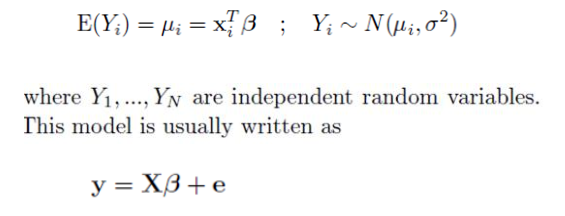
# 第一节课

建模的策略：

**1 设立模型 根据收集的数据 需要对数据有一定的认识**

正太线性模型 normal linear models



**2 估计未知的参数**

常用的方法有：最小二乘法Least squares 极大似然法Maximum likelihood

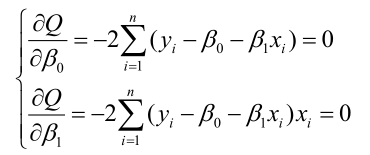


我设了一个Q

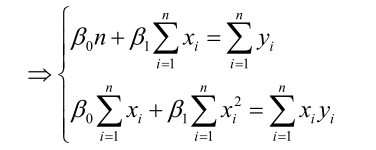


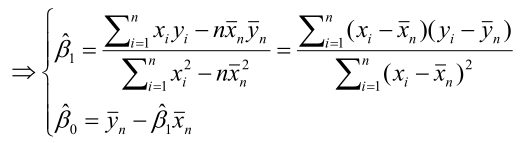
我让Q（残差平方和）最小就对了

为了让Q最小 对β0和β1求导



注意第二个式子求和符号外的xi不能直接消掉





**极大似然法**

为了使likelihood function最大



就得让最小

得出的结果和最小二乘法是一样的

区别：

最小二乘法不需要知道yi怎么分布

极大似然法需要知道

**3 模型诊断：检测模型的充分性**

1：检查残差是不是正态分布

残差标准化：残差除以标准差应该是标准正态分布的

标准正态分布的情况下，绝对值大于1.96的不超过5%，绝对值超过2.58的不超过1%

要用到Q-Q plot

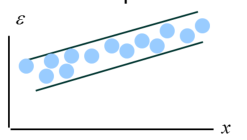
如果Y不是正态分布的 QQ图就不会是对角线

图上还会有个置信区间 如果大多数点在置信区间内 就是成立的

2：残差里有没有东西

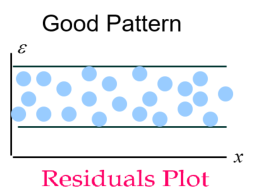
这个残差要plot against所有可能的解释变量 不应该呈现明显的pattern（比如二次函数）

如果这样



就说明x越大残差越大 所以不独立

下面的是好的



有没有其他可能的变量被我忽略了？

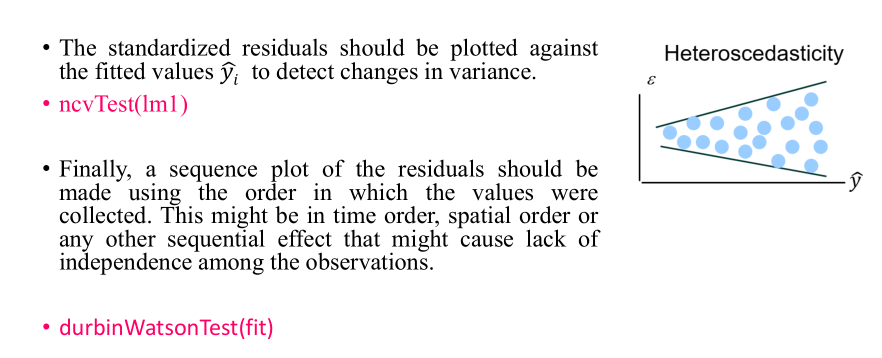
我应该把残差和其他可能的解释变量都画个图

3我们还假设了扰动项同方差 那么这个假设是真的吗？

把残差对y画图

方差描述的是变量波动的幅度

如果长下面这样



说明y越大波动（方差）越大

所以是异方差的

可以用ncvTest



P值小 代表存在异方差

4还有个假设是变量之间是独立的

这个也需要检验

如果两个变量是独立的，那就是不相关的

所以可以检验序列是否存在相关性

durbinWatson检验可以做到

独立同分布：每个数据对观测值的影响是差不多的 但是如果有个outlier 注意线性回归中不寻常的数据可能是一个信号：你没有用你的模型很好地描绘这堆数据的特点

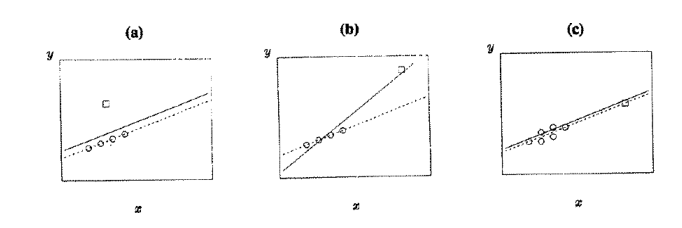
5 异常值检验

这样的值有什么特点呢？

A：加上它拟合是实线 不加上它是另外一条完美的虚线 对截距项产生影响 x的取值差不多 但是y的取值很不一样

B：x的取值和y的取值都非常不一样

C：x的取值差很多 但是对拟合线没啥影响



那我怎么知道一个数据对我的模型的影响？

比如对β和 对个别y（DFFITS 临界值如下）



一个值可能不知对一个y产生影响 可能对所有y都产生影响

此时我们用到Cook distance

，α=0.5

这个数如果特别大，就说明是真有影响

我想知道第i个数据对β的影响

DFBETA i =b - b(i)

