应用回归分析

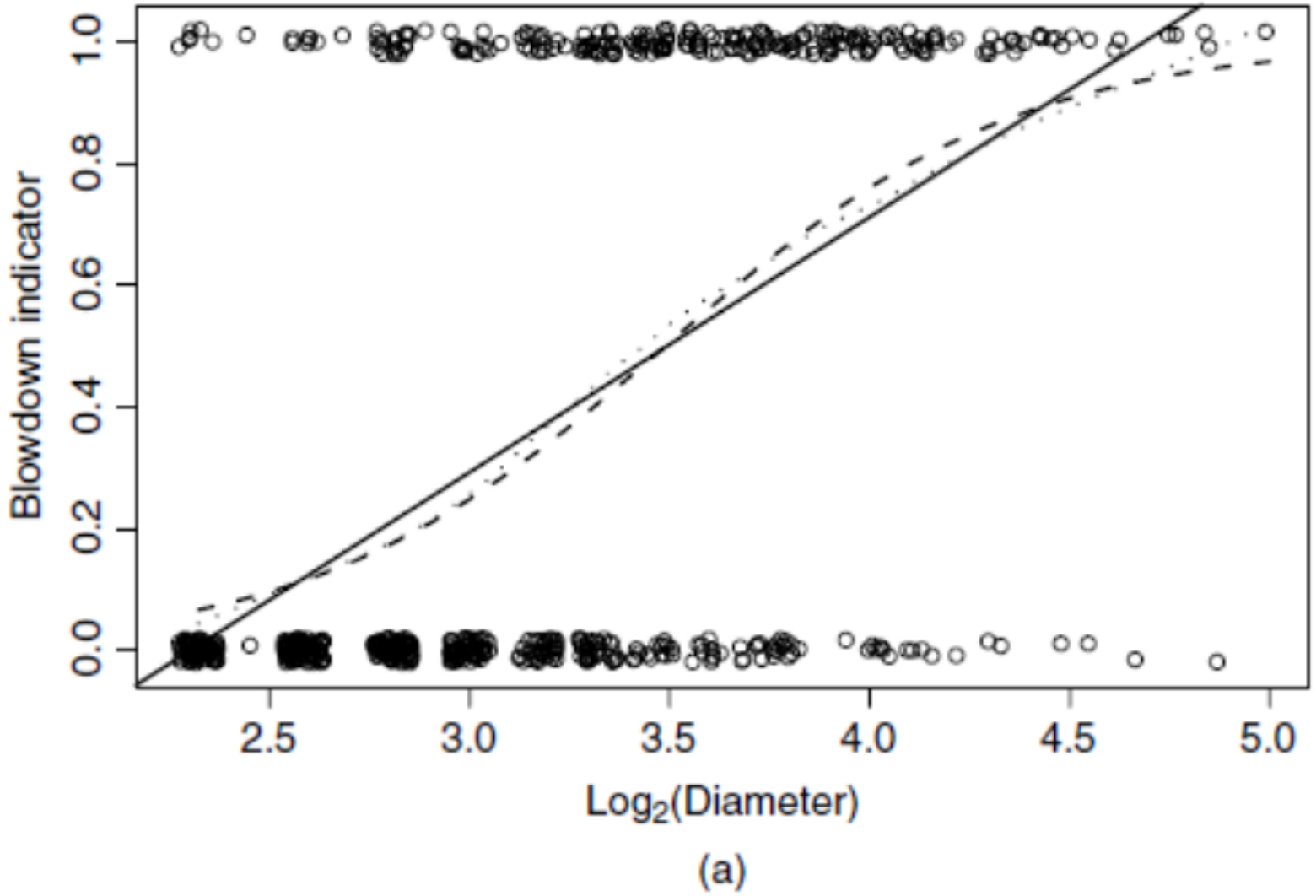
上海财经大学　统计与管理学院

第十二章逻辑回归

章节概括：

- 实例
- 逻辑回归
- 联系函数
- 广义线性模型

blowBF数据


Blowdown indicator
0.0
0.2
0.4
0.6
0.8
1.0
2.5
3.0
3.5
4.0
4.5
5.0
Log2(Diameter)
(a)

