: - 我们生成了符合正态分布的数据 - 使用 `glm()` 函数拟合了 GLM, 指定 `family = gaussian()` 表示使用正态分布 - 输出显示了系数估计、标准误差和显著性 - 图形展示了数据点和拟合的线性关系

1.2.5 5.1.5 方差函数的重要性

理解方差函数对于选择合适的 GLM 模型至关重要: 1. **模型选择**: 不同的响应变量可能有不同的方差函数, 这指导我们选择合适的分布族。2. **诊断**: 通过检查残差和方差函数的关系, 我们可以评估模型拟合的好坏。3. **解释**: 方差函数帮助我们理解预测的不确定性如何随均值变化。

在接下来的章节中, 我们将看到更多不同分布的方差函数及其应用。

1.3 5.2 准似然 (Quasi-likelihood)

1.3.1 5.2.1 引言

在某些情况下, 我们可能遇到这样的数据: 它们的均值-方差关系似乎遵循某种已知的模式, 但不完全符合任何标准的概率分布。这就是准似然方法发挥作用的地方。

1.3.2 5.2.2 准似然的概念

准似然是一种在不完全指定概率分布的情况下进行统计推断的方法。它只需要指定:

1. 均值和方差之间的关系（即方差函数）
2. 一个连接函数

这比完整的似然函数需要的信息要少, 因此更加灵活。

1.3.3 5.2.3 为什么需要准似然?

考虑以下情况: - 泊松分布假设均值等于方差 ($V(\mu) = \mu$) - 二项分布假设方差是均值和样本大小的函数 ($V(\mu) = \mu(1-\mu)/n$)

但在实际数据中, 我们可能观察到: - 方差比均值大（过度离散）- 方差比均值小（欠离散）

在这些情况下, 标准的 GLM 可能不够灵活, 这时准似然方法就变得非常有用。

1.3.4 5.2.4 准似然模型的例子

两个常见的准似然模型是:

