属性数据分析

第一章

- 在属性数据分析中**响应变量**为属性数据
  - **响应变量** Y 是属性变量
  - **解释变量** X 可以是连续也可以是属性
- 分类变量的两种类型：
  - **名义**（Nominal）： 无序的类别
  - **有序**（Ordinal）： 有序的类别
- 二项分布：
  - $$P(Y=k)=\binom{n}{k}\pi^k(1-\pi)^{n-k},\quad k=0,1,\ldots,n$$
  - 该分布为试验次数为n、试验成功概率为π的二项分布，均值为$n\pi$、方差为$n\pi(1-\pi)$。当n足够大时，Y的分布近似为$N(n\pi,n\pi(1-\pi))$。
  - Rcode *dbinom(0,3,0.6)* 表示$P(X=0)$，$X\sim B(3,0.6)$ 。
  - $X\sim B(n,p)$，大n
    $$P(X\le k)\approx\Phi\left(\frac{k-np}{\sqrt{np(1-p)}}\right)$$
  - $X\sim B(n,p)$，Yates（1934），有修正
    $$P(X\le k)\approx\Phi\left(\frac{k+1/2-np}{\sqrt{np(1-p)}}\right)$$
- 多项分布：
  - $$P(X_1=n_1,\ldots,X_c=n_c)=\frac{n!}{n_1!\cdots n_c!}\pi_1^{n_1}\cdots\pi_c^{n_c}$$
  - $X\sim M(N,p_1,\ldots,p_n)$
  - 如果记$(X_{i1},\ldots,X_{ic})$为第i次试验的结果，其中
    $$X_{ij}=\begin{cases}1, & \text{如果第}i\text{次试验结果为}A_j;\\ 0, & \text{否则}.\end{cases}$$

显然，$X_{i1},\ldots,X_{ic}$中恰好有一个为1而其余的为0，且n次试验中A_j发生的总次数为$X_j=\sum_{i=1}^n X_{ij}$。那么易由二项分布的定义知$X_j\sim B(n,\pi_j),j=1,\ldots,c$。从而$X_j$的均值和方差分别为$n\pi_j$和$n\pi_j(1-\pi_j)$。另外，当$1\le j_1\le j_2\le n$时，

$$cov(X_{j_1},X_{j_2})=cov(\sum_{i=1}^n X_{ij_1},\sum_{i=1}^n X_{ij_2})=\sum_{i=1}^n cov(X_{ij_1},X_{ij_2})=-n\pi_{j_1}\pi_{j_2}$$

- 二项分布的显著性检验
  - 对于给定的π_0，考虑原检验$H_0:\pi=\pi_0$，通常采用大样本检验的方法，相应的Z检验统计量为

$$Z=\frac{\hat{\pi}-\pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

显然，Z在H_0下的极限分布（称为零极限分布）为标准正态分布，检验的P-值近似为$2[1-\Phi(|z|)]$，其中z为Z的样本观测值。

- P-值定义：当原假设为真时，比所得到的样本观察结果更极端的结果出现的概率，P值越小越有利于对立假设

$$\text{P-值}=P_{H_0}(|T|>\tau)$$

- 二项比例的置信区间
  - 二项比例的显著性检验统计量可以写成

$$Z^{'}=\frac{\hat{\pi}-\pi_0}{\text{SE}},\text{其中 SE}=\sqrt{\hat{\pi}(1-\hat{\pi})/n}$$

由Slutsky定理知$Z^{'}$的零极限分布为标准正态分布，故相应的近似水平α的检验接受域为

$$\{|Z^{'}|\le u_{\alpha/2}\}。$$

即

$$P_{\pi_0}(|Z^{'}|\le u_{\alpha/2})\approx 1-\alpha,$$

表明$\hat{\pi}\pm u_{\alpha/2}\text{SE}$为$\pi$的近似水平$1-\alpha$的置信区间。该区间是标准的Z-置信区间。

- 离散数据的更多统计推断
  对于任意参数β的显著性检验，针对$H_0:\beta=\beta_0$

- Wald统计量
  基于Z-统计量的双边检验等价于基于Z-统计量的平方

$$W := Z^2 = \frac{(\hat{\beta} - \beta_0)^2}{\widehat{var}(\hat{\beta})}$$

  Z^2的零极限分布为$\mathcal{X}_1^2$ 分布。对于p维的β，Wald统计量为

$$W = (\hat{\beta} - \beta_0)^T \widehat{cov}(\hat{\beta})^{-1} (\hat{\beta} - \beta_0)$$

  当$\hat{\beta}$取为极大似然估计时，$\hat{\beta}$的协方差矩阵可以取为Fisher信息阵的逆

$$W = -(\hat{\beta} - \beta_0)^T E_{\beta_0} \ddot{\ell}(\beta_0) (\hat{\beta} - \beta_0)$$

  当Fisher信息阵$-E_{\beta_0} \ddot{\ell}(\beta_0)$不好求时，可以改用样本Fisher信息阵$-\ddot{\ell}(\beta_0)$。
- 似然比检验统计量

$$LR = 2 \left[\sup_{\beta \in B} \ell(\beta) - \sup_{\beta \in B_0} \ell(\beta) \right]$$

  B_0维度为s，B维度维t，则LR近似为$\mathcal{X}_{t-s}^2$（这里不一定正确）
- 得分（Score）检验统计量

$$S := \dot{\ell}^T(\beta_0) [-E_{\beta_0} \ddot{\ell}(\beta_0)]^{-1} \dot{\ell}(\beta_0)$$

  当$-E_{\beta_0} \ddot{\ell}(\beta_0)$不方便计算时可以使用样本Fisher信息阵$-\ddot{\ell}(\beta_0)$，可以得到

$$S^{'} = \dot{\ell}^T(\beta_0) [-\ddot{\ell}(\beta_0)]^{-1} \dot{\ell}(\beta_0)$$

  S与$S^{'}$的零极限分布为$\mathcal{X}_p^2$。

- 二项参数的Wald，得分和似然比检验
  - Wald检验统计量（$\pi = X/n$）

$$W = \frac{(\hat{\pi} - \pi)^2}{\pi_0 (1 - \pi_0)/n}$$

  - 似然比检验统计量

$$LR = 2\left[X\log\frac{\hat{\pi}}{\pi_0} + (n-X)\log\frac{1-\hat{\pi}}{1-\pi_0}\right]$$

- 得分检验统计量

$$S = \frac{(\frac{X}{\pi_0} - \frac{n-X}{1-\pi_0})^2}{\frac{X}{\pi_0^2} - \frac{n-X}{(1-\pi_0)^2}}$$

注意当$\pi_0 = 0.5$时，S=W。

- 小样本二项推断
  - 当$n\pi \geq 5$且$n(1-\pi) \geq 5$时，大样本双边Z检验和置信区间有好的效果。对于小样本情形，直接使用二项分布计算P-值更安全。
  - 假设检验问题$H_0: \pi = \pi_0 \leftrightarrow H_a: \pi > \pi_0$，取检验拒绝域为

$$\{\hat{\pi} > \tau\} = \{X \geq \tau^{'}\}$$

记X的观测值为X_o，其独立于X，相应的P-值为

$$P_{H_0}(X \geq X_o \mid X_o) = \sum_{i=X_o}^{n}\binom{n}{i}\pi_0^i(1-\pi_0)^{n-i}.$$

对于双边假设检验问题$H_0: \pi = \pi_0 \leftrightarrow H_a: \pi \neq \pi_0$，取双侧拒绝域

$$\{X \leq \tau_1 或 X \geq \tau_2\}$$

相应的P-值为

$$\sum_{\substack{i:0\leq i\leq n \\ P_{H_0}(X=i)\leq P_{H_0}(X=X_o|X_o)}} P_{H_0}(X=i)$$