如果没啥影响 那b（β的估计值）和bi的差距应该不是很大

所以我想知道DFBETA i是不是接近0

把DFBETA i标准化 临界值



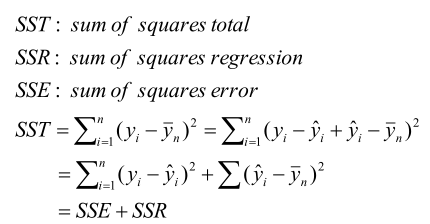
R方：模型能解释的数据的百分比

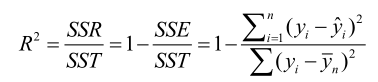
SST是全部平方和

SSE是残差平方和

SSR是我用x预测和不用的差异的平方和 也就是衡量x到底多重要

如果x能很好的解释y 那SSR应该比较大

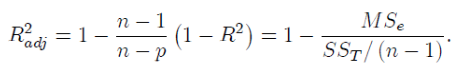




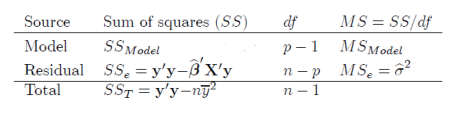


如何解决：R方会随着变量个数的增加而变大？我要让它和模型参数联系取来

调整R方



P是变量的个数



SSmodel就是SSR；df是自由度 SSR和SST独立 所以相加自由度是二者自由度相加

MS的M是取平均的意思 就是除以自由度

F检验



两个自由度：



如果这个统计量比较大 那说明β不同时为0

找一个上α分位数 （F分布就是有这个分位数）

原假设是β都是0

备择假设是β不是0

F就是原假设和备择假设的残差平方和的差异

就是有的时候 我还想检验有p个参数的模型中的q个是不是0



SSe1是全模型的残差平方和，自由度为n-p

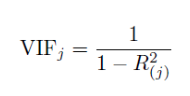
SSe2是删掉了一些参数的模型的残差平方和，自由度为n-p-q

这样的话 SSe2-SSe1就是一个自由度为q的卡方分布

共线性：

如果在一个多重线性的模型里 好几个解释变量是相关的 那这就叫共线性，或者多重共线性

我可以用一个variance inflation factor来检测共线性



可以对每个解释变量做这件事

R方是其他所有变量对第j个变量进行回归的R方

如果R方特别大（接近于1） 那VIF就会特别大 所以共线性越大 VIF越大

如果不存在线性关系，Rj就是等于0的 ，VIF=1 证明没有共线性

当VIF>5 那么就认为VIF很大 这时候可能存在多重共线性

**4 推断：**

就比如Var（β）这种

T检验

置信区间

## 第二节 方差分析

**方差分析 aov**

用来检验不同组的均值是否相等 ，这些组有一定的相关性

这里的所有解释变量都是分类变量 虚拟变量

比如检验不同专业的学生毕业后薪酬是否相同

H0：均值向等

对比T test

T test用来检验两组数据的均值是否相等



比如一组施肥 一组不施肥 看看植物的高度一不一样 就是T test的范畴

T=均值估计的差比标准差 看是不是标准正太分布

ANOVA：n组数据的均值是否相等



如果1组不施肥 2组施20%的肥…… 8组施100%的肥 看八组的植物高度 就是anova检验

Factor A 是植物的不同对待条件

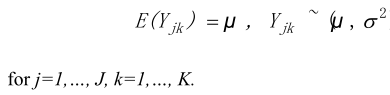
J levels是说植物有多少种不同的对待条件

在j这个条件下 Y1到Ynj有着同样的均值 却可能等于不同的数

就是说吧 这里的每个观测值 它都是一个以μ为中心的正太分布

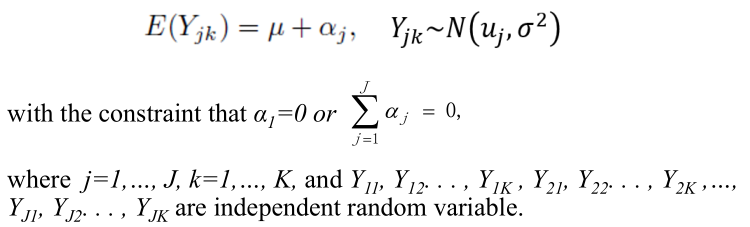
如果假设H0成立，那就意味着

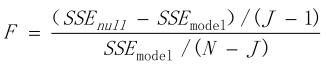
Null model：

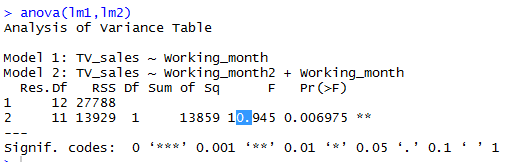


如果假设H1成立，那就意味着

Specified model：



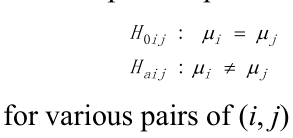


****

你就看最后这个显不显著 显著的话就说这个均值是不一样的

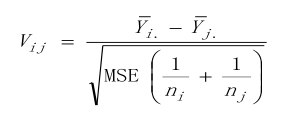
那我已经知道了 哦 不是说大伙的均值都一样 那我还想知道更多 到底是谁不一样呢

**多重检验**

****

有三种方法可以使用

第一种 LSD



MSE是mean squares error

认为μi≠μj

如果



但是逐对检验不是一个好方法 当对数特别大的时候你总是会拒绝原假设

第二种 Bonferroni correction

就是认为μi≠μj

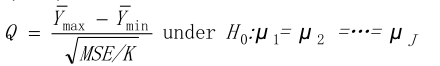
如果



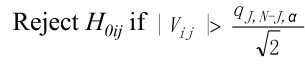
就是调整了拒绝的规则

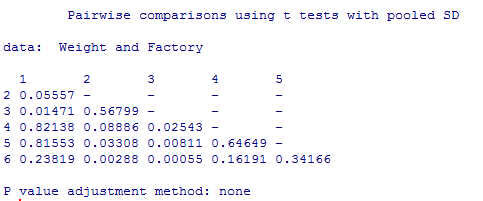
第三种 Tukey的检验

Bonferroni的方法在m特别大的时候不好用



和这个分布的1-α分点比





前面的都是只有一种factor

如果有两个factor

Factor A有三种情况 Factor B有两种情况

一共就是六种情况 六个subgroup

每个subgroup的expected value是一样的

这个时候可能出现四种情况：

Factor A有影响 Factor B也有影响 没有相互影响

对应的模型是：HI 没有交互作用



Factor A有影响 Factor B也有影响 有相互影响

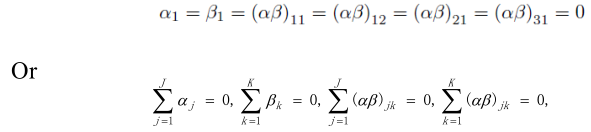
如果有相互影响 那模型就会变为 Saturated model



这里有1+j+k+jk个参数

但是缺只有jk个expected value

这时候我们增加限制条件



Factor A有影响 Factor B没有影响 HB



Factor A没有影响 Factor B有影响 HA



**Analysis of covariance**

**协方差分析**

模型中的一些解释变量是用来代表factor level的哑变量

另外一些是被成为协变量的持续变量

就是比如说我测试三种训练方法对足球队的成绩的提高的影响

这里有一个自变量是足球队原来的水平

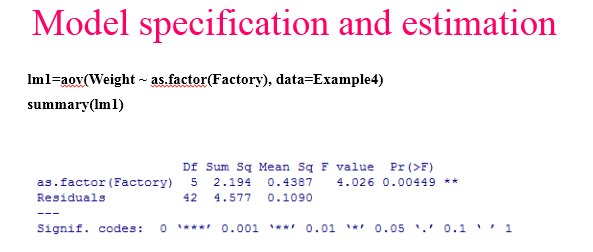
那协方差分析 就得把原来的水平进行调整

全模型：



Reduced model：





## 第三节 线性回归和方差分析上机

以线性分析为例的上机操作

1 看看数据怎么样

Summary（data1）

Cor(data1)

Plot(y~x,data=data1)