1. **准泊松模型**:
    - 方差函数: $V(\mu) = \mu$

- 但允许 $\phi \neq 1$（标准泊松分布假设 $\phi = 1$）

2. **准二项模型**:
   - 方差函数: $V(\mu) = \mu(1-\mu)$
   - 同样允许 $\phi \neq 1$

1.3.5 5.2.5 准似然的优势

1. **灵活性**: 可以处理不完全符合标准分布的数据
2. **稳健性**: 对分布假设的偏离不太敏感
3. **简单性**: 不需要完整指定概率分布

1.3.6 5.2.6 例子：Noisy Miner 数据

让我们看一个实际的例子, 使用 Noisy Miner 数据集。这个数据集来自一项研究澳大利亚噪声矿工鸟的生态学研究。

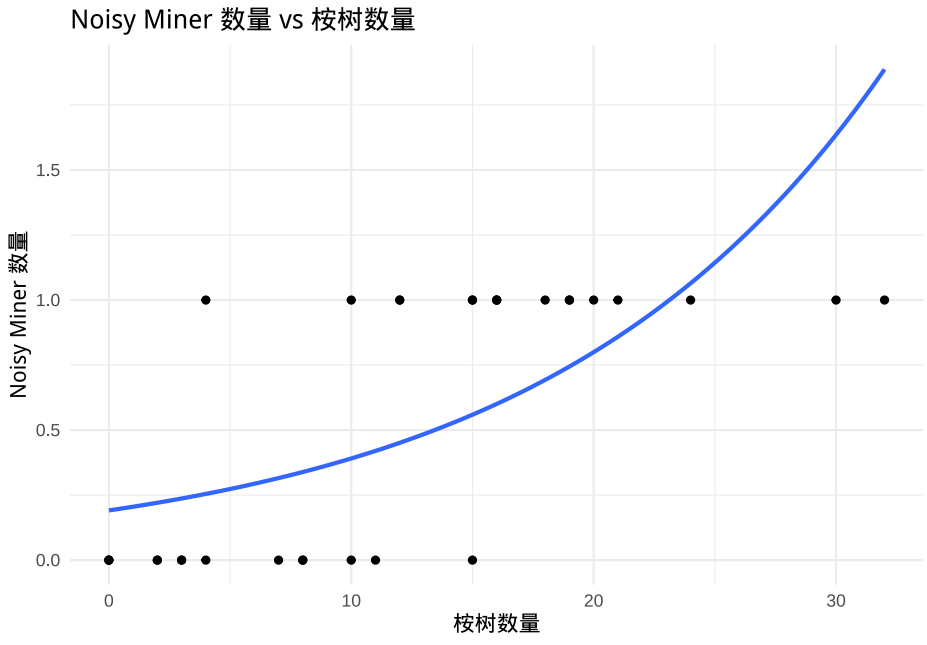
```
data(nminer)
# 修正后的 GLM 模型
glm_poisson <- glm(Miners ~ Eucs, data = nminer, family = poisson())

# 查看模型摘要
summary(glm_poisson)
```

```
##
## Call:
## glm(formula = Miners ~ Eucs, family = poisson(), data = nminer)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.65485    0.53051  -3.119  0.00181 **
## Eucs         0.07156    0.02683   2.667  0.00764 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 20.426  on 30  degrees of freedom
## Residual deviance: 13.473  on 29  degrees of freedom
## AIC: 51.473
##
## Number of Fisher Scoring iterations: 5
```

```r
# 绘制图表
library(ggplot2)
ggplot(nminer, aes(x = Eucs, y = Miners)) +
  geom_point() +
  geom_smooth(method = "glm", method.args = list(family = "poisson"), se = FALSE) +
  labs(title = "Noisy Miner 数量 vs 桉树数量",
       x = " 桉树数量", y = "Noisy Miner 数量") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



在这个例子中:

1. 我们首先拟合了一个标准的泊松 GLM。
2. 然后拟合了一个准泊松 GLM。
3. 比较两个模型的摘要, 你会注意到:
   - 系数估计是相同的
   - 但准泊松模型的标准误差和 p 值不同
   - 准泊松模型估计了一个离散参数 ϕ

准泊松模型考虑了数据的过度离散性, 因此提供了更可靠的标准误差和 p 值估计。

1.3.7 5.2.7 如何选择使用准似然

1. **检查离散度**: 如果泊松或二项 GLM 的离散度明显大于 1, 考虑使用准似然。
2. **理论考虑**: 如果你有理由相信数据可能存在额外的变异源。
3. **模型诊断**: 如果标准 GLM 的残差图显示明显的模式。

准似然方法为我们提供了一种处理不完全符合标准分布假设的数据的强大工具, 特别是在处理计数数据和比例数据时。

1.4 5.3 Gamma 分布

1.4.1 5.3.1 Gamma 分布介绍

Gamma 分布是一个非常灵活的连续概率分布, 常用于建模正值连续数据, 特别是当数据显示出方差随均值增加而增加的趋势时。

1.4.2 5.3.2 Gamma 分布的性质

概率密度函数:

对于形状参数 $s > 0$ 和速率参数 $r > 0$, Gamma 分布 $Y \sim \Gamma(r, s)$ 的概率密度函数为:

$$p(y|r,s) = \frac{\exp(-ry)r^s y^{s-1}}{\Gamma(s)}$$

其中 $y > 0$, $\Gamma(s)$ 是 gamma 函数。

重要性质: - 均值: $E(Y) = \frac{s}{r}$ - 方差: $Var(Y) = \frac{s}{r^2}$ - 方差函数: $V(\mu) = \mu^2$

1.4.3 5.3.3 Gamma 分布的应用场景

Gamma 分布适用于许多实际情况, 例如: 1. 等待时间或服务时间 2. 降雨量或河流流量 3. 保险索赔金额 4. 修理时间 5. 生存时间分析

这些情况通常涉及正值且右偏的数据分布。

1.4.4 5.3.4 在 GLM 中使用 Gamma 分布

在 R 中, 我们可以使用 `family = Gamma(link = "log")` 来指定 Gamma 分布的 GLM。对数链接函数通常是一个很好的选择, 因为它确保了预测值始终为正。

让我们用一个实际例子来说明 Gamma GLM 的使用:

