第 4 章指数分散模型 (EDMs) 详细笔记

小狗

目录

1 指数分散模型 (Exponential Dispersion Models, EDMs)

1.1 1. 引言

在统计学中, 我们经常需要处理各种类型的数据和分布。指数分散模型 (Exponential Dispersion Models, EDMs) 是一类重要的统计模型, 它为许多常见的概率分布提供了统一的框架。本章我们将深入学习 EDMs 的概念、特性和应用。

1.2 2. 动机 (Motivation)

1.2.1 2.1 为什么需要 EDMs?

在统计学习的过程中, 你可能已经接触过正态分布、泊松分布、二项分布等。这些分布看似不同, 但实际上它们有一些共同的特性。EDMs 就是为了捕捉这些共同特性而提出的。

EDMs 的一个重要特性是: 对于独立同分布的观测值, 其最大似然估计 (MLE) 恰好等于样本均值。这个特性在实际应用中非常有用, 因为它简化了参数估计的过程。

1.2.2 2.2 示例: 估计共同均值

让我们通过一个具体的例子来理解这一点。假设我们有一组数据, 我们想估计它的均值。我们可以用两种方法:

1. **样本均值**: 这是我们最常用的方法, 简单直接。$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
2. **最大似然估计 (MLE)**: 这是一种更复杂但更强大的方法。$\hat{\mu}_{MLE} = \arg\max_{\mu} \sum_{i=1}^{n} \log(p(x_i|\mu))$

在理想情况下, 这两种方法应该给出相同的结果。但事实是否如此呢? 让我们用 R 代码来验证:

```r
set.seed(123)
n <- 1000

# 生成三种不同分布的数据
normal_data <- rnorm(n, mean = 5, sd = 2)
poisson_data <- rpois(n, lambda = 5)
t_data <- rt(n, df = 5)

# 计算样本均值
mean_normal <- mean(normal_data)
mean_poisson <- mean(poisson_data)
```

```r
mean_t <- mean(t_data)

# 定义对数似然函数
log_likelihood_normal <- function(mu) sum(dnorm(normal_data, mean = mu, sd = sd(normal_
log_likelihood_poisson <- function(mu) sum(dpois(poisson_data, lambda = mu, log = TRUE)
log_likelihood_t <- function(mu) sum(dt((t_data - mu)/sd(t_data), df = 5, log = TRUE))

# 计算 MLE
mle_normal <- optimize(log_likelihood_normal, interval = c(0, 10), maximum = TRUE)$maxi
mle_poisson <- optimize(log_likelihood_poisson, interval = c(0, 10), maximum = TRUE)$ma
mle_t <- optimize(log_likelihood_t, interval = c(-5, 5), maximum = TRUE)$maximum

# 创建结果表格
results <- data.frame(
  Distribution = c("Normal", "Poisson", "t"),
  Sample_Mean = c(mean_normal, mean_poisson, mean_t),
  MLE = c(mle_normal, mle_poisson, mle_t)
)

kable(results, caption = " 样本均值与 MLE 的比较", digits = 4)
```

表 1: 样本均值与 MLE 的比较

Distribution	Sample_Mean	MLE
Normal	5.0323	5.0323
Poisson	4.9870	4.9870
t	-0.0364	-0.0310

从上表我们可以看出:

1. 对于正态分布和泊松分布, 样本均值和 MLE 非常接近。这两种分布都是 EDMs 的例子。
2. 对于 t 分布, 样本均值和 MLE 有明显差异。t 分布不是 EDM。

这个例子说明了 EDMs 的一个重要特性: 样本均值和 MLE 的一致性。这种一致性使得 EDMs 在统计建模中特别有用。

1.3 3. EDM 的定义 (Definition of an EDM)

现在, 让我们正式定义什么是指数分散模型。

1.3.1 3.1 EDM 的规范形式

指数分散模型的概率密度函数 (或概率质量函数) 可以写成以下规范形式:

$$p(y|\theta,\phi) = a(y,\phi)\exp\left(\frac{\theta y - b(\theta)}{\phi}\right)$$

这个公式看起来可能有点复杂, 让我们逐项解释:

- y 是观测值。
- θ 是**规范参数 (canonical parameter)**。它决定了分布的位置 (如均值)。
- ϕ 是**分散参数 (dispersion parameter)**。它控制分布的尺度 (如方差)。
- $b(\theta)$ 是**累积量函数 (cumulant function)**。它在确定分布的性质中起着关键作用。
- $a(y,\phi)$ 是**归一化函数**。它确保概率密度函数的积分为 1。