第二章

- 列联表的概率结构
  - 两个属性变量X和Y，如果都只有有限个取值，则可以用列联表来刻画。
  - 有关属性变量的概率计算及MLE均与其具体取值无关，但是均值、方差等数字特征与X和Y的取值有关。
- $\pi_{ij} = P(X=i, Y=j), \pi_{i+} = P(X=i), \pi_{+j} = P(Y=j)$

- $\pi_{j|i}^{Y|X}=P(Y=j \mid X=i)=\pi_{ij}/\pi_{i+}$表示给定X条件下Y=j的概率，如果都是对同一Y=j，可以简记为π_i。
- $n_{ij}=\sum_{k=1}^{n} I(X_k=i, Y_k=j)$，$n_{ij} \sim B(n_{i+}, \frac{\pi_{ij}}{\pi_{i+}})$，其中$\frac{\pi_{ij}}{\pi_{i+}}$可以由$\frac{n_{ij}}{n_{i+}}$估计。
- X为真实疾病状态（1=患某病；2=未患某病），Y为诊断结果（1=阳性；2=阴性），则称
  - **敏感度**=$P(Y=1 \mid X=1)$
  - **特异度**=$P(Y=2 \mid X=2)$
  - 敏感度和特异度越高，则诊断效果越好。
- 独立性
  - 两个属性变量X和Y的独立性定义：

$$\pi_{ij}=\pi_{i+}\pi_{+j}, \quad i=1,\ldots,I \text{ 及 } j=1,\ldots,J$$

- 二项抽样和多项抽样
  - 对X和Y都是随机的情形，$(n_{11}, n_{12}, \ldots, n_{IJ})$服从参数为n和$(\pi_{11}, \pi_{12}, \ldots, \pi_{IJ})$的多项分布，特别的，$n_{ij}$服从参数为n和$\pi_{ij}$的二项分布。
  - 给定$\{X=i\}$（此时n_{i+}非随机），$(n_{i1}, \ldots, n_{iJ})$服从参数为n_{i+} 和 π_{ij}/π_{i+}的二项分布。
- 2×2列联表的比较
  - X和Y均为二分变量（dichotomous variable），$\pi_1:=\pi_{1|1}^{Y|X}$ 和 $\pi_2:=\pi_{1|2}^{Y|X}, n_1=n_{1+}, n_2=n_{2+}$
  - **比例差**（proportion difference）或者**风险差**（risk difference）

$$\delta:=\pi_1-\pi_2$$

比例差的MLE为$\hat{\pi}_1-\hat{\pi}_2$， 其中 $\hat{\pi}=n_{11}/n_1$ 和 $\hat{\pi}_2=n_{21}/n_2$。
$H_0: \delta=0$的Z统计量

$$Z=\frac{\hat{\pi}_1-\hat{\pi}_2}{SE}, \text{ 其中 SE}=\sqrt{\hat{\pi}_1(1-\hat{\pi}_1)/n_1+\hat{\pi}_2(1-\hat{\pi}_2)/n_2}$$

$$\operatorname{Var}(\hat{\pi}_1-\hat{\pi}_2)=\operatorname{Var}(\hat{\pi}_1)+\operatorname{Var}(\hat{\pi}_2)=\hat{\pi}_1(1-\hat{\pi}_1)/n_1+\hat{\pi}_2(1-\hat{\pi}_2)/n_2$$

相应的δ的近似水平$1-\alpha$（Wald）置信区间可取为(CLT)

$$[\hat{\pi}_1-\hat{\pi}_2-u_{\alpha/2}\mathrm{SE}, \hat{\pi}_1-\hat{\pi}_2+u_{\alpha/2}\mathrm{SE}].$$

  - **相对风险**（relative risk）

$$RR = \frac{\pi_1}{\pi_2}$$

- 疫苗保护率：$\frac{\pi_1 - \pi_2}{\pi_1} = 1 - \frac{1}{RR}$。
- 当两组比例都很低时，比例差很小，不容易看出差异，这个时候使用相对风险。
- 相对风险的置信区间：先构造$\log(RR)$的置信区间（使用delta方法构造），然后再变回来。
- Delta方法：

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \Sigma)$$

则

$$\sqrt{n}[g(\hat{\theta}) - g(\theta)] = \sqrt{n}\dot{g}(\theta^\star)^T(\hat{\theta} - \theta) + o_p(1) \xrightarrow{d} N(0, \dot{g}(\theta)^T \Sigma \dot{g}(\theta))$$

- $\log(RR)$的极大似然估计为$\log(\hat{\pi}_1/\hat{\pi}_2)$，所以方差为

$$\begin{aligned}
\operatorname{var}(log\hat{\pi}_1 - \log \hat{\pi}_2) &\approx \frac{\operatorname{var}(\hat{\pi}_1)}{\hat{\pi}_1^2} + \frac{\operatorname{var}(\hat{\pi}_2)}{\hat{\pi}_2^2} \\
&\approx \frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{n_1 \hat{\pi}_1^2} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{n_1 \hat{\pi}_2^2} \\
&= \frac{1 - \hat{\pi}_1}{n_1 \hat{\pi}_1} + \frac{1 - \hat{\pi}_2}{n_1 \hat{\pi}_2} := \hat{\sigma}^2
\end{aligned}$$

因此，$\log(RR)$的近似水平$(1 - \alpha)$的置信区间为

$$[\log(\hat{\pi}_1/\hat{\pi}_2) - u_{\alpha/2}\hat{\sigma}, \log(\hat{\pi}_1/\hat{\pi}_2) + u_{\alpha/2}\hat{\sigma}]$$

从而RR的近似水平$(1 - \alpha)$的置信区间为

$$[\hat{\pi}_1/\hat{\pi}_2 \exp\{-u_{\alpha/2}\hat{\sigma}\}, \hat{\pi}_1/\hat{\pi}_2 \exp\{u_{\alpha/2}\hat{\sigma}\}].$$

- Rcode，epitools包函数riskratio可以计算相对风险RR。

- **优势比**
  - **优势**或**比值**（odds）：

$$\text{odds} = \frac{\pi}{1 - \pi}$$

  - 两个概率π_1 和 π_2的优势之比为优势比（odds ratio，OR）

$$\theta = \mathrm{OR} = \frac{\mathrm{odds}_1}{\mathrm{odds}_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

- 优势比与相对风险的关系：

$$\mathrm{OR} = 1 \Longleftrightarrow \mathrm{RR} = 1,$$

$$\mathrm{OR} > 1 \Longleftrightarrow \mathrm{RR} > 1,$$

$$\mathrm{OR} < 1 \Longleftrightarrow \mathrm{RR} < 1.$$

- 优势比越偏离1，代表概率差别越大，也就是有越强的关联性。
- $$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$
- 样本优势比$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$.
- 样本对数优势比$\log(\hat{\theta})$的分布更接近钟形分布（即正态分布），由delta法估计得到标准差$\mathrm{SE} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$.
  证明如下：

$$\begin{aligned} \mathrm{var}[\log n_{11} - \log n_{12}] &= \mathrm{var}[\log n_{11} - \log(n_{1+} - n_{11})] \\ &\approx (1/n_{11} + 1/n_{12})^2 \mathrm{var}(n_{11}) \\ &= 1/n_{11} + 1/n_{12}. \end{aligned}$$

