# 第四章 二项分布的统计推断

# The binomial trial

- The setup:
  1. There are a fixed number of observations in the sample,m.
  2. The m observations are all independent.
  3. There are only two possible outcomes for each observation: **success**(1)or **failure**(0).
  4. The probability of success,p,is constant for each observation.
- Then,the number(count)of success,Y,has a **Binomial** distribution with parameters m and p. We say

$$Y \sim B(m, p)$$

# The binomial distribution

- The **pmf** of Y at a value y is given by

$$f_Y(y) = \binom{m}{y} p^y (1-p)^{m-y}.$$

- The **moment generating function** of Y is

$$M_Y(t) = (1 - p + pe^t)^m$$

- Thus, the **mean** is

$$E(Y) = mp,$$

  and the variance is

$$var(Y) = mp(1-p).$$

# Large sample normal approximations

- Asymptotically for large m, we have that

$$\frac{Y - E(Y)}{\sqrt{var(Y)}} = \frac{Y - mp}{\sqrt{mp(1-p)}},$$

  has a N(0,1) distribution, for any fixed value of p.

- Proof: Use the standard central limit theorem noting that $Y/m = \sum_{i=1}^{m} W_i/m$, where $W_i$ are independent B(1,p) RVs (Bernoulli RVs).

# Large sample normal approximations

- a better approximation for smaller values of m is to use the continuity correction:

  We use the fact that

$$Pr(Y \le y) \to \Phi\left(\frac{y - mp + \frac{1}{2}}{\sqrt{mp(1-p)}}\right)$$

  and

$$Pr(Y \ge y) \to 1 - \Phi\left(\frac{y - mp - \frac{1}{2}}{\sqrt{mp(1-p)}}\right)$$

  as $m \to \infty$ for **fixed** p.

思考题：利用矩母函数证明当$n \to \infty$时，二项分布可以利用正态分布近似.

证明: 假定$X$ 服从二项分布$B(n,p)$, 则其矩母函数为

$$
\begin{aligned}
M_X(t) =& E[e^{tX}] = \sum_{x=0}^{n} C_n^x p^x q^{n-x} e^{tx} \\
=& \sum_{x=0}^{n} C_n^x (pe^t)^x q^{n-x} \\
=& (q+pe^t)^n \sum_{x=0}^{n} C_n^x \left(\frac{pe^t}{q+pe^t}\right)^x \left(\frac{q}{q+pe^t}\right)^{n-x} \\
=& (q+pe^t)^n
\end{aligned}
$$

# 思考题1

令$Y=\frac{X-np}{\sqrt{npq}}$, 可以得到$Y$ 的矩母函数

$$
\begin{aligned}
M_Y(t) =&E[e^{tY}] = E[e^{t(\frac{X-np}{\sqrt{npq}})}] \\
=&e^{-\frac{npt}{\sqrt{npq}}}E[e^{\frac{tX}{\sqrt{npq}}}] = e^{-\frac{npt}{\sqrt{npq}}}(q+pe^{\frac{t}{\sqrt{npq}}})^n \\
=&(qe^{-\frac{pt}{\sqrt{npq}}}+pe^{\frac{qt}{\sqrt{npq}}})^n \\
=&\left[q(1-\frac{pt}{\sqrt{npq}}+\frac{pt^2}{2nq})+p(1+\frac{qt}{\sqrt{npq}}+\frac{qt^2}{2np}+o(\frac{t^2}{n}))\right]^n \\
=&\left[1+\frac{t^2}{2n}+o(\frac{t^2}{n})\right]^n
\end{aligned}
$$

# 思考题1

$$
\begin{aligned}
\lim_{n\to+\infty} M_Y(t) &= lim_{n\to+\infty}[1+\frac{t^2}{2n}+o(\frac{t^2}{n})]^n \\
&= lim_{n\to+\infty}[1+\frac{t^2}{2n}]^{\frac{2n}{t^2}\cdot\frac{t^2}{2}} \\
&= e^{\frac{t^2}{2}}
\end{aligned}
$$

当$n\to\infty$ 时, $M_Y(t)\to e^{\frac{t^2}{2}}$, 即证二项分布近似正态分布.

# 思考题2

思考题：利用密度函数证明当$n \to \infty$时，二项分布近似正态分布.

证明：假定$X$服从二项分布$B(n,p)$，则

$$P(X=m)=C_n^m p^m q^{n-m}$$

不妨假设$m=np+d$，则有

$$\begin{aligned} P(X=m) =& P(X=np+d) \\ =& C_n^{np+d} p^{np+d} q^{n-(np+d)} \\ =& C_n^{np+d} p^{np+d} q^{nq-d} \\ =& \frac{n!}{(np+d)!(nq-d)!} p^{np+d} q^{nq-d} \end{aligned}$$

# 思考题2

对上式利用Stirling公式$n! = \sqrt{2\pi n}n^n e^{-n}$, 可得

$$
\begin{aligned}
P(X = np + d) = & \frac{\sqrt{2\pi n}n^n e^{-n}}{\sqrt{2\pi(np+d)}(np+d)^{np+d}e^{-(np+d)}} \\
& \cdot \frac{1}{\sqrt{2\pi(nq-d)}(nq-d)^{nq-d}e^{-(nq-d)}} \cdot p^{np+d}q^{nq-d} \\
= & \frac{1}{\sqrt{2\pi}} \cdot [\frac{n}{(np+d)(nq-d)}]^{\frac{1}{2}} \cdot (\frac{n}{np+d})^{np+d} \\
& \cdot (\frac{n}{nq-d})^{nq-d} \cdot p^{np+d}q^{nq-d} \\
= & \frac{1}{\sqrt{2\pi}} \cdot [\frac{(np+d)(nq-d)}{n}]^{-\frac{1}{2}} \cdot (\frac{np+d}{n})^{-(np+d)} \\
& \cdot (\frac{nq-d}{n})^{-(nq-d)} \cdot p^{np+d}q^{nq-d}
\end{aligned}
$$

# 思考题2

$$
\begin{aligned}
&=\frac{1}{\sqrt{2\pi npq}}[(1+\frac{d}{np})(1-\frac{d}{nq})]^{-\frac{1}{2}}(p+\frac{d}{n})^{-(np+d)} \\
&\cdot(q-\frac{d}{n})^{-(nq-d)}p^{np+d}q^{nq-d} \\
&=\frac{1}{\sqrt{2\pi npq}}(1+\frac{d}{np})^{-\frac{1}{2}}(1-\frac{d}{nq})^{-\frac{1}{2}}(1+\frac{d}{np})^{-(np+d)}(1-\frac{d}{nq})^{-(nq-d)}
\end{aligned}
$$

由于$ln(1+x)=x-\frac{x^2}{2}+\frac{x^3}{3}-\frac{x^4}{4}+\cdots$

限制$|\,d\,|\le c\cdot\sqrt{n}\ (c>0)$，当$n\to\infty$时，$\frac{d}{np}\to 0$, $\frac{d}{nq}\to 0$, 则对上式的后半部分取对数并展开可得

# 思考题2

$$
\begin{aligned}
& ln\left[(1+\frac{d}{np})^{-\frac{1}{2}}(1-\frac{d}{nq})^{-\frac{1}{2}}(1+\frac{d}{np})^{-(np+d)}(1-\frac{d}{nq})^{-(nq-d)}\right] \\
=& -\frac{1}{2}ln(1+\frac{d}{np})-\frac{1}{2}ln(1-\frac{d}{nq})-(np+d)ln(1+\frac{d}{np}) \\
& -(nq-d)ln(1-\frac{d}{nq}) \\
=& -\frac{d}{2np}+\frac{d}{2nq}-(np+d)\left[\frac{d}{np}-\frac{1}{2}(\frac{d}{np})^2+\cdots\right] \\
& -(nq-d)\left[-\frac{d}{nq}-\frac{1}{2}(\frac{d}{nq})^2-\cdots\right] \\
=& -\frac{d^2}{2npq}+\frac{1}{6}\frac{d^3}{n^2p^2}-\frac{1}{6}\frac{d^3}{n^2q^2}-\frac{d}{2np}+\frac{d}{2nq}+\cdots
\end{aligned}
$$

从而可得

$$P(X=m)=P(X=np+d)\approx\frac{1}{\sqrt{2\pi npq}}e^{-\frac{d^2}{2npq}}$$

# 思考题3

思考题：Prove that when $n \to \infty$ ,binomial distribution is close to Poisson distribution.

证明: Assume that $X \sim B(n,p)$, $np = \lambda$

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{n(n-1)...(n-k+1)}{k!} (\frac{\lambda}{n})^k (1-\frac{\lambda}{n})^{n-k}$$

$$= \frac{\lambda}{k!}(1-\frac{1}{n})(1-\frac{2}{n})...(1-\frac{k-1}{n})(1-\frac{\lambda}{n})^{n-k}$$

for the given $k$

$$\lim_{n\to+\infty}(1-\frac{\lambda}{n})^{n-k} = e^{-\lambda},$$

$$\lim_{n\to+\infty}(1-\frac{1}{n})...(1-\frac{k-1}{n}) = 1.$$

So,

$$\lim_{n\to+\infty} \binom{n}{k} p^k (1-p)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}.$$

# 思考题3

The equation holds when $np \to +\lambda$,so when we calculate the binomial distribution $b(n,p)$,when $n \to +\infty$ and $p$ is small, $\lambda = np$.

$$\lim_{n \to +\infty} \binom{n}{k} p^k (1-p)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}$$

is established.

# 思考题3

采用另外一种方法证明: Using the Characteristic Function

The Characteristic Function of the binomial distribution with $P$ is

$$f_n(t) = (p_n e^{it} + q_n)^n = [1 + \frac{np_n(e^{it} - 1)}{n}].$$

when $n \to +\infty, \; np_n \to +\lambda$

$$\lim_{n \to +\infty} f_n(t) = exp[\lambda(e^{it} - 1)] = f(t).$$

# Remarks on the normal limit

- To use the normal approximation for a fixed sample size m, we note that the approximation is best for p=1/2.
- Useful rule of thumb: use the normal approximation when

$$mp \geq 5, and \quad m(1-p) \geq 5.$$

(some people use 10 instead of 5)

# Poisson limit

- Let $Y \sim B(m, p)$.
- Suppose that $m \to \infty$ and $p \to 0$ such that $mp$ converges to (or is always equal to) some fixed constant $\lambda$.
- Then Y converges in distribution to a Poisson RV with parameter $\lambda$.

# Inference for a binomial proportion

- Let $Y \sim B(m, p)$ and let the sample proportion be

$$\hat{p} = \frac{Y}{m}.$$

- Then, $\hat{p}$ is an unbiased estimator of p, with

$$var(\hat{p}) = \frac{p(1-p)}{m}.$$

- Asymptotically for large m, $\hat{p}$ is normally distributed.
- The standard $100(1-\alpha)\%$ confidence interval(CI) for p is

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{m}},$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of a N(0,1) RV.

# Inference for p in practice

- In practice we do not know p in the variance for $\hat{p}$.
- Some solutions:
  1. Use p=1/2. The estimated variance in this case is $(4m)^{-1/2}$.(Intervals are conservative).
  2. Estimate the variance by plugging in $\hat{p}$ for p:

     $$\widehat{var}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{m}}.$$

     (Intervals are anti-conservative).
  3. We "add two successes and add two failures". Let

     $$\tilde{p} = \frac{Y+2}{m+2+2} = \frac{Y+2}{m+4}.$$

     Then the adjusted CI is

     $$\tilde{p} \pm z_{1-\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{m}}$$

# Inference for p in practice(cont.)

4. In the Wilson CI we solve the equation

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/m}} = \pm z_{\alpha/2}.$$

   Letting $z = z_{\alpha/2}$, the resulting $100(1-\alpha)\%$ CI for p is

$$\frac{Y + z^2/2}{m + z^2} \pm \sqrt{\frac{m\hat{p}(1-\hat{p}) + z^2/4}{(m+z^2)^2}}.$$

   (Intervals oscillate between conservative and anticonservative with increasing m).