blowBF数据

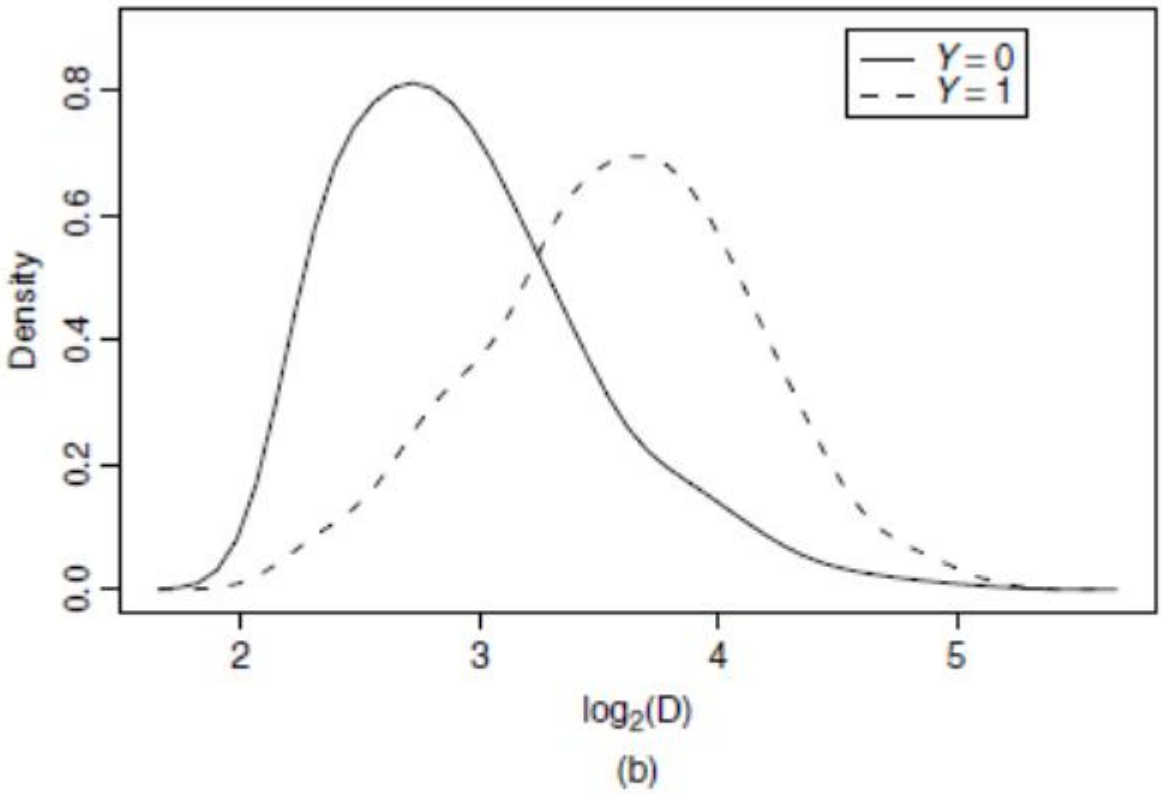


FIG. 12.1 Blowdown data for Balsam Fir. (a) Scatterplot of Y versus $\log(D)$. The solid line is the OLS line. The dotted line is the fit of a smoothing spline. The dashed line is the logistic regression fit. Data have been jittered in both variables to minimize overprinting. (b) Separate density estimates for $\log(D)$ for survivors, $Y = 0$, and blowdown, $Y = 1$.

blowBF数据

- Y = 0 / 1, 成功/ 失败，有限制，
- Y = 0 / 1 的密度函数估计

  成正态分布，有重合， Y=1相对偏右偏低

  从左到右，Y=1的概率增加，

  Log(D) = 3.3 分界线
- $\Pr(Y = 1|X = x)$ 条件概率替代均值函数
- $\theta(\log(D))$ 拟合函数

二项回归

- 二项分布

$$y \sim \text{Bin}(m, \theta)$$

$$\Pr(y = j) = \binom{m}{j} \theta^j (1-\theta)^{(m-j)}$$

$$\text{E}(y) = m\theta; \quad \text{Var}(y) = m\theta(1-\theta)$$

- 二项回归

$$(Y|X = \mathbf{x}_i) \sim \text{Bin}(m_i, \theta(\mathbf{x}_i)), i = 1, \ldots, n$$

$$\text{E}(y_i/m_i|\mathbf{x}_i) = \theta(\mathbf{x}_i)$$

$$\text{Var}(y_i/m_i|\mathbf{x}_i) = \theta(\mathbf{x}_i)(1-\theta(\mathbf{x}_i))/m_i$$

均值函数

- $\theta(\mathbf{x}_i)$ 决定均值和方差函数
- 线性： $\boldsymbol{\beta}'\mathbf{x}$

$$\theta(\mathbf{x}_i) = \mathrm{m}(\boldsymbol{\beta}'\mathbf{x}_i)$$

- 逻辑函数

$$\theta(\mathbf{x}_i) = \mathrm{m}(\boldsymbol{\beta}'\mathbf{x}_i) = \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}{1+\exp(\boldsymbol{\beta}'\mathbf{x}_i)} = \frac{1}{1+\exp(-\boldsymbol{\beta}'\mathbf{x}_i)}$$