2 建立回归模型

3 检测回归模型

Summary（lm1）

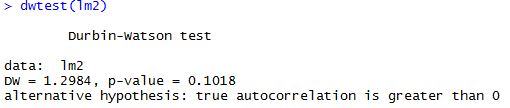
Plot（standresid~x，data=data1）

Plot（standresid~y，data=data1）

NcvTest检测异方差

DurbinWatson检验残差序列是否存在自相关性





检验残差是否为正态分布QQ-plot

检验是否存在影响大的异常值

Defbetas

Dffits

Cooks

用VIF检验多重共线性

## 第四章 0-1回归

* 1. 回 归Binary Regression 我关心的被解释变量只有两种可能的结果

指数分布族Exponential Family

我们平时接触的正态分布 是连续的 对称的 y的期望和解释变量是线性关系

但是有时候我们要解释的变量是离散的 不对称的，他们和解释变量的关系也不一定是线性的

如果被解释变量的密度函数如下



变形为



我们就称y的分布在指数分布族里

如果a（y）=y 我们管它叫经典分布 canonical

B（cita）是自然参数 natural parameter

其他的参数我们不感兴趣 叫冗余参数nuisance parameters

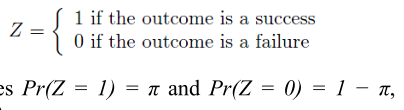
正太分布可以写成指数分布的形式





**伯努利分布**

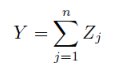
**Bernoulli**



**二项分布**

Binomial distribution

连续的几个伯努利分布



Y是实验成功的次数



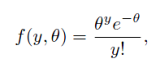
对上式进行整理





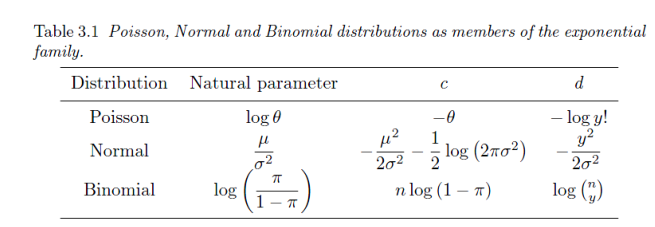
**泊松分布** Poisson

（比如一个时间段到餐厅用餐的人数；每页书有多少个错字）也可以表示成指数函数





如果一个随机变量服从泊松分布 那么它的均值=方差



这三个例子里a（y）都等于y 都是经典形式

**广义线性回归模型**

这样的模型是由一对独立随机变量构成 这些独立随机变量每一个的分布都服从指数族

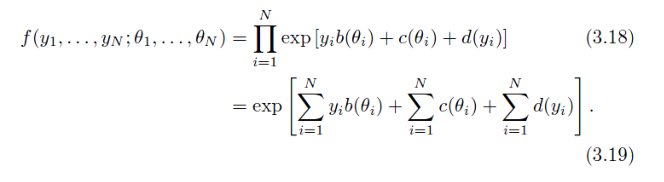
1. 先知道分布

每一个Y的分布都服从经典分布 但是并不是每个的cita都要一样



这些Y 要么全是normal distribution 要么全是Binomial 反正得一样

联合概率密度



我们并不感兴趣这里的cita

1. 再知道一个link function



这个link function对于μi是单调的

G应该是一个单调可导函数 要么是水平的 要么是增加的 要么是降低的



这个μ是cita的函数

对于正态分布来说





3、specification of the linear predictor

23两步通常同时完成

**逻辑回归**

我们感兴趣的是n次实验中成功的比例



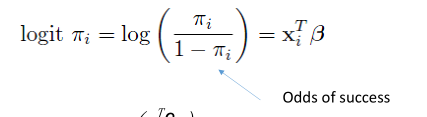




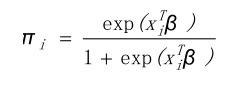
注意 π在（0，1）

G（π）存在的意义就是把范围拉到正负无穷

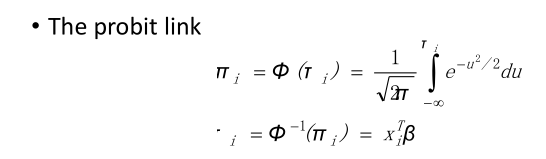
Logit函数的定义：



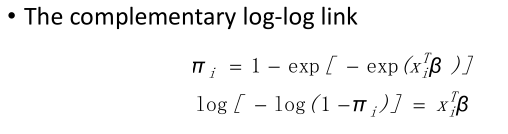
G函数（link function）就取得这种形式



其他的link function



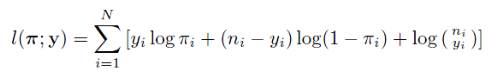
Probit和logit几乎没什么差



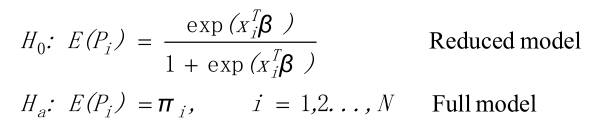
关于0对称

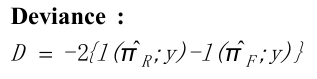
有了link function 接下来我想知道β

Maximaize the log-likelihood function



我们需要构造关于拟合优度的统计量

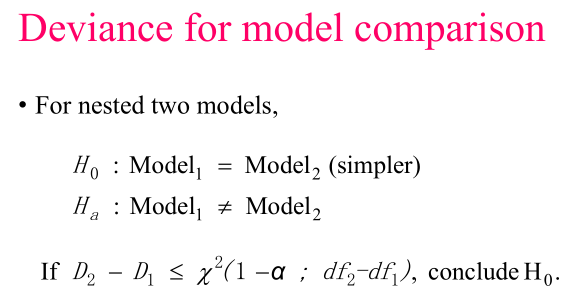




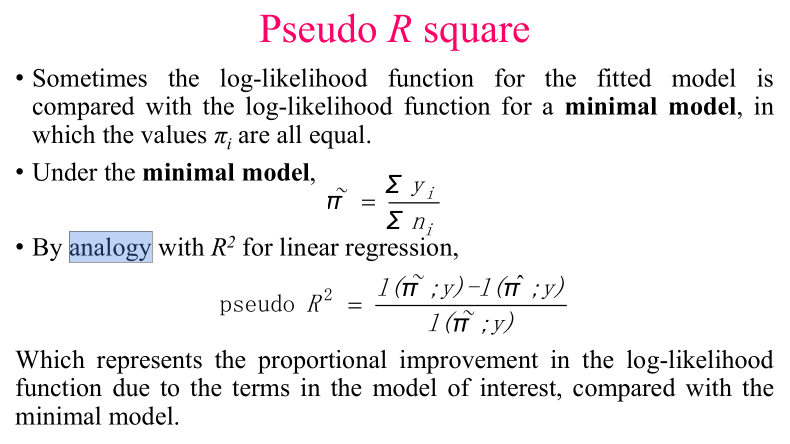
如果满足条件

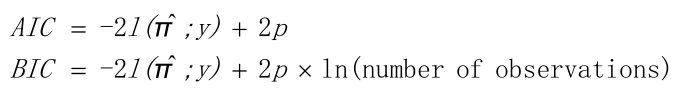


认为H0正确



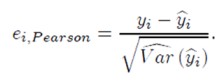
**调整R方**

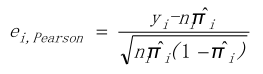
****

****

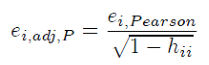
我们倾向于选择AIC和BIC比较小的

残差 pearson residual

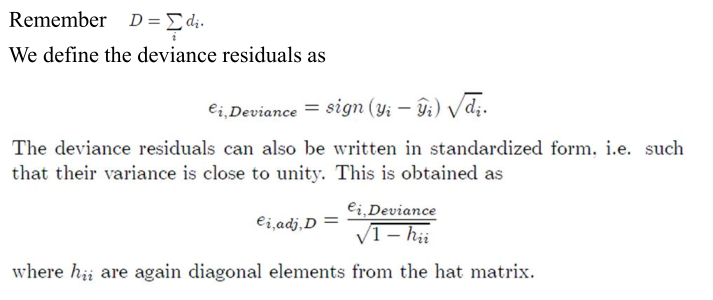




标准化的pearson residual



Deviance residuals



Residual plot

1. 对被解释变量
2. 对协变量 如果不对的话可能是因为link function不对 模型不对（遗漏二次项）
3. 对index

检测异常值（还是那三个手段）

VIF

过度分散 over dispersion 可能是因为观测值之间不独立 观测值之间存在集群

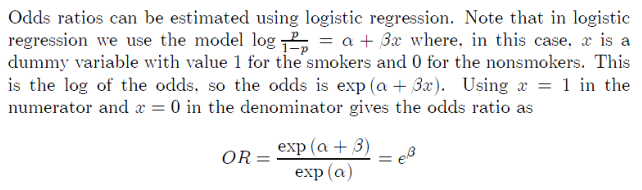
Deviance/df 这个值大于1就表示存在过度分散

**Odds ratio**

Odds =p/(1-p)