5. You can ignore the normal limit, and use the exact binomial distribution of Y instead. (Intervals tends to be conservative).

思考题: 设$X_1, X_2, \cdots, X_n$ 为取自二点分布 $B(1, p)$ 的一个样本，其中 $0 \le p \le 1$. 求 $p$ 的置信水平为 $1-\alpha$ 的置信区间 $[\hat{p}_L, \hat{p}_U]$.

(1) 当样本量 $n$ 充分大时，由中心极限定理得出 $p$ 的置信水平近似为 $1-\alpha$ 的置信区间 $[\hat{p}_L, \hat{p}_U]$

$$\left[\bar{X} - U_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{\bar{X}(1-\bar{X})}}{\sqrt{n}},\ \bar{X} + U_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{\bar{X}(1-\bar{X})}}{\sqrt{n}}\right]. \qquad (1)$$

(2) 当样本精确分布已知且样本量 $n$ 较小时，可基于充分统计量 $T(X) = \sum_{i=1}^{n} X_i$ 的分布函数 $G(t, p)$ 构造 $p$ 的置信区间.

# 思考题

$p$ 的精确置信区间构造过程:

(1) $p$的MLE为 $\bar{X}$，且 $p$ 的充分统计量为 $T(X)=\sum_{i=1}^{n}X_i$，其分布函数为

$$G(t,p)=P_p(T\leq t)=\sum_{i=0}^{[\,t\,]}C_n^i\cdot p^i\cdot(1-p)^{n-i}.$$

其中 $[\,t\,]$ 表示 $t\ (0\leq t\leq n)$ 的整数部分.

(2) 判断分布函数 $G(t,p)$ 的单调性:

$$\sum_{i=0}^{k}C_n^i\cdot p^i\cdot(1-p)^{n-i}=\frac{\Gamma(n+1)}{\Gamma(k+1)\cdot\Gamma(n-k)}\cdot\int_p^1 u^k\cdot(1-u)^{n-k-1}du,$$

令 $k=[\,t\,]$，且 $Be(x|m,n)$ 表示 $Be(m,n)$ 的分布函数，则 $G(t,p)=1-Be(p|k+1,n-k)$，可见，$T(X)$ 的分布函数 $G(t,p)$ 是 $p$ 的连续、严格减函数。

(3) 如果 $G(t,\theta)$ 是 $\theta$ 的连续、严格减函数，那么 $\theta$ 的置信水平为 $1-\alpha$ 的置信区间为 $[\,\theta_L,\ \theta_U\,]$，且 $\theta_L$ 和 $\theta_U$ 分别是关于 $\theta$ 的方程 $G(T-0,\theta)=1-\alpha_1$ 和 $G(T,\theta)=\alpha_2$ 的解，其中 $\alpha_1+\alpha_2=\alpha$，且 $0\le\alpha\le 1$.

$$\begin{cases} \frac{\Gamma(n+1)}{\Gamma(k)\cdot\Gamma(n-k+1)}\cdot\int_0^{p_L} u^{k-1}\cdot(1-u)^{n-k}\,du=\alpha_1, \\ \\ \frac{\Gamma(n+1)}{\Gamma(k+1)\cdot\Gamma(n-k)}\cdot\int_0^{p_U} u^{k}\cdot(1-u)^{n-k-1}\,du=1-\alpha_2, \end{cases}$$

$$\begin{cases} Be(p_L|k,n-k+1)=\alpha_1, \\ \\ Be(p_U|k+1,n-k)=1-\alpha_2, \end{cases} \tag{2}$$

(4) 若随机变量 $B \sim Be(m,n)$，则 $F = \frac{B}{1-B} \cdot \frac{n}{m} \sim F(2m, 2n)$. 所以方程 (2) 可以等价变换为

$$\begin{cases} F(\frac{p_L}{1-p_L} \cdot \frac{n-k+1}{k} \mid 2k, 2(n-k+1)) = \alpha_1, \\ F(\frac{p_U}{1-p_U} \cdot \frac{n-k}{k+1} \mid 2(k+1), 2(n-k)) = 1-\alpha_2. \end{cases} \tag{3}$$

$$\begin{cases} \frac{p_L}{1-p_L} \cdot \frac{n-k+1}{k} = F_{\alpha_1}(2k, 2(n-k+1)), \\ \frac{p_U}{1-p_U} \cdot \frac{n-k}{k+1} = F_{1-\alpha_2}(2(k+1), 2(n-k)). \end{cases} \tag{4}$$

(5) 解方程 (4) 得 $p$ 的置信水平为 $1-\alpha$ 的置信区间 $[\,\widehat{p}_L,\ \widehat{p}_U\,]$

- 当 $0 < k < n$ 时,

$$\widehat{p}_L = \frac{k}{k + (n-k+1)\ \cdot\ F_{1-\alpha_1}(\,2(n-k+1),\ 2k\,)},$$

$$\widehat{p}_U = \frac{(k+1)\cdot F_{1-\alpha_2}(\,2(k+1),\ 2(n-k)\,)}{(k+1)\ \cdot\ F_{1-\alpha_2}(\,2(k+1),\ 2(n-k)) + (n-k)}.$$

- 当 $k = 0$ 时,

$$\widehat{p}_L = 0,\ \widehat{p}_U = \frac{F_{1-\alpha_2}(\,2,\ 2n\,)}{F_{1-\alpha_2}(\,2,\ 2n\,) + n}.$$

- 当 $k = n$ 时,

$$\widehat{p}_U = 1,\ \widehat{p}_L = \frac{n}{n + F_{1-\alpha_1}(\,2,\ 2n\,)}.$$

# 思考题

- **Example:**

  从一批产品中随机抽查63件，发现有 3 件不合格品，求这批产品的不合格品率 $p$ 的 0.90 置信区间.

- **Solution:**

$$
\begin{gathered}
n=63,\ k=3,\ \hat{p}=0.048,\ \alpha=0.1. \\
F_{0.95}(122,\ 6)=3.704,\ F_{0.95}(8,\ 120)=2.016. \\
\widehat{p}_L=\frac{3}{3+61\times F_{0.95}(122,\ 6)}=0.013, \\
\widehat{p}_U=\frac{4}{60+4\times F_{0.95}(8,\ 120)}=0.119.
\end{gathered}
$$

故这批产品的不合格品率的 0.90 置信区间为 $[0.013,\ 0.119]$.

# Further problems with inference for p: An example

- Suppose a marketing analyst carries out a study to determine the proportion of people in a city who drink coffee regularly. In her random sample of 50 people, the analyst finds that 2 regularly drink coffee.
- What is a 95% CI for the population proportion of those who regularly drink coffee?

# The logit transformation

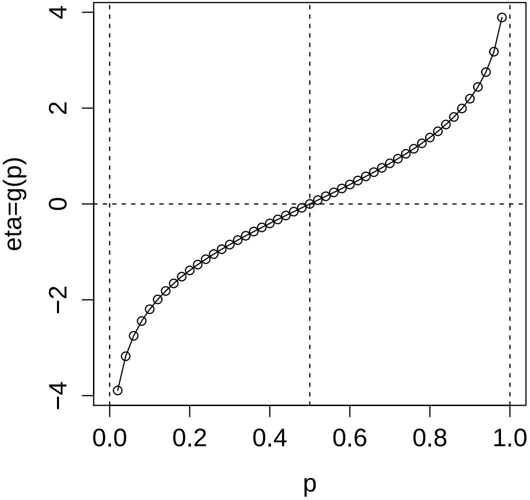
- For certain combinations of m and y, regardless of the CI used, we obtain intervals that may not lie entirely in the region [0,1].
- To evade this problem, we define a 1-1 transformation of the probability p which maps the interval(0,1) to the whole real line. Consider the **logit transformation**. Define a new variable $\eta$ by

$$\eta = g(p) = log(\frac{p}{1-p}).$$

- **The key idea**: build a CI for $\eta$ and then transform back to obtain a CI for p.
- The inverse function of g(.) is

$$p = h(\eta) = \frac{e^{\eta}}{1+e^{\eta}}.$$

# A plot of the transformation



# Likelihood inference (revision)

- The binomial log likelihood is

$$l(p) = log \begin{pmatrix} m \\ y \end{pmatrix} + ylog(p) + (m-y)log(1-p).$$

- The first derivative with respect to p is

$$\frac{dl}{dp} = \frac{y}{p} - \frac{m-y}{1-p}.$$

- Solving the score equations, the MLE of p is $\hat{p} = \frac{y}{m}$

# Likelihood inference (revision)

- The second derivative with respect to p is

$$\frac{d^2l}{dp^2} = -\frac{y}{p^2} - \frac{m-y}{(1-p)^2}.$$

and thus the Fisher information is

$$-E(\frac{d^2l}{dp^2}) = \frac{mp}{p^2} + \frac{m-mp}{(1-p)^2} = m(\frac{1}{p} + \frac{1}{1-p}).$$

# Using the Fisher information

- The inverse of the Fisher information is

$$I(p)^{-1} = [-E(\frac{d^2 l}{dp^2})]^{-1} = \frac{p(1-p)}{m}.$$

- Using standard results for MLEs, an approximate $100(1-\alpha)\%$ CI for p is as before,

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{m}},$$

- If we replace the Fisher information by the observed information the CI becomes

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{m}}.$$

# Using the delta method

- On the previous page we used the result that, as $m \to \infty$,

$$\sqrt{m}(\hat{p} - p) \xrightarrow{d} N(0, I(p)^{-1}).$$

- Using the Delta method it follows that

$$\sqrt{m}(g(\hat{p}) - g(p)) \xrightarrow{d} N(0, I(p)^{-1}(g^{'}(p))^2),$$

as $m \to \infty$, where $g^{'}(.)$ is the derivative of the logit transform given by

$$g^{'}(p) = (\frac{p}{1-p})^{-1}\frac{1}{(1-p)^2} = \frac{1}{p(1-p)}.$$

Calculating we find that

$$I(\eta)^{-1} \equiv I(p)^{-1}(g^{'}(p))^2 = \frac{p(1-p)}{m}(\frac{1}{p(1-p)})^2 = \frac{1}{mp(1-p)}.$$

# CI for the logit

- Now

$$g(\hat{p}) = log(\frac{\hat{p}}{1-\hat{p}}) = log(\frac{y}{m-y}) = \hat{\eta};$$

$$g(p) = log(\frac{p}{1-p}) = \eta.$$

$$I(\eta)^{-1} = \frac{(1+e^{\eta})^2}{me^{\eta}}.$$

- An estimate of $I(\eta)^{-1}$ is

$$I(\hat{\eta})^{-1} = \frac{m}{y(m-y)} = \frac{1}{y} + \frac{1}{m-y}.$$

- Thus, an approximate $100(1-\alpha)\%$ CI for $\eta$ is

$$log(\frac{y}{m-y}) \pm z_{1-\alpha/2}\sqrt{\frac{1}{y} + \frac{1}{m-y}}.$$

  that is, $[\eta_L, \eta_U]$, say.

