统计基础部分

张 霜

中国科学院生态环境研究中心

2021-05-08

北京凌云翼数据科技有限公司

统计模型：让我们爱恨交加

Statistics matter greatly in biology, whether we like it or not. Much of the time, the growing level of complexity and sophistication of the models we put to use in ecology and evolution have led to more appropriate analyses of our data.

Arnqvist, G. 2020. Mixed models offer no freedom from degrees of freedom. Trends in Ecology & Evolution **35**:329-335.

统计中的关键问题

- 想好你的科学问题
- 充分了解你的数据
- 选择合适的模型
- 对模型结果进行正确的解读

你的科学问题是？

- Y的值是多少？（如中国的人均收入是多少？）
- X对Y的影响？（性别是否影响收入？）
- Y和哪些X有关？（众多因素中，哪些因素会影响收入？）
- 某些处理（X）是否会对Y产生影响？（（不同）因素对收入的影响程度如何？）

本课程中，统一把因变量记为Y, 自变量记为X

我们为什么需要统计分析？

ID	收入
1	10.26
2	3.76
3	19.15
4	20.57
5	2.01
6	1.12
7	3.95
8	18.44
9	1.98
10	5.36

观测数据有变异，结论无法100%确定。如果数据无任何变异，则无需统计

我们为什么需要统计分析？

ID	性别	收入
1	男	10.26
2	男	3.76
3	男	19.15
4	男	20.57
5	男	2.01
6	女	1.12
7	女	3.95
8	女	18.44
9	女	1.98
10	女	5.36

观测数据有变异，结论无法100%确定。如果数据无任何变异，则无需统计

统计的核心目标：分析y发生变异的原因

- 数据越“烂”（变异来源复杂），则对统计的要求越高！
- 调查数据（observational）的复杂性往往大于实验（experimental）数据
- 统计的难点在于：如何找到能达到自己的目的，且和数据的特点相匹配的方法。

统计方法的选择

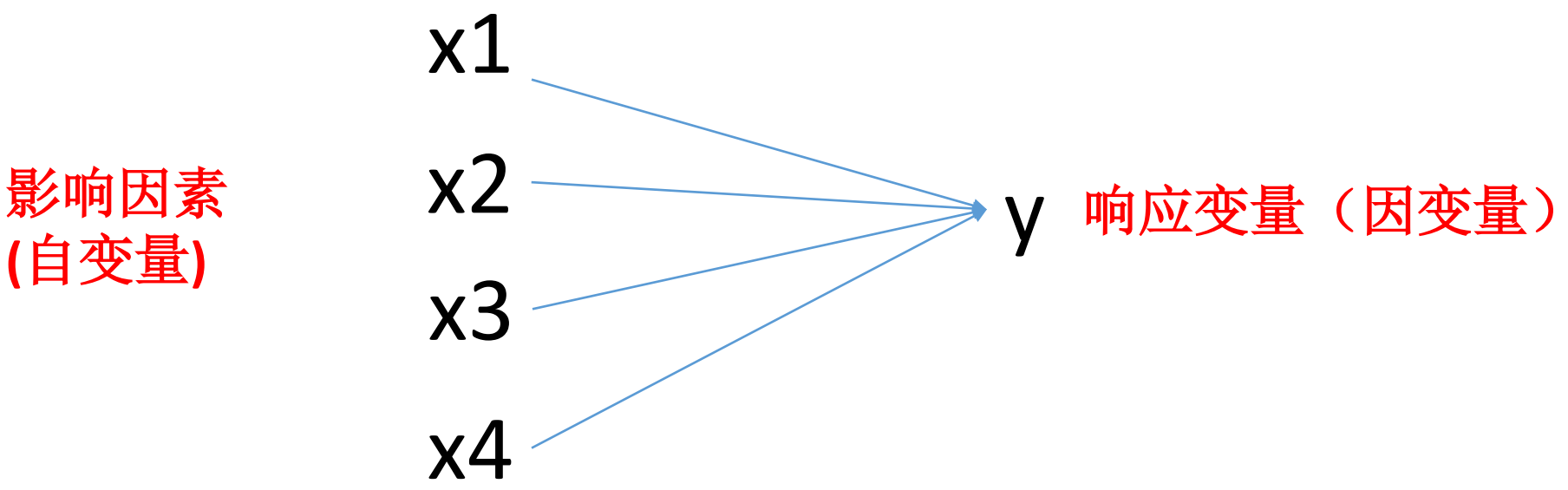
- 一切统计方法都要服务于科学问题！
- 一切统计模型的应用对数据都有一定的前提要求（适用的问题与数据类型）！
- 越简单的模型，对数据质量的要求往往越严格！
- 在数据满足模型前提要求的情况下，统计模型当然越简单越好！

统计方法的分类

- **参数统计：** 能够定量的给出y和x之间的关系，
  如回归，相关，方差分析等（x可以是连续变量也是，也可以是分类变量）
- **非参数统计：** 秩和检验，卡方检验等
  如对孟德尔豌豆实验的卡方检验

当前绝大多数的统计分析的目的都在于寻找y和x之间的定量关系，即参数统计，因为可以依据这种定量关系对未来进行预测。

回归(regression)：最为常用的参数统计分析方法



$$y=b_0+b_1x_1+b_2x_2+\ldots+b_jx_j+\varepsilon$$

- 同一个模型中，无论自变量x有多少个，因变量y只能有一个
- 其核心在于，依据一定的原理分别对公式中的参数b和error进行估计

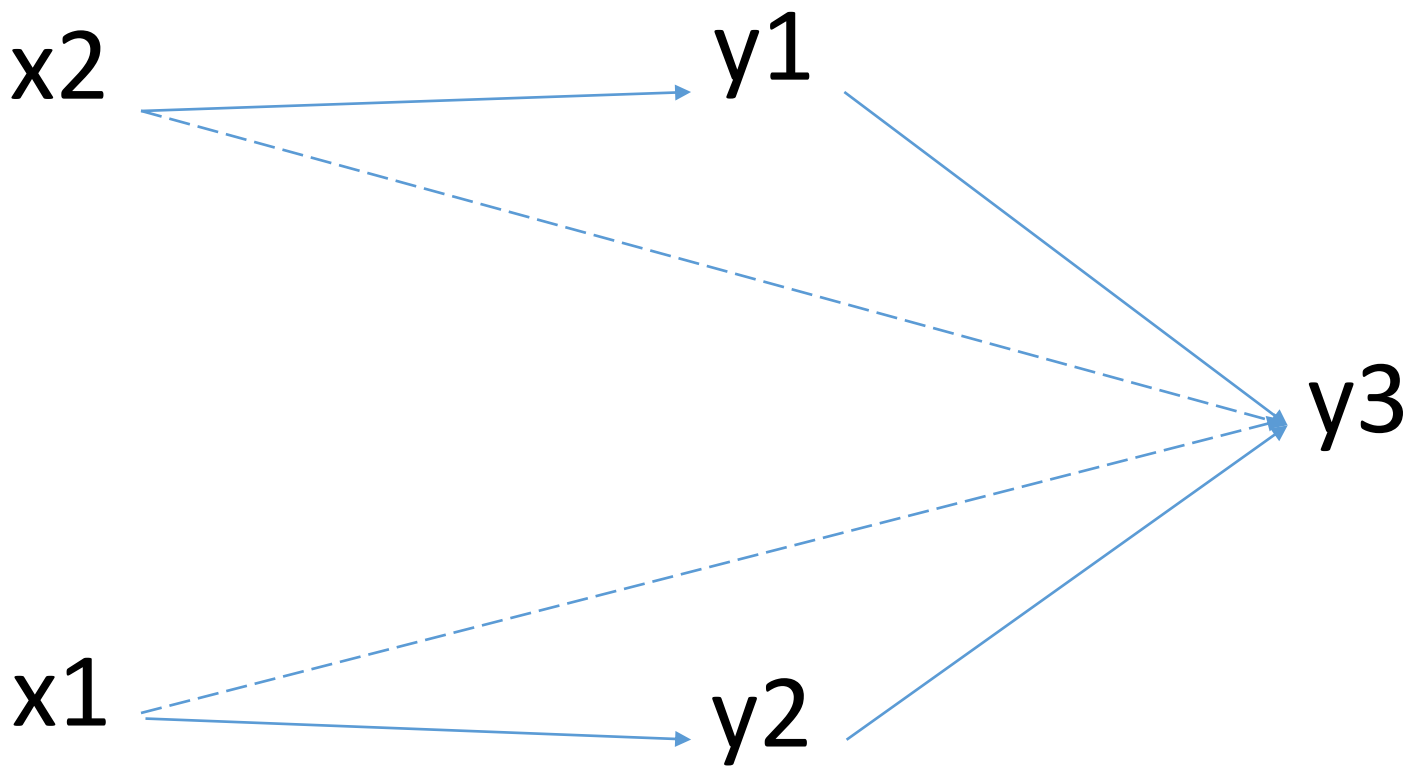
回归能告诉我们什么？

- 找出y的均值及其变异范围
- 找出y和x之间的关系
- 判定哪些x与y有关，哪些x与y无关
- 衡量x对于y的决定程度（R2）
- 对未来进行预测（依据新的x值对y进行预测）

回归公式

$$y=b_0+b_1x_1+b_2x_2+\ldots+b_jx_j+\varepsilon$$

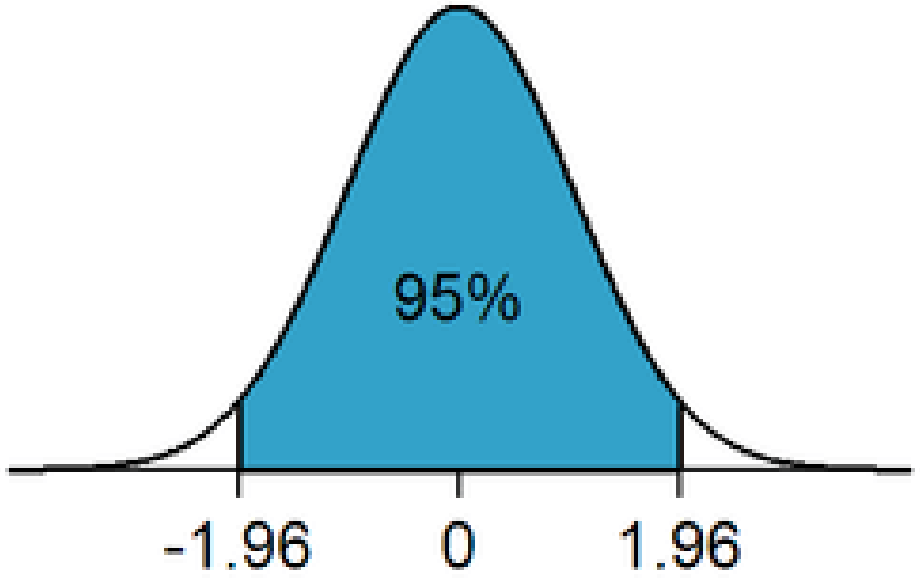
如果同一个模型中，y有多个且存在彼此影响，那么请选择 结构方程模型

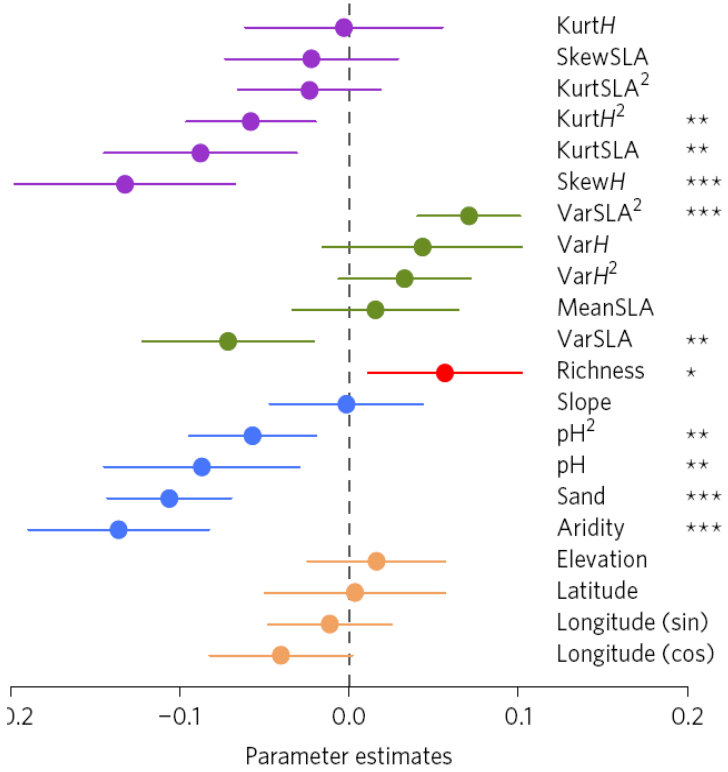


回归中的参数估计与置信区间

$$y=b_0+b_1x_1+b_2x_2+\ldots+b_jx_j+\varepsilon$$

- SE：均值的SD，所以定义均值的置信区间，必须用SE
- 依据中心极限定理，理论上，所有参数的均值都应该符合正态分布，所以可以依据mean±1.96 *SE，来判定参数**均值的95%置信区间**





(Gross et al. 2017, Nature Ecology & Evolution 1:0132.)

回归分析—
从一般线性模型到广义线性模型

一般线性模型—最为简单的回归模型

$$y=b_0+b_1x_1+b_2x_2+\ldots+b_jx_j+\varepsilon;\ \ \varepsilon\sim N(0,\sigma^2)$$

简单是条件的（与ANOVA相同）

1）不同因子对y的作用是相加的

2）观测值y彼此独立（残差彼此独立）

3）残差的方差具有同质性

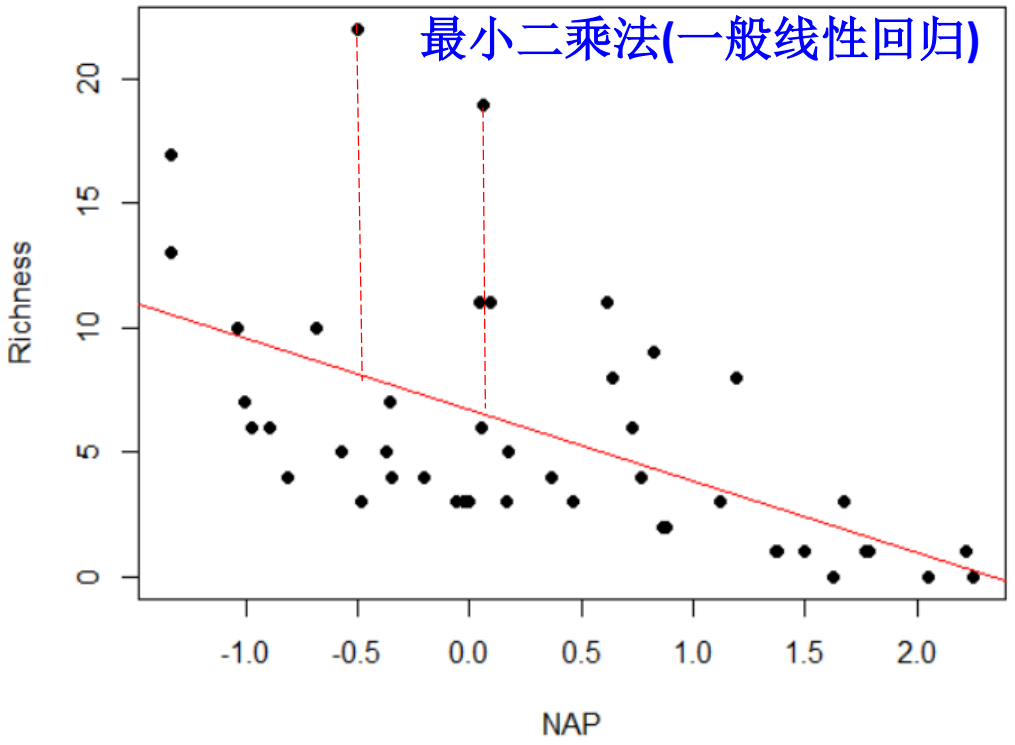
4）残差符合正态分布

回归中参数估计的方法

$y=b_0+b_1x+\varepsilon$, $\varepsilon \sim N(0, \sigma^2)$ 如何求得公式中b_0, b_1, σ^2

- **最小二乘法(一般线性回归)**
- 最大似然法（ML, maximum likelihood，混合模型）
- 限制性最大似然法（REML,
  restricted maximum likelihood，混合模型）
- Laplace approximation （广义线性混合模型）
- Penalized quasi-likelihood （广义线性混合模型）
- 贝叶斯法(Markov chain Monte Carlo) （任意模型都可以）

.....