OR可以手算



## 第六节课

**数据解释：**

Y: 被污染的水源的个数

N区域内水源的个数

Landuse：水源使用

Dotplot：点图

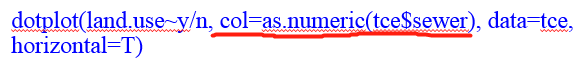
sewer ：是否有下水道

nitrate 含n化学物质的量

chloride含c化学物质的量

**整个思路：先看分类是否合理，在比较不同分类的model；在考虑交互性建模，再比较模型；再考虑是否存在数据过度离散，即是否符合分布上binomial的假设**

* 首先通过点图观察数据特征

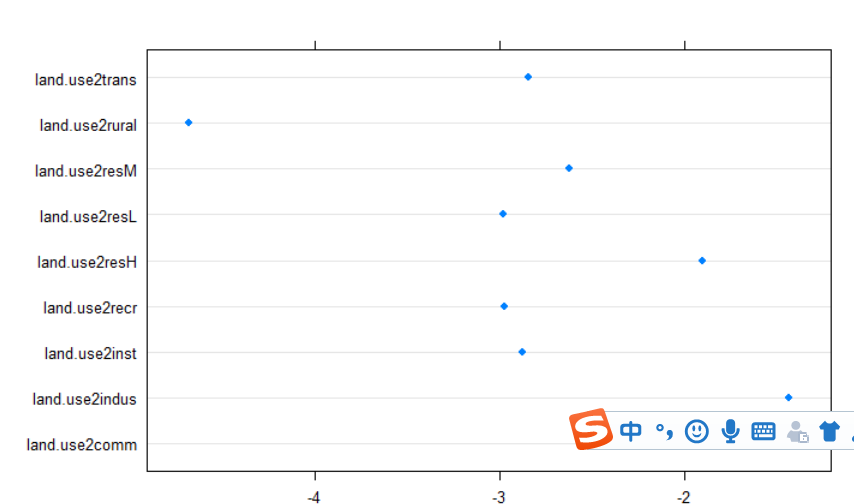


* 在R中glm广义线性回归默认用逻辑回归 建模

model1 <- glm(cbind(y,n-y) ~ land.use + sewer, data=tce, family=binomial)

因为agri是a开头，默认把之设为比较基准（假设检验时用的比较：它对水污染的影响是否与agri对水污染的影响有差异）

* 发现有一项的方差贼大 认为glm可能不好使 为了发现这一项到底怎么了
* tapply(tce$y, tce$land.use,sum):对y按landuse分类求和
* tce$land.use2 <- factor(ifelse(tce$land.use %in% c('undev', 'agri'), 'rural', as.character(tce$land.use))):将agri的数据和所定义的udev的数据合并，定义为一个新的变量land.use2，放在tce中；用ifelse条件语句；c是指若数据是undev或agri，则定义为rural；所以土地分类变为九类，再做一次模型拟合
* model2 <- glm(cbind(y,n-y) ~ land.use2 + sewer-1, data=tce, family=binomial) # to compare the nine landuses' effects on logits of contamination：“-1”指去除截距项
* xlab='Log odds of contamination横轴命名为



* 从上图，可观察到-3上方有很多点都接近-3，所以考虑进一步缩减分类

【模型比较】

sapply(list(model2, model3, model4), AIC)列出3个模型的ALC，比较模型的拟合好坏

sapply(list(model2, model3), logLik)比较模型的极大似然函数

anova(model2, model3, test='Chisq')比较两个模型的deviances

model5 <- glm(cbind(y,n-y) ~ land.use4\*sewer, data=tce, family=binomial)加入交互项，两项相乘

若anova，p值显示没有显著差异（p值大于显著性水平），则选择较简单的模型

【过度离散】

model4$deviance/model4$df.residual 若结果大于1说明有过度离散的问题

有一种方法是这样的

model6<- glm(cbind(y,n-y)~land.use4 + sewer, data=tce, family=quasibinomial)把family由binomial改为quasibinomial，再做逻辑回归即可