  同理可以得到

$$\mathrm{var}[\log n_{21} - \log n_{22}] \approx 1/n_{21} + 1/n_{22}$$

  再有独立性知

$$\begin{aligned} \mathrm{var}[\log(\hat{\theta})] &= \mathrm{var}[\log n_{11} - \log n_{12}] + \mathrm{var}[\log n_{21} - \log n_{22}] \\ &\approx \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \end{aligned}$$

- $\log(\theta)$的渐进水平$1-\alpha$的置信区间为

$$[\log(\hat{\theta}) - u_{\alpha/2}\mathrm{SE}, \log(\hat{\theta}) + u_{\alpha/2}\mathrm{SE}]$$

- θ的渐进水平$1-\alpha$的置信区间为

$$[\hat{\theta}\exp\{-u_{\alpha/2}\mathrm{SE}\}, \hat{\theta}\exp\{u_{\alpha/2}\mathrm{SE}\}]$$

- Rcode，epitools包oddsratio函数。

- $$\mathrm{OR} = \mathrm{RR} \times \frac{1-\pi_2}{1-\pi_1}.$$

  - 如果π_1 和 π_2都很小，则$\mathrm{OR} \approx \mathrm{RR}$。

- 病例-对照研究
  在流行病学研究中，有一类回溯性的研究方法。例如先找一些心肌梗死病人，然后找一些与心肌梗死病人年龄性别等潜在混淆因素相当的正常人，通过问卷调查的方式了解受试者的既往吸烟史，这样的回溯性研究称为病例-对照研究。因为抽样一般是严重有偏的，即不能认为疾病状态的抽样是随机的，所以不能使用比例差和相对风险。但是吸烟情况是随机抽样的，因此可以估计患病者吸烟的概率$P(X=1 \mid Y=1)$，和正常者的吸烟概率$P(X=1 \mid Y=2)$。所以可以估计优势比。

- Pearson的卡方检验
  - 考虑属性变量只有两个类别的情形，其取值为第一个类别的概率记为p，两个类别的观测频数分别记为X和Y。设感兴趣的原检验是：

$$H_0 : \pi = \pi_0$$

$$\mathcal{X}^2 = \sum_{i=1}^{k} \frac{(n_i - n\pi_{i0})^2}{n\pi_{i0}} \sim \mathcal{X}^2_{k-1}, \text{其中k为种类数量}$$

如果原假设中概率π_{i0}是r个独立参数$\theta_1, \ldots, \theta_r$的函数：

$$\pi_{i0} = \pi_{i0}(\theta_1, \ldots, \theta_r),$$

记θ_j的估计量为$\hat{\theta}_j = \theta_j(n_1, \ldots, n_k)$,，代入$\pi_{i0}$的表达式得到其估计量$\hat{\pi}_{i0} := \pi_{i0}(\hat{\theta}_1, \ldots, \hat{\theta}_r)$,相应的检验统计量改成：

$$\mathcal{X}^2 = \sum_{i=1}^{k} \frac{(n_i - n\hat{\pi}_{i0})^2}{n\hat{\pi}_{i0}}$$

直观上自由度为k-r-1。

- 列联表的独立性检验
  - $I \times J$列联表的独立性假设为

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}, i = 1, \ldots, I \text{ 和 } j = 1, \ldots, J.$$

  - Pearson的拟合优度检验统计量

$$\mathcal{X}^2 = \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

  其中$\hat{\mu}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = \frac{n_{i+}n_{+j}}{n}$，服从$\mathcal{X}^2_{(I-1)(J-1)}$。水平$\alpha$的拒绝域可以取成

$$\{\mathcal{X}^2 \geq \mathcal{X}^2_{(I-1)(J-1)}(\alpha)\}$$

  - 似然比检验统计量（P39）

$$G^2 = 2\sum_{i,j} n_{ij} \log \frac{n_{ij}}{\hat{\mu}_{ij}}$$

  可以证明，G^2与X^2是渐进等价的。在小样本下，X^2有更好的表现。
  - Pearson残差

$$e_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}}$$

  - 标准化残差

$$\frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1 - \hat{\pi}_{i+})(1 - \hat{\pi}_{+j})}}$$

  每个单元格的标准化残差近似服从标准正态分布，如果标准化残差绝对值过大（大于2或3），表明该单元格拟合不佳。
- 卡方统计量的分解（P44或书P33）
- 有序数据的独立性检验（广义线性模型会讲或书P34）
- 小样本的精确推断（课上没讲，书P39）
  - 当样本量较小时，有些理论频数可能低于5，此时卡方检验不太适用，可以改用超几何分布分布的Fisher精确检验法。
  - Rcode，fisher.test函数实现。

- 三项列联表的关联性
  - 在研究X与Y之间的关联性的研究中，需要控制混淆因素Z的影响，对Z进行分层处理是一种常用的控制其影响的重要方法。
  - 假设X与Y条件不关联（给定Z），但Z同时与X和Y关联时，X与Y之间会有边际关联性，这种关联性称为伪关联性。（P49）
  - **部分表**：对Z的每一个类别，X与Y各类别组合计数构成的列联表称为部分表
  - **边缘表**：各个部分表计数相加得到的表称为XY边缘表。
  - **辛普森悖论**（Simpson Paradox）：边缘关联和条件关联的结论相反的情况称为辛普森悖论。
  - **条件优势比**：部分表相应的优势比称为条件优势比。
  - **边缘优势比**：边缘表对应的优势比称为边缘优势比。
  - 齐次关联性
    - 记Z的类别数为K，当X和Y均为二分变量时，如果条件优势比相等：

      $$\theta_{(1)} = \theta_{(2)} = \cdots = \theta_{(K)} = \theta$$

      则称X与Y是齐次关联的。特别地，如果公共的条件优势比θ等于1，则X与Y是条件独立的。
    - 对于$I \times J \times K$表，称X与Y是齐次关联的，如果X与Y的任意类别组合(i_1, i_2, j_1, j_2)的条件优势比不依赖于Z的类别。（P58）
    - XY有齐次关联性，则ZY和ZX也有齐次关联性。（P59）
    - 当XY有齐次关联性的时候，即X对Y的效应不受Z的影响或等价地Z对Y的效应不受X影响，称Z与X对Y的效应没有交互效应。
    - 当没有齐次关联性时，两个变量的条件优势比随第三个变量的改变而改变。

第三章

- 广义线性模型的构成部分
  - **随机部分**：识别响应变量Y并假设其概率分布
  - **系统部分**：指定模型线性预测函数中用到的解释变量（自变量）
  - **联系函数**：指定Y的期望关于X的函数
- 随机部分
  - 设$Y_1, \ldots, Y_n$为Y的独立观测
    - 如果Y表示试验成功次数（试验总次数$n \geq 1$非随机），可以假定其服从二项分布。
    - 如果Y代表计数，可以假定其服从泊松分布（没有超散布）或者负二项分布（超散布）。
    - 如果Y连续，可以假定服从正态分布。
- 系统部分
  - GLM的系统部分指定公式（线性预测量）

$$\alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

　　这里的$x_1, ..., x_k$可以是原始变量$z_1, z_2, ...$的函数

- 联系函数
  - 联系函数（link function）是将$\mu := EY$与线性预测量联系起来的函数$g(\cdot)$：

$$g(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k.$$

　　联系函数连接了随机部分和系统部分。

- 常见联系函数
  - **恒等联系**（常用于连续响应变量）：

$$\mu = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

　　如普通的线性回归模型

  - **对数联系**（常用于计数响应变量）：

$$\log(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

　　相应的模型称为对数线性模型

  - **逻辑联系**（适用于二分响应变量）：

$$\log(\mu/(1-\mu)) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

　　相应的模型称为逻辑回归模型

- 如果响应变量服从指数分布

$$f(y; \theta, \phi) = \exp\{[y\theta - b(\theta)]/a(\phi) + c(y, \phi)\}$$

　则称θ为自然参数。
  - 正态分布的自然参数就是均值本身
  - 二项分布的自然参数是成功概率的logit变换
  - 取自然参数等于$g(\mu)$时，相应的联系函数称为典则联系。
- 正态GLM（P7或书P57）
- 二分数据的广义线性模型