一般线性模型—最为简单的回归模型

$$y=b_0+b_1x_1+b_2x_2+\ldots+b_jx_j+\varepsilon;\ \ \varepsilon\sim N(0,\sigma^2)$$

一般线性模型—最为简单的回归模型

$$y=b_0+b_1x_1+b_2x_2+\ldots+b_jx_j+\varepsilon;\quad \varepsilon \sim N(0, \sigma^2)$$

模型可以只包含截距

即只对Y的均值和变异范围进行估计

如：某一地区居民的平均收入是多少？

```
lm1<- lm(y~ 1, data=d1)
```

一般线性模型—最为简单的回归模型

$$y=b_0+b_1x_1+b_2x_2+\ldots+b_jx_j+\varepsilon;\ \ \varepsilon\sim N(0,\sigma^2)$$

通常，除截距外，模型还包括一些对y可能具有影响的因素x

如：某一地区居民的收入和哪些因素有关？

这些因素可以是分类变量（如男，女）也可以是数值型变量（如年龄，身高，体重）

lm2<- lm(y~x_1+x_2+...+x_j, data=d1)

回归与方差分析：
均以寻找y和x之间的定量关系为核心目的

- 回归（**定量关系**）：x与y的定量关系$y=b_0+b_1x_1+b_2x_2+\ldots+\varepsilon$
- 方差分析（**影响程度**）：x对y的影响是否显著，对y的变异的贡献有多少？
- 二者分别回答了同一个问题的两个方面

回归和方差分析的应用非常广泛
实际应用中，二者往往结合使用

回归与方差分析（anvoa）--高度统一

```
lm0<-lm(Richness~NAP,data=RIKZ)
```

```
> summary(lm0)

Call:
lm(formula = Richness ~ NAP, data = RIKZ)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0675 -2.7607 -0.8029  1.3534 13.8723

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.6857     0.6578  10.164 5.25e-13 ***
NAP          -2.8669     0.6307  -4.545 4.42e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.16 on 43 degrees of freedom
Multiple R-squared:  0.3245,    Adjusted R-squared:  0.3088
F-statistic: 20.66 on 1 and 43 DF,  p-value: 4.418e-05
```

```
> anova(lm0)
Analysis of Variance Table

Response: Richness
           Df Sum Sq Mean Sq F value    Pr(>F)
NAP         1 357.53  357.53   20.66 4.418e-05 ***
Residuals 43 744.12   17.31
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$Total\ SS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n}$$

$$\text{Among group } SS = \sum_{i=1}^{p}\sum_{j=1}^{ni}\left(Y_{ij} - \bar{Y}\right)^2$$

$$\text{Error (within group SS)} = Total\ SS \text{- Among group } SS$$

$$F = \frac{(Between\ group\ SS)/(p-1)}{Error\ SS/(n-1)}$$

模型的公式化表达：

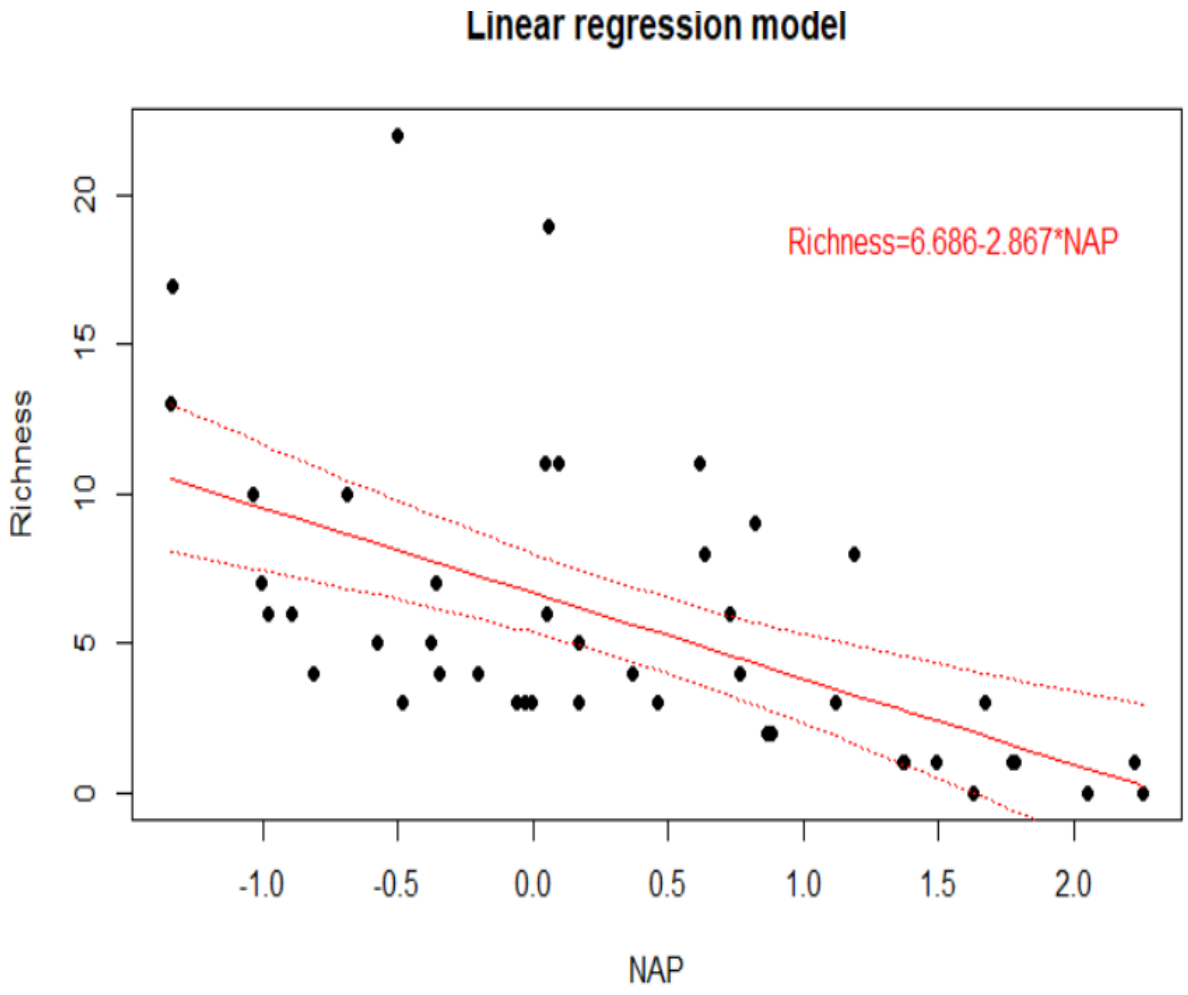
$y=b_0+b_1x+\varepsilon$ ； $\varepsilon \sim N(0, \sigma^2)$ $\sigma = 4.16*\sqrt{n}$

Richness=6.6857 – 2.8669*NAP

$$R^2 = \frac{357.53}{357.53+744.12} = 0.3245$$

一般线性回归分析、与结果制图

```
RIKZ <- read.delim("RIKZ.txt")
plot(Richness~NAP,data=RIKZ, pch=19)
##一般线性回归
RIKZ<-RIKZ[order(RIKZ$NAP),]
lm3<-lm(Richness~NAP,data=RIKZ)
summary(lm3)
##调用ciTools包，得到拟合线的置信区间
library(ciTools)
lm_ci<-add_ci(RIKZ, lm3, alpha = 0.05,names = c("lwr", "upr"))
lines(lm_ci$NAP, lm_ci$pred, col = "red", lwd = 1,lty="solid")
lines(lm_ci$NAP, lm_ci$lwr, col = "red", lwd = 1,lty="dotted")
lines(lm_ci$NAP, lm_ci$upr, col = "red", lwd = 1,lty="dotted")
title("Linear regression model")
###mtext("Richness=6.686-2.867*NAP")
temp <- locator(1)
text(temp,"Richness=6.686-2.867*NAP",col="red")
```



回归与方差分析的结合使用

```
> job<- read.csv("job.csv")
> levels(as.factor(job$education_level))
[1] "college"    "school"     "university"
```

```
> lm1<-lm(score~education_level, data=job)
> summary(lm1)

Call:
lm(formula = score ~ education_level, data = job)

Residuals:
     Min       1Q   Median       3Q      Max
-2.32900 -0.32936  0.03053  0.34789  1.15100

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 6.3495     0.1397  45.448  < 2e-16 ***
education_levelschool      -0.7574     0.1976  -3.833 0.000327 ***
education_leveluniversity   2.4995     0.1951  12.812  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.609 on 55 degrees of freedom
Multiple R-squared:  0.8482,    Adjusted R-squared:  0.8427
F-statistic: 153.7 on 2 and 55 DF,  p-value: < 2.2e-16
```

```
> anova(lm1)
Analysis of Variance Table

Response: score
                 Df  Sum Sq Mean Sq F value    Pr(>F)
education_level   2 113.999  57.000   153.7 < 2.2e-16 ***
Residuals        55  20.396   0.371
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

anova：告诉我们x对y的一个整体性结果
每个x在anova结果中占据一行

判定某因素x对指标y影响是否显著的其他方法

```
> lm1<-lm(score~education_level, data=job)
> lm2<-lm(score~1, data=job)
```

```
> ## 1) anova
> anova(lm1, lm2)
Analysis of Variance Table

Model 1: score ~ education_level
Model 2: score ~ 1
  Res.Df     RSS Df Sum of Sq     F    Pr(>F)
1     55  20.396
2     57 134.396 -2      -114 153.7 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

注意: 模型间必须为嵌套关系

判定某因素x对指标y影响是否显著的其他方法

构造一个不包含x基准模型，与含x的模型进行AIC值对比，或Log likelihood 差异性检验（lmtest::lrtest）

```
> lm1<-lm(score~education_level, data=job)
> lm2<-lm(score~1, data=job)
```

```
> ## 2)AIC值对比
> AIC(lm1, lm2)
    df       AIC
lm1  4 111.9819
lm2  2 217.3370
```

依据自由度的变化，
对残差的差异程度进行检验

```
> ## 3)likelihood ratio test
> library(lmtest)
> lrtest(lm1, lm2)
Likelihood ratio test

Model 1: score ~ education_level
Model 2: score ~ 1
  #Df   LogLik Df  Chisq Pr(>Chisq)
1   4  -51.991
2   2 -106.668 -2 109.36  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

根据自由度的变化，
对对数似然值log likelihood 的差异程度进行检验

不同x类型下，回归模型结果的解读

1）数值变量

2）分类变量

3）数值+分类

4）交互作用

其实质：依据x的值对y进行预测

X为分类变量时，回归结果的解读

用emmeans包轻松获得不同x水平下，对应的y值

```
> job<- read.csv("job.csv")
> levels(job$education_level)
[1] "college"    "school"     "university"
```

```
> jm1<-lm(score~education_level, data=job)
> summary(jm1)

Call:
lm(formula = score ~ education_level, data = job)

Residuals:
     Min       1Q   Median       3Q      Max
-2.32900 -0.32936  0.03053  0.34789  1.15100

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 6.3495     0.1397  45.448  < 2e-16 ***
education_levelschool      -0.7574     0.1976  -3.833 0.000327 ***
education_leveluniversity   2.4995     0.1951  12.812  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.609 on 55 degrees of freedom
Multiple R-squared:  0.8482,    Adjusted R-squared:  0.8427
F-statistic: 153.7 on 2 and 55 DF,  p-value: < 2.2e-16
```

```
> library(emmeans)
> emmeans(jm1, "education_level")
 education_level emmean    SE df lower.CL upper.CL
 college           6.35 0.140 55     6.07     6.63
 school            5.59 0.140 55     5.31     5.87
 university        8.85 0.136 55     8.58     9.12

Confidence level used: 0.95
```

X为分类变量时，令其中一个水平为截距，
其他水平与之相比：

College= intercept=6.35

School= 6.35-0.7574

University= 6.35+2.50

```
> anova(jm1)
Analysis of Variance Table

Response: score
                Df  Sum Sq Mean Sq F value    Pr(>F)
education_level  2 113.999  57.000   153.7 < 2.2e-16 ***
Residuals       55  20.396   0.371
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

两因素，分类变量回归结果解读
（施肥类型(A,B)，种植密度(高、低)）对作物产量y的影响

因素组合

A， high	B， high
A， low	B， low

```
Call:
lm(formula = yield ~ fertilize + density, data = crop)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1600 -0.3055 -0.1028  0.4310  1.3642

Coefficients:
                Estimate Std. Error  t value Pr(>|t|)
(Intercept)    176.9932     0.1276 1387.072  < 2e-16 ***
fertilizeTypeB   0.1762     0.1473    1.196  0.23646
densitylow      -0.4724     0.1473   -3.206  0.00214 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5894 on 61 degrees of freedom
Multiple R-squared:  0.161,     Adjusted R-squared:  0.1335
F-statistic: 5.854 on 2 and 61 DF,  p-value: 0.004725
```

b1
b2
b3

参数	对应组合	估计值
b1 intercept	A， high	b1
b2	B	b2
b3	low	b3

预测各因素组合下产量

A， high y=b1=176.9	B， high: y=b1+b2=176.9+0.17
A， low y=b1+b3=176.9-0.47	B， low y=b1+b2+b3=176.9+0.17-0.47

两因素，分类变量回归：结果解读

```
> emmeans(an1, "fertilize")
 fertilize emmean     SE df lower.CL upper.CL
 typeA      176.8 0.1042 61    176.5    177.0
 TypeB      176.9 0.1042 61    176.7    177.1

Results are averaged over the levels of: density
Confidence level used: 0.95
```

```
> emmeans(an1, "fertilize", "density")
density = High:
 fertilize emmean     SE df lower.CL upper.CL
 typeA      177.0 0.1276 61    176.7    177.2
 TypeB      177.2 0.1276 61    176.9    177.4

density = low:
 fertilize emmean     SE df lower.CL upper.CL
 typeA      176.5 0.1276 61    176.3    176.8
 TypeB      176.7 0.1276 61    176.4    177.0

Confidence level used: 0.95
```

A， high	B， high:
A，low	B，low

两因素，分类变量+交互作用对y的影响