还有一种是使劲往模型里加解释变量

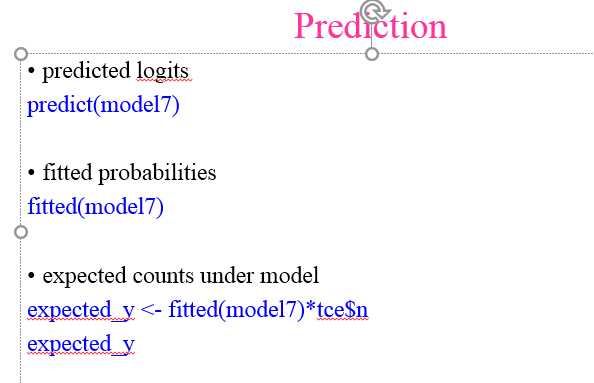
【模型诊断】

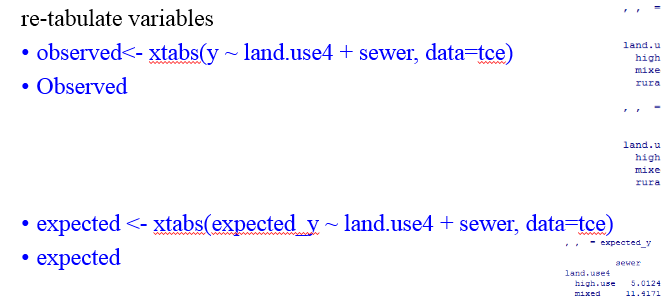
与线性回归模型基本相同

【模型预测】

函数predict

Fitted计算拟合的概率





## 第六讲 多分类回归polytomous logistic regression

目标变量是一种分类的变量

两种：

Multinomial logistic regression 被解释变量之间不存在顺序关系

Proportional odds model

当解释变量之间存在顺序的时候 用proportional odds model比较好

**Multinomial logistic regression**

被解释变量没有次序

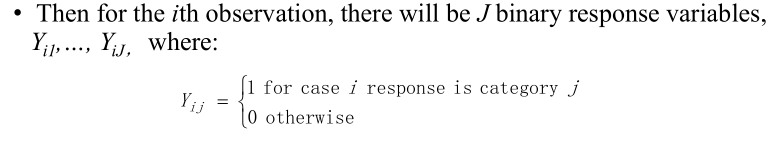
三个以上没有次序的分组

有j种被解释变量

有i个观测

认为有j个二项的变量

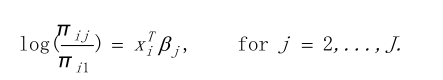
如果第ij项的确等于1 那么说明它恰好在j类别里



对Yij求和=1

J个类别里会有一个rreference category

假设j=1是baseline



我们的目标：

有j个专业 每个被解释变量（专业）有一个概率 我想知道这个概率受什么影响

我可以两个两个地拿出来进行二项回归

我把一个被解释变量作为base category

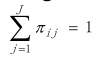
和其它的挨个组合 做二元回归

Πi2/πi1 可以看成是只把它们两个拿出来做了一个二项回归

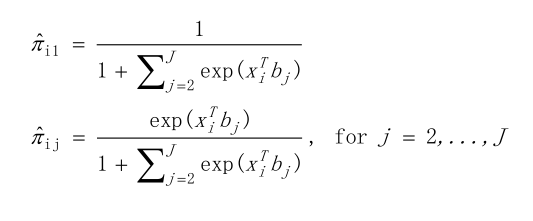
一共有j-1个β向量

做j-1个逻辑回归

加上

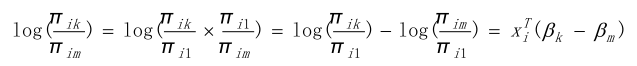


一共是j个方程



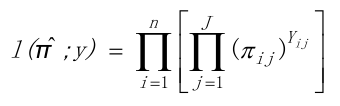
这是方程的解

如果是为了比较两种情况的概率



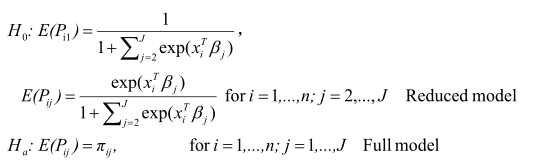
得到结果有两种方法 一种是挨个和baseline做二项回归

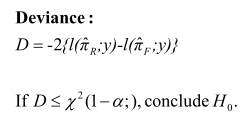
一种是用全模型的likelihood estimate

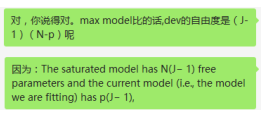


**模型诊断**

与binomial的诊断类似

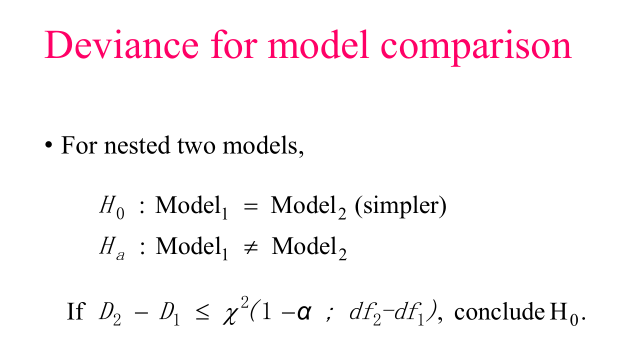




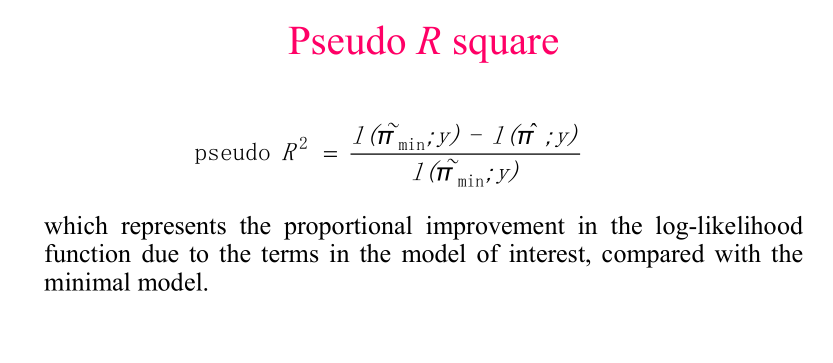


这里说的max model就是这页PPT上面那个模型

这个模型里有多少个未知的量呢？



这是为了衡量哪个模型更好



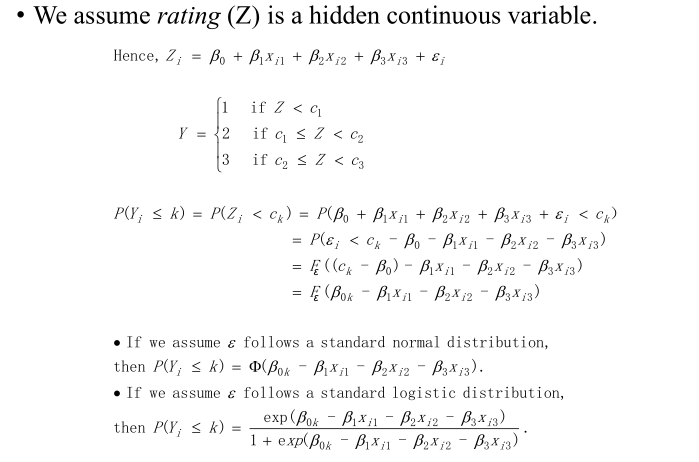
这是伪R方

分子是极大似然估计

AIC和BIC一样适用

多分类的maximum value代表啥？是参数的估计值么？这里的minimal model和maximal model都是啥？

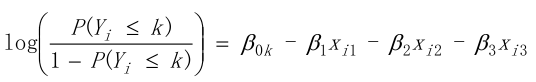
**对于有次序的分类变量 proportional odds model**



从P到F那里是概率分布函数

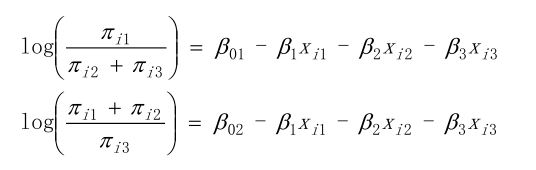
如果e是遵从标准正太分布的话 那P就是对应分为点的那个函数值

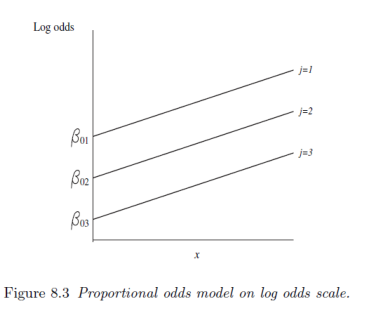
如果e是standard logistic distribution



它的含义是

X改变一单位 给所有的response categorie的odds ratio带来的变化是相同的





## 第七章 多分类回归上机

**多分类无序**

1 导入数据 数据描述

可用with(data1, table(ses, prog))

表示按照ses和prog两项制表

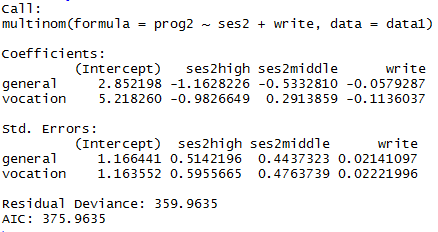
tapply(data1$write, data1$prog, sd) 探索选择不同项目的小孩的writing成绩都怎么样

2建模

先做数据处理 把数据都变成层级数据

test <- multinom(prog2 ~ ses2 + write, data = data1)

这行打出来如果弹出“converged” 意思就是“收敛的”



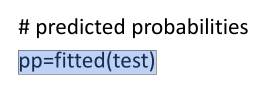
含义：write增加一个单位 general prog的log odds比academic program的log odds减少0，058个单位