- Bernoulli分布是二分响应变量Y（不妨记取值为0或1）的分布，概率$\pi = P(Y=1)$ 和 $1-\pi = P(Y=0)$确定，均值为$EY=\pi$。如果π依赖于解释变量的取值x，则用$\pi(x)$来代替π。
- 线性概率模型

  $$\pi(x) = \alpha + \beta x$$

  该模型采用的是恒等联系，尽管可以直接采用最小二乘法估计，但是缺点就是拟合值经常会超出合理的范围。
- Logistic回归模型
  当发现$\pi(x)$ 与 x非线性关系，实际中采用Logistic回归。

  $$\operatorname{logit}[\pi(x)] = \alpha + \beta x$$

  其中logit函数为联系函数

  $$\operatorname{logit} = \log \frac{t}{1-t}$$
- Probit回归模型

  $$\operatorname{probit}[\pi(x)] = \alpha + \beta x$$

  其中$\operatorname{probit}(x) = \Phi^{-1}(t)$。
  主要针对

  $$Y = \begin{cases} 1, & T \le X \\ 0, & T > X \end{cases}$$

  如果$T \sim N(\mu, \sigma^2)$，则

  $$\pi(x) := P(Y=1 \mid X=x) = \Phi((x-\mu)/\sigma),$$

  从而

  $$\Phi^{-1}(\pi(x)) = (x-\mu)/\sigma$$

  其中$\alpha = -\mu/\sigma, \beta = 1/\sigma$。

- 计数数据的广义线性模型
  - 对计数数据，最简单的分布假定是泊松。泊松分布的基本性质：
    - 取值为非负整数；
    - 均值和方差相等；
    - 分布右偏，且μ越小右偏越厉害；
    - 当均值小时，与正态分布差异较大，二当均值较大时差异很小。
  - 泊松对数线性模型

$$\log(\mu) = \alpha + \beta x$$

  - 超散布性：超出预期的变异性
    - 计数数据的方差大于均值，这种现象称为超散布性。
    - 超散布性是由于个体的异质性造成的。
    - 处理超散布性的一个方法是改用方差大于均值的离散分布，比如负二项分布和广义泊松分布。
- 统计推断和模型检验
  - 在GLM中，参数估计采用MLE。记模型参数β的MLE为$\hat{\beta}$。则其$(1-\alpha)100\%$的Wald置信区间为

$$\hat{\beta} \pm u_{\alpha/2}\mathrm{SE},$$

    其中SE为$\hat{\beta}$的标准差估计（常采用Fisher信息量的逆开方）。
  - $$H_0 := \beta = \beta_0$$

    Wald检验统计量为

$$W = (\hat{\beta} - \beta_0)^2/\mathrm{SE}^2,$$

    其在H_0之下的零极限分布为$\mathcal{X}_1^2$，水平为α的检验拒绝域可取为$\{W \geq \mathcal{X}_1^2(\alpha)\}$。
  - 似然比检验相比于Wald检验在小样本情形下更稳健。检验统计量为

$$\mathrm{LR}(\beta_0) = 2[\sup_{\alpha,\beta} \log L(\alpha, \beta) - \sup_{\alpha} \log L(\alpha, \beta_0)],$$

    在H_0之下的零极限分布为$\mathcal{X}_1^2$，因此检验的拒绝域为

$$\{\mathrm{LR}(\beta_0) > \mathcal{X}_1^2(\alpha)\}$$

相应的$1-\alpha$似然比置信区间为

$$\{\beta : \mathrm{LR}(\beta) \leq \mathcal{X}_1^2(\alpha)\}$$

- 离差（P39或书P72）
- 任意两个嵌套模型都可以通过比较离差看是否有显著差异。
- 比较观测和模型拟合的残差
    - $$\text{Pearson残差} = e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{\mathrm{var}}(y_i)}}.$$
    - 对于泊松GLM，第i各观测的皮尔逊残差等于

    $$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}.$$

    - Pearson残差在0附近波动，且在μ_i很大（大样本）时近似服从正态分布。
    - 标准化残差

    $$\text{标准化残差} = \frac{y_i - \hat{\mu}_i}{\mathrm{SE}}$$

    其中$\mathrm{SE} = [\widehat{\mathrm{var}}(y_i)(1-h_i)]^{1/2}$。标准化残差绝对值大于2或3表示偏离模型假定。
- 指数族Exponential dispersion family
    - N个Y的对立观察值$(y_1, ..., y_N), y_1$的密度函数：

    $$f(y_i;\theta_i,\phi) = \exp\{[y_i\theta_i - b)\theta_i]/a(\phi + c(y_i,\phi))\}$$

    ϕ散步参数dispersion parameter。θ_i自然参数。如果ϕ已知，则上面的参数族称为自然指数族。
    - 均值$\mu_i = E(Y_i) = b^{'}(\theta_i)$.
    - 方差$\mathrm{var}(Y_i) = b^{''}(\theta_i)a(\phi)$.
    - 具体建模看chapter3，P51

第四章

- 线性近似解释
    - $\pi(x)$随x的变化趋势：
        - $\beta = 0$，$\pi(x)$不随x变化；
        - $\beta > 0$，$\pi(x)$随x呈S形增长趋势；

        - $\beta < 0$，$\pi(x)$随x呈S形递减趋势。
    - 当$\beta \neq 0$时，$\pi(x)$不是线性的，但是因为$\pi(x)$是光滑函数，局部可以用线性逼近，且在x附近的增长速率为$\pi(x)(1-\pi(x))\beta$。
- X根据Y的情况服从不同的正态分布（P20）
    - 设$X \mid Y=j \sim N(\mu_j, \sigma^2)$，其pdf为$\phi((x-\mu_j)/\sigma)/\sigma(j=0,1)$，则Y服从logistic回归模型：

$$
\begin{aligned}
& P(Y=1 \mid X=x) \\
& =\frac{P(X=x \mid Y=1) P(Y=1)}{P(X=x \mid Y=1) P(Y=1)+P(X=x \mid Y=0) P(Y=0)} \\
& =\frac{\phi((x-\mu_1)/\sigma) f}{\phi((x-\mu_1)/\sigma) f+\phi((x-\mu_0)/\sigma)(1-f)} \\
& =\frac{\phi((x-\mu_1)/\sigma) f/[\phi((x-\mu_1)/\sigma)(1-f)]}{\phi((x-\mu_1)/\sigma) f/[\phi((x-\mu_1)/\sigma)(1-f)]+1}=\frac{e^{\alpha+\beta x}}{e^{\alpha+\beta x}+1},
\end{aligned}
$$