- An approximate $100(1-\alpha)\%$ CI for p is then

# The coffee example revisited

- A 95% CI for $\eta$ is given by

$$log(\frac{y}{m-y}) \pm z_{0.975}\sqrt{\frac{1}{y}+\frac{1}{m-y}}$$

- Hence, a 95% CI for p is

$$[h(\eta_L), h(\eta_U)] = [\frac{e^{\eta_L}}{1+e^{\eta_L}}, \frac{e^{\eta_U}}{1+e^{\eta_U}}].$$

# Generalized linear models for binary data

- Previously We considered statistical inference on a binomial proportion, p, for $Y \sim B(m, p)$.
- In this section we will consider **binary data**:
  -Each RV $Y_i$ only has two possible values: 0 or 1.
  -We assume that $Pr(Y_i = 0) = 1 - p_i$ and $Pr(Y_i = 1) = p_i$.
  -For each observation, $i$, we have a vector of **covariates** or **explanatory** variables
  $$x_i{}^T = (x_{i,1}, \ldots, x_{i,p})^T.$$
  -The RVs, $\{Y_1, \ldots, Y_n\}$ are independent.
- Our aim is to model the relationship between $p_i$ and the explanatory variables $x_i{}^T$

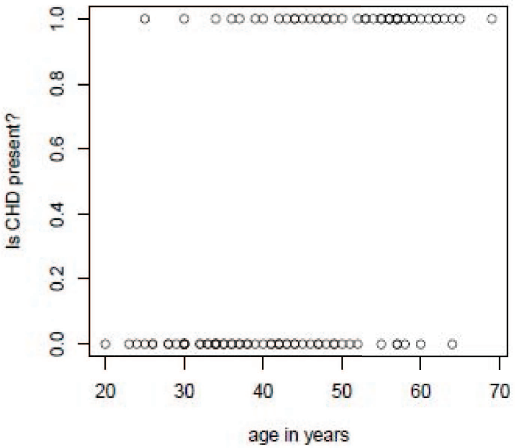
# CHD example

- (Taken from Hosmer and Lemeshow (2000), "Applied Logistic Regression: Second Edition", Copyright John Wiley & Sons).
- In a study of Coronary Heart Disease the following data were collected on 100 individuals:
  1. Patient identification code.
  2. Age
  3. Whether they have Coronary Heart Disease (0=absent, 1=present).
- Investigators are interested in modeling the relationship between age and the presence or absence of Coronary Heart Disease.

$E(y|x)$

$g(\mu) = X\beta$

# Plotting the age versus CHD



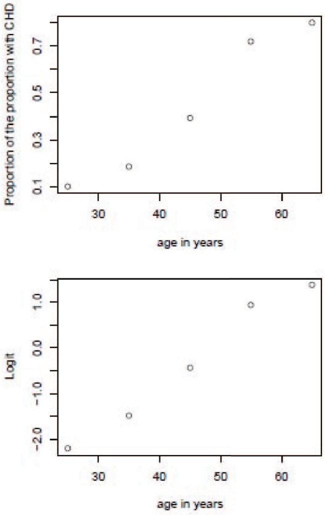
# Examining CHD proportion and age

- Instead of age, consider age groups, (e.g.,20-29,30-39,. . . )
- Count the number of 0's and 1's in each age group.
- Now tabulate the proportion of 1's for each age group.

| Age group | CHD absent | CHD present | CHD Proportion |
|---|---|---|---|
| 20-29 | 9 | 1 | 0.100 |
| 30-39 | 22 | 5 | 0.185 |
| 40-49 | 17 | 11 | 0.393 |
| 50-59 | 7 | 18 | 0.720 |
| 60-69 | 2 | 8 | 0.800 |

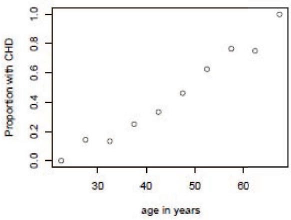
- Plot the midpoint of each age group versus the proportion.
- We can also plot the midpoint of each group versus the **logit** of the proportion.

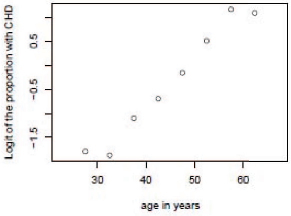
# Examining CHD proportion and age (cont.)



# Examining CHD proportion and age (cont.)

- Now consider a different choice of age groups (every 5 years of age):



# The simple linear logistic regression model

- Suppose that $Y_i(i=1,\ldots,n)$ are n independent $B(1,p_i)$ RVs. We let
$$\eta_i = \log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_i,$$
Where $x_i(i=1,\ldots,n)$ is some explanatory variable.
- For our example, $Y_i = CHD$ present/absent, and $x_i = age$.
- The above model equivalent to fitting $Y_i \sim B(1,p_i)$ where
$$p_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$$
- This is a generalized linear model for binomial data with a logit link function.

- To fit the simple linear logistic model:

```
tdata=read.table("file:///F:/Xu_WL/chdage.dat.txt",
      header=T)
chd.model=glm(formula=CHD~AGE,
     family=binomial(link = "logit"),data=tdata)
summary(chd.model)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9718  -0.8456  -0.4576   0.8253   2.2859

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.30945    1.13365  -4.683 2.82e-06 ***
AGE          0.11092    0.02406   4.610 4.02e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 136.66  on 99  degrees of freedom
Residual deviance: 107.35  on 98  degrees of freedom
AIC: 111.35

Number of Fisher Scoring iterations: 4
```

- We use the command **anova(chd.model)** to obtain the analysis of deviance table (here in R):

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: CHD

Terms added sequentially (first to last)

     Df Deviance Resid. Df Resid. Dev
NULL                    99     136.66
AGE   1     29.31       98     107.35
```

- The observed value of $D(\hat{y}, \hat{\mu})$, is the “deviance residual".

# The model test

- Asymptotically for large n, $D(y, \hat{\mu})$ has a $\chi_1^2$ distribution.
- So the P-value for this model is

$$Pr(D(y, \hat{\mu}) > 29.31) < 0.001$$

- To see this, note we obtain the P-value as follows.

$$1 - pchisq(29.31, 1)$$

$$6.167658e - 08$$

- Conclusion:

# Statistical inference

- **Example**: What is the estimated probability that a randomly chosen 50 year old subject has CHD present?
- **Example**: Produce a 95% CI for the estimated probability that a randomly chosen 50 year old subject has CHD present.

# Using other link functions

- The logit is the most commonly used link function.
- Historically, it was not the first link function to be used in binomial regression. The probit link is defined to be

$$g_1(p) = \phi^{-1}(p),$$

where $\phi^{-1}()$ is the inverse cumulative distribution function for a standard normal random variable.

- Used traditionally for bioassay experiments (Finney,1973).
——Experiments in which a biological organism is used to test for chemical toxicity.

# Beeetle example

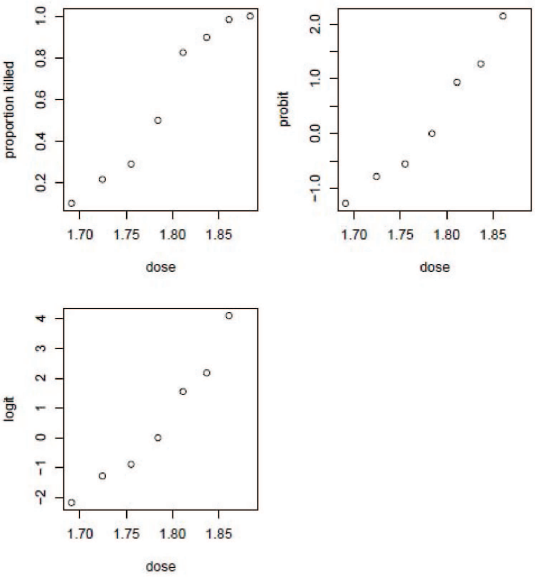
- (Taken from Table 7.2 of Dobson,2000).
- The number of beetles that died after five hours exposure to a different number of concentrations of carbon disulfide(doses).
- Concentration units are $\log_{10} CS_2$ mg/l.

1.6907

| Dose | Total Number | Number Killed |
|---|---|---|
| 1.6907 | 59 | 6 |
| 1.7242 | 60 | 13 |
| 1.7552 | 62 | 18 |
| 1.7842 | 56 | 28 |
| 1.8113 | 63 | 52 |
| 1.8369 | 59 | 53 |
| 1.8610 | 62 | 61 |
| 1.8839 | 60 | 60 |

- What link do we choose?

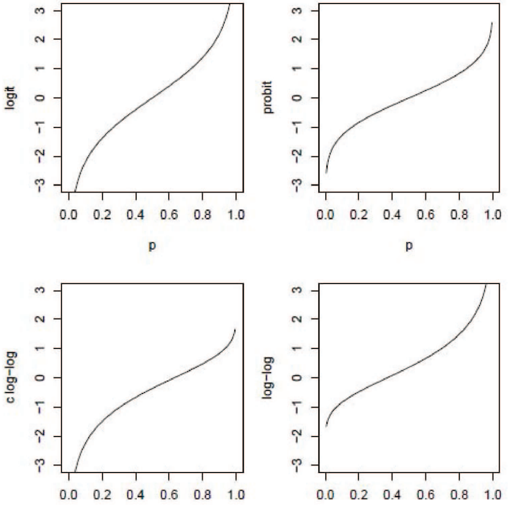
# Examining proportion and dose



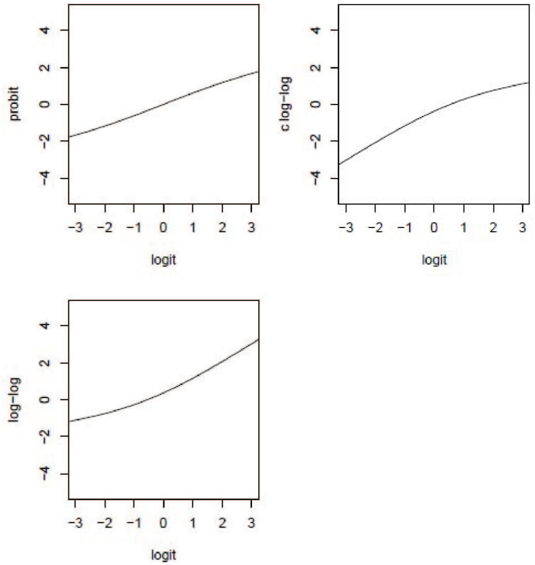
# Common choices of link functions

- The four hours commonly used link functions(from most common to least common)are:
  1.Logit: $g_1(p) = \log(p/(1-p))$
  2.Probit: $g_2(p) = \Phi^{-1}(p)$
  3.Complementary log-log: $g_3(p) = \log(-\log(1-p))$
  4.Log-log: $g_4(p) = -\log(-\log p)$

# Plots of the links



# Comparing links(on the logit scale)



# Comments on the links

- All the functions are increasing, continuous, and differentiable over $0 < p < 1$.
- The logit and probit are almost linearly related over the internal $p \in [0.1, 0.9]$.
- For small $p$, complementary log-log close to logit.
- For large $p$, log-log close to logit.
- The complementary log-log approaches infinity slower than any other link function.

- The data is represented in a compact form of counts for each combination of dose.
- We do not need to expand the dataset into 0s and 1s.
- Instead, we fit the model using the **weights** command:

```
beetles=read.table("file:///F:/Xu_WL/beetles.dat.txt",
       header=T)
prob=beetles$killed/beetles$number
beetle.probit<-glm(formula=prob~dose,
  family=binomial(link="probit"), data=beetles,
  weights=number)
summary(beetle.probit)
anova(beetle.probit)
```

# Comments on the code

- **Weights** declares the number of individuals which make up each proportion and each factor combination.
- **Binominal(link="probit")** declares a binominal glm model with a probit link function.
  (we can also use**logit** or **cloglog-log** is not available in R)

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5714  -0.4703   0.7501   1.0632   1.3449

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -34.935      2.648  -13.19   <2e-16 ***
dose          19.728      1.487   13.27   <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: 284.20  on 7  degrees of freedom
Residual deviance:  10.12  on 6  degrees of freedom
AIC: 40.318

Number of Fisher Scoring iterations: 4
```

## Comparing analysis of deviance tables

- ```
  Analysis of Deviance Table

  Model: binomial, link: probit

  Response: prob

  Terms added sequentially (first to last)

       Df Deviance Resid. Df Resid. Dev
  NULL                      7     284.20
  dose  1   274.08          6      10.12
  ```
- ```
  Analysis of Deviance Table

  Model: binomial, link: logit

  Response: prob

  Terms added sequentially (first to last)

       Df Deviance Resid. Df Resid. Dev
  NULL                      7    284.202
  dose  1   272.97          6     11.232
  ```
- Which link do would you choose?