```
# 加载 lime 数据
data(lime)

# 修正后的 Gamma GLM 模型
glm_gamma <- glm(Foliage ~ DBH + Age, data = lime, family = Gamma(link = "log"))

# 查看模型摘要
summary(glm_gamma)
```

```
##
## Call:
## glm(formula = Foliage ~ DBH + Age, family = Gamma(link = "log"),
##     data = lime)
```

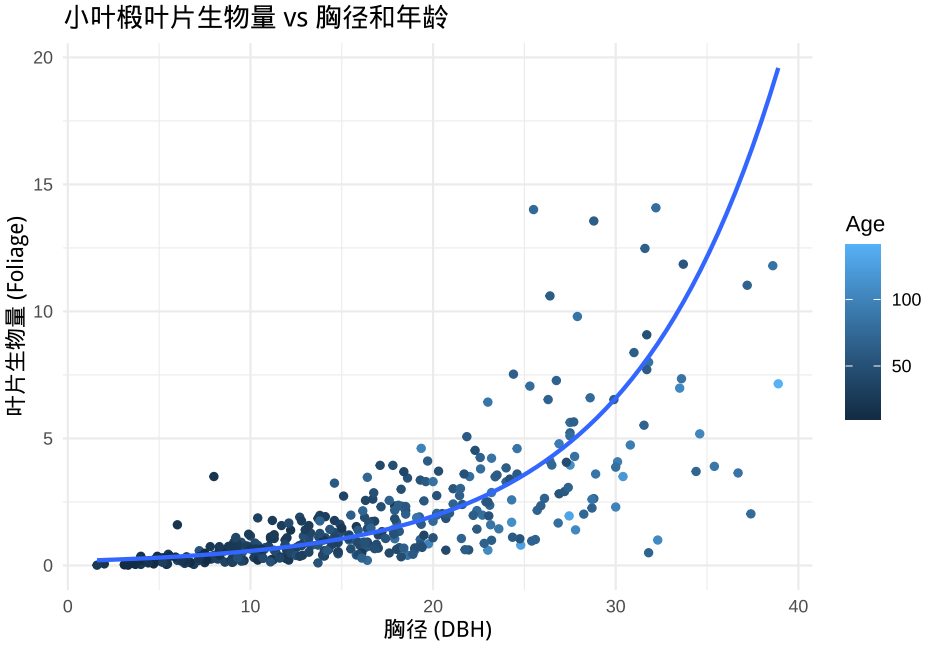
```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.743541   0.090141 -19.342   <2e-16 ***
## DBH          0.133088   0.008175  16.279   <2e-16 ***
## Age         -0.004419   0.002777  -1.591    0.112
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.5374689)
##
##     Null deviance: 508.48  on 384  degrees of freedom
## Residual deviance: 188.29  on 382  degrees of freedom
## AIC: 830.84
##
## Number of Fisher Scoring iterations: 5
```

```r
# 绘制图表
library(ggplot2)
ggplot(lime, aes(x = DBH, y = Foliage, color = Age)) +
  geom_point() +
  geom_smooth(method = "glm",
              method.args = list(family = Gamma(link = "log")),
              se = FALSE) +
  labs(title = " 小叶椴叶片生物量 vs 胸径和年龄",
       x = " 胸径 (DBH)", y = " 叶片生物量 (Foliage)") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: The following aesthetics were dropped during statistical transformation:
## colour.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
```