    其中$\beta=(\mu_1-\mu_0)/\sigma^2, \alpha=\frac{\mu_0^2-\mu_1^2}{2\sigma^2}+\log\frac{f}{1-f}$。
    - 对于非正态的预测变量，可以考虑引进平方项等以便更好地拟合模型。
- Logistic回归的推断
    - **未分组二分数据**：在原始资料（raw data）中，每个个体有一行数据（含响应变量和预测变量值），称为未分组二分数据
    - **分组二分数据**：如果预测变量也是属性数据，可以将原始资料总结成列联表的形式，称为分组二分数据。
    - 效应的置信区间：可以构造Wald置信区间和似然比置信区间。（P22）
    - 显著性检验
        - 回归系数β的显著性检验对应的原假设为$H_0: \beta=0$。可以采用Z检验，等价的可以用Wald检验。对小样本可以改用似然比检验
    - 概率的置信区间
        - $\pi(x)$的置信区间的构造可以基于估计量$\hat{\alpha}$和$\hat{\beta}$的联合渐近正态性，可以先构造前者的置信区间，再变换得到后者的置信区间。
        - 算法如下：
            1. 首先，估计$\hat{\alpha}+\hat{\beta}x$的方差

$$
\operatorname{var}(\hat{\alpha}+\hat{\beta} x)=\operatorname{var}(\hat{\alpha})+x^2 \operatorname{var}(\hat{\beta})+2 x \operatorname{cov}(\hat{\alpha}, \hat{\beta}),
$$

            其中涉及到三个方差/协方差可以采用样本Fisher信息阵的逆矩阵中相应分量估计，这

样得到方差的估计$\widehat{\text{var}}(\hat{\alpha}+\hat{\beta}x)$，进而得到$\hat{\alpha}+\hat{\beta}x$的标准差的估计$\text{SE}(x) := \sqrt{\hat{\alpha}+\hat{\beta}x}$。

2. 接着构造$\alpha+\beta x$的Wald置信区间：$(\hat{\alpha}+\hat{\beta}x) \pm u_{\alpha/2}\text{SE}(x)$。
3. 最后得到$\pi(x)$的Wald置信区间

$$\left[\frac{\exp\{\hat{\alpha}+\hat{\beta}x-u_{\alpha/2}\text{SE}(x)\}}{1+\exp\{\hat{\alpha}+\hat{\beta}x-u_{\alpha/2}\text{SE}(x)\}}, \frac{\exp\{\hat{\alpha}+\hat{\beta}x+u_{\alpha/2}\text{SE}(x)\}}{1+\exp\{\hat{\alpha}+\hat{\beta}x+u_{\alpha/2}\text{SE}(x)\}}\right].$$

- 属性预测变量的logistic回归
  - 对属性预测变量X，可以采用指示变量/哑变量。
  - 记属性预测变量X的水平个数为k：$1,\ldots,k$，定义k个哑变量：$X_j = I(X=j), j = 1,\ldots,k$。在相对的哑变量定义中需要一个基准水平，默认按字母顺序排在最前面的一个作为基准水平，例如k，此时$X_1,\ldots,X_{k-1}$进入模型：

    $$P(Y=1 \mid X_1=x_1,\ldots,X_{k-1}=x_{k-1}) = \text{expit}(\alpha+\beta_1 x_1+\cdots+\beta_{k-1}x_{k-1}).$$

    截距项α即基准水平相应的对数优势：

    $$\frac{P(Y=1 \mid X=k)}{P(Y=0 \mid X=k)} = \frac{P(Y=1 \mid X_1=\cdots=X_{k-1}=0)}{P(Y=0 \mid X_1=\cdots=X_{k-1}=0)} = \exp(\alpha),$$

    而x_j的系数β_j等于水平j相对于水平k的对数优势比：

    $$\frac{P(Y=1 \mid X=1)/P(Y=0 \mid X=1)}{P(Y=1 \mid X=k)/P(Y=0 \mid X=k)} = \frac{\exp(\alpha+\beta_1)}{\exp(\alpha)} = \exp(\beta_1)$$
  - 在绝对的哑变量定义中，所有k个哑变量进入模型（模型没有截距项）：

    $$P(Y=1 \mid X_1=x_1,\ldots,X_k=x_k) = \text{expit}(\beta_1 x_1+\cdots+\beta_k x_k),$$

    其中x_j的系数等于水平j相对应的对数优势β_j：

    $$\frac{P(Y=1 \mid X=1)}{P(Y=0 \mid X=1)} = \exp(\beta_1)$$
  - 检验一般的线性假设$H_0: H\beta = \beta_0$(β_0可以是多维的)，此时的Wald检验统计量为

    $$(H\hat{\beta}-\beta_0)^T[H\widehat{\text{cov}}(\hat{\beta})H^T]^{-1}(H\hat{\beta}-\beta_0)$$

更一般的$H_0: g(\theta)=0$，Wald统计量为：

$$W_n = n[g(\hat{\theta}_n)]^T \left\{ \frac{\partial g(\hat{\theta}_n)}{\partial \theta^T} [I(\hat{\theta}_n)]^{-1} \left[\frac{\partial g(\hat{\theta}_n)}{\partial \theta^T} \right]^T \right\}^{-1} g(\hat{\theta}_n)$$

此处$\hat{\theta}_n$为MLE。

- $2\times 2\times K$列联表的CMH检验
  设控制变量Z有K个类别（不妨用Z=k表示Z属于第k个类别），可以引进如下模型：

  $$\operatorname{logit}[P(Y=1 \mid X=x, Z=k)] = \alpha_k + \beta x,$$

  其中$\exp(\beta)$是给定Z=k之下XY的公共优势比。可以通过检验$H_0:\beta=0$验证条件独立性。下面构造检验$H_0:\beta=0$的Cochran-Mantel-Haenszel（CMH）检验。

  - $$\mu_{11k} = E(n_{11k}) = \frac{n_{1+k} n_{+1k}}{n_{++k}}$$
  - $$\operatorname{var}(n_{11k}) = \frac{n_{1+k} n_{2+k} n_{1+k} n_{+2k}}{n_{++k}^2 (n++k-1)}$$
  - $$\mathrm{CMH} = \frac{[\sum_{k=1}^K (n_{11k} - \mu_{11k})]^2}{\sum_{k=1}^K \operatorname{var}(n_{11k})}$$
  - CMH的零极限分布为$\mathcal{X}_1^2$分布。
- 多元logistic回归
  考虑k个预测变量的logistic回归模型：

  $$\operatorname{logit}[P(Y=1)] = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

  参数β_i是控制其他预测变量时x_i vs. Y条件对数优势比

  $$\beta_i = \log \left[\frac{P(Y=1 \mid x_i = t+1, x_{(i)}) / P(Y=0 \mid x_i = t+1, x_(i))}{P(Y=1 \mid x_i = t, x_{(i)}) / P(Y=0 \mid x_i = t, x_{(i)})} \right],$$

  其中$x_{(i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$。上述模型蕴含着齐次关联假定。
- 通过模型比较确认某项是否必要（P44或书P99）
- 有序预测变量的定量化处理（P45或书P100）

- 容许交互效应（P46或书P101）
- Logistic回归效应的概括（P48或书P101）

第六章

- 名义响应变量的logit模型
  记J为响应变量Y的类别个数，记$\pi_j = P(Y=j)(j=1,\ldots,J)$，令$Y_1,\ldots,Y_n$为简单随机样本，$n_j = \sum_{i=1}^{n} I(Y_i=j)(j=1,\ldots,J)$，则$(n_1,\ldots,n_J)$服从参数为n和$(\pi_1,\ldots,\pi_J)$的多项分布。
  - 基线-类别logit
    - 取最后一个类别（J）作为基线（baseline）时，基线-类别logit为

$$\log\frac{\pi_j}{\pi_J}, \quad j=1,\ldots,J-1$$

      如果给定响应只落在类别j和J中，则$\log\{\pi_j/\pi_J\}$就是响应为j的对数优势，从而基线-类别logit是二分情形的推广。事实上，记

$$\pi_j^{'} = P(Y=j \mid Y=j \text{ 或 } J) = \frac{P(Y=j)}{P(Y=j \text{ 或 } J)} = \frac{\pi_j}{\pi_j+\pi_J},$$