```
> an2<-lm(yield~fertilize+density+fertilize:density, data=crop)
> summary(an2)

Call:
lm(formula = yield ~ fertilize + density + fertilize:density,
    data = crop)

Residuals:
     Min       1Q   Median       3Q      Max
-1.19171 -0.33385 -0.08319  0.34971  1.28295

Coefficients:
                          Estimate Std. Error  t value Pr(>|t|)
(Intercept)             177.07449    0.14708 1203.967  < 2e-16 ***
fertilizeTypeB            0.01365    0.20800    0.066  0.94790
densitylow               -0.63489    0.20800   -3.052  0.00338 **
fertilizeTypeB:densitylow  0.32504    0.29415    1.105  0.27356

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5883 on 60 degrees of freedom
Multiple R-squared:  0.1778,    Adjusted R-squared:  0.1366
F-statistic: 4.324 on 3 and 60 DF,  p-value: 0.007956
```

参数	对应因素水平	估计值
b1 （intercept）	A, high	b1
b2	B	b2
3	low	b3
4 （交互）	B, low	b4

预测各因素组合下的产量

A， high y=b1	B， high: y=b1+b2
A， low y=b1+b3	B， low y=b1+b2+b3+b4

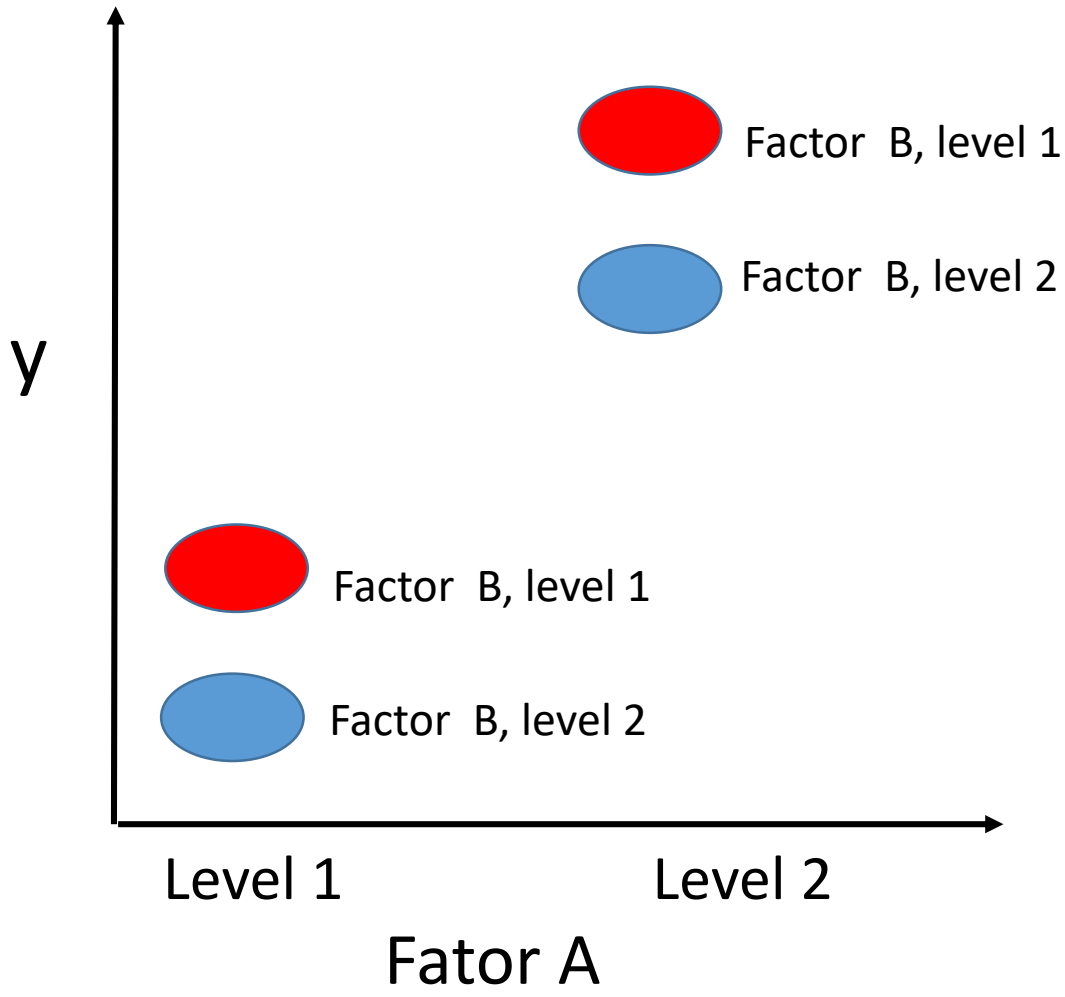
两因素，分类变量+交互作用对y的影响：结果解读

```
> emmeans(an2, "fertilize", "density")
density = High:
 fertilize emmean     SE df lower.CL upper.CL
 typeA      177.1 0.1471 60    176.8    177.4
 TypeB      177.1 0.1471 60    176.8    177.4

density = low:
 fertilize emmean     SE df lower.CL upper.CL
 typeA      176.4 0.1471 60    176.1    176.7
 TypeB      176.8 0.1471 60    176.5    177.1

Confidence level used: 0.95
```

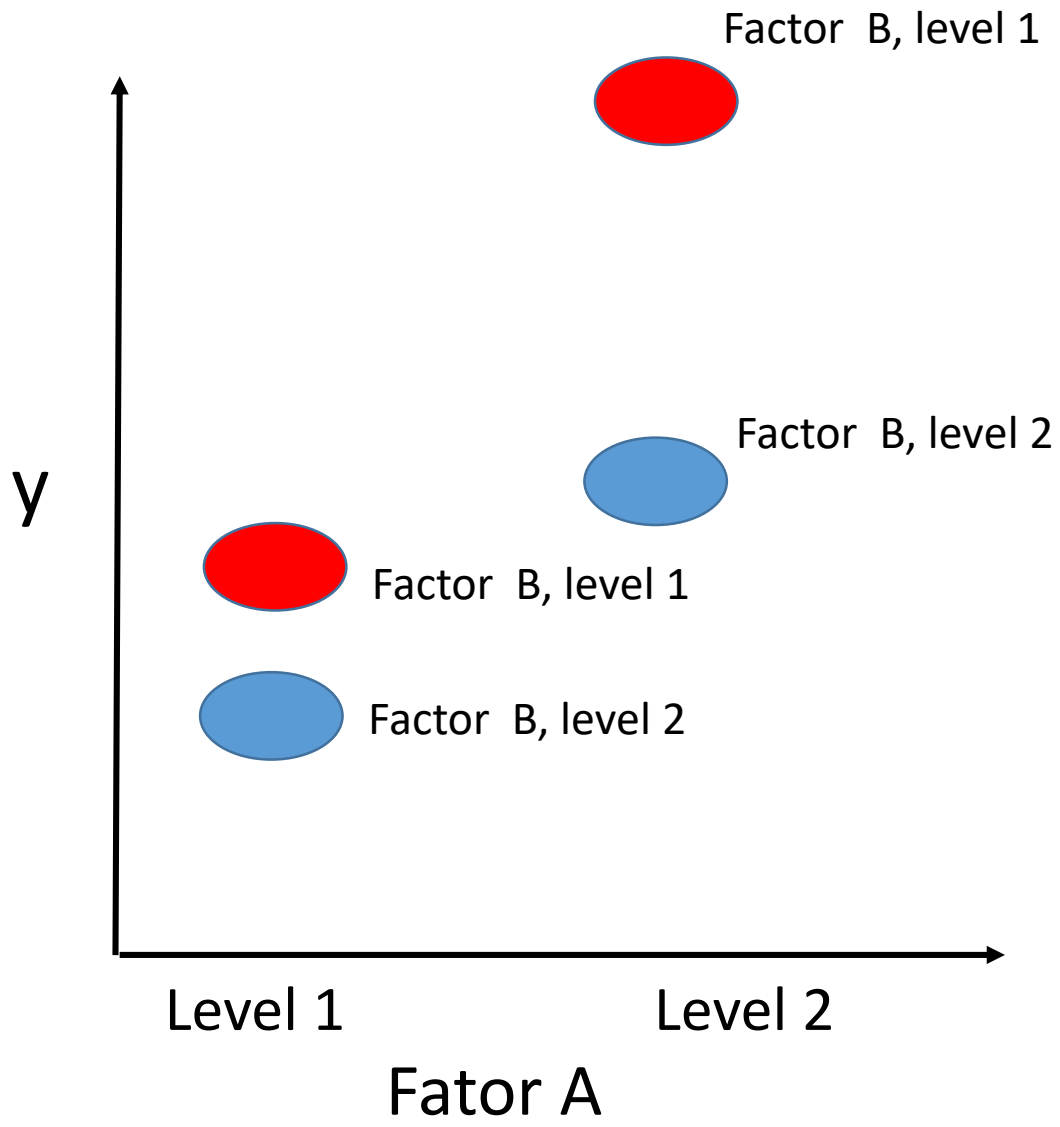
A， high y=b1	B， high: y=b1+b2
A，low y=b1+b3	B，low y=b1+b2+b3+b4

交互作用：两个分类变量之间的交互

Factor A 和Factor B之间无交互



Factor A 和Factor B之间有交互



如果x既包括连续变量又包括分类变量，那么ancova(协方差分析)

an3<-lm(mpg~hp*am, data=input)
summary(an3)

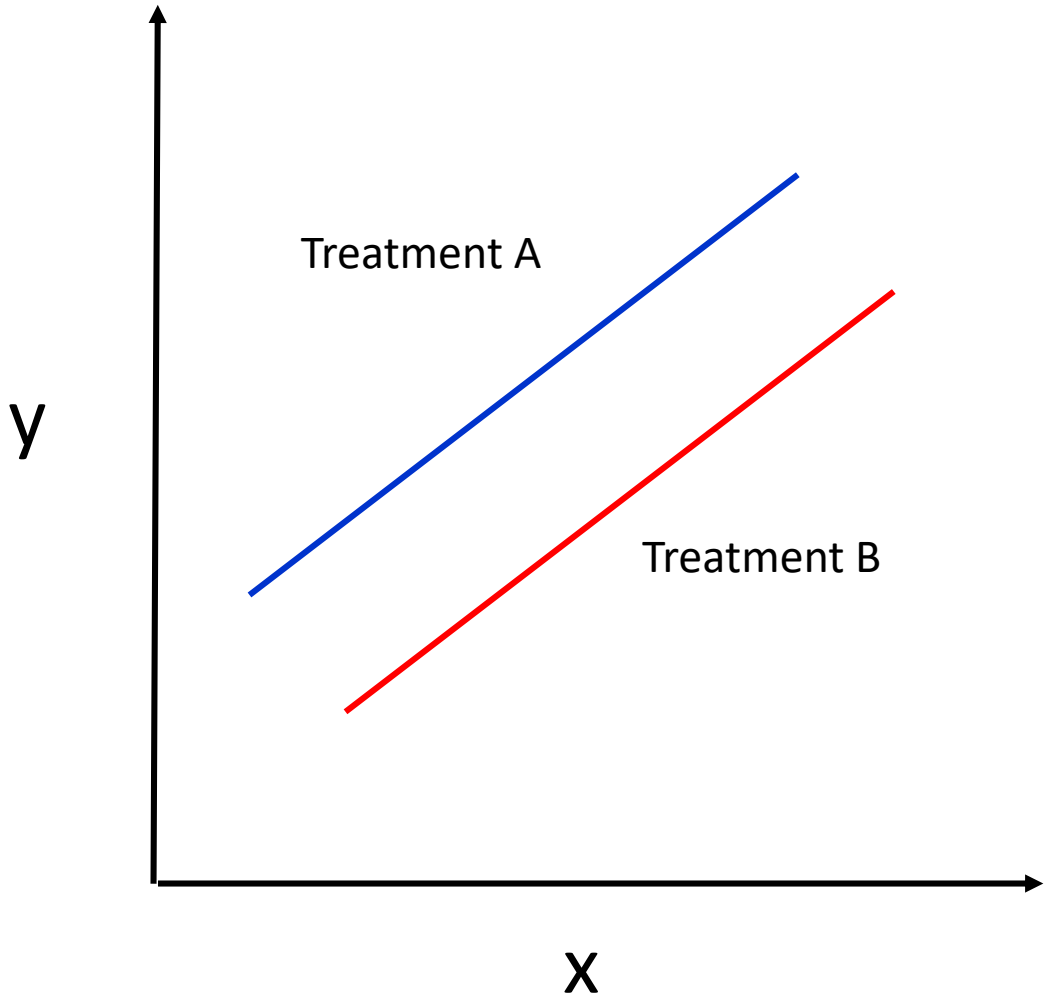
```
> summary(an3)

Call:
lm(formula = mpg ~ hp * am, data = input)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3818 -2.2696  0.1344  1.7058  5.8752

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.6248479  2.1829432  12.197 1.01e-12 ***
hp          -0.0591370  0.0129449  -4.568 9.02e-05 ***
amB          5.2176534  2.6650931   1.958   0.0603 .
hp:amB       0.0004029  0.0164602   0.024   0.9806
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.961 on 28 degrees of freedom
Multiple R-squared:  0.782,     Adjusted R-squared:  0.7587
F-statistic: 33.49 on 3 and 28 DF,  p-value: 2.112e-09
```

```
> anova(an3)
Analysis of Variance Table

Response: mpg
          Df Sum Sq Mean Sq F value    Pr(>F)
hp         1 678.37  678.37 77.3912 1.505e-09 ***
am         1 202.24  202.24 23.0717 4.749e-05 ***
hp:am      1   0.01    0.01  0.0006    0.9806
Residuals 28 245.43    8.77
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

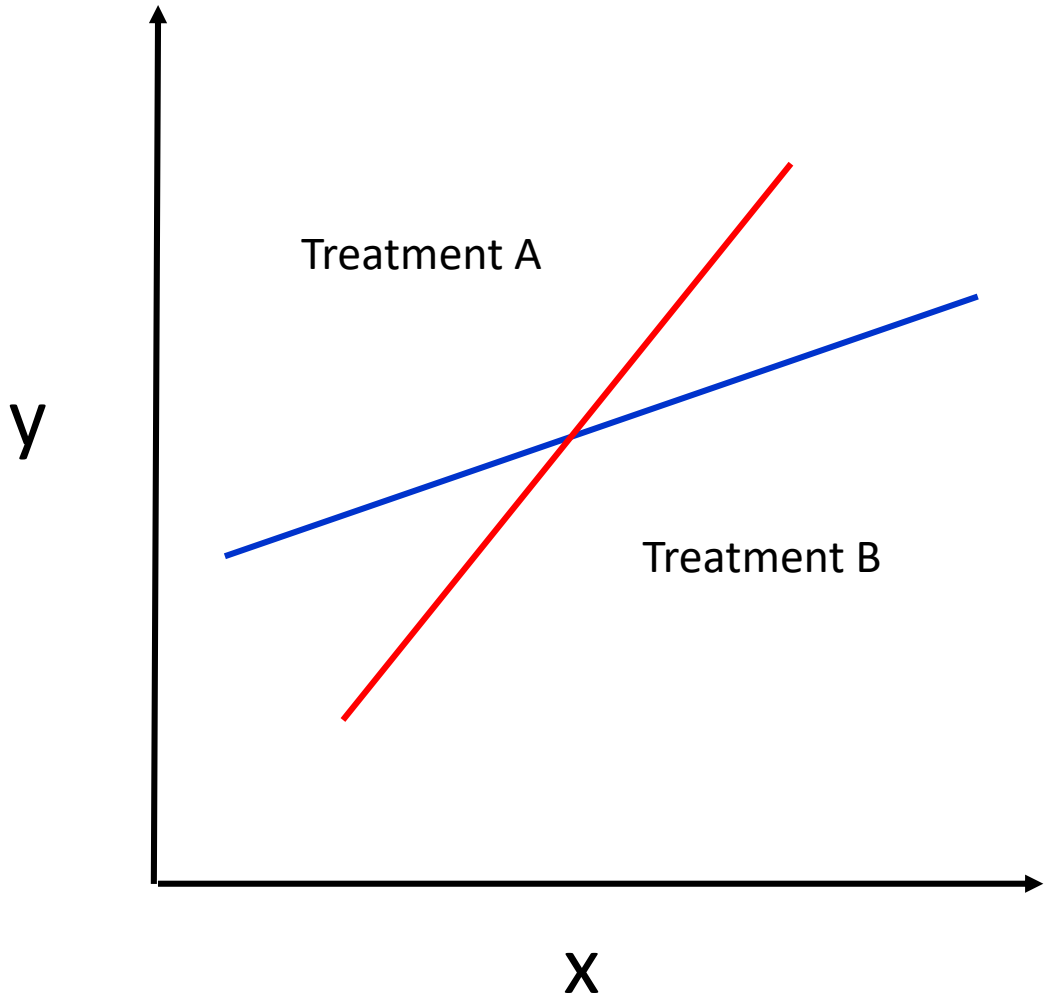
当am=A时， y=26.6-0.059*hp
当am=B时， y=26.6+5.21+（0.0004-0.059）*hp

交互作用：一个分类变量和一个数值变量之间的交互

X和treatment 之间无交互作用时



X和treatment 之间有交互作用时



不同处理水平下，y~x的斜率不同，这种斜率的不同就是交互，即两条线一定会相交

线性回归的前提条件

- X（自变量）间的共线性
- 正态性（**所指何物？**）
- 方差齐次性
- y_i相互独立

那么这些问题如何解决？

样本量

We are just emphasizing that, just **as you never have enough money**, because perceived needs increase with resources, your inferential needs will increase with your sample size.