1.3.2 3.2 EDM 的例子

让我们看几个常见分布的 EDM 形式:

1. **正态分布** $N(\mu,\sigma^2)$: $\theta = \mu$, $\phi = \sigma^2$, $b(\theta) = \frac{\theta^2}{2}$
2. **泊松分布** $Poisson(\lambda)$: $\theta = \log(\lambda)$, $\phi = 1$, $b(\theta) = e^\theta$
3. **二项分布** $Binomial(n,p)$: $\theta = \log(\frac{p}{1-p})$, $\phi = 1$, $b(\theta) = n\log(1+e^\theta)$

1.3.3 3.3 支持集

不同的 EDM 有不同的支持集 (即 y 可能的取值范围):

- 正态分布: $S = \mathbb{R}$ (所有实数)
- 泊松分布: $S = \{0, 1, 2, ...\}$ (非负整数)
- 二项分布: $S = \{0, 1, ..., n\}$ (0 到 n 的整数)

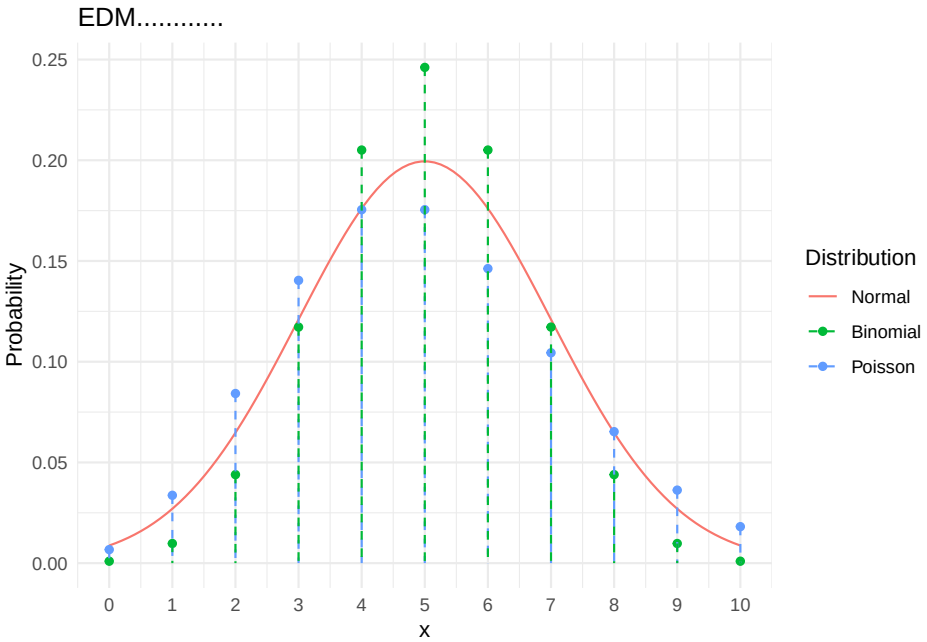
让我们用 R 代码来可视化这些分布:

```r
# 定义分布函数
normal_pdf <- function(x) dnorm(x, mean = 5, sd = 2)
poisson_pmf <- function(x) dpois(x, lambda = 5)
binomial_pmf <- function(x) dbinom(x, size = 10, prob = 0.5)

# 创建数据框
x_normal <- seq(0, 10, length.out = 100)
x_discrete <- 0:10

df <- data.frame(
  x = c(x_normal, x_discrete, x_discrete),
  y = c(normal_pdf(x_normal), poisson_pmf(x_discrete), binomial_pmf(x_discrete)),
  Distribution = factor(rep(c("Normal", "Poisson", "Binomial"), c(100, 11, 11)))
)

# 绘图
ggplot(df, aes(x = x, y = y, color = Distribution)) +
  geom_line(data = subset(df, Distribution == "Normal")) +
  geom_point(data = subset(df, Distribution != "Normal")) +
  geom_segment(data = subset(df, Distribution != "Normal"),
               aes(xend = x, yend = 0), linetype = "dashed") +
  labs(title = "EDM 分布示例", x = "x", y = "Probability") +
  theme_minimal() +
  scale_x_continuous(breaks = 0:10)
```