# Interpreting glm models with different links

- Interpretation of the coefficients in each model follows by considering the inverse link function.

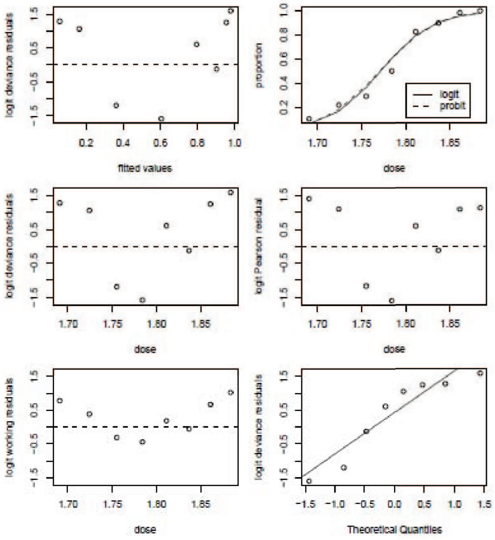
  1.Logit: $h_1(\eta) = e^\eta/(1 - e^\eta)$

  2.Probit: $h_2(\eta) = \Phi(\eta)$

  3.Complementary log-log: $h_3(\eta) = 1 - \exp(-e^\eta)$

  4.Log-log: $h_4(\eta) = \exp(-e^\eta)$

# Residuals analysis



# Residuals analysis(cont.)

- Conclusions?

# 思考题

思考题：$X$ 和$Y$ 均为二分类变量，$X \in \{0,1\}, Y \in \{0,1\}$, 则列联表分析的卡方检验和logistic模型对于系数$\beta_1$的检验是否一致？

对于获取的样本$X_i, Y_i (i = 1, 2, \cdots, n)$, 整理可得

| X/Y | 0 | 1 |
|---|---|---|
| 0 | $n_{00}$ | $n_{01}$ |
| 1 | $n_{10}$ | $n_{11}$ |

二维列联表的独立性检验

设有$X$，$Y$二个离散型随机变量，分别取$r$个值和$c$个值，对应地有二维列联表$r \times c$，作$n$次观测，在$(i,j)$格的观测频数为$n_{ij}, i = 1, 2, ..., r; j = 1, 2, ..., c$. 观测值落入$(i,j)$格的概率为$p_{ij}$，观测频数服从多项分布，其概率密度为

$$\frac{n!}{\prod_{i=1}^{r} \prod_{j=1}^{c} n_{ij}!} \cdot \prod_{i=1}^{r} \prod_{j=1}^{c} p_{ij}^{n_{ij}}$$

由于$\sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} = 1$，所以参数空间的独立参数个数为$r \cdot c - 1$，$p_{ij}$的$MLE$ 为$\hat{p}_{ij} = n_{ij}/n$

# 思考题

考虑列联表的独立性检验，原假设$H_0: p_{ij}=p_{i.}\cdot p_{.j}, i=1,2,...,r; j=1,2,...,c$，其中$p_{i.}$和$p_{.j}$分别是$X$和$Y$的边缘分布。原假设成立时，$p_{ij}=p_{i.}\cdot p_{.j}$,所以参数空间的独立参数个数为$r+c-2$个。

这时，$p_{ij}$的$MLE$为$\hat{p_{i.}}\cdot\hat{p_{.j}}=(n_i/n)\cdot(n_j/n)$

该检验问题的似然比统计量为：

$$\Lambda(X)=\frac{\prod_{i=1}^{r}\prod_{j=1}^{c}(\frac{n_{ij}}{n})^{n_{ij}}}{\prod_{i=1}^{r}\prod_{j=1}^{c}(\frac{n_{i.}}{n}\cdot\frac{n_{.j}}{n})^{n_{ij}}}$$

由于$(r\cdot c-1)-(r+c-2)=(r-1)(c-1)$，故在原假设成立时，$2ln\Lambda$的极限分布为$\chi^2((r-1)(c-1))$。在$2ln\Lambda\geq\chi^2_{1-\alpha}((r-1)(c-1))$时，拒绝原假设。

logistic模型系数检验:

$Y_i \sim B(1, p_i)$，密度函数为：$f(y_i, p_i) = p_i^{y_i}(1-p_i)^{1-y_i}$.

logistic 模型：$p_i = \frac{exp(\beta_0+\beta_1 x_i)}{1+exp(\beta_0+\beta_1 x_i)}$.

$H_0 : \beta_1 = 0$ $H_1 : \beta_1 \neq 0$，似然比检验统计量：

$$\Lambda = \frac{\prod_{i=1}^{n}\left(\frac{exp(\hat{\beta}_0+\hat{\beta}_1 x_i)}{1+exp(\hat{\beta}_0+\hat{\beta}_1 x_i)}\right)^{y_i}\left(\frac{1}{1+exp(\hat{\beta}_0+\hat{\beta}_1 x_i)}\right)^{1-y_i}}{\prod_{i=1}^{n}\left(\frac{exp(\tilde{\beta}_0)}{1+exp(\tilde{\beta}_0)}\right)^{y_i}\left(\frac{1}{1+exp(\tilde{\beta}_0)}\right)^{1-y_i}}$$

其中$\hat{\beta}_0$ 和$\hat{\beta}_1$ 是在饱和模型下参数的极大似然估计值。$\tilde{\beta}_0$是在原假设下参数的极大似然估计值，$2ln\Lambda \sim \chi^2(1)$。

# 思考题

证明两种方法的一致性

先计算$\hat{\beta}_0$ 和$\hat{\beta}_1$:

$$L=\prod_{i=1}^{n}\left(\frac{exp(\beta_0+\beta_1x_i)}{1+exp(\beta_0+\beta_1x_i)}\right)^{y_i}\left(\frac{1}{1+exp(\beta_0+\beta_1x_i)}\right)^{1-y_i}$$

$$lnL=\sum_{i=1}^{n}\left(y_iln\frac{exp(\beta_0+\beta_1x_i)}{1+exp(\beta_0+\beta_1x_i)}+(1-y_i)ln\frac{1}{1+exp(\beta_0+\beta_1x_i)}\right)$$

$$\frac{\partial lnL}{\partial\beta_0}=\sum_{i=1}^{n}\left(y_i\frac{1}{1+exp(\beta_0+\beta_1x_i)}+(1-y_i)\frac{-exp(\beta_0+\beta_1x_i)}{1+exp(\beta_0+\beta_1x_i)}\right)$$

令$\frac{\partial lnL}{\partial\beta_0}=0$，等价于

$$\sum_{i=1}^{n}\frac{y_i}{1+exp(\beta_0+\beta_1x_i)}=\sum_{i=1}^{n}\frac{(1-y_i)exp(\beta_0+\beta_1x_i)}{1+exp(\beta_0+\beta_1x_i)}$$

$$n_{01}\cdot\frac{1}{1+exp(\beta_0)}+n_{11}\cdot\frac{1}{1+exp(\beta_0+\beta_1)}=n_{00}\cdot\frac{exp(\beta_0)}{1+exp(\beta_0)}+n_{10}\cdot\frac{exp(\beta_0+\beta_1)}{1+exp(\beta_0+\beta_1)}$$

# 思考题

$$\frac{\partial lnL}{\partial\beta_1}=\sum_{i=1}^{n}\left(y_i\frac{x_i}{1+exp(\beta_0+\beta_1x_i)}+(1-y_i)x_i\frac{-exp(\beta_0+\beta_1x_i)}{1+exp(\beta_0+\beta_1x_i)}\right)$$

令 $\frac{\partial lnL}{\partial\beta_1}=0$，等价于

$$\sum_{i=1}^{n}\frac{y_ix_i}{1+exp(\beta_0+\beta_1x_i)}=\sum_{i=1}^{n}\frac{(1-y_i)x_iexp(\beta_0+\beta_1x_i)}{1+exp(\beta_0+\beta_1x_i)}$$

$$n_{11}\cdot\frac{1}{1+exp(\beta_0+\beta_1)}=n_{10}\cdot\frac{exp(\beta_0+\beta_1)}{1+exp(\beta_0+\beta_1)}$$

$n_{11}=n_{10}\cdot exp(\beta_0+\beta_1)$，结合上面的结果，得到：

$$exp(\hat{\beta}_0+\hat{\beta}_1)=\frac{n_{11}}{n_{10}},\quad exp(\hat{\beta}_0)=\frac{n_{01}}{n_{00}}$$

# 思考题

将$\hat{\beta}_0$ 和$\hat{\beta}_1$ 代入Λ分子部分的似然函数中：

$$
\begin{aligned}
L &= \prod_{i=1}^{n}\left(\frac{exp(\hat{\beta}_0+\hat{\beta}_1 x_i)}{1+exp(\hat{\beta}_0+\hat{\beta}_1 x_i)}\right)^{y_i}\left(\frac{1}{1+exp(\hat{\beta}_0+\hat{\beta}_1 x_i)}\right)^{1-y_i} \\
&= (\frac{1}{1+exp(\hat{\beta}_0)})^{n_{00}}\cdot(\frac{exp(\hat{\beta}_0)}{1+exp(\hat{\beta}_0)})^{n_{01}}\cdot(\frac{1}{1+exp(\hat{\beta}_0+\hat{\beta}_1)})^{n_{10}}\cdot(\frac{exp(\hat{\beta}_0+\hat{\beta}_1)}{1+exp(\hat{\beta}_0+\hat{\beta}
\end{aligned}
$$

$$
= (\frac{n_{00}}{n_{0.}})^{n_{00}}\cdot(\frac{n_{01}}{n_{0.}})^{n_{01}}\cdot(\frac{n_{10}}{n_{1.}})^{n_{10}}\cdot(\frac{n_{11}}{n_{1.}})^{n_{11}}.
$$

# 思考题

计算$\tilde{\beta}_0$：

$$L=\prod_{i=1}^{n}\left(\frac{exp(\beta_0)}{1+exp(\beta_0)}\right)^{y_i}\left(\frac{1}{1+exp(\beta_0)}\right)^{1-y_i}$$

$$\frac{\partial lnL}{\partial\beta_0}=\sum_{i=1}^{n}\left(y_i\frac{1}{1+exp(\beta_0)}+(1-y_i)\frac{-exp(\beta_0)}{1+exp(\beta_0)}\right)$$

令$\frac{\partial lnL}{\partial\beta_0}=0$，等价于$\sum_{i=1}^{n}y_i=\sum_{i=1}^{n}(1-y_i)exp(\beta_0)$