逻辑函数

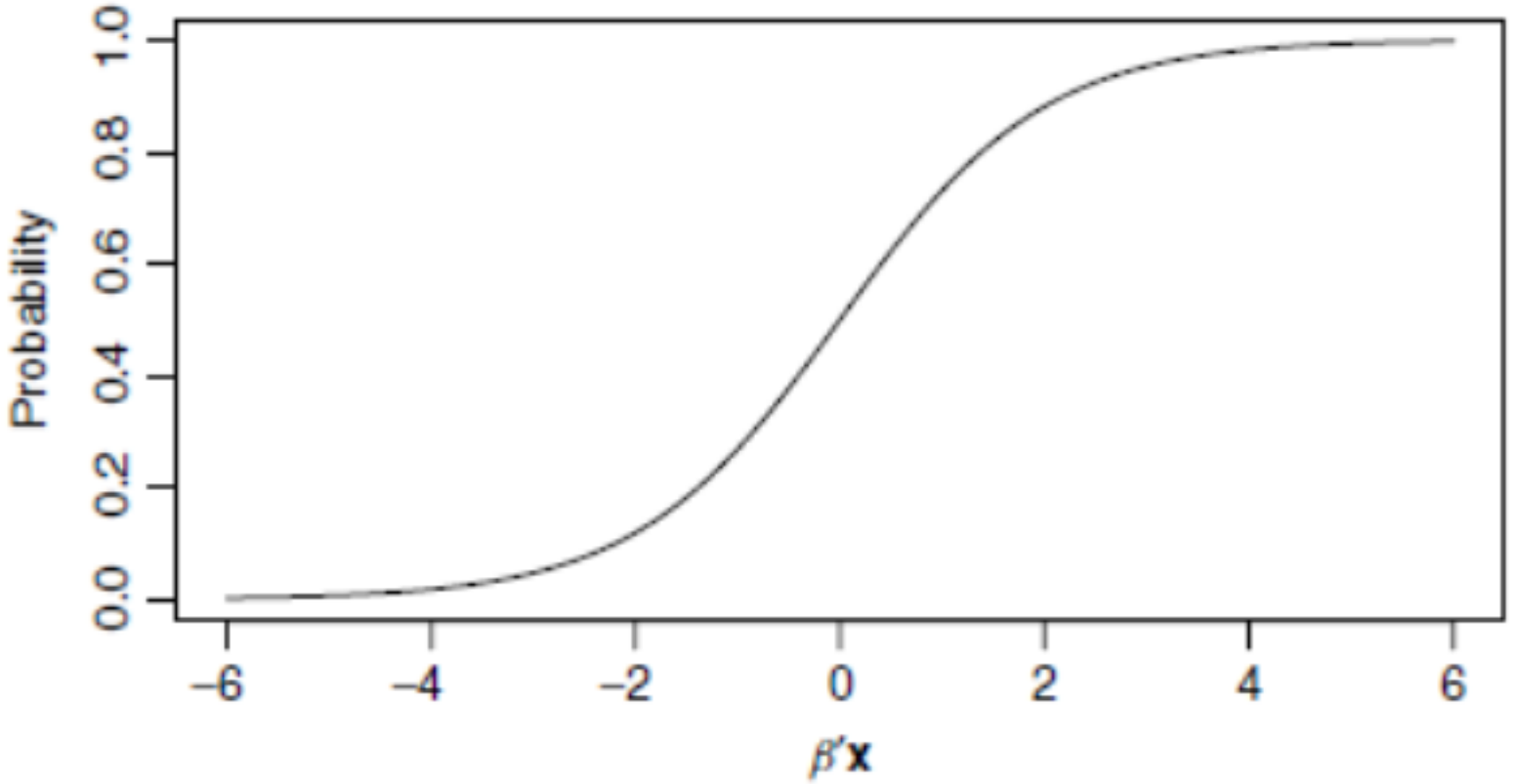


FIG. 12.2 The logistic kernel mean function.

联系函数

- link function, logit 函数

$$\log\left(\frac{\theta(\mathbf{x})}{1-\theta(\mathbf{x})}\right)=\boldsymbol{\beta}'\mathbf{x}$$