这个图展示了三种不同 EDM 的概率分布。注意正态分布是连续的, 而泊松和二项分布是离散的。

1.4 4. 加权 EDMs(Weighted EDMs)

在实际应用中, 我们经常遇到需要对不同观测赋予不同权重的情况。加权 EDMs 就是为了处理这种情况而引入的。

1.4.1 4.1 加权 EDM 的定义

假设我们有独立的观测值 $Y_1, ..., Y_n$, 它们来自同一个 EDM, 有相同的规范参数 θ, 但可能有不同的分散参数。加权 EDM 的定义为:

$$p(y_i|\theta, \phi) = a(y_i, \frac{\phi}{w_i}) \exp\left(\frac{w_i[\theta y_i - b(\theta)]}{\phi}\right)$$

这里 w_i 是已知的权重。

1.4.2 4.2 加权 EDM 的解释

- 权重 w_i 可以看作是观测 y_i 的重要性或可靠性的度量。
- 较大的权重意味着该观测在模型中有更大的影响。
- 在实践中, 权重可能来自样本设计、测量精度或先验知识。

1.4.3 4.3 加权样本均值和 MLE

让我们通过 R 代码来比较加权样本均值和加权 MLE:

```r
set.seed(123)
n <- 1000

# 生成数据和权重
y <- rnorm(n, mean = 5, sd = 2)
weights <- runif(n, 0.5, 1.5)

# 计算加权样本均值
weighted_mean <- sum(weights * y) / sum(weights)

# 定义加权对数似然函数
weighted_log_likelihood <- function(mu) {
  sum(weights * dnorm(y, mean = mu, sd = sd(y), log = TRUE))
}

# 计算加权 MLE
weighted_mle <- optimize(weighted_log_likelihood, interval = c(0, 10), maximum = TRUE)$

# 输出结果
cat(" 加权样本均值:", weighted_mean, "\n")
```

```
## 加权样本均值: 5.046039
```

```r
cat(" 加权 MLE:", weighted_mle, "\n")
```

```
## 加权MLE: 5.046039
```

我们可以看到, 加权样本均值和加权 MLE 非常接近, 这再次验证了 EDMs 的特性。

1.5 5. 累积量函数 (Cumulants for EDMs)

累积量函数 $b(\theta)$ 是 EDM 中的核心概念, 它决定了分布的许多重要性质。

1.5.1 5.1 累积量函数的性质

对于 EDM, 我们有以下重要性质:

1. **期望**: $E(Y|\theta) = b'(\theta)$
2. **方差**: $Var(Y|\theta) = \phi b''(\theta)$

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

1.5.2 5.2 累积量函数的例子

让我们看几个常见分布的累积量函数:

1. **正态分布**: $b(\theta) = \frac{\theta^2}{2}$
2. **泊松分布**: $b(\theta) = e^\theta$
3. **二项分布**: $b(\theta) = n\log(1 + e^\theta)$

我们可以用 R 来可视化这些函数及其导数:

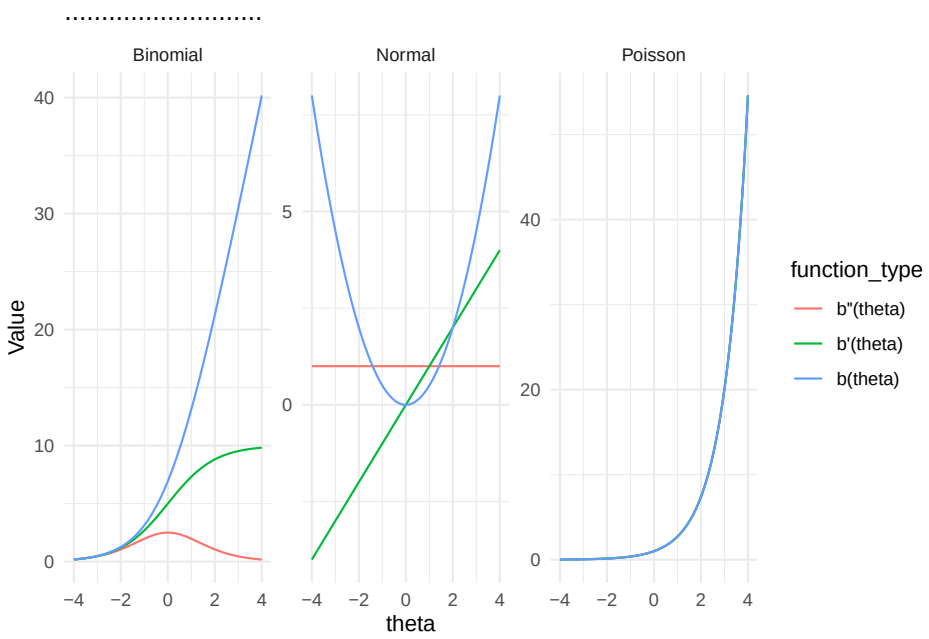
```r
# 定义累积量函数及其导数
b_normal <- function(theta) theta^2 / 2
b_prime_normal <- function(theta) theta
b_double_prime_normal <- function(theta) rep(1, length(theta))
```

```r
b_poisson <- function(theta) exp(theta)
b_prime_poisson <- function(theta) exp(theta)
b_double_prime_poisson <- function(theta) exp(theta)

b_binomial <- function(theta) 10 * log(1 + exp(theta))  # 假设 n=10
b_prime_binomial <- function(theta) 10 * exp(theta) / (1 + exp(theta))
b_double_prime_binomial <- function(theta) 10 * exp(theta) / (1 + exp(theta))^2

# 创建数据框
theta <- seq(-4, 4, length.out = 100)
df <- data.frame(
  theta = rep(theta, 9),
  value = c(b_normal(theta), b_prime_normal(theta), b_double_prime_normal(theta),
            b_poisson(theta), b_prime_poisson(theta), b_double_prime_poisson(theta),
            b_binomial(theta), b_prime_binomial(theta), b_double_prime_binomial(theta)),
  function_type = rep(rep(c("b(theta)", "b'(theta)", "b''(theta)"), each = 100), 3),
  distribution = rep(c("Normal", "Poisson", "Binomial"), each = 300)
)

# 绘图
ggplot(df, aes(x = theta, y = value, color = function_type)) +
  geom_line() +
  facet_wrap(~ distribution, scales = "free_y") +
  labs(title = " 累积量函数及其导数", x = "theta", y = "Value") +
  theme_minimal()
```



这个图展示了三种分布的累积量函数及其导数。注意它们的形状差异, 这反映了不同分布的特性。

1.6 6. 规范链接函数 (The canonical link)

规范链接函数是连接 EDM 的均值参数和规范参数的桥梁。

1.6.1 6.1 定义

对于每个 EDM, 存在一个函数 g 将均值 μ 映射到规范参数 θ:

$$g(\mu) = \theta$$

这就是规范链接函数。它的逆函数 $h = g^{-1}$ 被称为规范均值函数:

$$h(\theta) = \mu = b'(\theta)$$

1.6.2 6.2 常见分布的规范链接函数

1. **正态分布**: $g(\mu) = \mu$ (恒等链接)
2. **泊松分布**: $g(\mu) = \log(\mu)$ (对数链接)
3. **二项分布**: $g(\mu) = \log(\frac{\mu}{1-\mu})$ (logit 链接)

1.6.3 6.3 规范链接函数的重要性

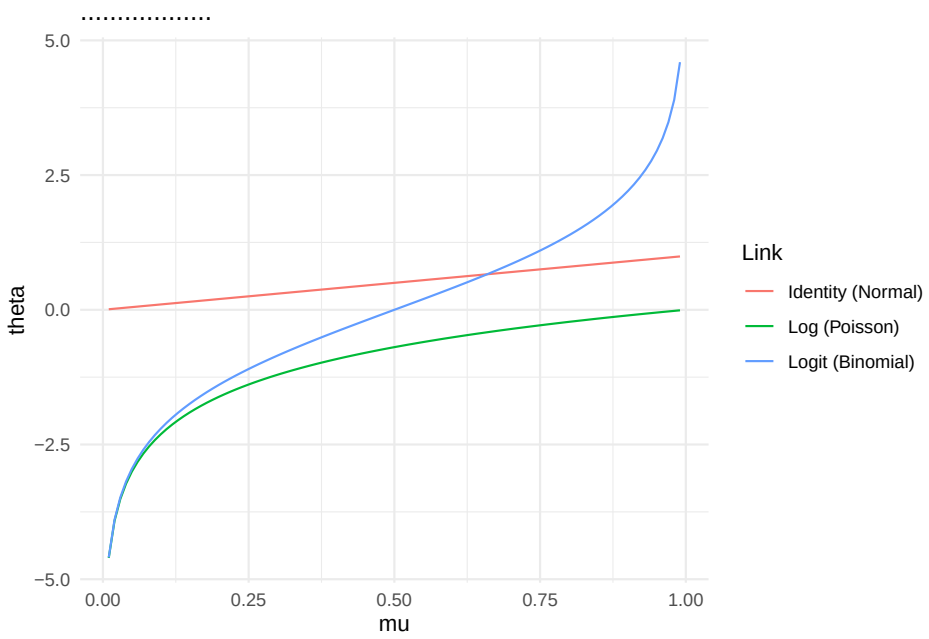
规范链接函数在广义线性模型 (GLM) 中扮演着重要角色。它们提供了一种自然的方式来连接线性预测器和响应变量的均值。使用规范链接函数通常可以简化计算并提高模型的解释性。