Gelman, A. and J. Hill. 2007. Data analysis using regression and multilevel/hierarchical models. Cambridge, UK: Cambridge University Press. P438

模型包括的自变量x越多，对样本量的要求越大，一般要求n/k>10，
其中k为模型中参数的个数（截距+误差+X）

Harrison, X. A. *et al.* A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ* 6, e4794, doi:10.7717/peerj.4794 (2018).

样本量越大，参数估计越准确

假如 已知男人的体重为均值为125，女人的体重为100斤，二者的SD都为10斤，那要么weight~sex模型中

N=10

```
> female<-rnorm(10,100, 10)
> male<-rnorm(10,125,10)
> weight<-(c(female,male))
> d2<-as.data.frame(weight)
> d2<-(c(female,male))
> d2<-as.data.frame(d2)
> d2$sex[c(0:10,0)]<-"female"
> d2$sex[c(11:20,0)]<-"male"
> lm3_10<-lm(weight~sex-1,data=d2)
> summary(lm3_10)

Call:
lm(formula = weight ~ sex - 1, data = d2)

Residuals:
    Min      1Q  Median      3Q     Max
-17.331  -9.453  -2.055   9.330  23.272

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
sexfemale   96.996      3.799   25.53 1.37e-15 ***
sexmale    123.565      3.799   32.53  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.01 on 18 degrees of freedom
Multiple R-squared:  0.9896,    Adjusted R-squared:  0.9884
F-statistic: 854.9 on 2 and 18 DF,  p-value: < 2.2e-16
```

SD=SE*$\sqrt{N}$=3.799*$\sqrt{10}$=12.01349

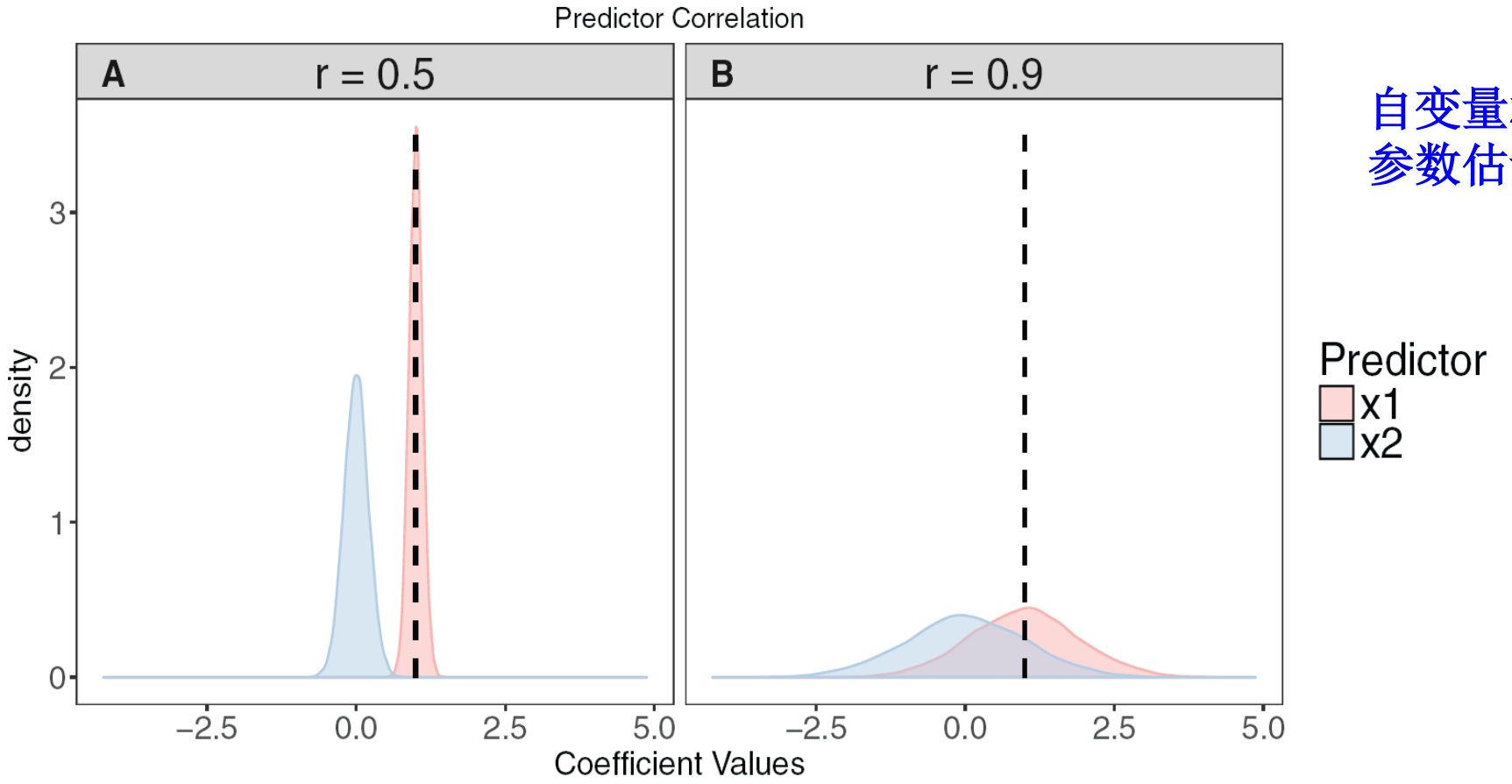
N=50

```
> female<-rnorm(50,100, 10)
> male<-rnorm(50,125,10)
> weight<-(c(female,male))
> d2<-as.data.frame(weight)
> d2<-(c(female,male))
> d2<-as.data.frame(d2)
> d2$sex[c(0:50,0)]<-"female"
> d2$sex[c(51:100,0)]<-"male"
> lm3_50<-lm(weight~sex-1,data=d2)
> summary(lm3_50)

Call:
lm(formula = weight ~ sex - 1, data = d2)

Residuals:
     Min       1Q   Median       3Q      Max
-26.5110  -5.8994   0.0929   6.2630  28.5493

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
sexfemale   99.401      1.448   68.62   <2e-16 ***
sexmale    125.108      1.448   86.37   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.24 on 98 degrees of freedom
Multiple R-squared:  0.992,     Adjusted R-squared:  0.9918
F-statistic:  6085 on 2 and 98 DF,  p-value: < 2.2e-16
```

SD=SE*$\sqrt{N}$=1.448*$\sqrt{50}$=10.23891

回归对x的要求，首先要消除不同x之间的共线性

$$y=b_0+b_1x_1+b_2x_2+\ldots+\varepsilon$$



自变量相关性过高，会导致二者的参数估计不可区分！

(Harrison et al. 2018, PeerJ 6:e4794.)

共线性去除的依据：
相关系数r或方差膨胀因子(VIF)

```
lm4 <- lm(mpg ~ wt + cyl + gear + disp, data = mtcars)
library(performance)
result <- check_collinearity(lm4)
plot(result)
```

1) 对于相关性>0.7的x变量，应仅保留一个或者提取x的主成分进行分析，或者求x的VIF(variation inflate factor), 删掉VIF值较大（如3）的因子。
3) 想要直接比较各x对y影响的大小，可先将x标准化（如mean=0 ,SD=1）,以得 到标准化回归系数

案例来源 https://easystats.github.io/see/articles/performance.html

Harrison, X. A. *et al.* A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ* 6, e4794, doi:10.7717/peerj.4794 (2018).

共线性去除的依据：
相关系数r或方差膨胀因子(VIF)

```
lm4 <- lm(mpg ~ wt + cyl + gear + disp, data = mtcars)
library(performance)
result <- check_collinearity(lm4)
plot(result)
```

1) 对于相关性>0.7的x变量，应仅保留一个或者提取x的主成分进行分析，或者求x的VIF(variation inflate factor), 删掉VIF值较大（如3）的因子。
3) 想要直接比较各x对y影响的大小，可先将x标准化（如mean=0 ,SD=1）,以得 到标准化回归系数

案例来源 https://easystats.github.io/see/articles/performance.html

Harrison, X. A. *et al.* A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ* 6, e4794, doi:10.7717/peerj.4794 (2018).

共线性去除的依据:取消VIF>5的变量

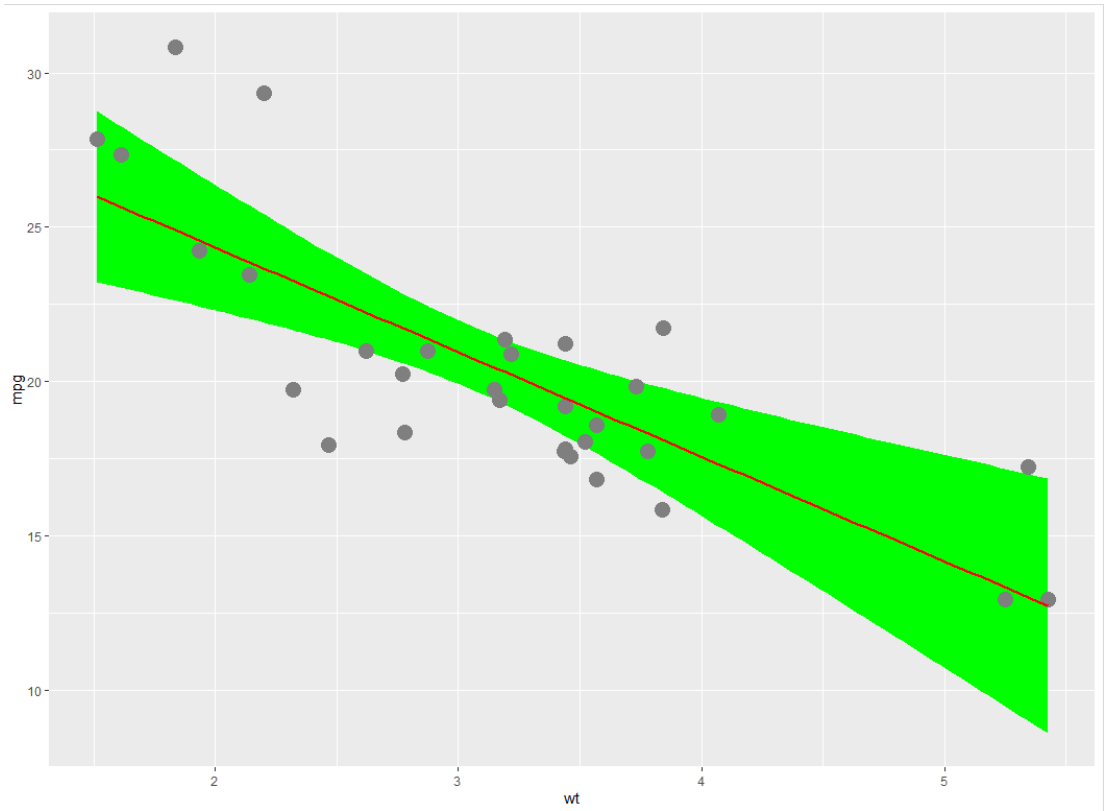
```
lm4_1 <- lm(mpg ~ wt + cyl + gear, data = mtcars)
library(performance)
result <- check_collinearity(lm4_1)
plot(result)
```

OK

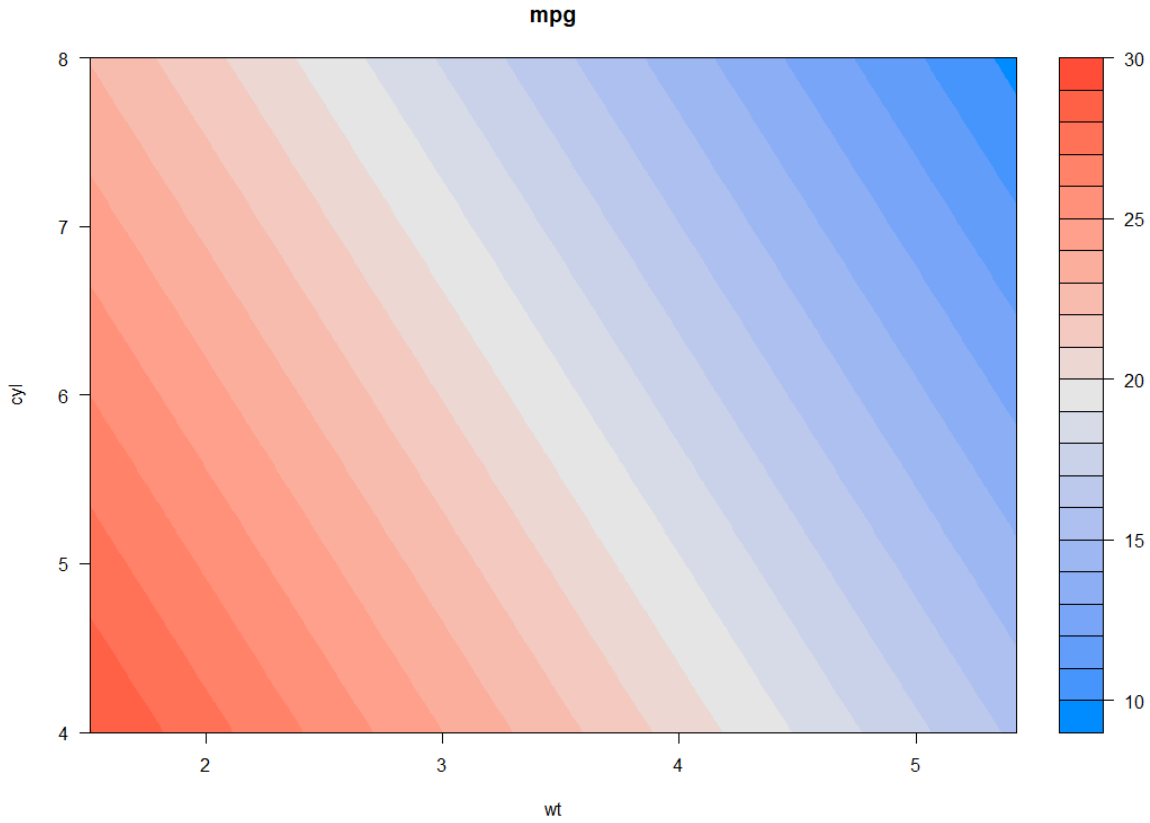
多元回归的作图

lm4_1 <- lm(mpg ~ wt + cyl + gear, data = mtcars)

```
> library(visreg)
> visreg(lm4_1, "wt", gg=TRUE,line=list(col="red"),
+        fill=list(fill="green"),
+        points=list(size=5, pch=16))
```