即$n_{.1}=n_{.0}exp(\beta)$，$\tilde{\beta}_0=\frac{n_{.1}}{n_{.0}}$

将$\tilde{\beta}_0$ 代入Λ分母部分的似然函数：

$$L=\prod_{i=1}^{n}\left(\frac{n_{.1}}{n}\right)^{y_i}\left(\frac{n_{.0}}{n}\right)^{1-y_i}=\left(\frac{n.1}{n}\right)^{n_{11}+n_{01}}\cdot\left(\frac{n.0}{n}\right)^{n_{00}+n_{10}}$$

# 思考题

$$
\begin{aligned}
\Lambda &= \frac{(\frac{n_{00}}{n_{0.}})^{n_{00}} \cdot (\frac{n_{01}}{n_{0.}})^{n_{01}} \cdot (\frac{n_{10}}{n_{1.}})^{n_{10}} \cdot (\frac{n_{11}}{n_{1.}})^{n_{11}}}{(\frac{n_{.1}}{n})^{n_{11}+n_{01}} \cdot (\frac{n_{.0}}{n})^{n_{00}+n_{10}}} \\
&= (\frac{n_{00} \cdot n}{n_{0.} \cdot n_{.0}})^{n_{00}} \cdot (\frac{n_{01} \cdot n}{n_{0.} \cdot n_{.1}})^{n_{01}} \cdot (\frac{n_{10} \cdot n}{n_{1.} \cdot n_{.0}})^{n_{10}} \cdot (\frac{n_{11} \cdot n}{n_{1.} \cdot n_{.1}})^{n_{11}} \\
&= \frac{(\frac{n_{00}}{n})^{n_{00}} \cdot (\frac{n_{01}}{n})^{n_{01}} \cdot (\frac{n_{10}}{n})^{n_{10}} \cdot (\frac{n_{11}}{n})^{n_{11}}}{(\frac{n_{0.}}{n} \cdot \frac{n_{.0}}{n})^{n_{00}} \cdot (\frac{n_{0.}}{n} \cdot \frac{n_{.1}}{n})^{n_{01}} \cdot (\frac{n_{1.}}{n} \cdot \frac{n_{.0}}{n})^{n_{10}} \cdot (\frac{n_{1.}}{n} \cdot \frac{n_{.1}}{n})^{n_{11}}} \\
&= \frac{\prod_{i\in\{0,1\}} \prod_{j\in\{0,1\}} (\frac{n_{ij}}{n})^{n_{ij}}}{\prod_{i\in\{0,1\}} \prod_{j\in\{0,1\}} (\frac{n_{i.}}{n} \cdot \frac{n_{.j}}{n})^{n_{ij}}}
\end{aligned}
$$

证毕。

# 问题

- 当$X$为四分类变量，$X \in \{1, 2, 3, 4\}$.
- 例如，$X$表示螃蟹的颜色，$X = 1$表示light medium，$X = 2$表示medium，$X = 3$表示dark medium，$X = 4$表示dark.
- 如何检验$X$作为一个整体，对$Y$是否有影响.

# 问题

检验1：列联表

似然比统计量为：

$$\lambda(X) = \frac{\prod_{i=1}^{n} p(X_i; \hat{\theta})}{\prod_{i=1}^{n} p(X_i; \theta_0)}$$

$$2ln\lambda(X) \sim \chi^2(3)$$

# 问题

检验2：logistic模型似然比检验

引入3个虚拟变量构建logistic模型

$$logit(p_i) = \beta_0 + X_1\beta_1 + X_2\beta_2 + X_3\beta_3$$

原假设$H_0$：$\beta_1 = \beta_2 = \beta_3 = 0$

似然比统计量：

$$\frac{\prod_{i=1}^{n} f(y_i; \hat{p}_i)}{\prod_{i=1}^{n} f(y_i; p_i)}$$

分子：饱和模型似然函数；分母：$H_0$成立下模型的似然函数

$2ln\lambda(X) \sim \chi^2(3)$

# More complicated binomial GLM models

- So far we have considered binomial GLMs in which we wish to model the dependence between a binomial proportion and a continuous explanatory variable.
- We shall now consider models which
  - incorporate factors.
  - incorporate multiple terms.
  - allow for interactions.

# Horse shoe crabs example

(Taken from Agresti,1996).

- Data from a study of nesting horse shoe crabs.
- Response variable:number of satellites (number of males residing nearby).
- Explanatory variables of interest:
  - color of the crab (1:light medium,2:medium,3:dark medium,4:dark).
  - spine condition (1:both good,2:one worn or broken,3:both worn or broken).
  - width of the crab (in cm).
  - weight of the crab (in kg).

# A binomial response

- Instead of counting the number of satellites, consider a binomial response:
  has satellite (0:no males reside nearby, 1:at least one male resides nearby).
- Scientific question of interest:
  Can we build a model to relate the probability of having a satellite nearby to the explanatory variables of interest?

# Incorporating factors

- We start by relating the probability of having a satellite nearby to whether color is dark or not. Recode color as a variable ‘is.dark’: 0 not dark (light medium or medium), 1 dark (dark medium or dark).
- Summarizing the counts we have:

| color | no satellites nearby | satellites nearby | total |
|---|---|---|---|
| not dark | 29 | 78 | 107 |
| dark | 33 | 33 | 66 |

- Expressed as proportions:

| color | no satellites nearby | satellites nearby |
|---|---|---|
| not dark | 0.271 | 0.729 |
| dark | 0.500 | 0.500 |

- Initial thoughts:

Estimating these proportions using a binomial GLM

- Here is the code to fit the binomial glm model:

```
crabs=read.table("file:///F:/Xu_WL/crabs.dat.txt",
    header=T)
has.satellite=c(crabs$satellite>=1)
is.dark=crabs$color>=3
crabs.glm=glm(has.satellite ~ is.dark,
 family=binomial(link="logit"), data=crabs)
summary(crabs.glm)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6159  -1.1774   0.7951   0.7951   1.1774

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.9894     0.2175   4.549 5.39e-06 ***
is.darkTRUE  -0.9894     0.3285  -3.012   0.0026 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 216.53  on 171  degrees of freedom
AIC: 220.53

Number of Fisher Scoring iterations: 4
```

# Odds and odds ratio

- Commonly used concepts when working with probabilities
- Odds of an event
  - The probability that the event happens=p
  - Odds in favor of the event is given as $odds = p/(1-p) = \exp[logit(p)]$
  - Examples: 病情复发的odds;
- The odds ratio is the ratio of the odds of an event occurring in one group to the odds of it occurring in another group
  - Example: Use odds ratio to compare the effects of the treatments

# Odds and odds ratio

- Let p denote the probability of relapse occurrence
- Odds of relapse occurrence is p/(1-p)
  - Smaller is better, indicating a relapse is less likely to occur
- The odds ratio of occurrence in drug A group against placebo group is used to compare the treatments
  - An odd ratio less than 1 indicates that relapse is less likely to occur in drug A group, i.e., drug A is better than placebo

# Odds and odds ratio-Interpret the parameter of a continuous predictor

- Assume a glm contains a single continuous predictor x. The mean structure is

$$logit(p) = \beta_0 + \beta_1 x$$

  - When $x = a$, the odds is $\exp(logit(p)) = \exp(\beta_0 + \beta_1 a)$
  - Increasing $x = a$ by one unit to $x = a+1$, the odds changes to $\exp(\beta_0 + \beta_1 a + \beta_1)$
- The odds ratio is

$$\frac{adds_{a+1}}{odds_a} = \frac{\exp(\beta_0 + \beta_1 a + \beta_1)}{\exp(\beta_0 + \beta_1 a)} = \exp(\beta_1)$$

- Then $[\exp(\beta_1) - 1] \times 100\%$ is the percent change in the odds of event occurrence when x is increased by one unit

# Odds and odds ratio-Interpret the parameter of a categorical predictor

- Assume a glm contains a single categorical predictor x with two levels (A and B). The mean structure is

$$logit(p_i) = \begin{cases} \beta_0, & \\ \beta_0 + \beta_B, & \text{x=B} \end{cases}$$

  - When $x = A$ the odds is $\exp(\beta_0)$
  - When $x = B$ the odds is $\exp(\beta_0 + \beta_B)$
- The odds ratio (B against A) is

$$\frac{odds_B}{odds_A} = \exp(\beta_B)$$

- Then $[\exp(\beta_B - 1] \times 100\%$ tells the percent change in the odds in group B compared to group A

# Understanding the model

- Our model is that $\{Y_{i,j} : i = 1, 2; j = i, \cdots, m_j\}$ are a set of independent RVs with $Y_{i,j} \sim B(1, p_i)$ and where $p_i$ satisfies

$$\log\left(\frac{p_i}{1 - p_i}\right) = \mu + \alpha_i$$

- In our example we have: $m_1 = 107$,$m_2 = 66$.
- To fit the model we assume that $\alpha_1 = 0$. This is the default in R.

# An equivalent model

- Equivalently $\{Y_i : i = 1, 2\}$ are two independent RVs with $Y_i \sim B(m_i, p_i)$ and where $p_i$ satisfies

$$\log\left(\frac{p_i}{1-p_i}\right) = \mu + \alpha_i$$

- We can fit this model using the **weights** option in **glm** (see last lecture).

# Interpreting the factor level effects

- Estimated logit for the two factors based glm model is
  not dark:
  dark:
- Estimated proportions are
  not dark:
  dark:
- Confidence intervals for these logits and proportions follow in a natural way.

# Differences of logits

- Example: What is the estimated difference between the ‘not dark’ and ‘dark’ logits? What is a 95% CI for this difference?

  (This difference is the **log odds ratio**).

- Here is the analysis of deviance table:

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: has.satellite

Terms added sequentially (first to last)

        Df Deviance Resid. Df Resid. Dev
NULL                      172     225.76
is.dark  1   9.2275       171     216.53
```

- Conclusions:
- Exercise: what is this deviance test equivalent to?

# More factor levels

- Now consider a model relating the probability of having a satellite and the color (which has 4 factor levels).
- The table of counts are:

| color | no satellites nearby | satellites nearby | total |
|---|---|---|---|
| light medium | 3 | 9 | 12 |
| medium | 26 | 69 | 9 |
| dark medium | 18 | 26 | 44 |
| dark | 15 | 7 | 22 |

- Initial summary:

## The code and model summary

- The code is:

```
crabs=read.table("file:///F:/Xu_WL/crabs.dat.txt",
  header=T)
has.satellite=c(crabs$satellite>=1)
crabs.color=glm(has.satellite~factor(color),
  data=crabs,family="binomial")
summary(crabs.color)
anova(crabs.color)
```

## The code and model summary

- The result is:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6651  -1.3370   0.7997   0.7997   1.5134

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.0986     0.6667   1.648   0.0994 .
factor(color)2  -0.1226     0.7053  -0.174   0.8620
factor(color)3  -0.7309     0.7338  -0.996   0.3192
factor(color)4  -1.8608     0.8087  -2.301   0.0214 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 212.06  on 169  degrees of freedom
AIC: 220.06

Number of Fisher Scoring iterations: 4
```