```
##   variable into a factor?
```



在这个例子中: 1. 我们使用了 `lime` 数据集, 这是关于小叶椴树木生物量的数据。2. 我们拟合了一个 Gamma GLM, 使用树干周长和树高作为预测变量。3. 模型摘要给出了系数估计、标准误差和显著性。4. 图表显示了生物量如何随树干周长变化, 并用颜色表示不同的树高。

1.4.5 5.3.5 解释 Gamma GLM 结果

1. **系数**: 由于我们使用了对数链接函数, 系数可以解释为对响应变量的乘性效应。例如, 如果树干周长的系数是 0.05, 这意味着树干周长每增加一个单位, 预期生物量会增加约 5%。
2. **显著性**: 通过 p 值我们可以判断每个预测变量的重要性。
3. **模型拟合**: 可以通过残差偏差和 AIC 等指标来评估模型拟合的好坏。
4. **诊断**: 应检查残差图, 以确保模型假设得到满足。

1.4.6 5.3.6 Gamma GLM 的优势和注意事项

优势: 1. 适合建模正值、右偏的连续数据。2. 方差随均值增加而增加, 这在许多实际问题中很常见。3. 比使用对数变换的线性回归更灵活。

注意事项: 1. 需要确保数据都是严格正的。2. 在极端值附近的拟合可能不太理想。3. 解释需要小心, 特别是当使用非恒等链接函数时。

1.5 5.4 逆高斯分布 (Inverse Gaussian Distribution)

1.5.1 5.4.1 逆高斯分布介绍

逆高斯分布, 也称为 Wald 分布, 是另一种用于建模正值连续数据的分布。它特别适用于高度偏斜的正值数据, 其中方差随均值的增加而迅速增加。

1.5.2 5.4.2 逆高斯分布的性质

概率密度函数:

对于均值参数 $\mu > 0$ 和形状参数 $\lambda > 0$, 逆高斯分布 $Y \sim IG(\mu, \lambda)$ 的概率密度函数为:

$$f(y|\mu, \lambda) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left(-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right)$$

其中 $y > 0$。

重要性质: - 均值: $E(Y) = \mu$ - 方差: $Var(Y) = \frac{\mu^3}{\lambda}$ - 方差函数: $V(\mu) = \mu^3$

1.5.3 5.4.3 逆高斯分布的应用场景

逆高斯分布在多个领域都有应用, 例如: 1. 金融风险分析 2. 可靠性工程 3. 生存分析 4. 极端事件建模 5. 某些物理过程, 如布朗运动的首次到达时间

1.5.4 5.4.4 在 GLM 中使用逆高斯分布

在 R 中, 我们可以使用 `family = inverse.gaussian(link = "log")` 来指定逆高斯分布的 GLM。同样, 对数链接函数通常是一个好选择。

让我们用一个实例来说明逆高斯 GLM 的使用:

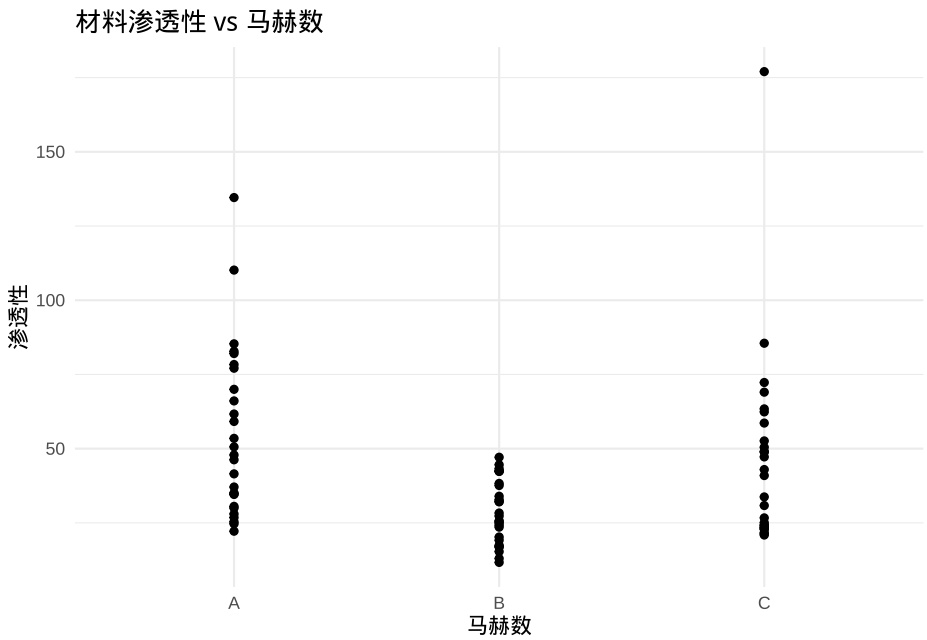
```
data(perm)
# 修正后的逆高斯 GLM 模型
glm_inverse_gaussian <- glm(Perm ~ Mach,
                            data = perm,
                            family = inverse.gaussian(link = "log"))