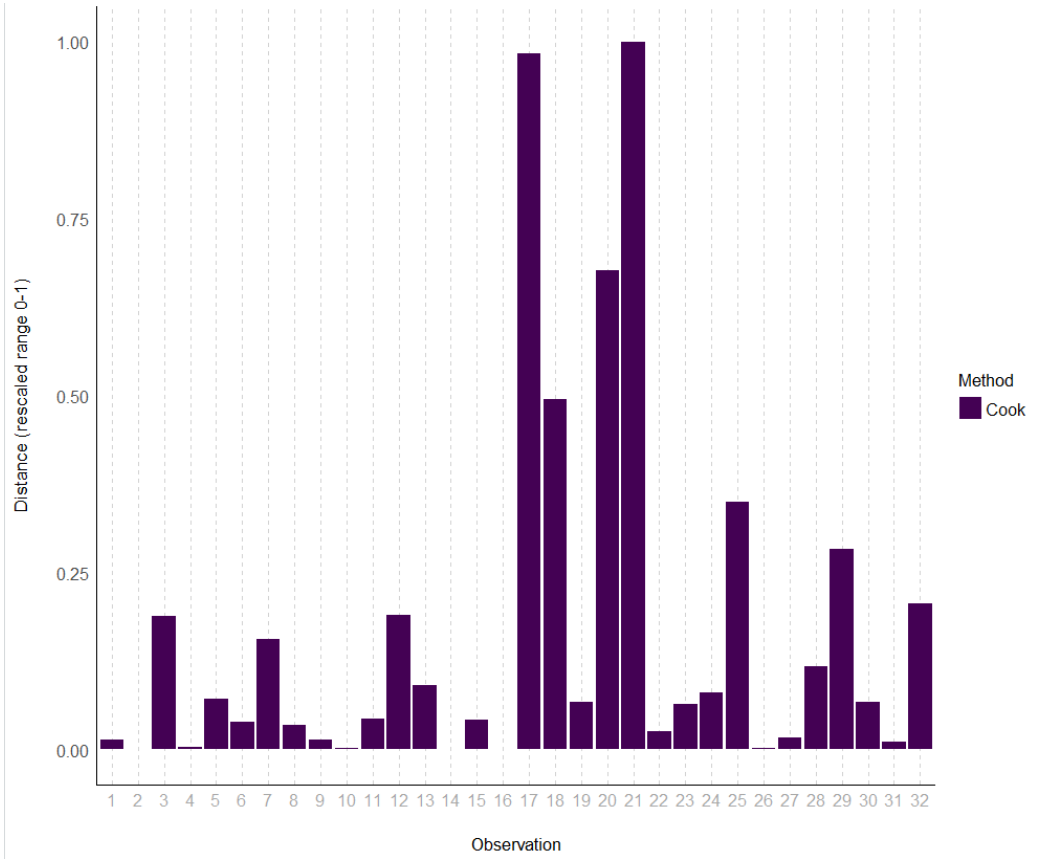

```
> visreg2d(lm4_1, "wt", "cyl")
```



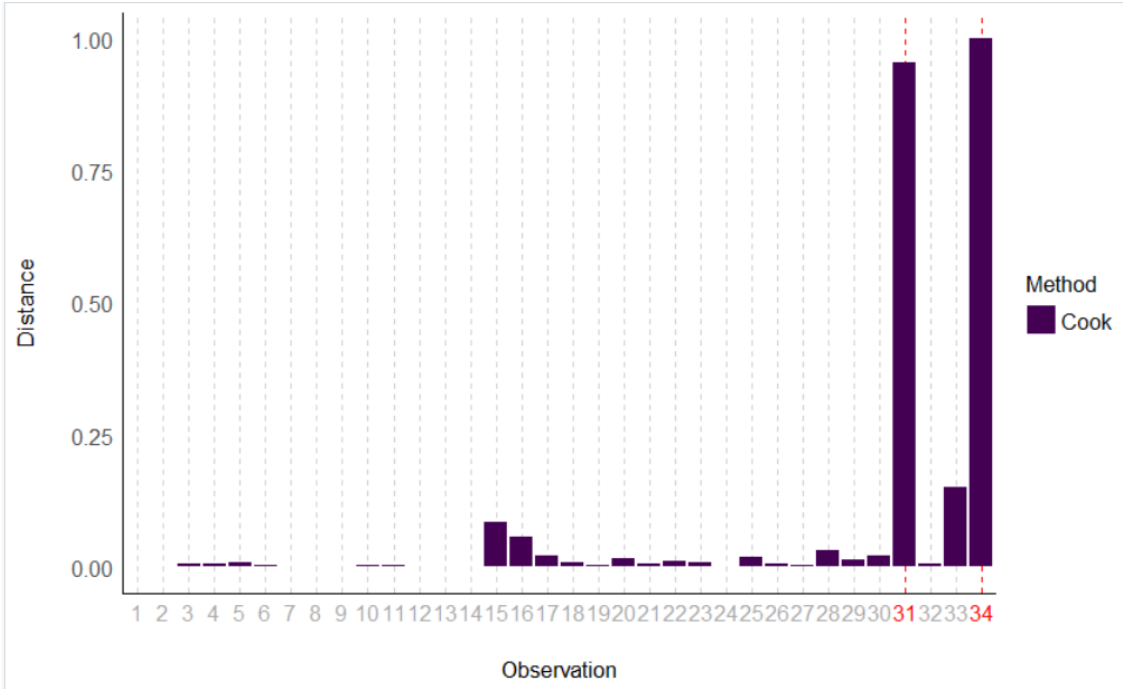
X1, X2, Y图

异常值识别

```
lm5 <- lm(mpg ~ wt + cyl + gear + disp, data = mtcars)
result <- check_outliers(lm5)
plot(result)
```




```
mt1 <- mtcars[, c(1, 3, 4)]
mt2 <- rbind(mt1, data.frame(mpg = c(37, 40), disp = c(300, 400), hp = c(110, 120)))
lm6<- lm(disp ~ mpg + hp, data = mt2)
result <- check_outliers(lm6)
plot(result)
```



案例来源https://easystats.github.io/see/articles/performance.html

异常值，共线性去除之后，即可进入回归模型

线性回归和anova的前提条件

- 正态性（**所指何物？**）
- 方差齐次性
- y_i相互独立

这些问题如何解决？

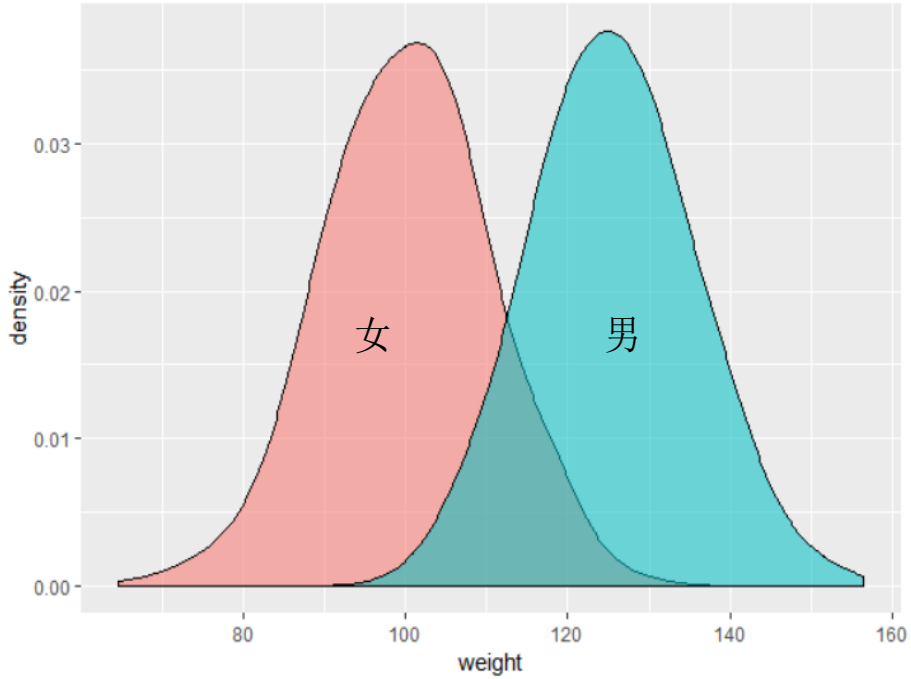
数据分析中的第一大误区—正态性

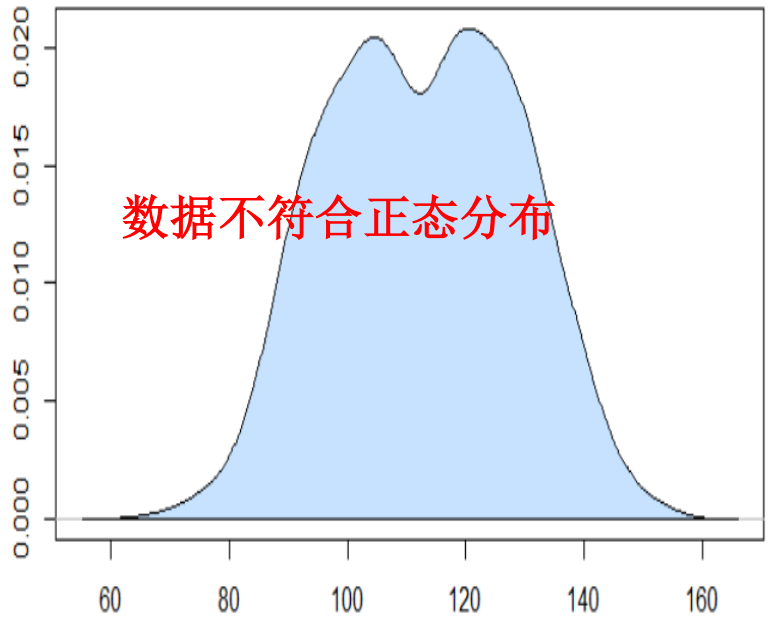
是否指所有原始数据的正态性？

正态性到底所指何物？

是否指全部数据Y要符合正态性？

分析性别对体重的影响






```
> ###对体重的正态性进行检验
> shapiro.test(d1$weight)

        Shapiro-Wilk normality test

data:  d1$weight
W = 0.98875, p-value = 2.198e-11
```

回归或方差分析就不能安排了？

正态性要求的实质是什么

线性模型回归公式$y = b_0 + b_1 x + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$

b0, 常数项，不影响y的分布类型，只影响分布的均值

b1, 常数项，乘以x后，影响y的分布

ε, 误差项，假定符合正态分布，模型才能成立

如果y不符合正态分布那么可能性有两种：

1）由因素x引起，控制x的影响之后，这时y的分布为正态，$\sim N(b_0, \sigma^2)$

2）残差不符合正态分布

- **所以事实上，我们只要求模型的残差符合正态分布，**
- **bx造成的变异,正是自变量x对y的影响，对其分布没有要求**

正态性要求的实质

贝叶斯式模型写法：$y \sim N(b_0 + b_1 x, \sigma^2)$

1) x固定在某一值时， 对应的y服从正态分布；
2) 这与回归模型的残差符合正态分布意义完全相同
3) x变化时，y服从一系列方差相同正态分布，而不是一个正态分布。

正态性假设的实质

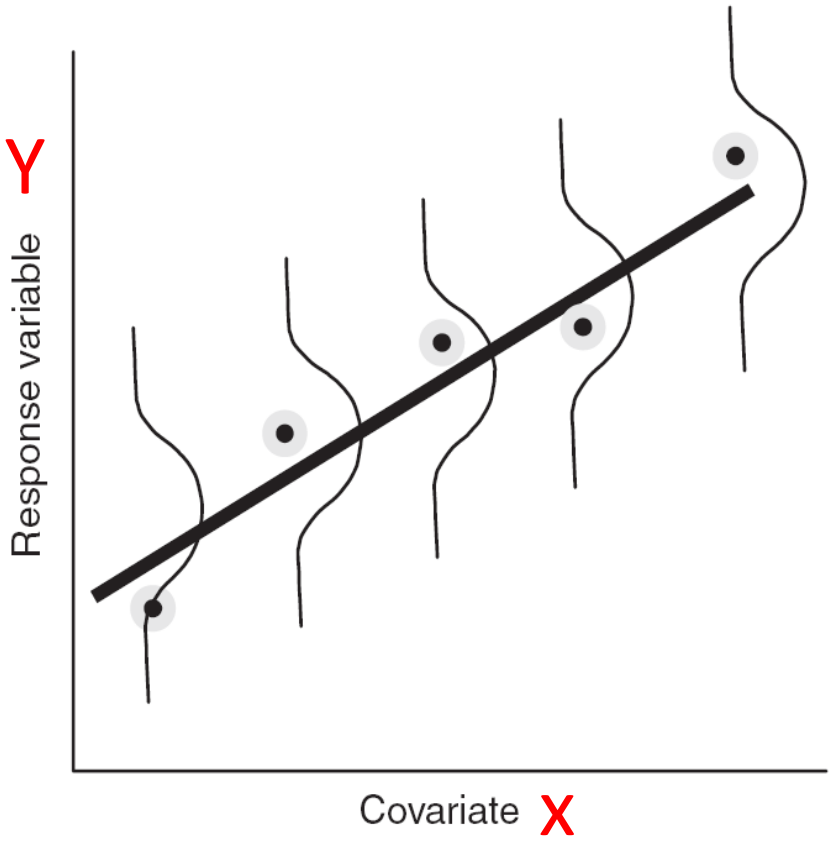
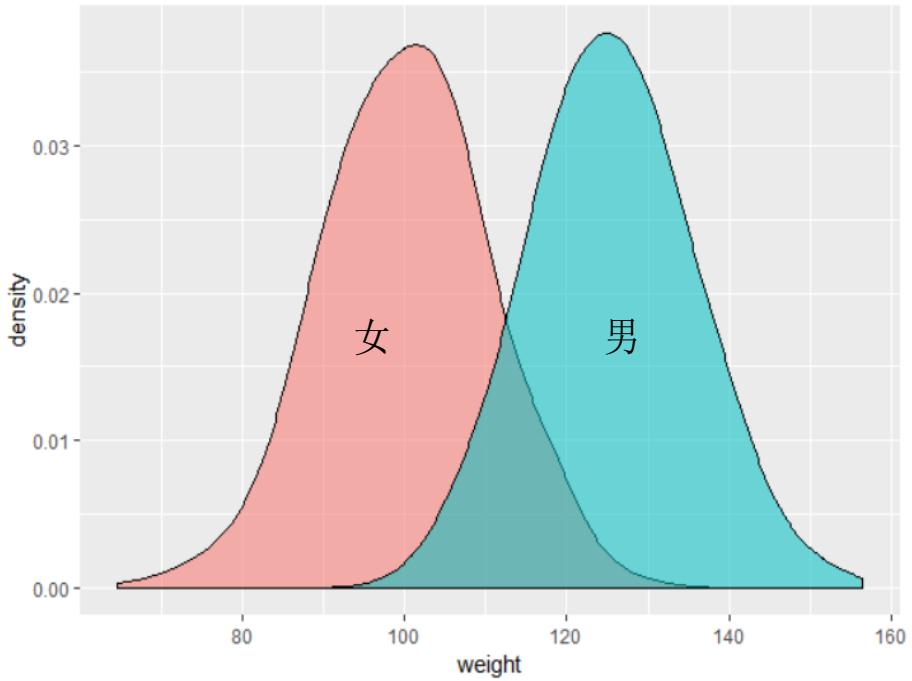