## Analysis of deviance:

- ```
  Analysis of Deviance Table

  Model: binomial, link: logit

  Response: has.satellite

  Terms added sequentially (first to last)

                Df Deviance Resid. Df Resid. Dev
  NULL                            172     225.76
  factor(color)  3   13.698       169     212.06
  ```
- Overall conclusions:
  ```

# A warning and a discussion of asymptotics

- The limiting distribution of $\lambda^2_{n-p}$ for the deviance and Pearson statistics are based on the following assumptions:

  1. The sample of $n$ observations are distributed independently with a $B\,(m_i, p_i)$ distribution.

  2. The overall sample size $n$ remains fixed, but $m_i \to \infty$, such that $m_i p_i (1-p_i) \to \infty$ for each $i$.

- if $n$ is large and $m_i p_i (1-p_i)$ remains bounded the theory breaks down:

  1. The $\lambda^2$ distribution no longer applies.

  2. $D(y, \hat{\mu})$ is not independent of $\hat{p}$.

# Light at the end of the tunnel

- Suppose we fit two glm models. In the first model assume that $\hat{\mu}_0$ is the estimate of $\mu$. In the second model we add one additional covariate to the first model. Let $\hat{\mu}_1$ be the estimate of $\mu$ in this case.
- Then $D(y;\hat{\mu}_0) - D(y;\hat{\mu}_1)$ has an asymptotic $\lambda_1^2$ distribution under either of following settings.

  1. $n \to \infty$.

  2. The overall sample size $n$ remains fixed, but $m_i \to \infty$, such that $m_i p_i(1-p_i) \to \infty$ for each $i$.

# Multiple terms in the binomial GLM

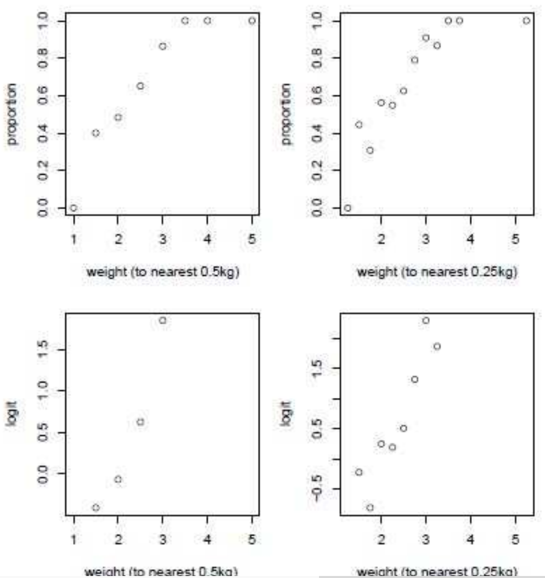
- We will continue our analysis of the horseshoe crabs dataset.
- We are now interested in **model building**, allowing for the possibility of incorporating more terms in the model.
  (We have already shown that the color is associated with the probability of having a satellite nearby).
- We start by producing **pairwise summaries** of the response variable (whether there is a satellite nearby) by each of the explanatory variables in the dataset (first the factor variables, and then the continuous variables).
- We could extend these comparisons to combinations of three or more variables.

# Summary of having a satellite by weight

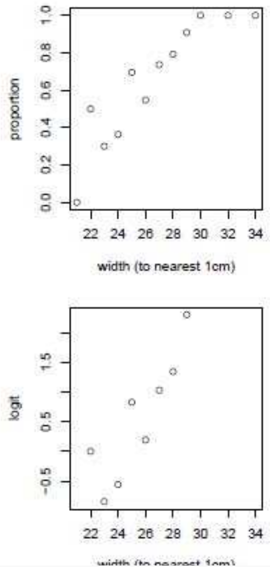
| color | counts: no satellites nearby | counts: satellites nearby | proportions: no satellites nearby | proportions: satellites nearby |
|---|---|---|---|---|
| light medium | 3 | 9 | 0.250 | 0.750 |
| medium | 26 | 69 | 0.274 | 0.726 |
| dark medium | 18 | 26 | 0.409 | 0.591 |
| dark | 15 | 7 | 0.682 | 0.318 |

| spine condition | counts: no satellites nearby | counts: satellites nearby | proportions: no satellites nearby | proportions: satellites nearby |
|---|---|---|---|---|
| both good | 11 | 26 | 0.297 | 0.703 |
| one broken | 8 | 7 | 0.533 | 0.467 |
| both broken | 43 | 78 | 0.355 | 0.645 |

# 2.2.Summary of having a satellite by weight



# Summary of having a satellite by width



# Our model building strategy

- We will consider a step-wise modeling strategy, starting with the simplest model and making our model more complicated.
- We are looking for a **parsimonious model**, which models/predicts the probability of having a satellite nearby.
- This is not the only way to choose a model!
- We use tests based on the analysis of deviance table to help us select a model.(We may also use residual plots, when appropriate).
- For brevity, I will demonstrate only a subset of the models we could consider.

# Modeling the probability of having a satellite nearby in terms of the width

```
crabs=read.table("file:///F:/Xu_WL/crabs.dat.txt",header=T)
has.satellite=c(crabs$satellite>=1)
crabs.width<-glm(has.satellite~width,
 data=crabs,family="binomial")
summary(crabs.width)
anova(crabs.width)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0281  -1.0458   0.5480   0.9066   1.6942

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.3508     2.6287  -4.698 2.62e-06 ***
width         0.4972     0.1017   4.887 1.02e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 194.45  on 171  degrees of freedom
AIC: 198.45

Number of Fisher Scoring iterations: 4
```

## The analysis of deviance

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: has.satellite

Terms added sequentially (first to last)

      Df Deviance Resid. Df Resid. Dev
NULL                    172     225.76
width  1   31.306       171     194.45
```

- We test:
- The P-value is:
- Conclusion:

# Adding color to the model

```
crabs=read.table("file:///F:/Xu_WL/crabs.dat.txt",header=T)
has.satellite=c(crabs$satellite>=1)
crabs.widcol<-glm(has.satellite~width+factor(color),data=crabs,family="binomial")
summary(crabs.widcol)
anova(crabs.widcol)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1124  -0.9848   0.5243   0.8513   2.1413

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -11.38519    2.87346  -3.962 7.43e-05 ***
width            0.46796    0.10554   4.434 9.26e-06 ***
factor(color)2   0.07242    0.73989   0.098    0.922
factor(color)3  -0.22380    0.77708  -0.288    0.773
factor(color)4  -1.32992    0.85252  -1.560    0.119
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 187.46  on 168  degrees of freedom
AIC: 197.46

Number of Fisher Scoring iterations: 4
```

## The analysis of deviance table

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: has.satellite

Terms added sequentially (first to last)

              Df Deviance Resid. Df Resid. Dev
NULL                            172     225.76
width          1  31.3059       171     194.45
factor(color)  3   6.9956       168     187.46
```

- We test:
- The P-value is:
- Conclusion:

# Adding weight to the width model

```
Call:
glm(formula = has.satellite ~ width + weight, family = "binomial",
    data = crabs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1127  -1.0344   0.5304   0.9006   1.7207

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.3547     3.5280  -2.652  0.00801 **
width         0.3068     0.1819   1.686  0.09177 .
weight        0.8338     0.6716   1.241  0.21445
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 192.89  on 170  degrees of freedom
AIC: 198.89

Number of Fisher Scoring iterations: 4
```

## The analysis of deviance table

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: has.satellite

Terms added sequentially (first to last)

       Df Deviance Resid. Df Resid. Dev
NULL                     172     225.76
width   1  31.3059       171     194.45
weight  1   1.5608       170     192.89
```

- We test:
- The P-value is:
- Conclusion:

Further steps in the modeling strategy

- Continuing in a similar way we find that:

  Spine condition is **not significant** when added to the model which predicts the probability of having a satellite nearby using the width variable;

  The variable 'is dark' **not significant** when added to the width model.

- Instead of our current definition of 'is dark', consider the variable 'is very dark' which is:

  0 if the color is light medium, medium, or dark medium;

  1 if the color is dark.

```
crabs=read.table("file:///F:/Xu_WL/crabs.dat.txt",header=T)
has.satellite=c(crabs$satellite>=1)
is.dark=crabs$color<=3
crabs.widdark<-glm(has.satellite~width+is.dark,data=crabs,family="binomial")
summary(crabs.widdark)
anova(crabs.widdark)
```

# Adding “is.very.dark" to the width model

```
Call:
glm(formula = has.satellite ~ width + is.dark, family = "binomial",
    data = crabs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0821  -0.9932   0.5274   0.8606   2.1553

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.9795     2.7272  -4.759 1.94e-06 ***
width         0.4782     0.1041   4.592 4.39e-06 ***
is.darkTRUE   1.3005     0.5259   2.473   0.0134 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 187.96  on 170  degrees of freedom
AIC: 193.96

Number of Fisher Scoring iterations: 4
```

## The analysis of deviance table

```
Analysis of Deviance Table

Model: binomial, link: logit

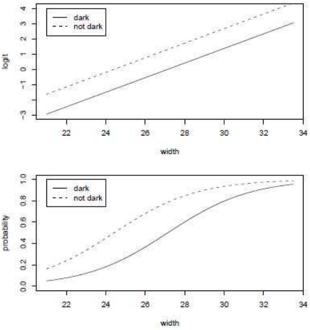
Response: has.satellite

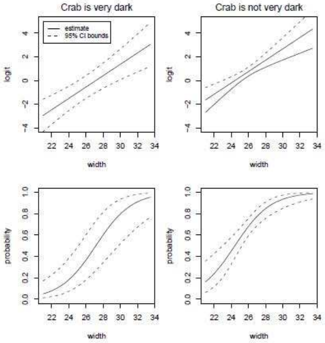
Terms added sequentially (first to last)

        Df Deviance Resid. Df Resid. Dev
NULL                      172     225.76
width    1  31.3059       171     194.45
is.dark  1   6.4948       170     187.96
```

- We test:
- The P-value is:
- Conclusion:

# Demonstrating the model predictions



# Incorporating interaction terms

- On the link scale (e.g., the logit), the **interpretation** of interactions is the same as for linear models.
  –The interpretation is more complicated on the original scale.
- For our example, suppose we want to fit a different slope parameter for the 'width' variable according to whether the crabs are 'very dark' or 'not very dark'.
- Arrange the data into two groups, letting $i = 1$ if a crab is dark and $i = 2$ if a crab is not dark. Let the index j denote the crabs in group i.
- Let $y_{ij}$ be 0 if no satellites nearby and 1 otherwise, with $Y_{ij}$ denoting the associated RVs.
- Let $w_{ij}$ be the widths of the $j$th crab in the $i$th color group.

# Comparing the model with and without the interaction

- **Without** an interaction term, the model is $Y_{ij} \sim B\left(1; p_{ij}\right)$ with

$$\eta_{ij} = g\left(p_{ij}\right) = \alpha + \beta w_{ij} + \gamma_i$$

  To fit this model we assume that $\gamma_1 = 0$.

- **Adding** an interaction term, the model becomes $Y_{ij} \sim B\left(1; p_{ij}\right)$ with

$$\eta_{ij} = g\left(p_{ij}\right) = \alpha + \beta w_{ij} + (\beta\gamma)_i \, w_{ij}$$

  We assume that $\gamma_1 = 0$ and $(\beta\gamma)_1 = 0$.

# The R model specifications

- **Without** an interaction term, the model is:

  has.satellite $\sim$ width $+$ weight.