# 查看模型摘要
summary(glm_inverse_gaussian)
```

```
##
## Call:
## glm(formula = Perm ~ Mach, family = inverse.gaussian(link = "log"),
##     data = perm)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.0011     0.1169  34.214  < 2e-16 ***
## MachB        -0.6390     0.1445  -4.421 3.14e-05 ***
## MachC        -0.1729     0.1587  -1.089    0.279
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for inverse.gaussian family taken to be 0.006755721)
##
##     Null deviance: 0.61746  on 80  degrees of freedom
## Residual deviance: 0.47701  on 78  degrees of freedom
## AIC: 699.89
##
```

```
## Number of Fisher Scoring iterations: 5
```

```
# 绘制图表
library(ggplot2)
ggplot(perm, aes(x = Mach, y = Perm)) +
  geom_point() +
  geom_smooth(method = "glm",
              method.args = list(family = inverse.gaussian(link = "log")),
              se = FALSE, color = "red") +
  labs(title = " 材料渗透性 vs 马赫数",
       x = " 马赫数", y = " 渗透性") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



在这个例子中: 1. 我们使用了 perm 数据集, 这是关于建筑材料渗透性的数据。2. 我们拟合了一个逆高斯 GLM, 使用压力作为预测变量。3. 模型摘要给出了系数估计、标准误差和显著性。4. 图表显示了渗透性如何随压力变化。

1.5.5 5.4.5 解释逆高斯 GLM 结果

1. **系数**: 由于使用了对数链接函数, 系数可以解释为对响应变量的乘性效应。
2. **显著性**: p 值告诉我们压力对渗透性的影响是否显著。
3. **模型拟合**: 可以通过残差偏差和 AIC 等指标来评估模型拟合的好坏。
4. **诊断**: 应检查残差图, 以确保模型假设得到满足。

1.5.6 5.4.6 逆高斯 GLM 的优势和注意事项

优势: 1. 适合建模高度偏斜的正值连续数据。2. 方差随均值的三次方增加, 适用于某些特定的物理或金融过程。3. 在某些情况下可能比 Gamma 分布提供更好的拟合。

注意事项: 1. 数据必须严格为正值。2. 对极端值很敏感, 需要谨慎处理异常值。3. 解释可能不如其他更常见的分布直观。

1.6 5.5 泊松分布与二项分布的极限 (Poisson Limit of the Binomial Distribution)

1.6.1 5.5.1 引言

泊松分布和二项分布是两种常用于离散数据建模的分布。在某些条件下, 二项分布会趋近于泊松分布。理解这种关系对于选择合适的模型非常重要, 特别是在处理稀有事件时。

1.6.2 5.5.2 泊松极限定理

考虑一个二项分布 $Y \sim Binomial(m, \mu)$, 其中: - m 是试验次数 - μ 是每次试验成功的概率

当 $m \to \infty$ 且 $m\mu \to \lambda$ (常数) 时, 二项分布趋近于参数为 λ 的泊松分布:

$$Y \sim Poisson(\lambda)$$

这个结果被称为泊松极限定理。

1.6.3 5.5.3 直观理解

这个定理可以这样理解: - 当事件非常罕见 (很小), 但观察次数很多 (m 很大) 时 - 总的事件发生次数 (m) 保持在一个固定水平 ( ) - 那么事件的发生可以用泊松分布很好地近似