1. 模型检验

计算z值 z值本身没有意义 是用来计算p的

计算p值

把log odds转变成odds： exp（coef(test)）



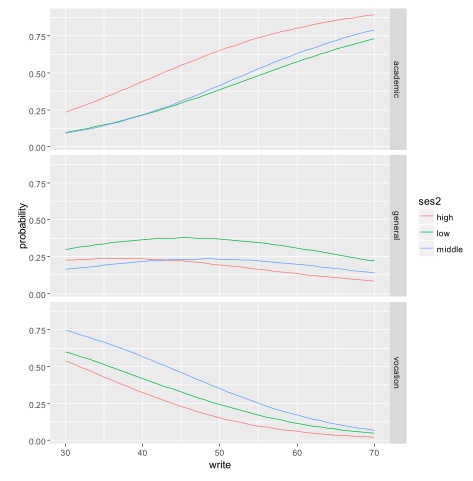
控制write不变在它的mean 探究另外一个解释变量ses和被解释变量的关系

dses <- data.frame(ses2= c("low", "middle", "high"), write = mean(data1$write))

可以在固定了writing的情况下根据ses的高低对选择哪个prog进行一波预测

然后再控制ses不变（虽说不变它还是有三种可能取值） 探究被解释变量和writing的关系

最后得到的是这样的图：



模型的诊断不好进行 可以两两用二元回归的logit检验

**多分类有序 ordinal logistic regression；the proportional odds model**

学生申请学校的意愿值为123

影响因素有父母受教育水平，原来学校私立功力，gpa

1 描述数据

2 建模

3 观察显著性

CI

Ci实际上是log（coef）

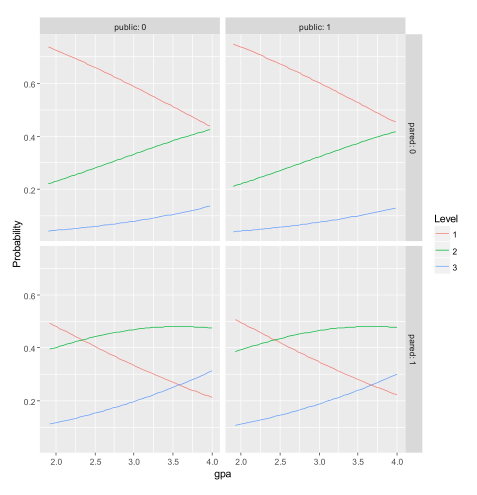
OR

就是odds ratio

回过头去看odds ratio咋回事它就是咋回事

注意理解如何解释OR

也可以画出来一个这样的图：



1. 模型比较

和null model比较

Null model是因变量~1

Anova检验

和multinomial logistic regression比较

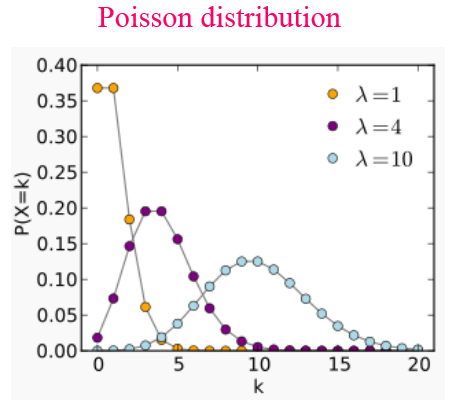
比较AIC

和probit模型比较

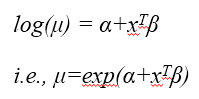
Anova

## 第八章 泊松回归

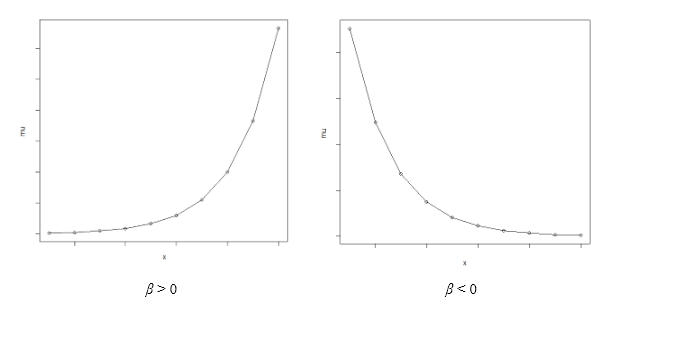
因变量是count



我们感兴趣的是



Exp(α)就是当x=0时 effect on the mean of Y



如果x从0增加到1 μ就会增加exp（β）

如果β=0 那就说明μ和解释变量没有关系

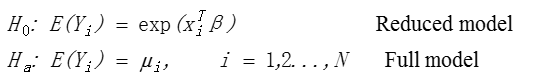
当β>0 那x增加 μ增加

极大似然估计：



这还是一个广义线性回归模型

**模型诊断**

****

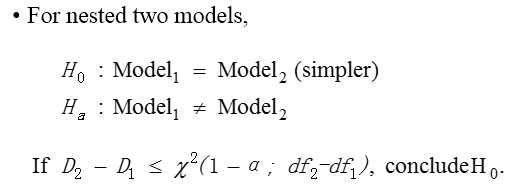


如果有



认为H0是真的

对于两个嵌套的模型



就是用于好几版模型的比较

调整R方

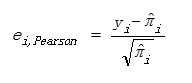
AIC和BIC

用残差的检验

定义Pearson 残差



在泊松分布的情况下





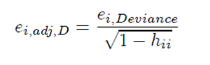
把这个东西标准化（标准正态分布）



还有deviance residuals



也可以把它标准化



-残差对拟合值画图

-残差对covariance画图

-残差按照观测顺序画图

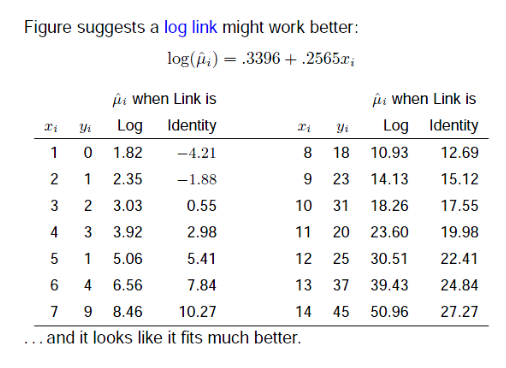
-QQ plot

看看有没有非常有影响的观测 还是那几种手段

Jesen inequality

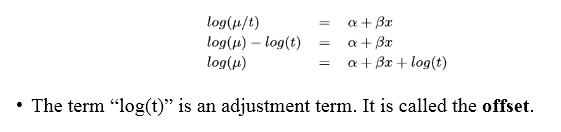
Eg(y)≠g(Ey)