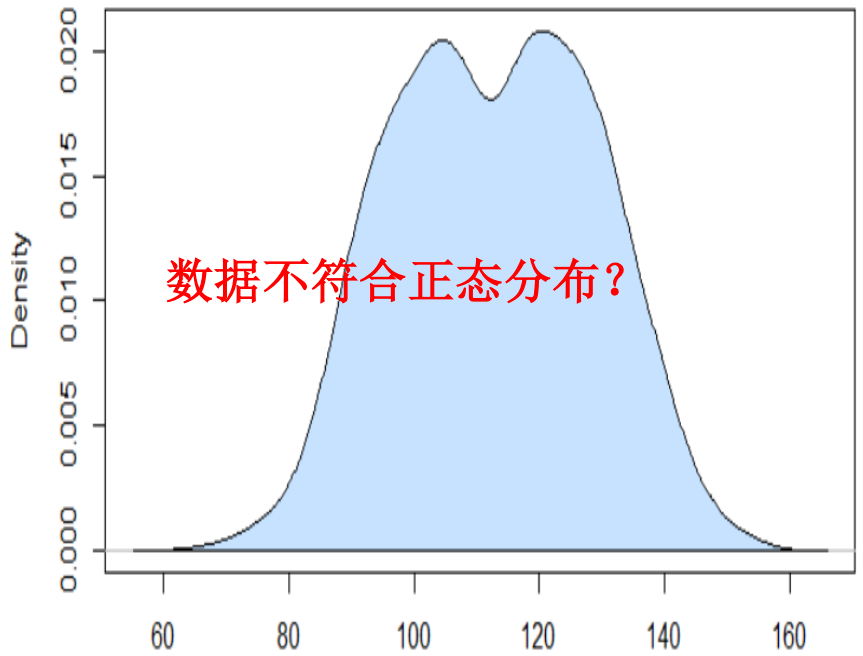
Fig. 6. Visualization of two underlying assumptions in linear regression: normality and homogeneity. The dots represent observed values and a regression line is added. At each covariate value, we assume that observations are normally distributed with the same spread (homogeneity). Normality and homogeneity at each covariate value cannot be verified unless many (>25) replicates per covariate value are taken, which is seldom the case in ecological studies. In practice, a histogram of pooled residuals should be made, but this does not provide conclusive evidence for normality. The same limitations holds if residuals are plotted vs. fitted values to verify homogeneity.

In linear regression, we actually assume normality of all the replicate observations at a particular covariate value.

Zuur, A. F., E. N. Ieno, and C. S. Elphick. 2010. A protocol for data exploration to avoid common statistical problems. Methods in Ecology and Evolution 1:3-14.

来自不同性别的数据，各自符合正态分布或回归模型残差符合正态分布即可





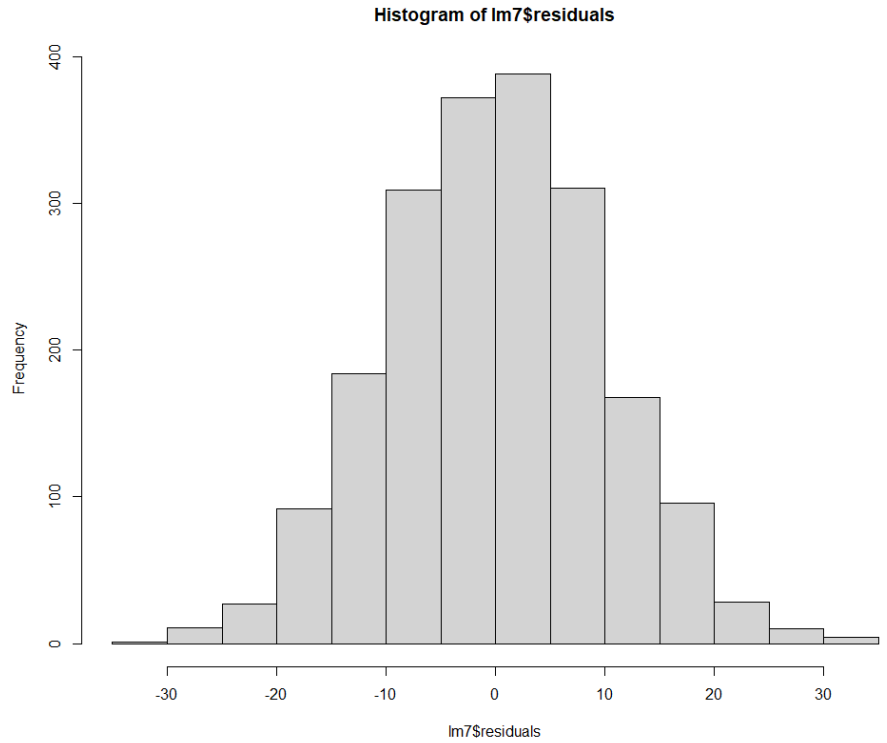
数据并不违反正态性的前提假设

```
lm7<-lm(weight~sex,data=d1)
```

模型诊断--对线性回归模型残差正态性的检验

shapiro.test命令

```
hist(lm7$residuals)
##对lm7模型的残差的正态性检验
shapiro.test(lm7$residuals)
```



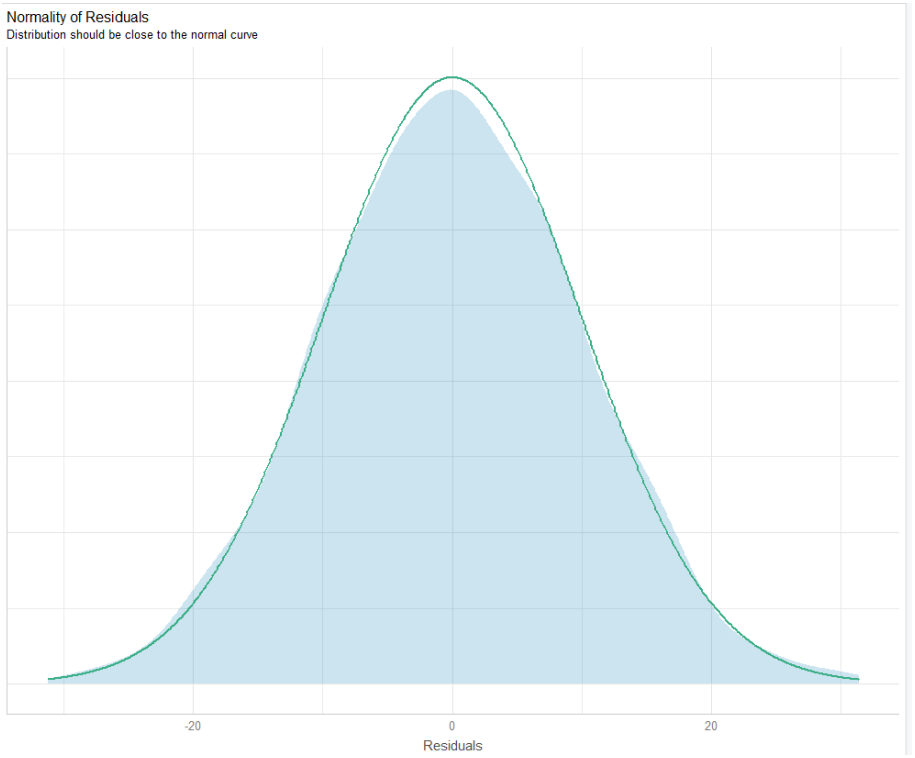

```
> shapiro.test(lm7$residuals)

        Shapiro-Wilk normality test

data:  lm7$residuals
W = 0.99932, p-value = 0.7057
```

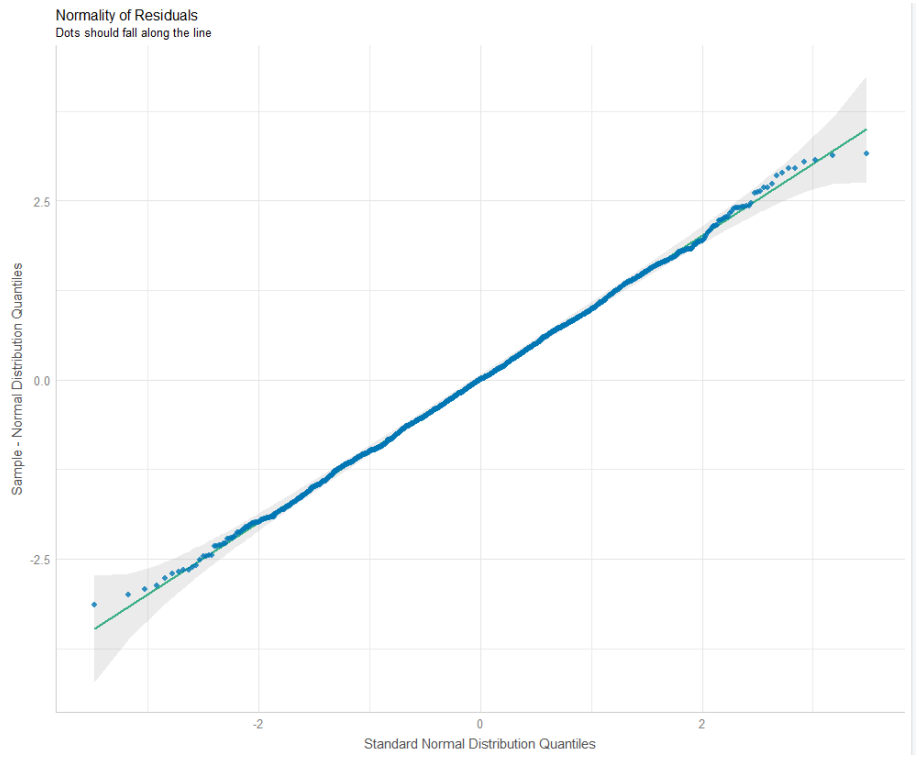
残差符合正态，模型没问题
原始数据的非正态性由性别造成

利用performance包的check_normality函数检验残差正态性

```
> library(performance)
> check_lm7<-check_normality(lm7)
OK: residuals appear as normally distributed (p = 0.706).
> plot(check_lm7)
```

```
plot(check_lm7, type="qq")
```





所以要牢记数据正态性指

模型残差的正态性

而非

所有的原始数据y的正态性

Distributions are often thought of as data summaries, but in the regression context they are more commonly applied to ε's. ***(Gelman and Hill, 2007, p13)***

问题二：方差（残差）异质性问题

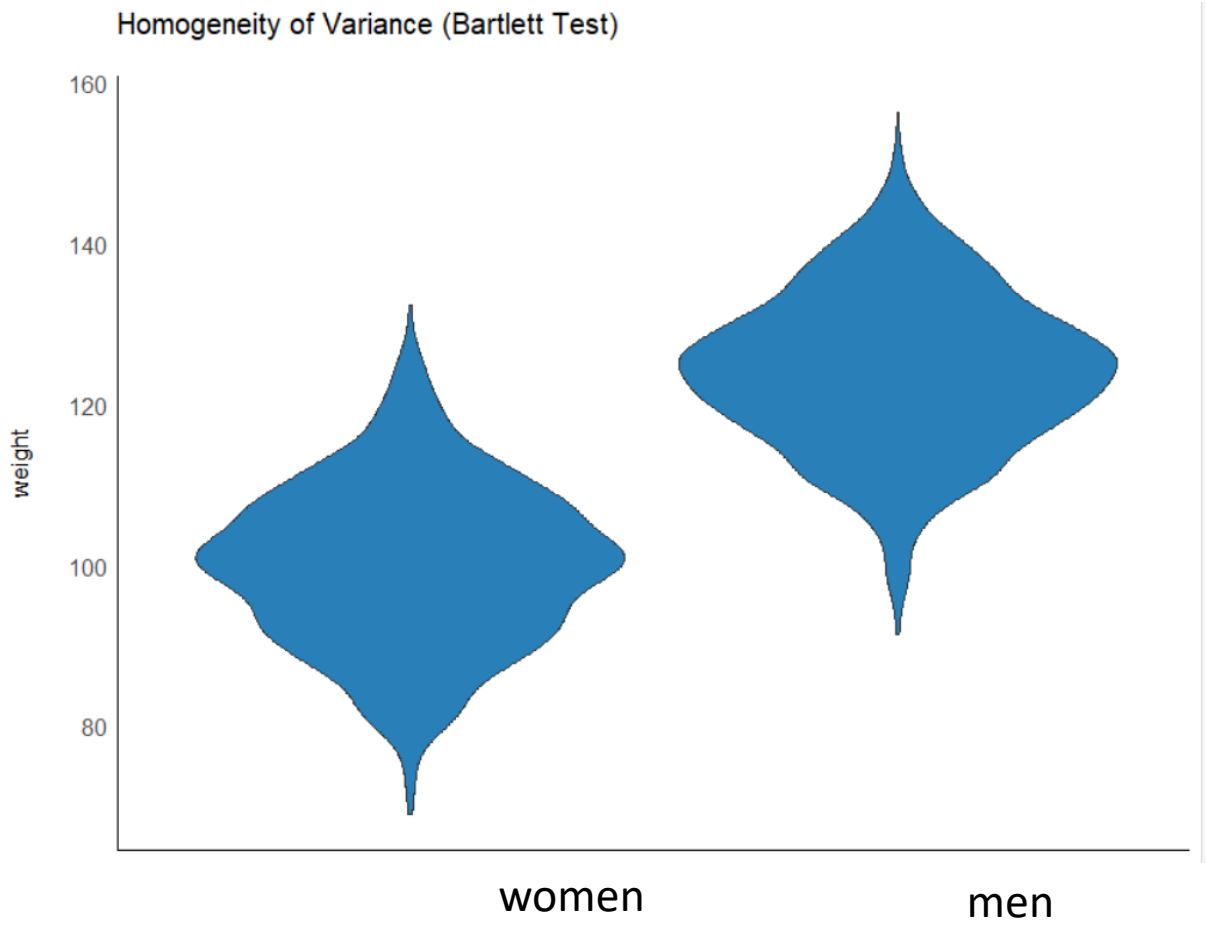
方差异质性（x为分类变量）--Bartlett test

```
> bartlett.test(weight~sex, data =d1)

        Bartlett test of homogeneity of
        variances

data:  weight by sex
Bartlett's K-squared = 1.5835, df = 1,
p-value = 0.2083
```