- 成功几率，odds of success

$$\theta(\mathbf{x})/(1-\theta(\mathbf{x}))$$

逻辑回归估计

- 12.3.2 极大似然估计，迭代法
- 系数的解释
- z-检验，非t-检验
- 偏差Deviance / 皮尔逊Pearson's X^2

blowBF数据

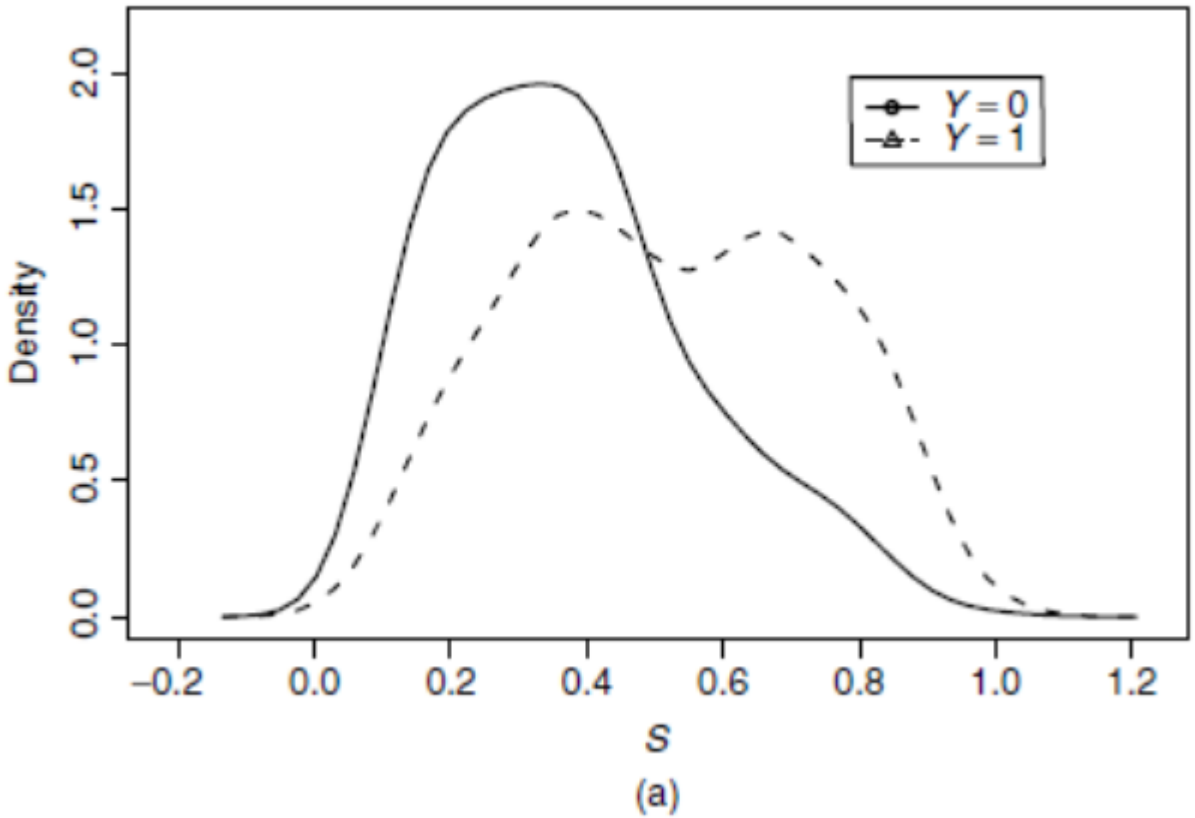
TABLE 12.1 Logistic Regression Summary for the Balsam Fir Blowdown Data