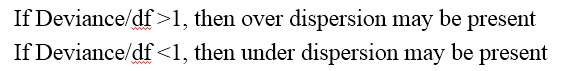
相等只当g是线性函数



Identity是线性回归

然后发现log之后拟合的更好 就是那个预期跟y更相近





## 第九章 泊松回归上机

例子：

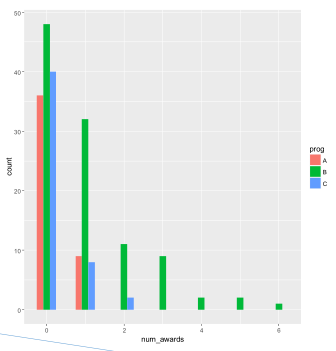
Y：某校学生总得奖数目

X：学生的项目 平均成绩

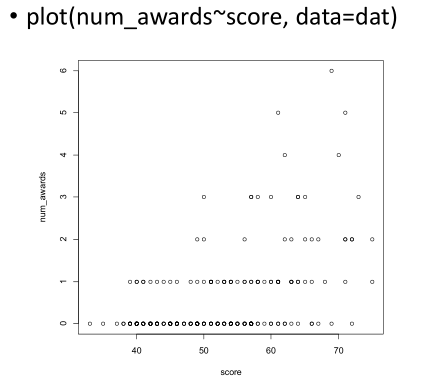
1 描述数据

Tapply mean&sd

Ggplot



Plot 得奖数目~score

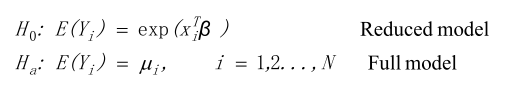


2 建模

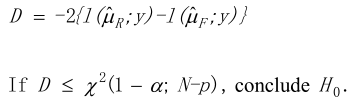
3 看模型怎么样

Deviance：

我用各种x去推测y的这个叫reduced model



Deviance是怎么算出来的

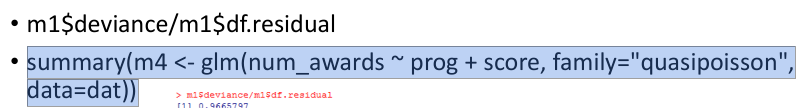


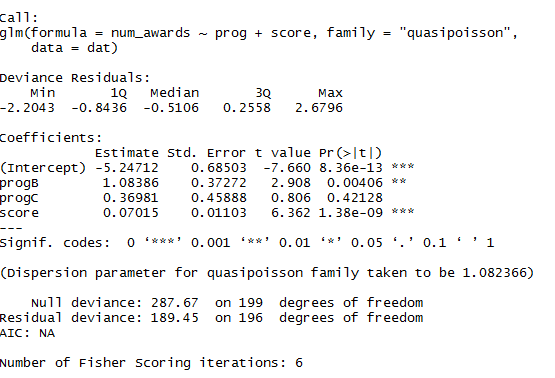
和别的模型比较

Anova和比较AIC

可以去掉不显著的项 或者增加交叉项

看有没有overdispersion

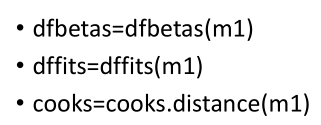




还有其他的诊断方法

把各种residual对y和x plot



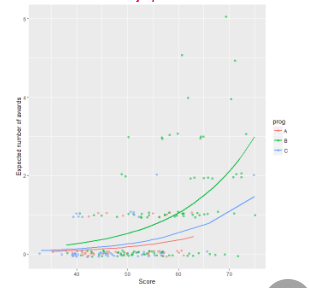


1. 对结果的解读



When program is fixed, for a 10-unit increase in score, the expectation number of awards increase by a factor of exp (0.07015\*10)= 2.016776.