- **Adding** an interaction term, the model becomes:

  has.satellite $\sim$ width $*$ is.very.dark

  Another way to write this is:

  has.satellite $\sim$ width $+$ is.very.dark $+$ width:is.very.dark

# The resulting model summary

```
Call:
glm(formula = has.satellite ~ width + is.dark + width * is.dark,
    family = "binomial", data = crabs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1366  -0.9344   0.4996   0.8554   1.7753

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -5.8538     6.6939  -0.874    0.382
width               0.2004     0.2617   0.766    0.444
is.darkTRUE        -6.9578     7.3182  -0.951    0.342
width:is.darkTRUE   0.3217     0.2857   1.126    0.260

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 186.79  on 169  degrees of freedom
AIC: 194.79

Number of Fisher Scoring iterations: 4
```

## Analysis of Deviance Table

```
Analysis of Deviance Table

Model: binomial, link: logit

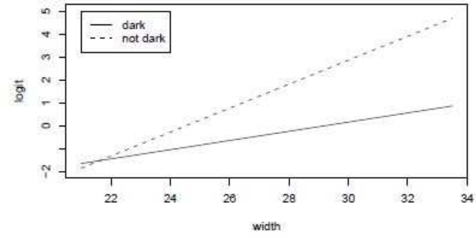
Response: has.satellite

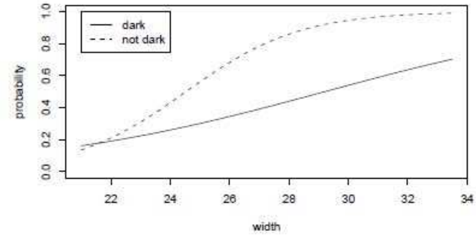
Terms added sequentially (first to last)

              Df Deviance Resid. Df Resid. Dev
NULL                            172     225.76
width          1  31.3059       171     194.45
is.dark        1   6.4948       170     187.96
width:is.dark  1   1.1715       169     186.79
```

- We test:
- The P-value is:
- Conclusion:

# Demonstrating the effect of the interactions



# 思考题1

思考题：Logistic回归中，参数假设检验$H_0:\beta_1=0$，$H_1:\beta_1\neq 0$，似然比检验与两因素t检验的关系。

- t统计量：根据Y的值将X分成两组，当$y=1$时，$x_i\sim N(\mu_1,\sigma^2),i=1,\ldots,n_1$，$\bar{X}_1$服从$N(\mu_1,\frac{\sigma^2}{n_1})$；当$y=0$时，$x_j\sim N(\mu_2,\sigma^2),j=1,\ldots,n_2$，$\bar{X}_2$服从$N(\mu_2,\frac{\sigma^2}{n_2})$。
- 等价于检验$H_0:\mu_1=\mu_2$，$H_a:\mu_1\neq\mu_2$。
- 当$H_0$成立时，$t=\frac{\bar{X}_1-\bar{X}_2}{\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})\hat{\sigma}^2}}\sim t(n-2)$。

# 思考题1

- 似然比检验：当$H_0$成立时，即$\beta_1 = 0$
- $p = \frac{e^{\beta_0}}{1+e^{\beta_0}}$，对数似然函数为：

$$
\begin{aligned}
l_0 &= \sum_{i=1}^{n}[y_i \log \frac{e^{\beta_0}}{1+e^{\beta_0}} + (1-y_i)\log \frac{1}{1+e^{\beta_0}}] \\
&= n_1 \log \frac{e^{\beta_0}}{1+e^{\beta_0}} + n_2 \log \frac{1}{1+e^{\beta_0}} \\
&= n_1\beta_0 - n\log(1+e^{\beta_0})
\end{aligned}
$$

- 关于$\beta_0$求导数：

$$\frac{dl_0}{d\beta_0} = n_1 - n\frac{e^{\beta_0}}{1+e^{\beta_0}} = 0$$

可以得到$\frac{\hat{\beta_0}}{1+e^{\hat{\beta_0}}} = \frac{n_1}{n}$，$l(\hat{\beta_0}) = n_1 \log \frac{n_1}{n} + n_2 \log \frac{n_2}{n}$

# 思考题1

- 似然比检验：当$H_0$不成立时，即$\beta_1 \neq 0$
- $p_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$，对数似然函数为：

$$
\begin{aligned}
l_1 &= \sum_{i=1}^{n}[y_i \log \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} + (1-y_i)\log \frac{1}{1+e^{\beta_0+\beta_1 x_i}}] \\
&= \sum_{i=1}^{n_1} \log \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} + \sum_{j=1}^{n_2} \log \frac{1}{1+e^{\beta_0+\beta_1 x_j}} \\
&= \beta_0 n_1 + \beta_1 \sum_{i=1}^{n_1} x_i - \sum_{i=1}^{n} \log(1+e^{\beta_0+\beta_1 x_i})
\end{aligned}
$$

- 关于$\beta_0$求偏导：

$$
\frac{\partial l_1}{\partial \beta_0} = n_1 - \sum_{i=1}^{n} \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} = 0
$$

$$
\frac{\partial l_1}{\partial \beta_1} = \sum_{i=1}^{n_1} x_i - \sum_{i=1}^{n} \frac{x_i e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} = 0
$$

可以得到$\sum_{i=1}^{n} \hat{p}_i = n_1$，$\sum_{i=1}^{n_1} x_i = \sum_{i=1}^{n} \hat{p}_i x_i$

# 思考题1

可以得到$\sum_{i=1}^{n}\hat{p}_i = n_1$，$\sum_{i=1}^{n_1} x_i = \sum_{i=1}^{n}\hat{p}_i x_i$

- 由此可以得到deviance为：

$$D = 2[l_1 - l_0] = 2[\sum_{i=1}^{n_1}\log\hat{p}_i + \sum_{j=1}^{n_2}\log(1-\hat{p}_j) - (n_1\log\frac{n_1}{n} + n_2\log\frac{n_2}{n})]$$

当$H_0$成立时，$D \sim \chi^2(1)$

# 思考题2,3

- 思考题：采用向前法、向后法和逐步回归法对逻辑回归进行变量选择
- 思考题：加入随机生成的四个随机变量，两个离散变量$X_5$、$X_6$，两个连续变量$X_7$、$X_8$，采用向前法、向后法和逐步回归法对逻辑回归进行变量选择

# 思考题2

多元线性回归模型：$Y = X\beta + \varepsilon$

原假设：$H_0 : \beta = 0$

$$W = \sqrt{n}(\hat{\beta} - 0)^T (Cov(\sqrt{n}\hat{\beta}))^{-1} \sqrt{n}(\hat{\beta} - 0)$$

$$F = \frac{\sum_{i=1}^{k}(\hat{y}_i - \bar{y})^2 / k}{\sum_{i=1}^{k}(y_i - \hat{y}_i)^2 / (n-k-1)}$$

判断F统计量与W的关系

当$Cov(\sqrt{n}\hat{\beta})$已知时，W渐近服从$\chi^2(k)$

当$Cov(\sqrt{n}\hat{\beta})$未知时，构造F统计量：

$$F = \sqrt{n}(\hat{\beta} - 0)^T (\hat{Cov}(\sqrt{n}\hat{\beta}))^{-1} \sqrt{n}(\hat{\beta} - 0) \frac{n-k-1}{k}$$

$$
\begin{aligned}
F &= \sqrt{n}(\hat{\beta}-0)^T(\hat{Cov}(\sqrt{n}\hat{\beta}))^{-1}\sqrt{n}(\hat{\beta}-0)\frac{n-k-1}{k} \\
&= \frac{\hat{\beta}^T(X^TX)\hat{\beta}/k}{\hat{\sigma}^T\hat{\sigma}/(n-k-1)} \\
&= \frac{Y^TX(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^TY/k}{\hat{\sigma}^T\hat{\sigma}/(n-k-1)} \\
&= \frac{Y^TX(X^TX)^{-1}X^TY/k}{\hat{\sigma}^T\hat{\sigma}/(n-k-1)} \\
&= \frac{Y^THY/k}{\hat{\sigma}^T\hat{\sigma}/(n-k-1)} \\
&= \frac{\sum_{i=1}^k \hat{y}_i^2/k}{\sum_{i=1}^k (y_i-\hat{y}_i)^2/(n-k-1)}
\end{aligned}
$$

对于$Y = \beta_0 + X_1\beta_1 + \cdots + X_k\beta_k$

令$A = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$ $U = \begin{bmatrix} A & X \end{bmatrix}$ 其中X为$n \times p$维矩阵

$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$ $\gamma = \begin{bmatrix} \beta_0 \\ \beta \end{bmatrix}$

则$Y = U\gamma$，令$M = [0, I_p]$，其中$I_p$为p维单位阵。

此时原假设为：$H_0 : \beta = 0$

构造F统计量：

$$F = \frac{(\hat{\gamma}-\gamma_0)^T M^T [M(U^T U)^{-1} M^T]^{-1} M(\hat{\gamma}-\gamma_0)/k}{\hat{\sigma}^T \hat{\sigma}/(n-k-1)}$$

其中

$$[M(U^T U)^{-1} M^T]^{-1} = X^T [I_n - \tfrac{1}{n} J] X$$
$$\hat{\beta} = M(\hat{\gamma} - \gamma_0) = (X^T [I_n - \tfrac{1}{n} J] X)^{-1} X^T [I_n - \tfrac{1}{n} J] Y$$

此时F统计量为：

$$F = \frac{Y^T [I_n - \frac{1}{n} J] X (X^T [I_n - \frac{1}{n} J] X)^{-1} X^T [I_n - \frac{1}{n} J] Y/k}{\hat{\sigma}^T \hat{\sigma}/(n-k-1)}.$$

# 思考题1

$X_1$、$X_2$是0-1变量，$Y$是连续随机变量，对$Y$与$X_1$、$X_2$建立线性模型，考虑$X_1$与$X_2$之间存在交互作用，求$\beta_{12}$的表达式，并解释$\beta_{12}$等于0所表示的意义。

| $X_1/X_2$ | 0 | 1 |
|---|---|---|
| 0 | $n_{00}$ | $n_{01}$ |
| 1 | $n_{10}$ | $n_{11}$ |

# 思考题1

线性模型无交互作用情况

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

$\epsilon \sim N(0, 1)$，似然函数：

$$\begin{aligned} L(\beta; y) &\propto \sum_{i=1}^{n} -\frac{(Y_i - EY_i)^2}{2} \\ &= -\sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i})^2}{2} \end{aligned}$$

似然函数求偏导：

$$\sum_{i=1}^{n} y_i - n\beta_0 - (n_{10}+n_{11})\beta_1 - (n_{01}+n_{11})\beta_2 = 0$$

$$\sum_{X_{1i}=1} y_i - (n_{10}+n_{11})\beta_0 - (n_{10}+n_{11})\beta_1 - n_{11}\beta_2 = 0$$

$$\sum_{X_{2i}=1} y_i - (n_{01}+n_{11})\beta_0 - n_{11}\beta_1 - (n_{01}+n_{11})\beta_2 = 0$$

# 思考题1

最小二乘方法：

$$
\begin{aligned}
\hat{\beta} &= (X^T X)^{-1} X^T Y \\
&= \begin{bmatrix} n & n_{10}+n_{11} & n_{01}+n_{11} \\ n_{10}+n_{11} & n_{10}+n_{11} & n_{11} \\ n_{01}+n_{11} & n_{11} & n_{01}+n_{11} \end{bmatrix}^{-1} \begin{bmatrix} \sum_i y_i \\ \sum_{x_1=1} y_i \\ \sum_{x_2=1} y_i \end{bmatrix}
\end{aligned}
$$

# 思考题1

线性模型有交互作用情况

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \epsilon$$

$\epsilon \sim N(0,1)$，似然函数：

$$L(\beta; y) \propto -\sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} - \beta_{12} X_{1i} X_{2i})^2}{2}$$

# 思考题1

似然函数求偏导：

$$\sum_{i=1}^{n} y_i - n\beta_0 - (n_{10}+n_{11})\beta_1 - (n_{01}+n_{11})\beta_2 - n_{11}\beta_{12} = 0$$

$$\sum_{X_{1i}=1} y_i - (n_{10}+n_{11})\beta_0 - (n_{10}+n_{11})\beta_1 - n_{11}\beta_2 - n_{11}\beta_{12} = 0$$

$$\sum_{X_{2i}=1} y_i - (n_{01}+n_{11})\beta_0 - n_{11}\beta_1 - (n_{01}+n_{11})\beta_2 - n_{11}\beta_{12} = 0$$

$$\sum_{X_{1i}=1, X_{2i}=1} y_i - n_{11}\beta_0 - n_{11}\beta_1 - n_{11}\beta_2 - n_{11}\beta_{12} = 0$$