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -7.8923     0.6301  -12.53   <2e-16
logD           2.2626     0.1907   11.86   <2e-16
---
Residual deviance: 655.24  on 657  degrees of freedom
    Pearson's X^2: 677.44  on 657  degrees of freedom
```

$$\hat{E}(Y|\log(D)) = \frac{1}{1+\exp[-(-7.8923+2.2626\log(D))]}$$

blowBF数据

- 增加一个自变量



blowBF数据

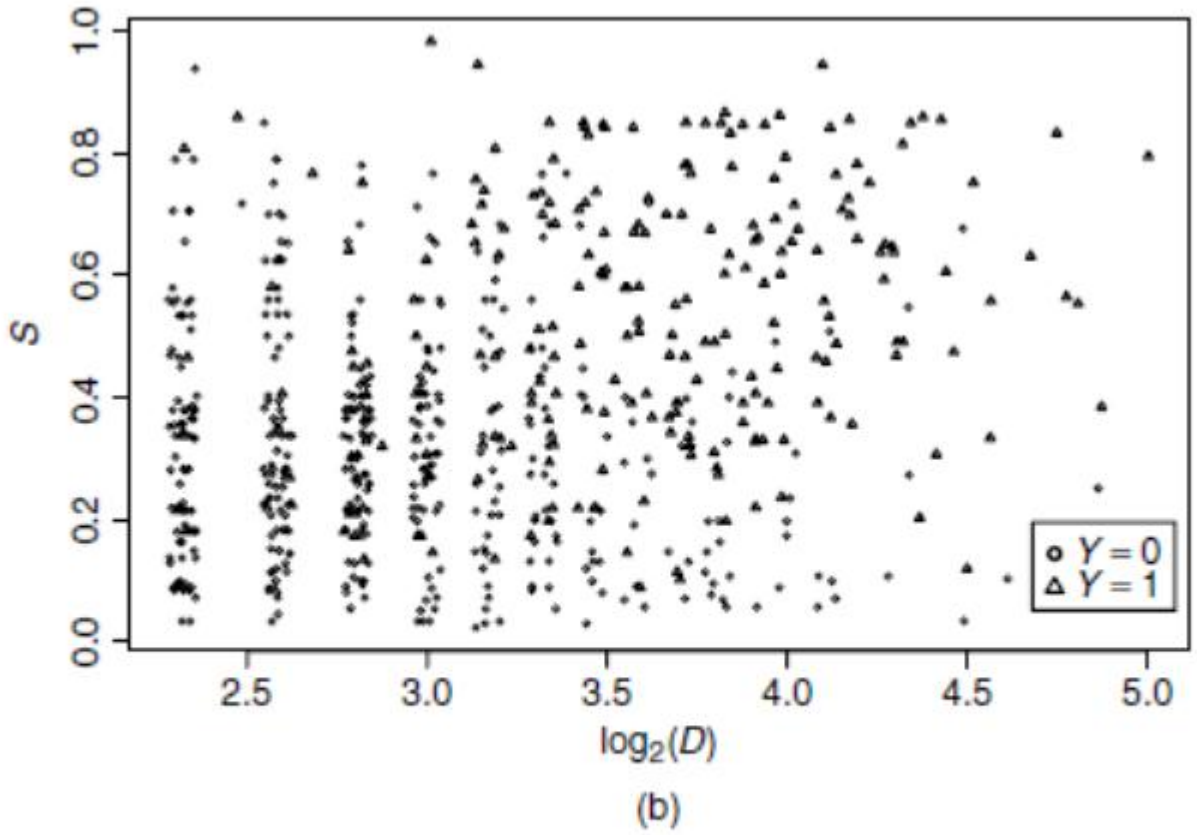
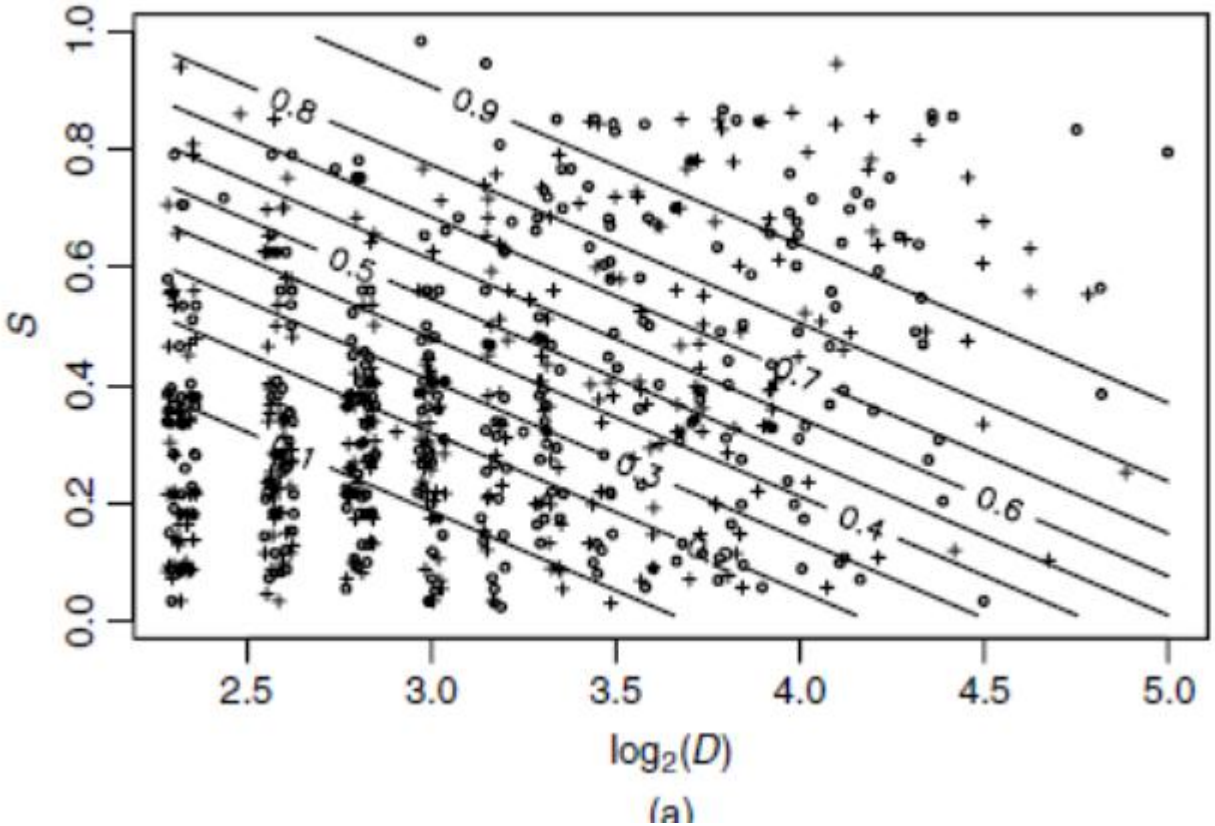


FIG. 12.3 (a) Density estimates for S for survivors, $Y = 0$ and blowdown, $Y = 1$, for the Balsam Fir data. (b) Plot of S versus $\log(D)$ with separate symbols for points with $Y = 1$ and $Y = 0$. The values of $\log(D)$ have been jittered.