      则$\pi_j^{'}$的优势（给定Y=j或J时类别为j的条件优势）为

$$\frac{\pi_j^{'}}{1-\pi_j^{'}} = \frac{\pi_j/(\pi_j+\pi_J)}{\pi_J/(\pi_j+\pi_J)} = \frac{\pi_j}{\pi_J}$$

    - 将二分响应变量的logit模型推广到多类别情形：

$$\log\frac{\pi_j}{\pi_J} = \alpha_j + \beta_j x, \quad j=1,\ldots,J-1$$

    - 模型不依赖于基线的选取。
    - 通过极大似然估计拟合，总共J-1个式子同时拟合。
    - 拟合方法可以参考P8
    - 拟合优度检验

$$X^2 = \sum\frac{(O-E)^2}{E}, \quad G^2 = 2\sum O\log\frac{O}{E},$$

      O为观测到的计数，而E则是在模型下的期望计数。要检查期望频数要均大于5。

- 离散选择模型（P13或书P152）
- 有序响应变量的累积logit模型
  - 对有序响应变量Y，Y=j表示取值排序第j列的类别，注意这一表示方式只代表顺序。
  - 累计概率

    $$P(Y \le j) = \pi_1 + \cdots + \pi_j, \quad j = 1, \ldots, J$$

    累积概率的logit（简称为累积logit）为

    $$\operatorname{logit}[P(Y \le j)] = \log \frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_J}, \quad j = 1, \ldots, J-1.$$

  - 有序变量经常是由一个潜在变量$Y^\star$诱导出来的：

    $$Y = j \text{ 当且仅当 } \tau_{j-1} < Y^\star \le \tau_j,$$

    其中$-\infty = \tau_0 < \tau_1 < \cdots < \tau_{J-1} = \infty$。如果假定$Y^\star$与x满足如下的线性模型：

    $$Y^\star = \alpha^\star - \beta x + e,$$

    其中误差项e独立于x且服从logistic分布，则Y确实满足模型：

    $$\begin{aligned} P(Y \le j) = P(Y^\star \le \tau_j) &= P(\alpha^\star - \beta x + e \le \tau_j) \\ &= P(e \le \tau_j - \alpha^\star + \beta x) = \operatorname{expit}(\tau_j - \alpha^\star + \beta x) \end{aligned}$$

  - 模型参数的推断
    - 记n个独立观测数据为(X_i, Y_i)，其中预测变量X_i可以是多维的，而Y_i的取值于$\{1, \ldots, J\}$。记模型参数向量为Θ，而累积（条件）概率$P(Y_j \le X_i)$依赖于X_i和Θ，记为$p_{ij}(\Theta)$，显然

      $$P(Y_i = j \mid X_j) = p_{ij}(\Theta) - p_{i,j-1}(\Theta) =: \pi_{ij}(\Theta), \text{ 其中} p_{i0}(\Theta) = 0$$

      例如，对政治形态数据，考虑部分成优势比例模型（取类别J=5为基准类别）

      $$\operatorname{logit}(P(Y \le j)) = \alpha_j + \beta_j x + \gamma z, \quad j = 1, \ldots, 4$$

      则$J = 5, \Theta = (\alpha_1, \ldots, \alpha_4, \beta_1, \ldots, \beta_4, \gamma)^T, X_i = (x_i, z_i)^T$,

      $$p_{ij}(\Theta) = \operatorname{expit}(\alpha_j + \beta_j x_i + \gamma z_i)$$

将Y_i视为随机变量，得到似然函数

$$L(\Theta)=\prod_{i=1}^{n}P(Y_i\mid X_i)=\prod_{i=1}^{n}\prod_{j=1}^{J}\pi_{ij}(\Theta)^{I(Y_i=j)}.$$

- 比较累计概率的解释（P24）
- 潜变量诱导（P26）
- 响应类别选择的不变形（P28）

假设潜变量模型

$$Y=j,\ \text{如果}a_{j-1}<Y^{\star}\leq a_j$$

成立，则合并任意类别所得的模型参数不变。

- 成对类比有序logit（P29或书P161）
  - 相邻类别logit
  - 连续比logit

第七章

- 双向表和三向表的对数线性模型
  - 考虑两个属性变量X和Y，不妨设它们的取值分别为$1,\ldots,I$ 和 $1,\ldots,J$，其联合分布为

    $$\pi_{ij}=P(X=i,Y=j),i=1,\ldots,I\text{ 和 }1,\ldots,J.$$

    令$(X_k,Y_k),k=1,\ldots,n$为(X,Y)的i.i.d.复制，则这个样本共有最多IJ种取值集合，相应的频数构成一个$I\times J$列联表。
  - 第(i,j)个单元格频数为$n_{ij}=\sum_{k=1}^{n}I(X_k=i,Y_k=j)$, 则 $(n_{11},n_{12},\ldots,n_{IJ})$服从参数为n和$(\pi_{11},\pi_{12},\ldots,\pi_{IJ})$的多项分布。
  - 记(i,j)单元格的期望频数为$\mu_{ij}=n\pi_{ij}$。
  - 假设X 和 Y独立，即

    $$\pi_{ij}=\pi_{i+}\pi_{+j},\quad i=1,\ldots,I\text{ 和 }j=1,\ldots,J,$$

    则(i,j)单元格的期望频数为

    $$\mu_{ij}=E(n_{ij})=n\pi_{i+}\pi_{+j}.$$

- 双向表的独立性对数线性模型

- 在独立性假定下有如下独立性对数线性模型：

  $$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y$$

  其中λ_i^X 和 λ_j^Y分别为行效应和列效应
- $\lambda_i^X(\lambda_j^Y)$越大，则第i行（j）列的期望频数越大。
- 独立检验统计量X^2 和 G^2适用于模型的拟合优度检验，通过和饱和模型的拟合优度进行比较。
- 模型对单元频数进行建模时把X 和 Y都看作响应变量，把单元频数看作是来自某个分布（比较典型的时泊松分布）的独立观测。
- 考虑$J=2$的情形，第i行概率$P(Y=1 \mid X=i)$的logit（对数优势）等于

  $$\begin{aligned}
  &\log[P(Y=1 \mid X=i)/P(Y=2 \mid X=i)] \\
  &= \log(\pi_{i1}/\pi_{i2}) = \log(\mu_{i1}/\mu_{i2}) \\
  &= (\lambda + \lambda_i^X + \lambda_1^Y) - (\lambda + \lambda_i^X + \lambda_2^Y) \\
  &= \lambda_1^Y - \lambda_2^Y
  \end{aligned}$$

  即$P(Y=1 \mid X=i)$的logit不依赖于X的水平：$\operatorname{logit}[P(Y=1 \mid X=i)] := \alpha$.这表明对于每一行，响应落入第一列的概率的优势都为$\exp(\alpha) = \exp(\lambda_1^Y - \lambda_2^Y)$。
- 独立模型中的参数$\{\lambda_i^X\}$中有一个是冗余的，可令其中一个为0，同理可以令$\{\lambda_i^Y\}$其中一个为0，一般是令最后一个为0，独立模型中总共有$I+J-1$个独立参数。