方差异质性（x为分类变量）

- 模型要求：Y的方差在不同的x值间，没有明显的异质性

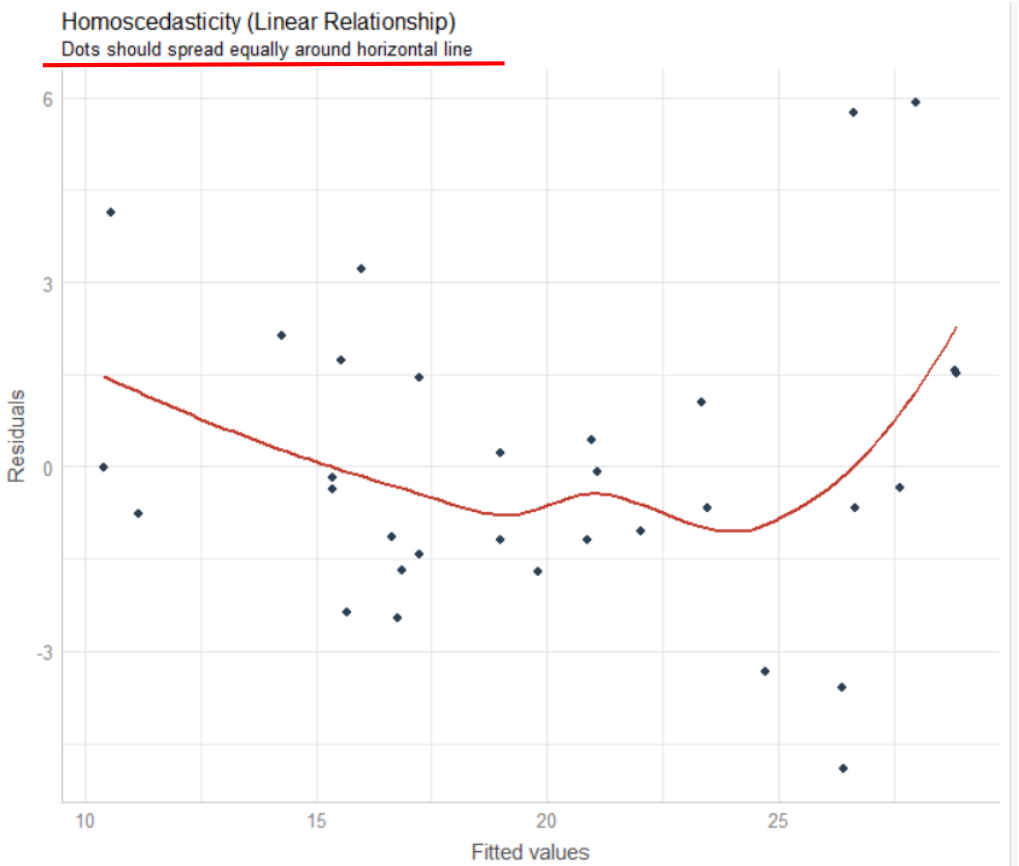
```
> lm7<-lm(weight~sex,data=d1)
> result<-check_homogeneity(lm7)
OK: Variances in each of the groups are the same (Bartlett Test, p = 0.208).
> plot(result)
```



方差（残差）异质性（x为数值变量）

要求：残差和拟合值之间无明显关系

```
> #### x为数值型变量时的残差异质性检验
> lm4_1 <- lm(mpg ~ wt + cyl + gear, data = mtcars)
> result<-check_heteroscedasticity(lm4_1)
Warning: Heteroscedasticity (non-constant error variance) detected (p = 0.047).
> plot(result)
`geom_smooth()` using formula 'y ~ x'
```



违反线性模型前提之一

违反正态性原则

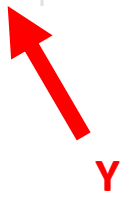
但是，模型残差符合正态分布通常只是一种理想情况，现实数据大多并不理想

那么我们到底该怎么办？

1) 对y进行转化，to improve the normality of residuals!
2) 有些数据天生无赖，无法将其(残差)转化为正态性
3) 认识你的数据的理论分布，据此选择模型，至关重要。

有的数据：天生无赖

Sample	Richness	Exposure	NAP	Beach
1	11	10	0.045	1
2	10	10	-1.036	1
3	13	10	-1.336	1
4	11	10	0.616	1
5	10	10	-0.684	1
6	8	8	1.190	2
7	9	8	0.820	2
8	8	8	0.635	2
9	19	8	0.061	2
10	17	8	-1.334	2
11	6	11	-0.976	3
12	1	11	1.494	3
13	4	11	-0.201	3
14	3	11	-0.482	3
15	3	11	0.167	3
16	1	11	1.768	4

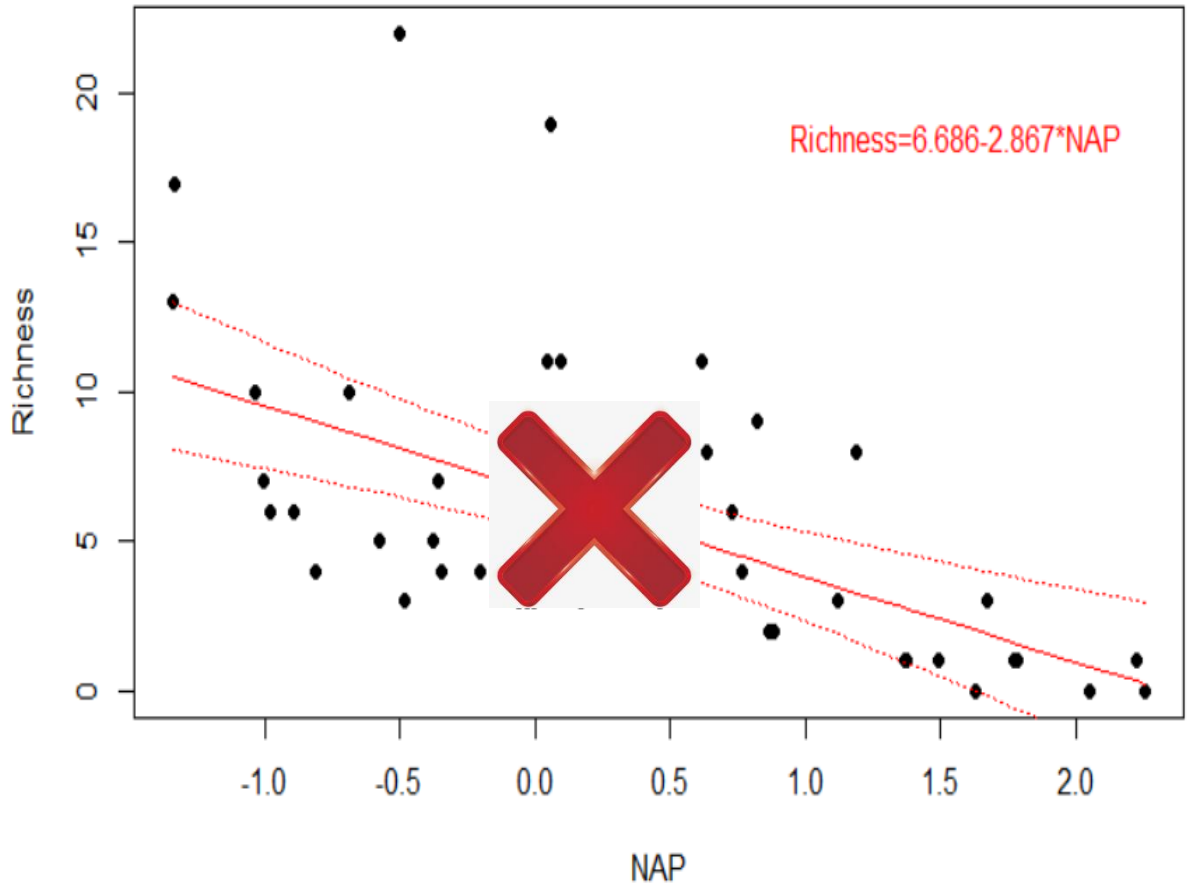


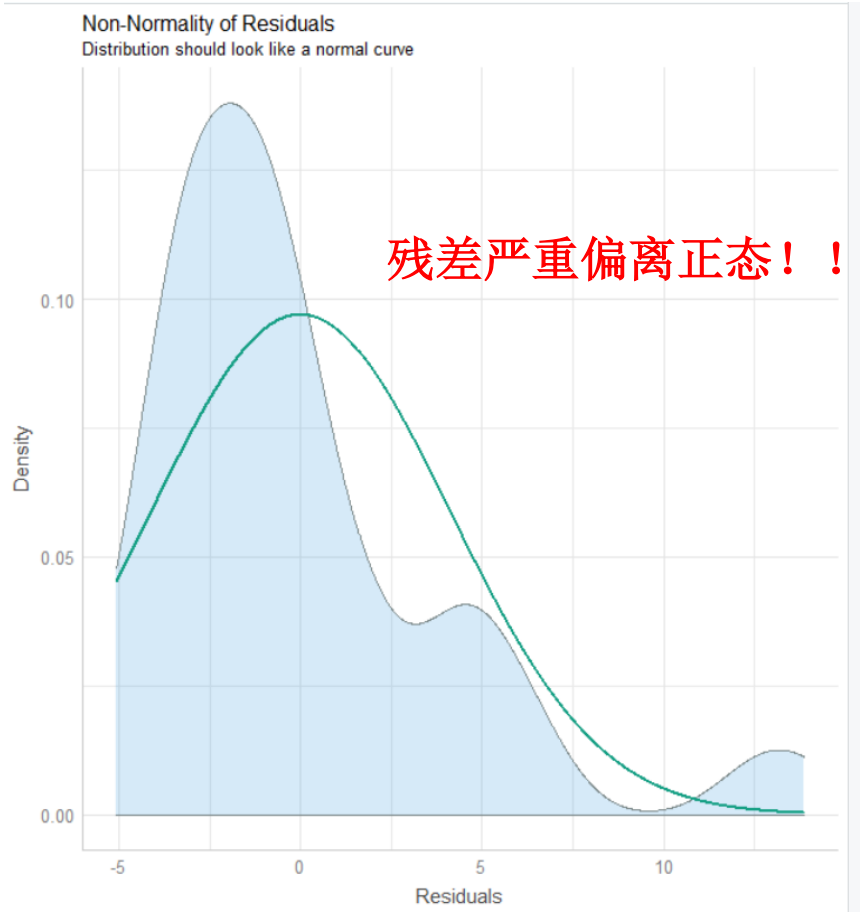
RIKZ.csv

如计数数据

有的数据：天生无赖

```
> lm0<-lm(Richness~NAP,data=RIKZ)
> plot(check_normality(lm0))
Warning: Non-normality of residuals detected (p < .001).
```





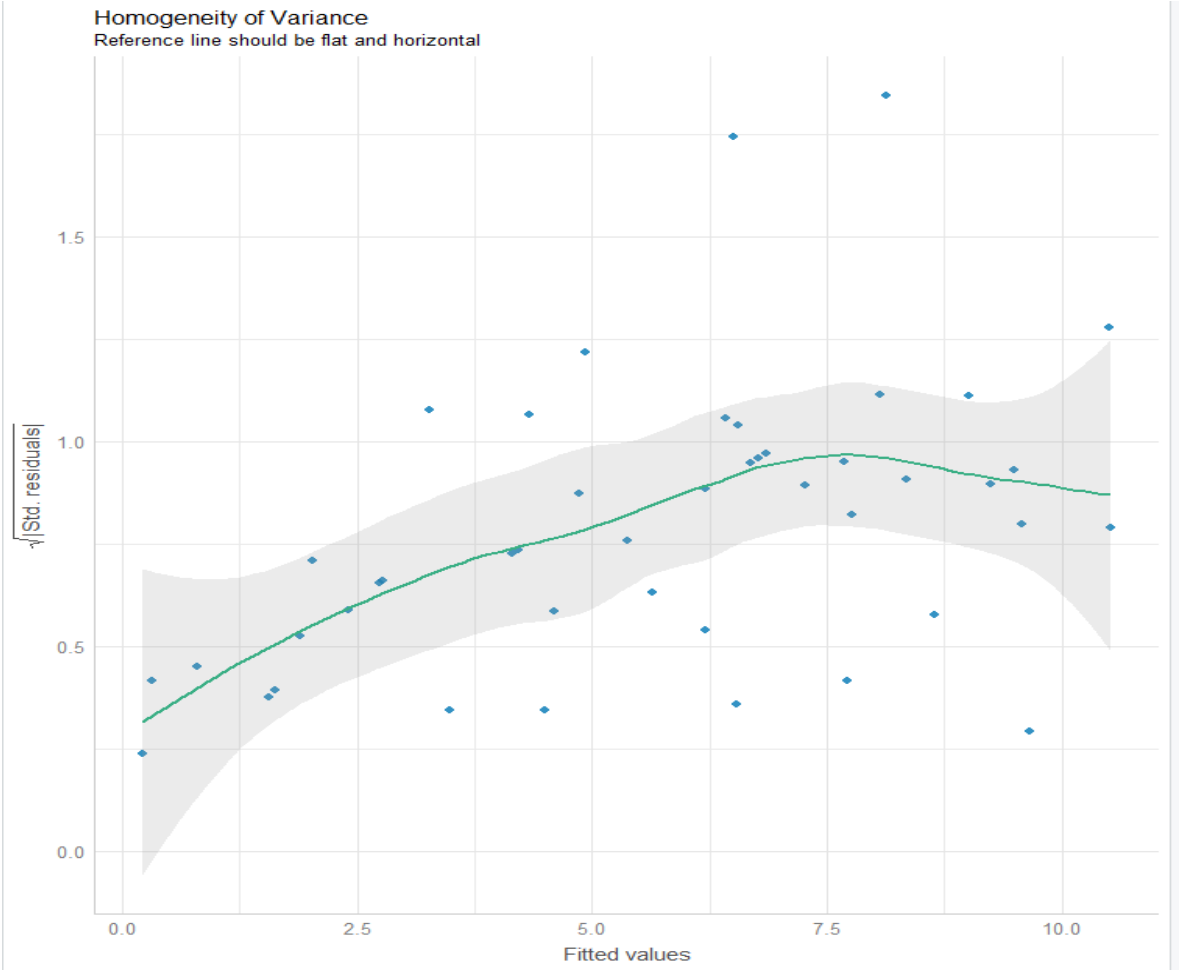
残差严重偏离正态！！！

有的数据：天生无赖

```
> check_heteroscedasticity(lm0)
Warning: Heteroscedasticity (non-constant error variance) detected (p = 0.015).
```

lm0 <- lm(Richness~NAP,data=RIKZ)