无交叉项



无交叉项

TABLE 12.2 Logistic Regressions for the Balsam Fir Data

(a) No interaction

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-9.5621	0.7499	-12.75	<2e-16	***
logD	2.2164	0.2079	10.66	<2e-16	***
S	4.5086	0.5159	8.74	<2e-16	***

Residual deviance: 563.9 on 656 degrees of freedom

Pearson's X^2: 715.3 on 656 degrees of freedom

有交叉项

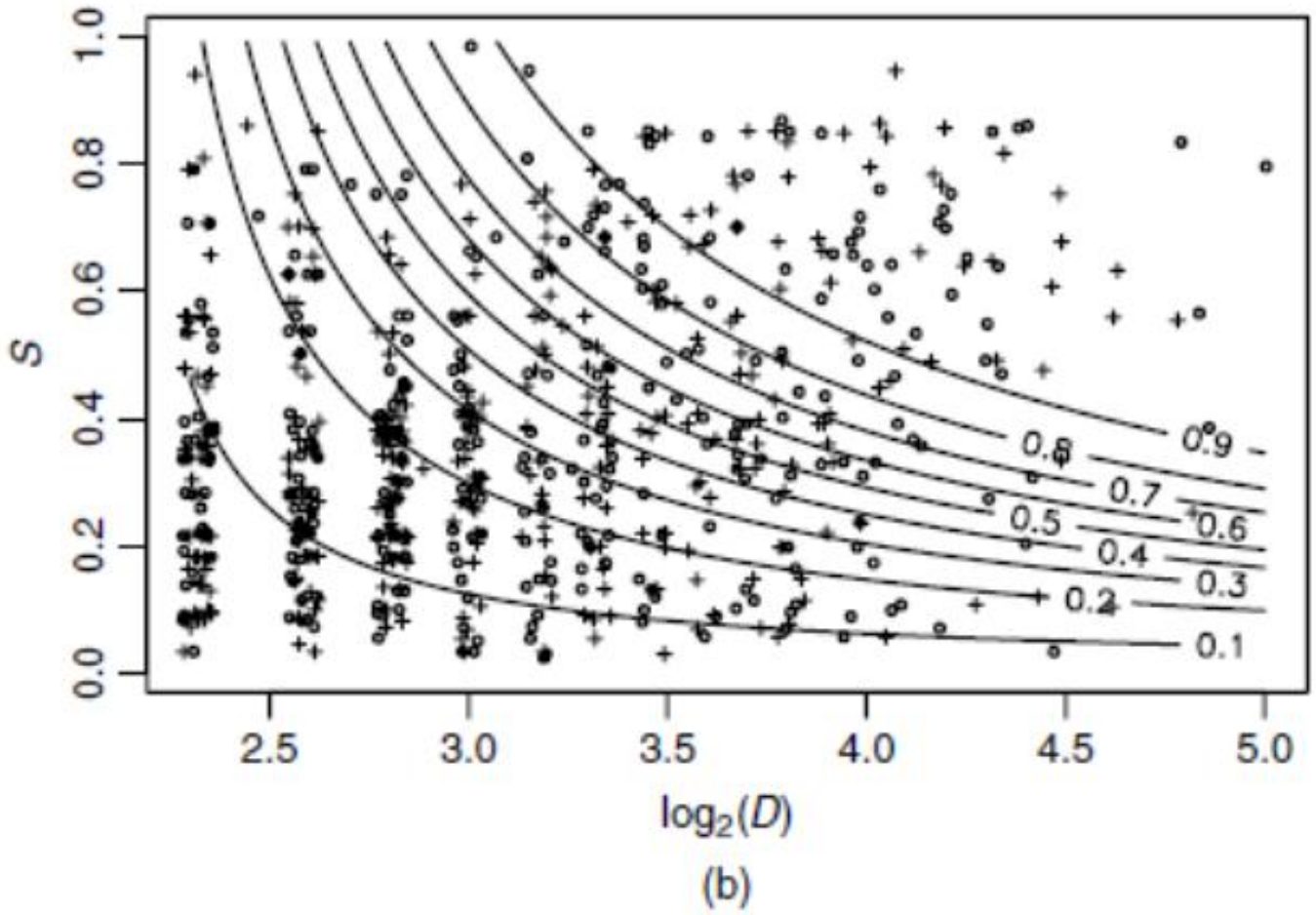


FIG. 12.4 Scatterplots with contours of estimated probability of blowdown. (a) No interaction mean function. (b) Mean function with interaction.

有交叉项