* 双向表的饱和模型
  - 考虑交互效应的对数线性模型

    $$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY},$$

    其中交互效应λ_{ij}^{XY}度量了偏离独立性的程度。对数优势比与λ_{ij}^{XY}有直接联系，比如对2×2表模型有

    $$\begin{aligned}
    \log \theta &= \log \frac{P(Y=1 \mid X=1)/P(Y=2 \mid X=1)}{P(Y=1 \mid X=2)/P(Y=2 \mid X=2)} \\
    &= \log \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \log \frac{\mu_{11}\mu_2}{\mu_{12}\mu_{21}} = \log \mu_{11} + \log \mu_{22} - \log \mu_{12} - \log \mu_{21} \\
    &= \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}.
    \end{aligned}$$

    故优势比由交互效应决定，交互效应全为0则独立性成立。

- 在饱和模型下，每一个计数相应有一个参数，因此总共有IJ个独立参数：

$$\begin{array}{ccccccccc} \text{总数} & & \text{截距项} & & \text{X主效应} & & \text{Y主效应} & & \text{XY交互效应} \\ IJ & = & 1 & + & (I-1) & + & (J-1) & + & (I-1)(J-1). \end{array}$$

  当某个交互效应非零时，关于主效应的解释变得比较复杂，此时需要结合交互效应一起解释。

- 三向表的对数线性模型
  - 但愿期望频数频数μ_{ijk}的对数线性模型

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

    XZ项λ_{ik}^{XZ}用于描述给定Y时X 与 Z的关联性。而XY的缺失反应了给定Z时X与Y的条件独立性。
  - 这类模型可以用它的高阶项来描述，比如上述模型可以记为(XZ, YZ)。
  - 更一般的模型：

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}.$$

    表示没有三阶交互效应，任意一对变量的条件关联性是齐次的，记为(XY, YZ, XZ)。

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}.$$

    该模型记为(XYZ)。
  - 模型解释与最高项有关。给定Z的XY对数优势比为

$$\begin{aligned} \log \theta_{ij|k}^{XY|Z} &= \log \frac{P(X=i, Y=j \mid Z=k)/P(X=I, Y=J \mid Z=k)}{P(X=i, Y=J \mid Z=k)/P(X=I, Y=j \mid Z=k)} \\ &= \log \frac{\mu_{ijk}\mu_{IJk}}{\mu_{iJk}\mu_{Ijk}} \\ &= (\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}) \\ &\quad + (\lambda + \lambda_I^X + \lambda_J^Y + \lambda_k^Z + \lambda_{IJ}^{XY} + \lambda_{Ik}^{XZ} + \lambda_{Jk}^{YZ}) \\ &\quad - (\lambda + \lambda_i^X + \lambda_J^Y + \lambda_k^Z + \lambda_{iJ}^{XY} + \lambda_{ik}^{XZ} + \lambda_{Jk}^{YZ}) \\ &\quad - (\lambda + \lambda_I^X + \lambda_j^Y + \lambda_k^Z + \lambda_{Ij}^{XY} + \lambda_{Ik}^{XZ} + \lambda_{jk}^{YZ}) \\ &= \lambda_{ij}^{XY} + \lambda_{IJ}^{XY} - \lambda_{iJ}^{XY} - \lambda_{Ij}^{XY}. \end{aligned}$$

    对数条件优势比不依赖于Z的取值，因此X与Y是齐次条件关联的。

    - 在饱和模型下给定Z=k时XY的对数条件优势比为

$$\log\theta_{ij|k}^{XY|Z}=\lambda_{ij}^{XY}+\lambda_{IJ}^{XY}-\lambda_{iJ}^{XY}-\lambda_{Ij}^{XY}+\lambda_{ijk}^{XYZ}+\lambda_{IJk}^{XYZ}-\lambda_{iJk}^{XYZ}-\lambda_{Ijk}^{XYZ}.$$

      一般来说此时条件关联是非齐次的。
- 对数线性模型的推断
    - 卡方拟合优度检验
        - $$G^2=2\sum_i n_i\log\frac{n_i}{\hat{\mu}_i},$$
        - $$X^2=\sum_i\frac{(n_i-\hat{\mu}_u)^2}{\hat{\mu}_i}.$$
        - 其中下标i取遍列联表单元格。检验统计量的零极限分布都是卡方分布，其自由度df为单元个数减模型中独立参数的个数。特别地，饱和模型下df=0。
- 条件关联的检验
    - 如果要检验给定M时A和C是否条件关联，可以采用离差（P20）。
    - 条件优势比的置信区间（P21或书P182）。
- 高维对数线性模型（书P183）
- 对数线性模型与logistic模型的联系（P23或书P186）
    - 区别
        - 对数线性模型：可以描述属性变量之间的关联性；
        - Logistic模型：描述解释变量与属性响应变量之间的关系，但不对解释变量之间的关联性建模。
    - 联系
        - 对数线性模型：可以对其中一个响应变量构造logit来解释模型；
        - Logistic模型：如果解释变量都是属性变量，则它有相对应的对数线性模型（可以对应多个对数线性模型）。
    - 利用logistic模型解释对数线性模型
        - 下面通过构造一个变量的logit来理解对数线性模型表达式的含义。以三向列联表的齐次关联模型为例说明：

$$\log\mu_{ijk}=\lambda+\lambda_i^X+\lambda_j^Y+\lambda_k^Z+\lambda_{ij}^{XY}+\lambda_{ik}^{XZ}+\lambda_{jk}^{YZ}.$$

          设Y是二分的，将其看作响应变量，则

$$
\begin{aligned}
&\operatorname{logit}[P(Y=1 \mid x+i, Z=k)] \\
&= \log \frac{P(Y=1 \mid X=i, Z=k)}{P(Y=2 \mid X=i, Z=k)} = \log \frac{\mu_{i1k}}{\mu_{i2k}} \\
&= (\lambda_1^Y - \lambda_2^Y) + (\lambda_{i1}^{XY} - \lambda_{i2}^{XY}) + (\lambda_{1k}^{YZ} - \lambda_{2k}^{YZ}) \\
&:= \alpha + \beta_i^X + \beta_k^Z .
\end{aligned}
$$

- 参考P24

- 模型选择策略：有多个变量的对数线性模型，通常先拟合只有单因子项的模型，如果拟合够好就停止，否则加入两因子项，依次类推，直到找到最合适的模型。
- 二元分类模型的衡量标准（第八章，model selection 2）
  - 真阳性（True Positive， TP）：预测为正，实际也为正的样本；
  - 真阴性（True Negative， TN）：预测为负，实际也为负的样本；
  - 假阳性（False Positive， FP）：预测为正，实际为负的样本；
  - 假阴性（False Negative， FN）：预测为负，实际为正的样本。

	实际为正类	**实际为负类**
预测为正类	真阳性（TP）	假阳性（FP）
预测为负类	假阴性（FN）	真阴性（TN）

- 正确率（Accuaracy）：$\frac{TP+TN}{TP+FP+FN+TN}$，即预测正确的样本数量占全部样本数的比例。（注意数据非平衡情况）。
- 精确率（Precision）：$\frac{TP}{TP+FP}$，即预测为正类的样本中实际也为正类预测正确的比例，又叫查准率。
- 召回率（Recall）：$\frac{TP}{TP+FN}$，即实际为正类的样本中正确预测为正类的比例，又叫查全率。
- 真正率（True Positive Rate，TPR）：$\frac{TP}{TP+FN}$，
- 假正率（False Positive Rate，FPR）：$\frac{FP}{FP+TN}$。
- ROC曲线：横坐标为FPR，纵坐标为TPR，通过遍历0到1之间所有阈值计算相应的混淆矩阵从而绘制而成。
- AUC：ROC曲线下的面积，越大越好。
- KS统计量：
  - 零假设H_0为两组数据的分布一致
  - $$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|,$$