1.6.4 5.5.4 数学证明

证明涉及到矩生成函数 (MGF) 的计算, 这里我们简要概述步骤:

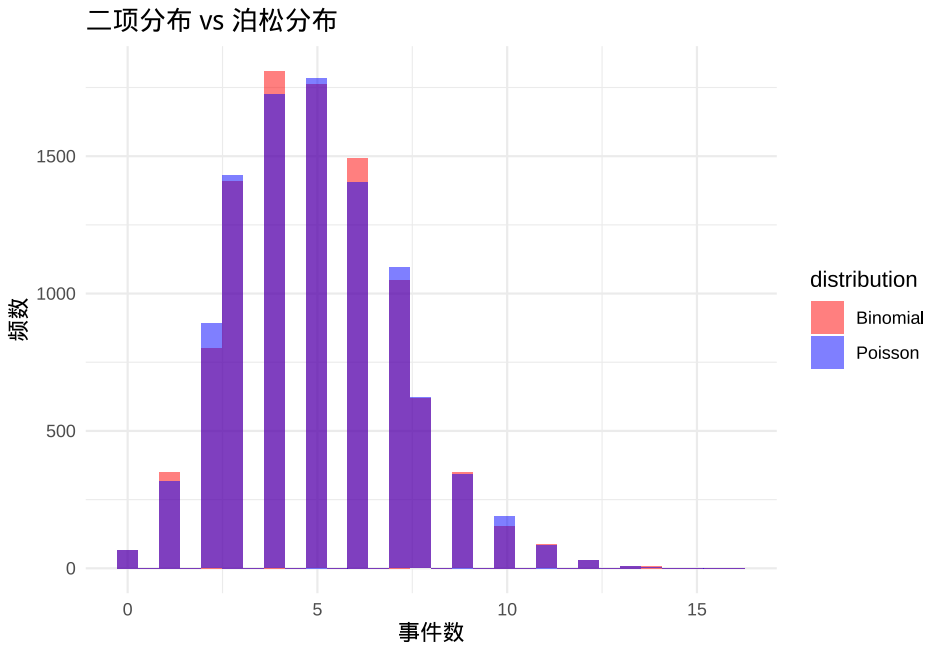
1. 写出二项分布的 MGF
2. 代入 $\mu = \lambda/m$
3. 取极限 $m \to \infty$
4. 得到的结果正是泊松分布的 MGF

1.6.5 5.5.5 R 实现：模拟二项分布与泊松分布的比较

让我们通过模拟来直观地看这个结果:

```
# 设置参数
set.seed(123)
n <- 10000  # 模拟次数
m <- 1000   # 二项分布的试验次数
lambda <- 5 # 期望事件数

# 生成数据
binom_data <- rbinom(n, size = m, prob = lambda/m)
poisson_data <- rpois(n, lambda = lambda)
```

```r
# 创建数据框
df <- data.frame(
  value = c(binom_data, poisson_data),
  distribution = rep(c("Binomial", "Poisson"), each = n)
)

# 绘制直方图
ggplot(df, aes(x = value, fill = distribution)) +
  geom_histogram(position = "identity", alpha = 0.5, bins = 30) +
  labs(title = " 二项分布 vs 泊松分布",
       x = " 事件数", y = " 频数") +
  theme_minimal() +
  scale_fill_manual(values = c("red", "blue"))
```



在这个例子中: 1. 我们模拟了二项分布和泊松分布的数据。2. 二项分布的参数设置使其接近泊松分布的条件。3. 图表显示了两种分布的直方图重叠, 视觉上很难区分。

1.6.6 5.5.6 实际应用

理解泊松极限对实际建模有重要意义:

1. **简化**: 在某些情况下, 使用泊松模型可能比二项模型更简单。
2. **计算效率**: 对于大样本, 泊松近似可能在计算上更高效。
3. **模型选择**: 在处理罕见事件时, 这个理论指导我们可以考虑使用泊松模型。