极大似然估计：

$$\widehat{\beta}_0 = \frac{1}{n_{00}} \sum_{X_{1i}=0, X_{2i}=0} y_i$$

$$\widehat{\beta}_1 = \frac{1}{n_{10}} \sum_{X_{1i}=1, X_{2i}=0} y_i - \frac{1}{n_{00}} \sum_{X_{1i}=0, X_{2i}=0} y_i$$

$$\widehat{\beta}_2 = \frac{1}{n_{01}} \sum_{X_{1i}=0, X_{2i}=1} y_i - \frac{1}{n_{00}} \sum_{X_{1i}=0, X_{2i}=0} y_i$$

$$\widehat{\beta}_{12} = \frac{1}{n_{11}} \sum_{X_{1i}=1, X_{2i}=1} y_i - \frac{1}{n_{10}} \sum_{X_{1i}=1, X_{2i}=0} y_i - \frac{1}{n_{01}} \sum_{X_{1i}=0, X_{2i}=1} y_i + \frac{1}{n_{00}} \sum_{X_{1i}=0, X_{2i}=0} y_i$$

# 思考题1

当没有交互作用项时，$\widehat{\beta}_{12}$=0

$$
\begin{aligned}
&\frac{1}{n_{11}}\sum_{X_{1i}=1,X_{2i}=1} y_i-\frac{1}{n_{10}}\sum_{X_{1i}=1,X_{2i}=0} y_i\\
=&\frac{1}{n_{01}}\sum_{X_{1i}=0,X_{2i}=1} y_i-\frac{1}{n_{00}}\sum_{X_{1i}=0,X_{2i}=0} y_i\\
&\frac{1}{n_{11}}\sum_{X_{1i}=1,X_{2i}=1} y_i-\frac{1}{n_{01}}\sum_{X_{1i}=0,X_{2i}=1} y_i\\
=&\frac{1}{n_{10}}\sum_{X_{1i}=1,X_{2i}=0} y_i-\frac{1}{n_{00}}\sum_{X_{1i}=0,X_{2i}=0} y_i\\
&\frac{1}{n_{11}}\sum_{X_{1i}=1,X_{2i}=1} y_i+\frac{1}{n_{00}}\sum_{X_{1i}=0,X_{2i}=0} y_i\\
=&\frac{1}{n_{01}}\sum_{X_{1i}=0,X_{2i}=1} y_i+\frac{1}{n_{10}}\sum_{X_{1i}=1,X_{2i}=0} y_i
\end{aligned}
$$

# 思考题1

方差分析

$$y_{ijl}=u+a_i+b_j+(ab)_{ij}+\epsilon_{ijl}, i=0,1; j=0,1; l=1,...,n_{ij}.$$

假设$\sum a_i=0,\quad \sum b_j=0,\quad \sum_i (ab)_{ij}=\sum_j (ab)_{ij}=0$。

假设$a_0=0,\quad b_0=0,\quad (ab)_{00}=(ab)_{01}=(ab)_{11}=0$，则是我们上面讨论的情况。

方差分析的估计：

$$\hat{u} = \overline{y}, \quad \hat{a}_i = \overline{y}_{i.} - \overline{y}, \quad \hat{b}_j = \overline{y}_{.j} - \overline{y},$$
$$\widehat{(ab)}_{ij} = \overline{y}_{ij} - \overline{y}_{i.} - \overline{y}_{.j} + \overline{y}$$

注意：无交互作用和有交互作用的估计值一致。

# 思考题2

$X_1$、$X_2$、$Y$ 均为二分类变量，令$Y_{ijl}$为$X_1 = i, X_2 = j$时Y的第l次观测,则$Y_{ijl} \sim B(1, p_{ij})$, 其中$i = 0, 1, j = 0, 1, l = 1, ...n_{ij}$.

令$Y_{ij}$为$X_1 = i, X_2 = j$时$Y = 1$的总次数，则$Y_{ij} \sim B(n_{ij}, p_{ij})$ 采用logistic模型，在无交互作用项时，有

$$logit(p_{ij}) = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2}$$

在有交互作用项时，有

$$logit(p_{ij}) = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_{12} x_{ij1} \cdot x_{ij2}$$

比较两个模型系数间的差异，及交互作用项的含义。

# 思考题2

由于$Y_{ij} \sim B(n_{ij}, p_{ij})$，采用极大似然估计，对数似然函数为：

$$L \propto \sum_i \sum_j y_{ij} ln p_{ij} + (n_{ij} - y_{ij}) ln(1 - p_{ij})$$

$$\frac{\partial l}{\partial p_{ij}} = \frac{y_{ij}}{p_{ij}} - \frac{n_{ij} - y_{ij}}{1 - p_{ij}} = 0$$

则$p_{ij}$ 的MLE为

$$\hat{p}_{ij} = \frac{y_{ij}}{n_{ij}}$$

# 思考题2

在无交互作用项时，$logit(p_{ij}) = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2}$

$$p_{ij} = \frac{exp(\beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2})}{1 + exp(\beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2})}$$

则有：

$$\begin{cases} \frac{y_{00}}{n_{00}} = \frac{exp(\beta_0)}{1+exp(\beta_0)} \\ \frac{y_{01}}{n_{01}} = \frac{exp(\beta_0+\beta_2)}{1+exp(\beta_0+\beta_2)} \\ \frac{y_{10}}{n_{10}} = \frac{exp(\beta_0+\beta_1)}{1+exp(\beta_0+\beta_1)} \\ \frac{y_{11}}{n_{11}} = \frac{exp(\beta_0+\beta_1+\beta_2)}{1+exp(\beta_0+\beta_1+\beta_2)} \end{cases}$$

化简得:

$$exp(\beta_0) = \frac{y_{00}}{n_{00} - y_{00}} \quad exp(\beta_0 + \beta_2) = \frac{y_{01}}{n_{01} - y_{01}}$$

$$exp(\beta_0 + \beta_1) = \frac{y_{10}}{n_{10} - y_{10}} \quad exp(\beta_0 + \beta_1 + \beta_2) = \frac{y_{11}}{n_{11} - y_{11}}$$

解得:

$$\begin{cases} exp(\beta_0) = \dfrac{y_{00}}{n_{00} - y_{00}} \\ exp(\beta_1) = \dfrac{y_{11}(n_{01} - y_{01})}{(n_{11} - y_{11})y_{01}} = \dfrac{y_{10}(n_{00} - y_{00})}{(n_{10} - y_{10})y_{00}} \\ exp(\beta_2) = \dfrac{y_{11}(n_{10} - y_{10})}{(n_{11} - y_{11})y_{10}} = \dfrac{y_{01}(n_{00} - y_{00})}{(n_{01} - y_{01})y_{00}} \end{cases}$$

## 思考题2

在无交互作用项时：

$$\begin{cases} exp(\beta_0) = \dfrac{p_{00}}{1-p_{00}} \\ exp(\beta_1) = \dfrac{p_{11}/(1-p_{11})}{p_{01}/(1-p_{01})} & = \dfrac{p_{10}/(1-p_{10})}{p_{00}/(1-p_{00})} \\ exp(\beta_2) = \dfrac{p_{11}/(1-p_{11})}{p_{10}/(1-p_{10})} & = \dfrac{p_{01}/(1-p_{01})}{p_{00}/(1-p_{00})} \end{cases}$$

# 思考题2

在有交互作用项时,

$$logit(p_{ij}) = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_{12} x_{ij1} \cdot x_{ij2}$$

$$p_{ij} = \frac{exp(\beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_{12} x_{ij1} \cdot x_{ij2})}{1 + exp(\beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_{12} x_{ij1} \cdot x_{ij2})}$$

则有:

$$\begin{cases} \frac{y_{00}}{n_{00}} = \frac{exp(\beta_0)}{1 + exp(\beta_0)} \\ \frac{y_{01}}{n_{01}} = \frac{exp(\beta_0 + \beta_2)}{1 + exp(\beta_0 + \beta_2)} \\ \frac{y_{10}}{n_{10}} = \frac{exp(\beta_0 + \beta_1)}{1 + exp(\beta_0 + \beta_1)} \\ \frac{y_{11}}{n_{11}} = \frac{exp(\beta_0 + \beta_1 + \beta_2 + \beta_{12})}{1 + exp(\beta_0 + \beta_1 + \beta_2 + \beta_{12})} \end{cases}$$

# 思考题2

化简得：

$$exp(\beta_0) = \frac{y_{00}}{n_{00} - y_{00}} \quad exp(\beta_0 + \beta_2) = \frac{y_{01}}{n_{01} - y_{01}}$$

$$exp(\beta_0 + \beta_1) = \frac{y_{10}}{n_{10} - y_{10}} \quad exp(\beta_0 + \beta_1 + \beta_2 + \beta_{12}) = \frac{y_{11}}{n_{11} - y_{11}}$$

解得：

$$\begin{cases} exp(\beta_0) = \dfrac{y_{00}}{n_{00} - y_{00}} \\ exp(\beta_1) = \dfrac{y_{10}(n_{00} - y_{00})}{(n_{10} - y_{10})y_{00}} \\ exp(\beta_2) = \dfrac{y_{01}(n_{00} - y_{00})}{(n_{01} - y_{01})y_{00}} \\ exp(\beta_{12}) = \dfrac{y_{11}(n_{10} - y_{10})(n_{01} - y_{01})y_{00}}{(n_{11} - y_{11})y_{10}y_{01}(n_{00} - y_{00})} \end{cases}$$

# 思考题2

在有交互作用项时：

$$
\begin{cases}
exp(\beta_0) = \dfrac{p_{00}}{1-p_{00}} \\
exp(\beta_1) = \dfrac{p_{10}/(1-p_{10})}{p_{00}/(1-p_{00})} \\
exp(\beta_2) = \dfrac{p_{01}/(1-p_{01})}{p_{00}/(1-p_{00})} \\
exp(\beta_{12}) = \dfrac{(p_{11}/(1-p_{11}))\times(p_{00}/(1-p_{00}))}{(p_{10}/(1-p_{10}))\times(p_{01}/(1-p_{01}))}
\end{cases}
$$

若无交互作用，则有$exp(\beta_{12}) = 1$,即：

$$(p_{11}/(1-p_{11}))\times(p_{00}/(1-p_{00})) = (p_{10}/(1-p_{10}))\times(p_{01}/(1-p_{01}))$$

# 比值比

$$exp(\beta_0) = \frac{p_{00}}{1-p_{00}} \quad exp(\beta_0+\beta_2) = \frac{p_{01}}{1-p_{01}}$$

$$exp(\beta_0+\beta_1) = \frac{p_{10}}{1-p_{10}} \quad exp(\beta_0+\beta_1+\beta_2+\beta_{12}) = \frac{p_{11}}{1-p_{11}}$$

则有：

$$exp(\beta_0) = OR_{00} = 1 \quad exp(\beta_2) = OR_{01}$$

$$exp(\beta_1) = OR_{10} \quad exp(\beta_{12}) = \frac{OR_{11}}{OR_{10} \times OR_{01}}$$

# 交互效应

$$logit(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 \cdot x_2$$

有：

$$logit(p) = \beta_0 + (\beta_1 + \beta_{12} x_2) x_1 + \beta_2 x_2$$

或：

$$logit(p) = \beta_0 + \beta_1 x_1 + (\beta_2 + \beta_{12} x_1) x_2$$

即：$X_1$对Y的作用被$X_2$影响或$X_2$对Y的作用被$X_1$影响，产生交互作用