```
(b) Mean function with interaction
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.6788     1.4209  -2.589  0.00963
logD           0.4009     0.4374   0.916  0.35941
S            -11.2026     3.6143  -3.100  0.00194
logD:S         4.9098     1.1319   4.338 1.44e-05
---
Residual deviance: 541.75  on 655  degrees of freedom
    Pearson's X^2: 885.44  on 655  degrees of freedom
```

系数解释

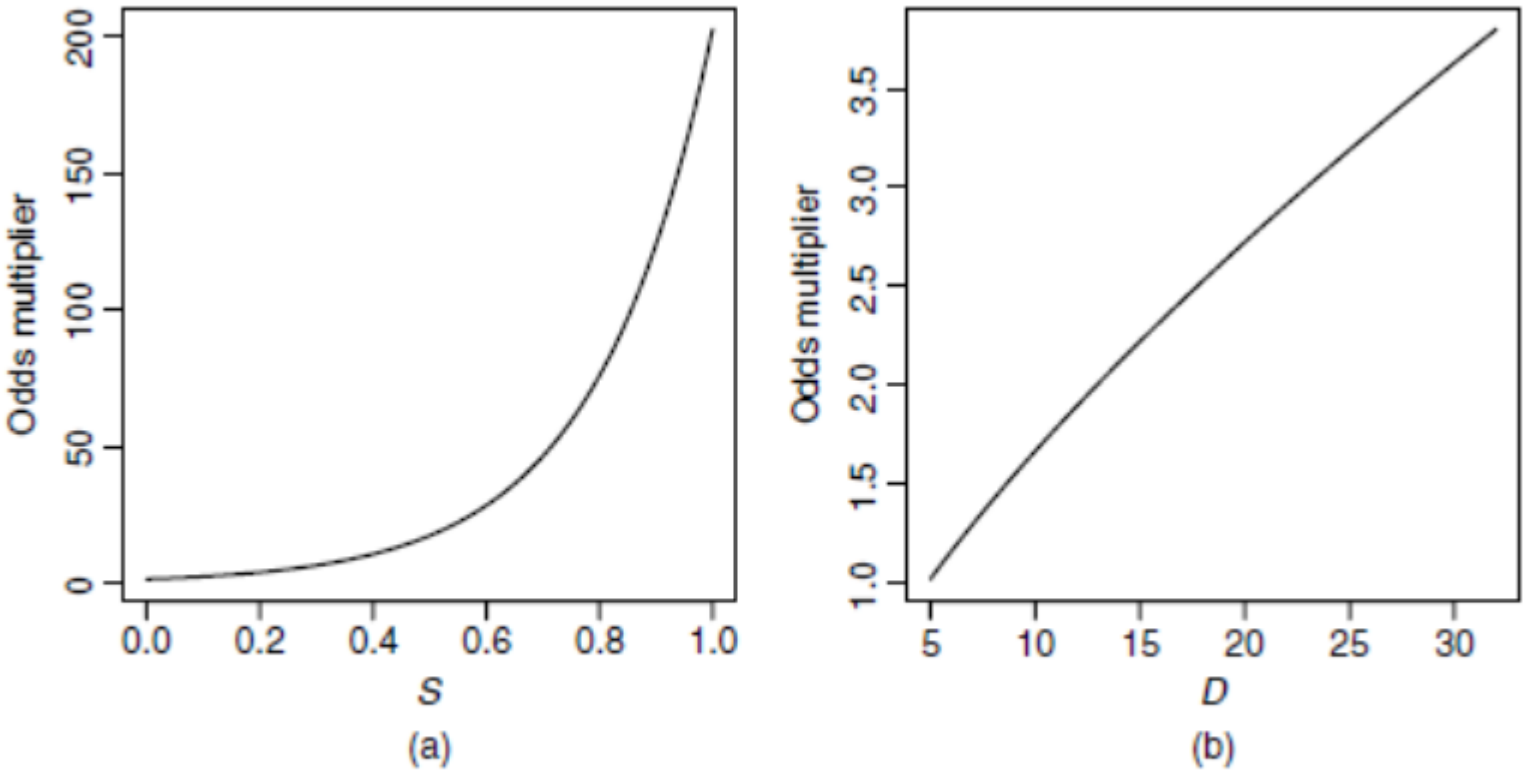


FIG. 12.5 Blowdown odds multiplier for (a) doubling the diameter of a Balsam Fir tree as a function of local severity, S, and (b) increasing S by 0.1 as a function of diameter.

Deviance偏差

- 残差平方和RSS 由偏差Deviance代替

$$G^2 = 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i - \hat{y}_i}\right)\right]$$

$$\hat{y}_i = m_i\hat{\theta}(\mathbf{x}_i)$$

模型比较

- 假设检验

$$\begin{aligned} \text{NH:} \quad & \theta(\mathbf{x}) = \mathrm{m}(\boldsymbol{\beta}_1' \mathbf{x}_1) \\ \text{AH:} \quad & \theta(\mathbf{x}) = \mathrm{m}(\boldsymbol{\beta}_1' \mathbf{x}_1 + \boldsymbol{\beta}_2' \mathbf{x}_2) \end{aligned}$$

- 检验统计量 $G_{NH}^2 - G_{AH}^2$

- 原假设成立时服从卡方分布
- 自由度 $\mathrm{df} = df_{\mathrm{NH}} - df_{\mathrm{AH}},$

blowBF数据

TABLE 12.3 Analysis of Deviance for Balsam Fir Blowdown Data

Terms	df	Deviance	Change in df	Deviance	$P(>\|Chi\|)$
1, log(D)	657	655.24			
1, log(D), S, $S \times$ log(D)	655	541.75	2	113.50	0.0000

TABLE 12.4 Sequential Analysis of Deviance for Balsam Fir Blowdown Data

Terms	df	Deviance	Change in df	Deviance	P(> |Chi|)
Intercept	658	856.21			
Add log(D)	657	655.24	1	200.97	0.0000
Add S	656	563.90	1	91.34	0.0000
Add interaction	655	541.75	1	22.16	0.0000

失拟检验

- 若 $m_i > 1$,且足够大
- 备择假设模型为饱和模型(saturated model)

G^2 在原NH成立时服从 χ^2_{n-p}

- Pearson's X^2 统计量

$$X^2 = \sum_{i=1}^{n} \left[(y_i - \hat{y}_i)^2 \left(\frac{1}{\hat{y}_i} + \frac{1}{m_i - \hat{y}_i} \right) \right]$$

$$= \sum_{i=1}^{n} \frac{m_i (y_i / m_i - \hat{\theta}(\mathbf{x}_i))^2}{\hat{\theta}(\mathbf{x}_i)(1 - \hat{\theta}(\mathbf{x}_i))}$$

Titanic数据

TABLE 12.5 Data from the Titanic Disaster of 1912. Each Cell Gives *Surv*/*M*, the Number of Survivors, and the Number of People in the Cell

	Female		Male	
Class	Adult	Child	Adult	Child
Crew	20/23	NA	192/862	NA
First	140/144	1/1	57/175	5/5
Second	80/93	13/13	14/168	11/11
Third	76/165	14/31	75/462	13/48

Titanic数据

TABLE 12.6 Fit of Four Mean Functions for the Titanic Data. Each of the Mean Functions Treats *Age*, *Sex*, and *Class* as Factors, and Fits Different Main Effects and Interactions