然而, 需要注意的是, 这种近似在实际应用中并不总是成立, 特别是当事件不是很罕见或样本量不够大时。

1.7 5.6 如何选择泊松与二项 GLM

1.7.1 5.6.1 引言

在处理离散数据时, 特别是计数数据, 我们经常需要在泊松 GLM 和二项 GLM 之间做出选择。这个决定对模型的适当性和结果的解释都有重要影响。

1.7.2 5.6.2 泊松 GLM vs 二项 GLM

泊松 GLM: - 用于建模**计数**或**速率**数据 - 假设事件发生是独立的 - 方差等于均值

二项 GLM: - 用于建模**比例**或**概率**数据 - 假设有固定数量的试验 - 方差是均值和试验次数的函数

1.7.3 5.6.3 选择标准

1. **数据类型**:
   - 如果数据是纯计数, 考虑泊松 GLM
   - 如果数据是成功/失败的比例, 考虑二项 GLM
2. **暴露量**:
   - 如果每个观察有不同的暴露时间或空间, 泊松 GLM 更合适

- 如果每个观察有固定的试验次数, 二项 GLM 更合适

3. **过度离散**:
   - 泊松分布假设均值等于方差
   - 如果数据显示过度离散, 考虑负二项分布或准泊松模型
4. **理论考虑**:
   - 考虑数据生成过程的本质
   - 某些领域可能有使用特定模型的传统

1.7.4 5.6.4 实例比较

让我们通过一个实例来比较泊松 GLM 和二项 GLM:

```r
# 生成模拟数据
set.seed(123)
n <- 1000
x <- rnorm(n)
exposure <- runif(n, 1, 10)

# 泊松过程
lambda <- exp(1 + 0.5 * x)
y_poisson <- rpois(n, lambda * exposure)

# 二项过程
p <- exp(1 + 0.5 * x) / (1 + exp(1 + 0.5 * x))
trials <- rpois(n, 20)  # 随机试验次数
y_binomial <- rbinom(n, size = trials, prob = p)

# 拟合模型
glm_poisson <- glm(y_poisson ~ x + offset(log(exposure)), family = poisson())
glm_binomial <- glm(cbind(y_binomial, trials - y_binomial) ~ x, family = binomial())

# 比较模型摘要
summary(glm_poisson)
```

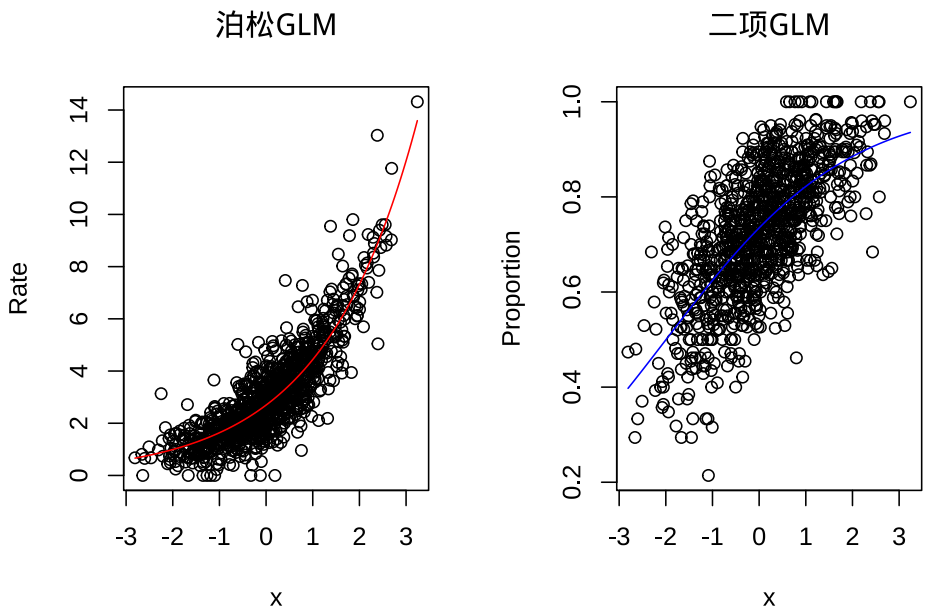
```
##
## Call:
## glm(formula = y_poisson ~ x + offset(log(exposure)), family = poisson())
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.991583   0.008671  114.35   <2e-16 ***
## x           0.499221   0.007560   66.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 5375.9  on 999  degrees of freedom
## Residual deviance: 1037.9  on 998  degrees of freedom
## AIC: 5374.7
##
## Number of Fisher Scoring iterations: 4
```

```
summary(glm_binomial)
```

```
##
## Call:
## glm(formula = cbind(y_binomial, trials - y_binomial) ~ x, family = binomial())
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.01935    0.01656   61.54   <2e-16 ***
## x            0.51088    0.01730   29.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 1972.0  on 999  degrees of freedom
## Residual deviance: 1034.6  on 998  degrees of freedom
## AIC: 4084.7
##
## Number of Fisher Scoring iterations: 4
```

```r
# 可视化
par(mfrow = c(1, 2))
plot(x, y_poisson/exposure, main = " 泊松 GLM", xlab = "x", ylab = "Rate")
curve(exp(coef(glm_poisson)[1] + coef(glm_poisson)[2] * x), add = TRUE, col = "red")

plot(x, y_binomial/trials, main = " 二项 GLM", xlab = "x", ylab = "Proportion")
curve(plogis(coef(glm_binomial)[1] + coef(glm_binomial)[2] * x), add = TRUE, col = "blu
```