让我们用 R 代码来可视化这些链接函数:

```r
# 定义链接函数
identity_link <- function(mu) mu
log_link <- function(mu) log(mu)
logit_link <- function(mu) log(mu / (1 - mu))

# 创建数据框
mu <- seq(0.01, 0.99, length.out = 100)
df <- data.frame(
  mu = rep(mu, 3),
  theta = c(identity_link(mu), log_link(mu), logit_link(mu)),
  Link = rep(c("Identity (Normal)", "Log (Poisson)", "Logit (Binomial)"), each = 100)
)

# 绘图
ggplot(df, aes(x = mu, y = theta, color = Link)) +
  geom_line() +
  labs(title = " 规范链接函数", x = "mu", y = "theta") +
  theme_minimal()
```



6. 规范链接函数 (The canonical link) 规范链接函数是连接 EDM 的均值参数和规范参数的桥梁。6.1 定义对于每个 EDM, 存在一个函数 g 将均值 μ 映射到规范参数 θ:

$$g(\mu) = \theta$$

这就是规范链接函数。它的逆函数 $h = g^{-1}$ 被称为规范均值函数:

$$h(\theta) = \mu = b'(\theta)$$

6.2 常见分布的规范链接函数

正态分布: $g(\mu) = \mu$ (恒等链接) 2. **泊松分布**: $g(\mu) = \log(\mu)$ (对数链接) 3. **二项分布**: $g(\mu) = \log(\frac{\mu}{1-\mu})$ (logit 链接)

1.6.4 6.3 规范链接函数的重要性

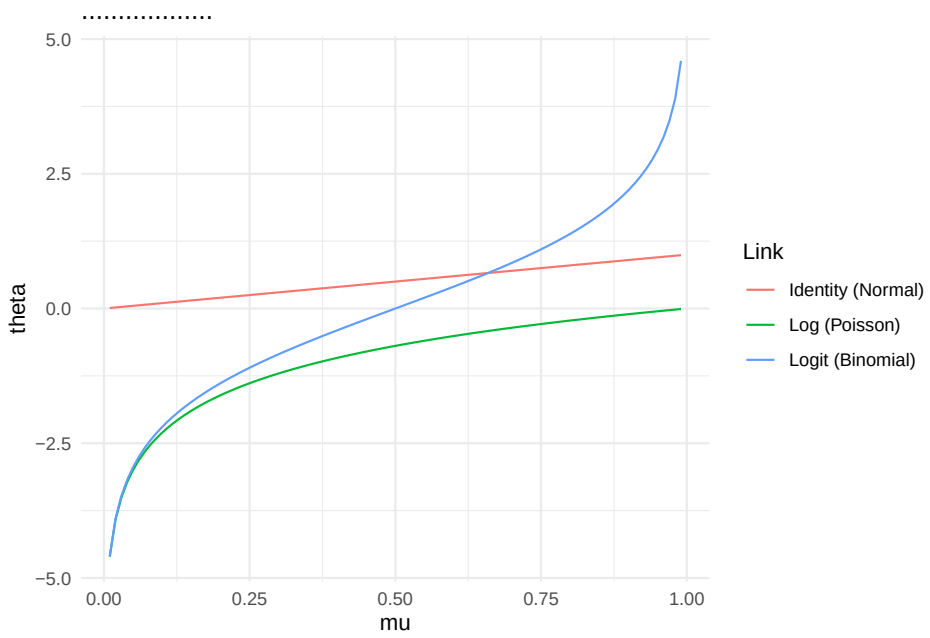
规范链接函数在广义线性模型 (GLM) 中扮演着重要角色。它们提供了一种自然的方式来连接线性预测器和响应变量的均值。使用规范链接函数通常可以简化计算并提高模型的解释性。