    其中$F_{1,n}(x)$ 和 $F_{2,m}(x)$分别为两组数据的经验分布函数，样本数为n和m，当$D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{n\cdot m}}$时拒绝零假设。

- $$\mathrm{KS}_n=\sup_{-\infty<t<\infty}\left\{\frac{1}{n_0}\sum_{y_i=0}I(S(x_i)\le t)-\frac{1}{n_1}\sum_{y_i=1}I(S(x_i)\le t)\right\}$$

  其中$n_0=\sum_{i=1}^n I(y_i=0), n_1=\sum_{i=1}^n I(y_i=1), S(x_i)$表示第i个客户的信用评分估计，一般认为和不会违约的条件概率$P(Y=1\mid X)$是正相关的。

牛顿迭代法

- $\beta_{k+1}=\beta_k+\Delta, 0=S_n(\beta_{k+1})=S_n(\beta_k+\Delta)\approx S_n(\beta_k)+\dot{S}_n(\beta_k)\Delta$
- $\Delta=-\left(\dot{S}_n(\beta_k)\right)^{-1}S_n(\beta_k)$
- 每次迭代修正$\beta_{k+1}=\beta_k+c\Delta$,选c使得$\|S_n(\beta_{k+1})\|<\|S_n(\beta_k)\|$.
- 停止条件$\|\beta_{k+1}-\beta_k\|\le\epsilon\|\beta_k\|$.
- 高等数理统计上的方法
  - Newton-Raphson算法
    - 记负对数似然函数的Hessian矩阵为$H(\theta)=-\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'}=-\frac{\partial s(\theta)}{\partial\theta'}$。
    - 第0步：令k=0，选初始值$\hat{\theta}^{(k)}$;
    - 第1步：$\hat{\theta}^{(k+1)}=\hat{\theta}^{(k)}+\left[H(\hat{\theta}^{(k)})\right]^{-1}s(\hat{\theta}^{(k)})$;
    - 第2步：若$|s(\hat{\theta}^{(k)})|$或$|\hat{\theta}^{(k+1)}-\hat{\theta}^{(k)}|$充分小，则停止；否则k=k+1，转到第1步。
  - Fisher得分算法
    - 将Newton-Raphson算法中$H(\theta)=-\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'}$换成$H^*(\theta)=E_\theta H(\theta)=-E_\theta\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'}$就得到Fisher得分算法，即

      $$\hat{\theta}^{(k+1)}=\hat{\theta}^{(k)}+\left[H^*(\hat{\theta}^{(k)})\right]^{-1}s(\hat{\theta}^{(k)}).$$

      即在迭代算法中，Newton-Raphson算法用的是观测Fisher信息矩阵而在Fisher得分算法中用的是期望Fisher信息矩阵。

Model Selection

- **Model uncertainty**:we are uncertain on many aspects of model.
  - uncertainty consists of three main types (Amini, 2012):
    - **theory uncertainty**: focusing on which principal determinants should be included in a model;
    - **heterogeneity uncertainty**: relating to whether or not the parameters are identical across countries;

- **functional form uncertainty**: relating to which regressors enter the model linearly and which ones enter nonlinearly.

- Akaike Information Criterion(AIC)
  - $\mathrm{AIC} = -2\log L(\hat{\theta} \mid y) + 2\mathrm{dim}(\theta)$
  - AIC estimates the expected Kullback-Leibler between the model generating the data and a fitted candidate model.
  - Kullback-Leibler (1951) information:
    K-L information between models f and g is defined as the intergral

$$I(f,g) = \int f(x)\log\frac{f(x)}{g(x \mid \theta)}dx.$$

    The notation $I(f,g)$ denotes the "information lost when g is used to approximate f".
  - An estiamte of $I(f, g(x \mid \theta_o))$:

$$I(f, g(x \mid \hat{\theta}(y))) = \int f(x)\log\frac{f(x)}{g(x \mid \hat{\theta}(y))}dx.$$

    Ignoring constant, the expected Kullback-Leibler is

$$-E_y E_x \log g(x \mid \hat{\theta}(y)).$$

    K-L information of the best approximating model in the class of models $g(x \mid \theta)$:

$$I(f, g(x \mid \theta_o)) = \int f(x)\log\frac{f(x)}{g(x \mid \theta_o)}dx.$$

  - Fisher information matrix $J = -E\frac{\partial^2 \log g(x|\theta)}{\partial\theta\partial\theta'}|_{\theta=\theta_0}$.
  - Considering candidate model k,

$$\sqrt{n}(\hat{\theta}_k - \theta_o) \to N(0, J^{-1}),\ \text{as}\ n \to \infty.$$

  - AIC in Normal Linear Regression Model:

$$\mathrm{AIC} = n\log\hat{\sigma}^2 + 2\mathrm{dim}(\theta), \quad \hat{\sigma}^2 = \|\hat{y} - y\|^2/n$$

- Bayesian Information Criterion
  - $\mathrm{BIC} = -2\log L(\hat{\theta} \mid y) + \mathrm{dim}(\theta)\log n$
  - Under linear model,

$$\mathrm{BIC} = n \log \hat{\sigma}^2 + \dim(\theta) \log n$$

- Mallows Criterion
  - $$C_I = \|y - \hat{\mu}\|^2 + 2\hat{\sigma}^2 k_m.$$

各种分布

- Binomial
  - B(n,p)
  - p.d.f: P(X=x)= $\binom{n}{x} p^x (1-p)^{n-x}$
  - Expectation: np
  - Variance: np(1-p)
- Poisson
  - $P(\theta)$
  - p.d.f: $P(X=x) = \theta^x e^{-\theta}/x!$
  - Expectation: θ
  - Variance: θ
- Geometric
  - G(p)
  - p.d.f: $P(X=x) = (1-p)^{x-1} p$
  - Expectation: 1/p
  - Variance: (1-p)/p^2
- Hypergeometric
  - HG(r,n,m)
  - p.d.f: $P(X=x) = \binom{n}{x}\binom{m}{r-x} / \binom{N}{r}, \quad x = 0, 1, \ldots, min\{r, n\}, \quad r - x \leq m$
  - Expectation: rn/N
  - Variance: rnm(N-r)/[N^2(N-1)]
- Negative binomial
  - NB(p,r)
  - p.d.f: $\binom{x-1}{p-1} p^r (1-p)^{x-r}, \quad x = r, r+1, \ldots$
  - Expectation: r/p
  - Variance: r(1-p)/p^2
- Exponential
  - $E(a, \theta)$

    - p.d.f: $\theta^{-1} e^{-(x-a)/\theta} I_{(a,\infty)}(x)$
    - Expectation: $\theta + a$
    - Variance: θ^2
- Chi-square
    - $\mathcal{X}_k^2$
    - p.d.f: $\frac{1}{\Gamma(k/2)2^{k/2}} x^{k/2-1} e^{-x/2} I_{(0,\infty)}(x)$
    - Expectation: k
    - Variance: 2k
- Gamma
    - $\Gamma(\alpha, \gamma)$
    - p.d.f: $\frac{1}{\Gamma(\alpha)\gamma^\alpha} x^{\alpha-1} e^{-x/\gamma} I_{(0,\infty)}(x)$
    - Expectation: $\alpha\gamma$
    - Variance: $\alpha\gamma^2$
- Logistic
    - $LG(\mu, \sigma)$
    - p.d.f: $\sigma^{-1} e^{-(x-\mu)/\sigma} / [1 + e^{-(x-\mu)/\sigma}]^2$
    - Expectation: μ
    - Variance: $\sigma^2\pi^2/3$