5.6.5 解释结果

1. **泊松 GLM**:
   - 系数直接解释为对数速率的变化
   - 使用 offset 项来考虑不同的暴露时间
2. **二项 GLM**:
   - 系数解释为对数优势比的变化

- 使用 cbind() 来指定成功和失败的次数

1.7.5 5.6.6 模型诊断

选择模型后, 进行适当的诊断非常重要:

1. **残差分析**:
   - 检查残差的正态性和同方差性
   - 寻找异常值或高杠杆点
2. **过度离散检验**:
   - 对于泊松模型, 检查离散度是否显著大于 1
   - 如果存在过度离散, 考虑使用准泊松或负二项模型
3. **连接函数检验**:
   - 确保选择的连接函数适合数据
4. **拟合优度**:
   - 使用 AIC 或 BIC 比较不同模型

1.7.6 5.6.7 实际应用建议

1. **数据探索**: 始终从探索性数据分析开始, 了解数据的分布和特征。
2. **理论指导**: 考虑研究问题的本质和领域知识。
3. **模型比较**: 尝试拟合多个模型并比较它们的性能。
4. **注意边界情况**: 特别是在处理罕见事件或小样本时。
5. **解释谨慎**: 记住模型的假设和限制, 在解释结果时保持谨慎。

1.8 总结

本章深入探讨了广义线性模型 (GLMs) 的核心概念和应用。我们详细讨论了:

1. **方差函数**的重要性及其在定义 GLMs 中的角色。

2. **准似然**方法如何扩展了标准 GLMs 的应用范围。
3. **Gamma 分布**和**逆高斯分布**在 GLMs 中的应用, 特别是对于正值连续数据。
4. **泊松分布**作为**二项分布**的极限, 及其在建模稀有事件中的应用。
5. 如何在**泊松 GLM** 和**二项 GLM** 之间做出选择。

关键要点:

- GLMs 提供了一个统一的框架来处理各种类型的响应变量。
- 选择合适的分布和连接函数对于成功应用 GLMs 至关重要。
- 模型诊断和适当的解释是 GLM 分析中不可或缺的步骤。
- 实际应用中, 要结合统计理论、领域知识和数据特征来选择最合适的模型。

通过掌握这些概念和技术, 你将能够更加灵活和准确地分析各种类型的数据, 从而在统计建模和数据科学领域取得更好的成果。

记住, 模型选择和应用是一个需要实践和经验的过程。鼓励你使用本章介绍的方法来分析实际数据集, 并持续深化对 GLMs 的理解。