Mean Function	df	G^2	X^2
Main effects only	8	112.57	103.83
Main effects + *Class* × *Sex*	5	45.90	42.77
Main effects + *Class* × *Sex* + *Class* × *Age*	3	1.69	1.72
Main effects + all two-factor interactions	2	0.00	0.00
Main effects, two-factor and three-factor interactions	0	0.00	0.00

极大似然估计

- 似然函数

$$L(\theta) = \binom{m}{y} \theta^{y}(1-\theta)^{(m-y)}$$

- 对数似然函数

$$\log\left(L(\theta)\right) = \log\binom{m}{y} + y\log(\theta) + (m-y)\log(1-\theta)$$

极大似然估计

- 求导得零

$$\frac{d\log(L(\theta))}{d\theta} = \frac{y}{\theta} - \frac{m-y}{1-\theta} = 0$$

- 极大似然估计

$$\hat{\theta} = \frac{y}{m} = \frac{\text{Observed number of successes}}{\text{Observed fixed number of trials}}$$

极大似然估计

- 方差

$$\mathrm{Var}(\hat{\theta}) = -\left[\mathrm{E}\left(\frac{\partial^2 \log(L(\theta))}{\partial\theta(\partial\theta)'}\right)\right]^{-1}$$

$$\left[-\mathrm{E}\left(\frac{d^2 \log(L(\theta))}{d\theta^2}\right)\right]^{-1} = \left[-\mathrm{E}\left(\frac{y}{\theta^2} - \frac{m-y}{(1-\theta)^2}\right)\right]^{-1}$$

$$= \left[\frac{m}{\theta(1-\theta)}\right]^{-1}$$

$$= \frac{\theta(1-\theta)}{m}$$

逻辑回极大归似然估计

- 似然函数

$$L = \prod_{i=1}^{n} \binom{m_i}{y_i} (\theta(\mathbf{x}_i))^{y_i} (1 - \theta(\mathbf{x}_i))^{m_i - y_i}$$

$$\propto \prod_{i=1}^{n} (\theta(\mathbf{x}_i))^{y_i} (1 - \theta(\mathbf{x}_i))^{m_i - y_i}$$

- 对数似然函数

$$\log(L) \propto \sum_{i=1}^{n} \left[y_i \log\left(\frac{\theta(\mathbf{x}_i)}{1 - \theta(\mathbf{x}_i)} \right) + m_i \log(1 - \theta(\mathbf{x}_i)) \right]$$

极大似然估计

- 似然函数

$$\log(L(\boldsymbol{\beta})) = \sum_{i=1}^{n} \left[(\boldsymbol{\beta}' \mathbf{x}_i) y_i - m_i \log(1 + \exp(\boldsymbol{\beta}' \mathbf{x}_i)) \right]$$

- 迭代估计(Newton-Raphson算法)

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}' \hat{\mathbf{W}} \mathbf{X})^{-1}$$

$\hat{\mathbf{W}}$ 对角元素为 $m_i \hat{\theta}(\mathbf{x}_i)(1 - \hat{\theta}(\mathbf{x}_i))$

联系函数

联系函数的几种主要类型

联系函数类型	形式	应用场合
Logit	$\log(\gamma/(1-\gamma))$	各类别均匀分布
Complementary log-log	$\log(-\log(1-\gamma))$	高层类别出现几率大
Negative log-log	$-\log(-\log(\gamma))$	低层类别出现几率大
Probit	$\Phi^{-1}(\gamma)$	正态分布
Cauchit (inverse Cauchy)	$\tan(\pi(\gamma-0.5))$	两端的类别出现几率大

广义线性模型

- Generalized Linear Model (GLM)

- 条件分布服从指数组分布
  正态分布，二项分布，poisson分布，gamma分布…
- 均值函数

$$\mathrm{E}(Y|X=\mathbf{x})=\mathrm{m}(\boldsymbol{\beta}'\mathbf{x})$$

Thank You !