1. 画图



## 第十章 生存数据分析

Y：某事件发生之前的时间

这个time就是survival time 所以叫生存数据分析

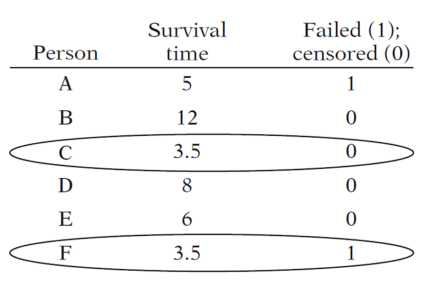
Event一般是failure 死亡疾病之类的

这类分析有一个问题 就是存在截尾数据censored data

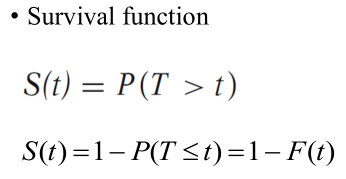
就比如说实验观察期结束了 但是人还没死

比如这个人被跟丢了

所以虽然这个人的“生存时间”没有继续增加了 但是可能不一定是因为failure 还有可能是因为censored

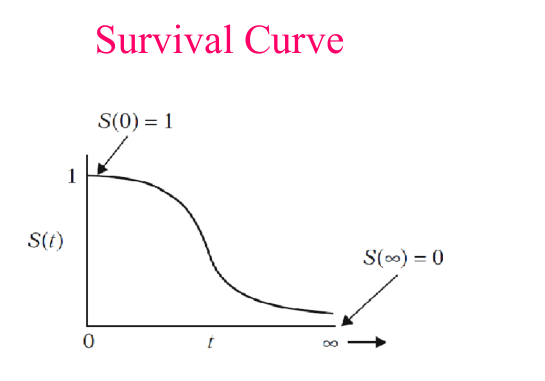


这里的T=suivival time



就是一个人活过那个时间点的概率

=1-在t时间内挂掉的概率



画这个survival curve用的是Kaplan-Meier Method

在survival function里根本没有参数 所以是一种凭借直觉的画法

我这个S（t）是如何来的？

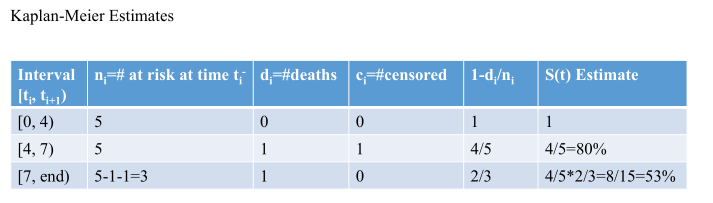
依靠观察

我现在有5个病号

1个4月份死了

1个6月份的时候消失了 不知死活

1个7月份的时候死了



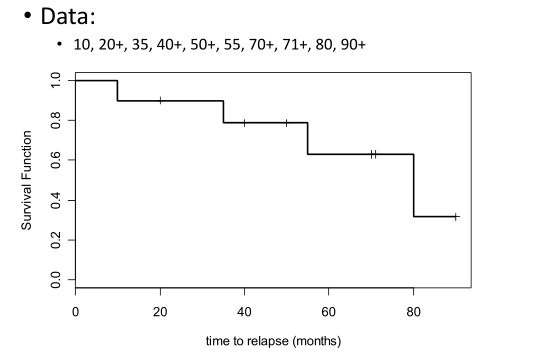
那在整个时期 存活概率就是1\*0.8\*2/3=0.53

在跟丢的人的处理上

我们不知道他会活多久

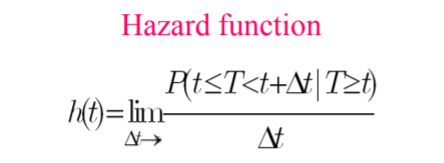
所以这个存活概率介于2/5和3/5之间

比较两组的存活率有没有差别



就像上面那个表一样算 就可以把这个图画出来

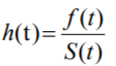
期末会考



T是生存时间

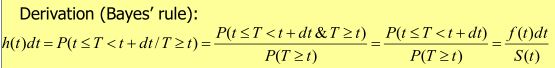
△t趋于0

H（t）是风险函数 即时死亡概率



分子是密度函数

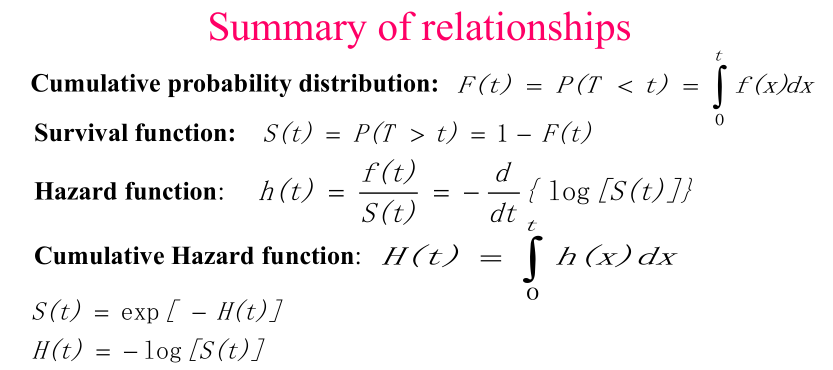
用贝叶斯准则得到

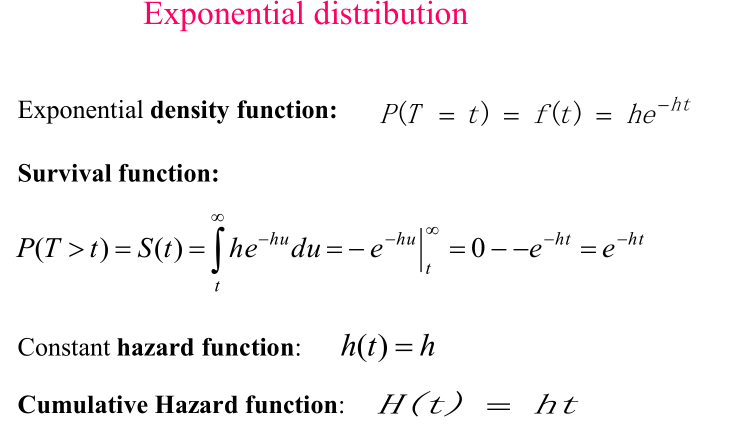


Hazard function和density function是有区别的

Density：今天出生的女婴 有1%的概率会在80岁死亡

Hazard：这个女婴她已经79岁了 她有80%的可能会在80岁死亡 这个“即时死亡”的概率是不断改变的

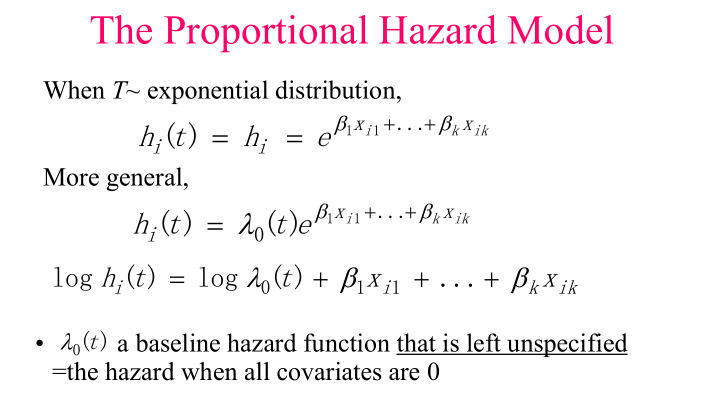


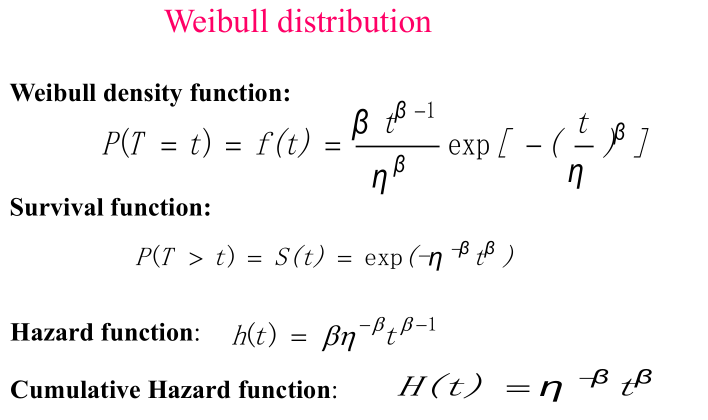


在exponential distribution的情况下 即时死亡的概率h（t）是个常数 不随时间变化

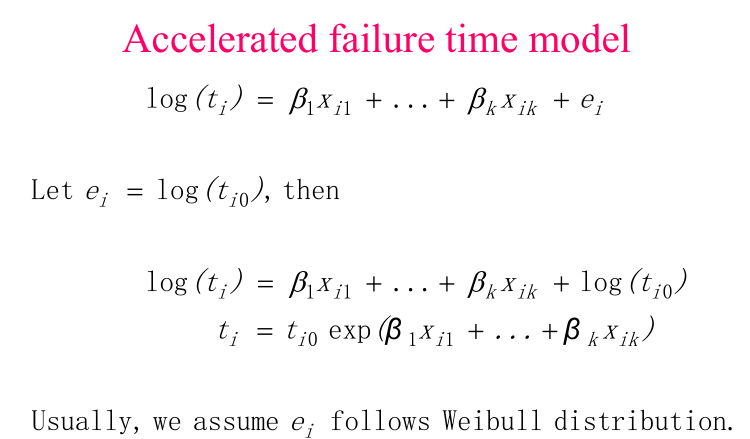
所以呈现出“缺失记忆”的特性

这个模型可以更一般化

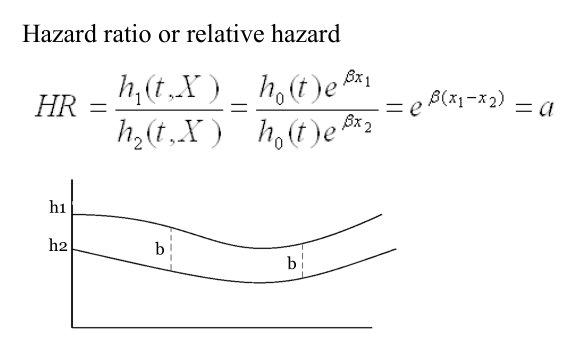




这个时候的即使死亡概率就不是不变的了

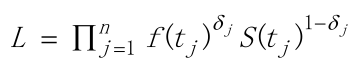


还有一种模型是proportional hazard



模型的估计：

其实就是需要知道maximum likelihood function是啥



如果没有被截断

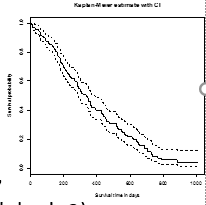
根据logH（t）对log（t）的plot 图来判断属于上述哪种分布

## 第十一讲 上机 生存数据分析

* **先使用Kaplan-Meier 方法**

建模

画出这样的图



一次看各个变量对survival probability的影响

**Log-rank-test看是否男女的survival function有显著差异**

Survdiff方法

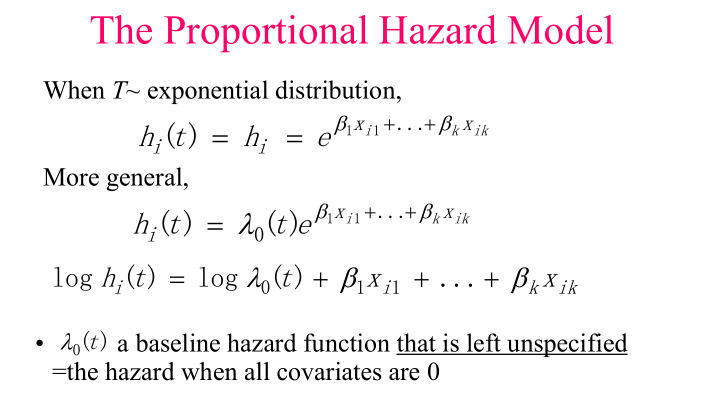
看p值

**Plot cumulative hazard**

先画水平的 再画log后的

**建模并对比AIC anova 选出一个最好的模型**

**测试proportional hazards assumption行不行**



先检测β是否随着时间变化

对每一个β都要检验 age的 sex的

通过画图的方式

**模型诊断**

把残差对index画图

Dfbeta找异常值