让我们用 R 代码来可视化这些链接函数:

```r
# 定义链接函数
identity_link <- function(mu) mu
log_link <- function(mu) log(mu)
logit_link <- function(mu) log(mu / (1 - mu))

# 创建数据框
mu <- seq(0.01, 0.99, length.out = 100)
df <- data.frame(
  mu = rep(mu, 3),
  theta = c(identity_link(mu), log_link(mu), logit_link(mu)),
  Link = rep(c("Identity (Normal)", "Log (Poisson)", "Logit (Binomial)"), each = 100)
)

# 绘图
ggplot(df, aes(x = mu, y = theta, color = Link)) +
  geom_line() +
  labs(title = " 规范链接函数", x = "mu", y = "theta") +
  theme_minimal()
```



这个图展示了三种常见分布的规范链接函数。注意它们如何将 μ 的值域映射到整个实数轴。

1.7 7. 偏差 (Deviance)

偏差是衡量模型拟合优度的重要指标。它基于似然比统计量, 比较当前模型与饱和模型（完全拟合数据的模型）的差异。

1.7.1 7.1 偏差的定义

对于 EDM, 总偏差定义为:

$$D(y, \mu) = \sum_{i=1}^{n} w_i d(y_i, \mu_i)$$

其中 $d(y_i, \mu_i)$ 是单位偏差, 定义为:

$$d(y, \mu) = 2\left[t(y, y) - t(y, \mu)\right]$$

这里 $t(y, \mu) = y\theta - b(\theta)$, 且 $\theta = g(\mu)$。

1.7.2 7.2 常见分布的单位偏差

1. **正态分布**: $d(y, \mu) = (y - \mu)^2$
2. **泊松分布**:
   - 如果 $y > 0$: $d(y, \mu) = 2\left[y \log(\frac{y}{\mu}) - (y - \mu)\right]$
   - 如果 $y = 0$: $d(0, \mu) = 2\mu$
3. **二项分布**: $d(y, \mu) = 2\left[y \log(\frac{y}{\mu}) + (1 - y) \log(\frac{1-y}{1-\mu})\right]$

1.7.3 7.3 偏差的应用

偏差有多种用途:

1. **模型比较**: 较小的偏差通常表示更好的拟合。
2. **模型诊断**: 偏差残差可用于检查模型假设。
3. **变量选择**: 偏差的变化可以帮助判断是否应该包含某个变量。

让我们用 R 代码来计算和比较不同分布的偏差:

```r
# 生成数据
set.seed(123)
n <- 1000
y_normal <- rnorm(n, mean = 5, sd = 2)
y_poisson <- rpois(n, lambda = 5)
y_binomial <- rbinom(n, size = 1, prob = 0.7)

# 计算偏差
deviance_normal <- function(y, mu) sum((y - mu)^2)
deviance_poisson <- function(y, mu) {
  2 * sum(ifelse(y == 0, mu, y * log(y/mu) - (y - mu)))
}
deviance_binomial <- function(y, mu) {
  2 * sum(y * log(y/mu) + (1-y) * log((1-y)/(1-mu)))
}

# 计算每个分布的偏差
dev_normal <- deviance_normal(y_normal, mean(y_normal))
dev_poisson <- deviance_poisson(y_poisson, mean(y_poisson))
dev_binomial <- deviance_binomial(y_binomial, mean(y_binomial))

# 创建结果表格
results <- data.frame(
  Distribution = c("Normal", "Poisson", "Binomial"),
  Deviance = c(dev_normal, dev_poisson, dev_binomial)
)

kable(results, caption = " 不同分布的偏差比较", digits = 2)
```

表 2: 不同分布的偏差比较

Distribution	Deviance
Normal	3929.90
Poisson	1050.48
Binomial	NaN

这个表格展示了三种不同分布的偏差。请注意, 这些偏差值本身并不能直接比较, 因为它们来自不同的分布。但在同一分布内, 我们可以用偏差来比较不同模型的拟合优度。

1.8 8. EDMs 在实践中的应用

EDMs 为许多统计方法提供了理论基础, 尤其是在广义线性模型 (GLM) 中。以下是一些 EDMs 在实际中的应用:

1. **线性回归**: 使用正态分布 EDM。
2. **逻辑回归**: 使用二项分布 EDM。
3. **泊松回归**: 使用泊松分布 EDM, 常用于计数数据。
4. **生存分析**: 某些生存模型可以表示为 EDM。
5. **方差分析 (ANOVA)**: 可以看作是正态 EDM 的特例。

让我们通过一个简单的 GLM 例子来说明 EDMs 的应用:

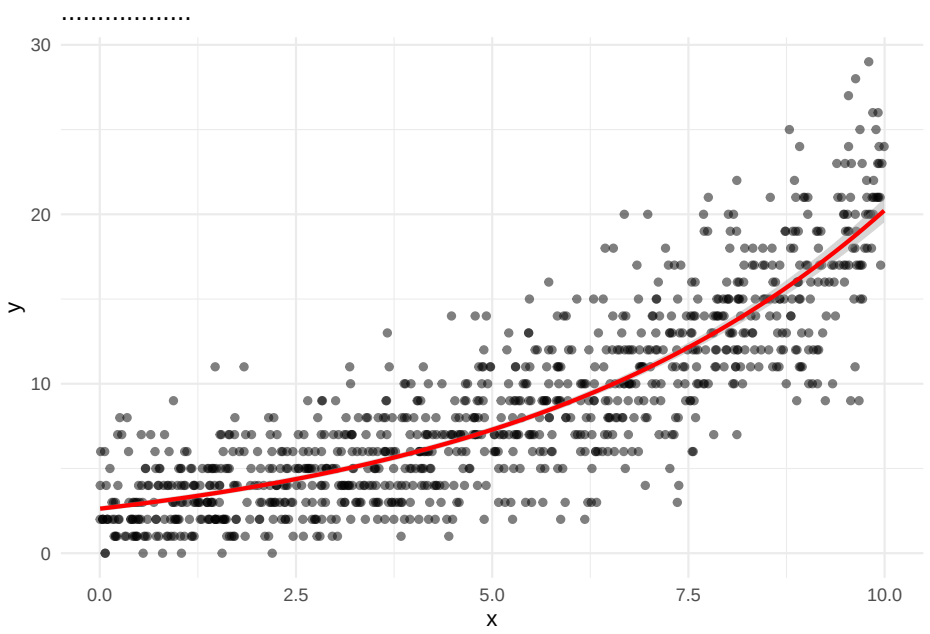
```r
# 生成模拟数据
set.seed(123)
n <- 1000
x <- runif(n, 0, 10)
lambda <- exp(1 + 0.2 * x)
y <- rpois(n, lambda)

# 拟合泊松回归模型
model <- glm(y ~ x, family = poisson)
```

```r
# 查看模型摘要
summary(model)
```

```
##
## Call:
## glm(formula = y ~ x, family = poisson)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.965772   0.029156   33.12   <2e-16 ***
## x           0.204178   0.004129   49.45   <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: 3743.8  on 999  degrees of freedom
## Residual deviance: 1056.4  on 998  degrees of freedom
## AIC: 4808.6
##
## Number of Fisher Scoring iterations: 4
```

```r
# 可视化结果
ggplot(data.frame(x = x, y = y), aes(x = x, y = y)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "glm", method.args = list(family = "poisson"), color = "red") +
  labs(title = " 泊松回归示例", x = "x", y = "y") +
  theme_minimal()
```



在这个例子中, 我们使用泊松分布 EDM 来建模计数数据。模型摘要给出了参数估计和显著性检验, 而图形展示了数据和拟合的模型。

1.9 9. 总结

指数分散模型 (EDMs) 是一个强大的统计工具, 它为多种常见分布提供了统一的框架。通过学习 EDMs, 我们可以:

1. 理解不同概率分布之间的联系。
2. 掌握参数估计的一般方法。
3. 理解广义线性模型的基础。
4. 学会如何评估模型拟合。

EDMs 的核心概念包括规范参数、分散参数、累积量函数、规范链接函数和偏差。这些概念不仅在理论上很重要, 在实际数据分析中也有广泛应用。

通过本章的学习, 你应该能够: - 识别常见分布是否属于 EDM 族。- 理解 EDM 的基本性质和参数。- 使用 R 进行基本的 EDM 相关计算和可视化。- 理解 EDMs 如何应用于实际问题, 特别是在广义线性模型中的应用。

在接下来的学习中, 我们将看到 EDMs 如何为更复杂的统计模型和方法奠定基础。继续深入学习, 你将发现 EDMs 在现代统计学和数据科学中的重要作用。