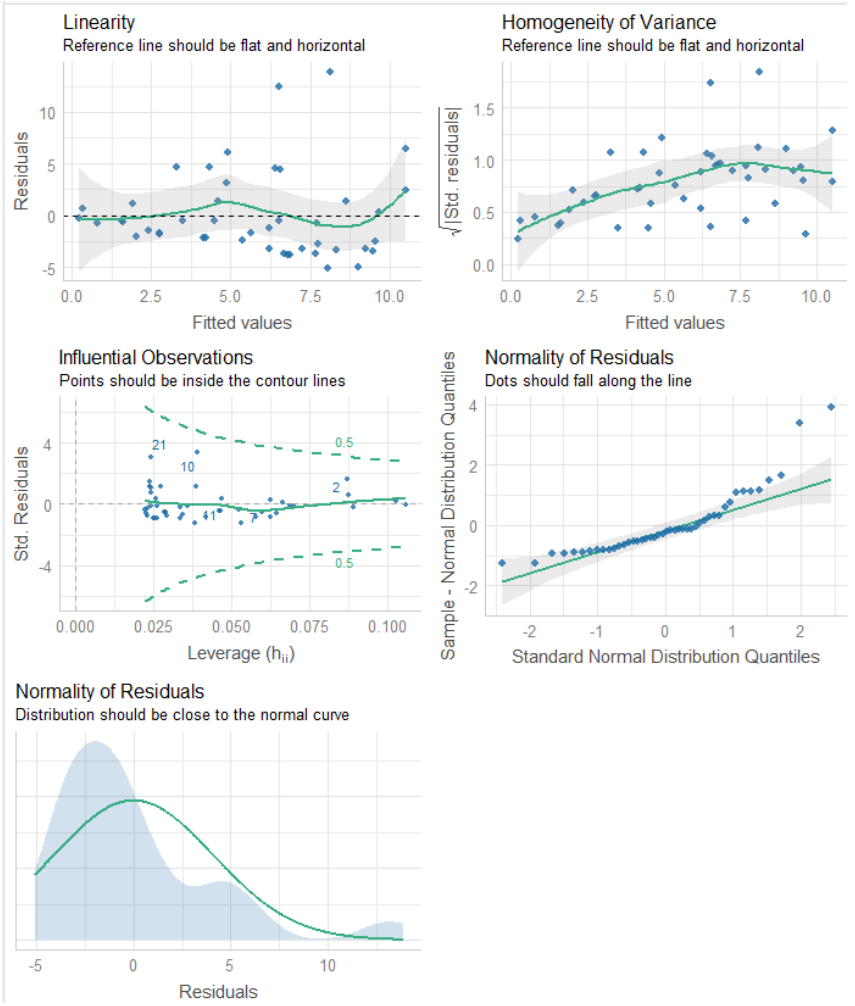
对于这个数据集来讲
线性模型不是一个好模型



有的数据：天生无赖

```
lm0 <- lm(Richness~NAP,data=RIKZ)

check_model(lm0)
```

lm0模型，不是个好模型！
怎么办？数据转换？



数据转换—是有争议的

- Warton, D. I. and F. K. C. Hui. 2011. *The arcsine is asinine: the analysis of proportions in ecology*. ***Ecology*** **92**:3-10.
- O'Hara, R. B. and D. J. Kotze. 2010. *Do not log-transform count data*. ***Methods in Ecology and Evolution*** **1**:118-122.
- Ives, A. R. 2015. *For testing the significance of regression coefficients, go ahead and log-transform count data*. ***Methods in Ecology and Evolution*** **6**:828-835.
- Xiao, X., E. P. White, M. B. Hooten, and S. L. Durham. 2011. *On the use of log-transformation vs. nonlinear regression for analyzing biological power laws*. ***Ecology*** **92**:1887-1894.

数据转换—是有争议的

- Warton, D. I. and F. K. C. Hui. 2011. *The arcsine is asinine: the analysis of proportions in ecology*. ***Ecology*** **92**:3-10.
- O'Hara, R. B. and D. J. Kotze. 2010. *Do not log-transform count data.* ***Methods in Ecology and Evolution*** **1**:118-122.
- Ives, A. R. 2015. *For testing the significance of regression coefficients, go ahead and log-transform count data*. ***Methods in Ecology and Evolution*** **6**:828-835.
- Xiao, X., E. P. White, M. B. Hooten, and S. L. Durham. 2011. *On the use of log-transformation vs. nonlinear regression for analyzing biological power laws*. ***Ecology*** **92**:1887-1894.

广义线性模型—专治各种天生无赖，
无需对y的原始数据进行转换

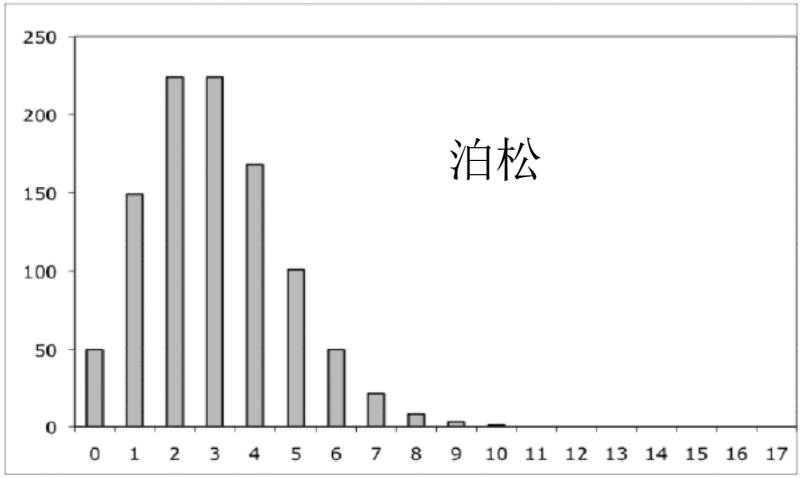
依据数据类型选取合适的回归模型

数据类型	案例	理论分布	对应模型
连续变量	如身高、体重、血压、作物产量等，可以对数据进行log，logit等装换，以促进其正态性	正态分布	一般线性模型
计数数据	物种个数，病人个数等，可能存在过度离散或零膨胀问题	泊松分布（均值=方差）	泊松回归 伪泊松回归 负二项回归 零膨胀泊松回归 零膨胀负二项回归
0,1数据	成功/失败，男/女，是否出现	二项分布	二项回归
由计数数据转化而来的比率数据	成功率，康复率（可能存在过度离散或零膨胀）	二项分布	二项回归 伪二项回归（过度离散） 零膨胀二项回归
非负数据，但0较多	每个人逛街的花费		Hurdle模型
连续数据型比率数据	土壤含水率，空气湿度	Beta分布	Beta回归（也可logit转换后线性回归）
连续变量[0，正无穷）	生物量(不可能小于0)	Gamma分布（均值越大，方差越大）	Gamma回归

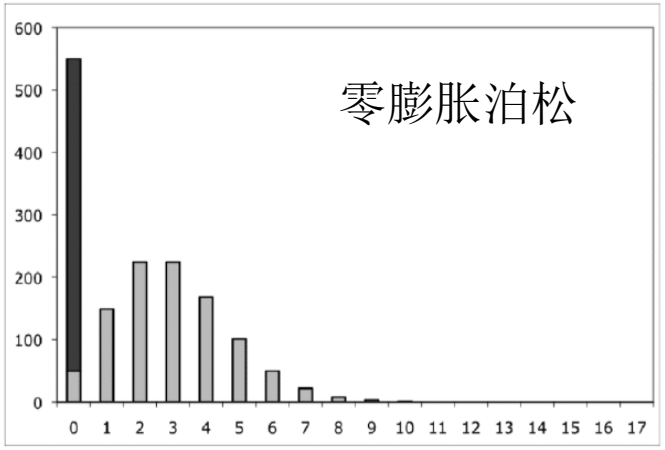
广义线性模型

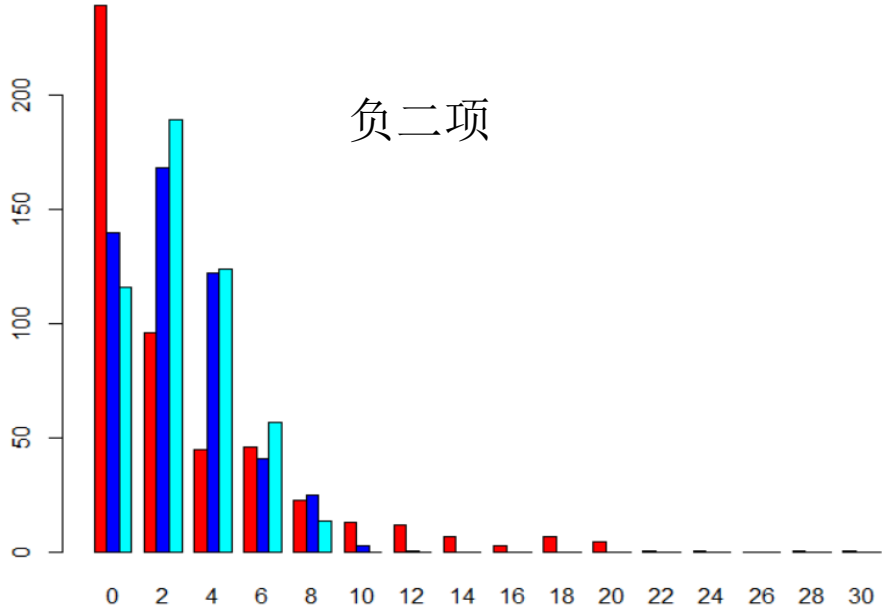
认识你的数据：计数数据

a) Data appropriate for Poisson model (n=1000, Mean=3, Variance=3)

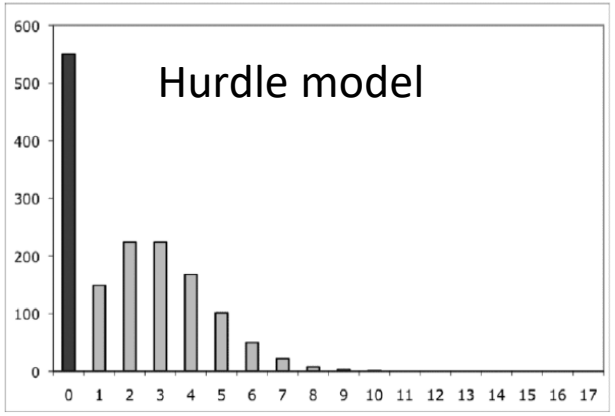


b) Data appropriate for Zero-Inflated Poisson model (n=1500, with 500 "structural" and 50 "sampling" zeros)



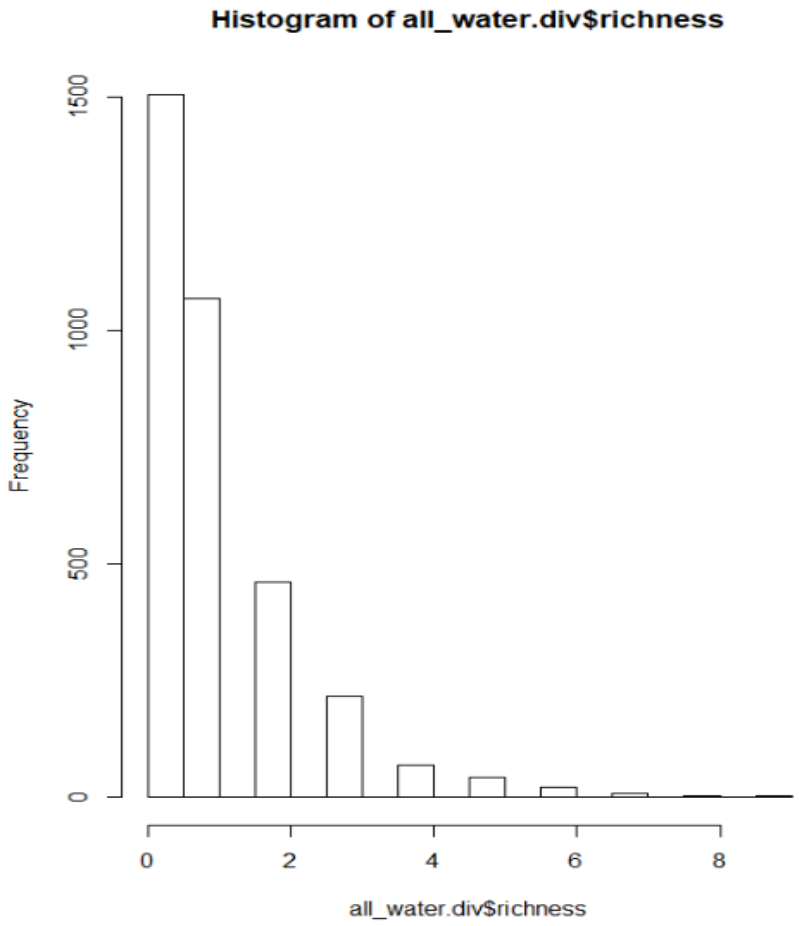


c) Data appropriate for Poisson Hurdle model (n=1500, with 550 "structural" and no "sampling" zeros)



Hu, M.-C., M. Pavlicova, and E. V. Nunes. 2011. Zero-Inflated and Hurdle Models of Count Data with Extra Zeros: Examples from an HIV-Risk Reduction Intervention Trial. The American Journal of Drug and Alcohol Abuse 37:367-375.

各种分布案例

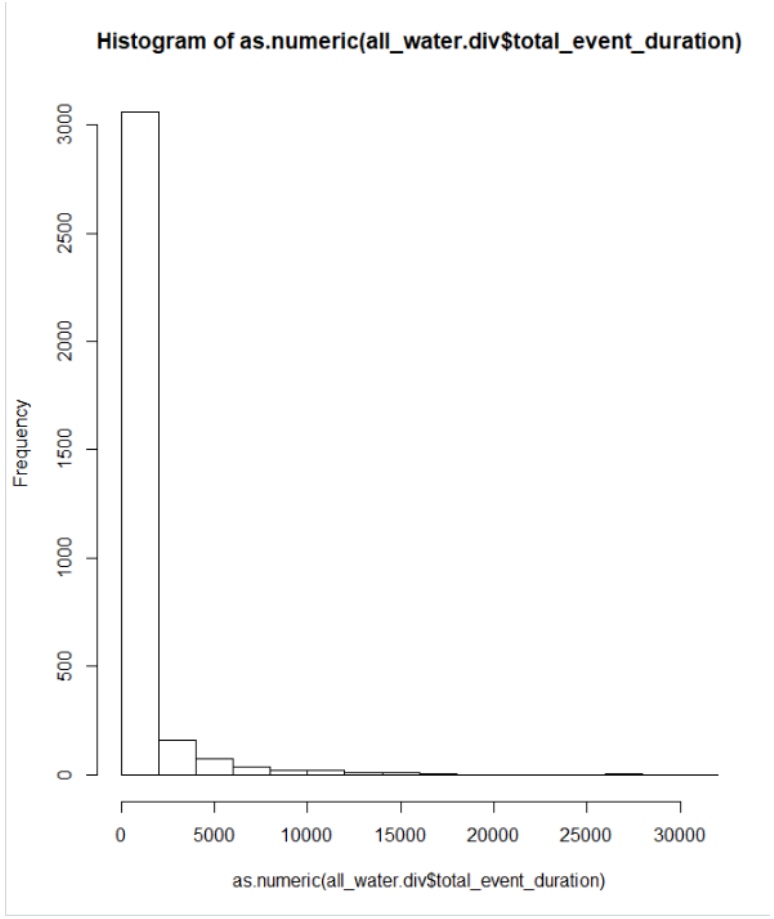


Zero inflated poisson

We chose distributions appropriate for each response variable

Nogué, S. et al. The human dimension of biodiversity changes on islands. Science 372, 488-491,doi:10.1126/science.abd6706 (2021).

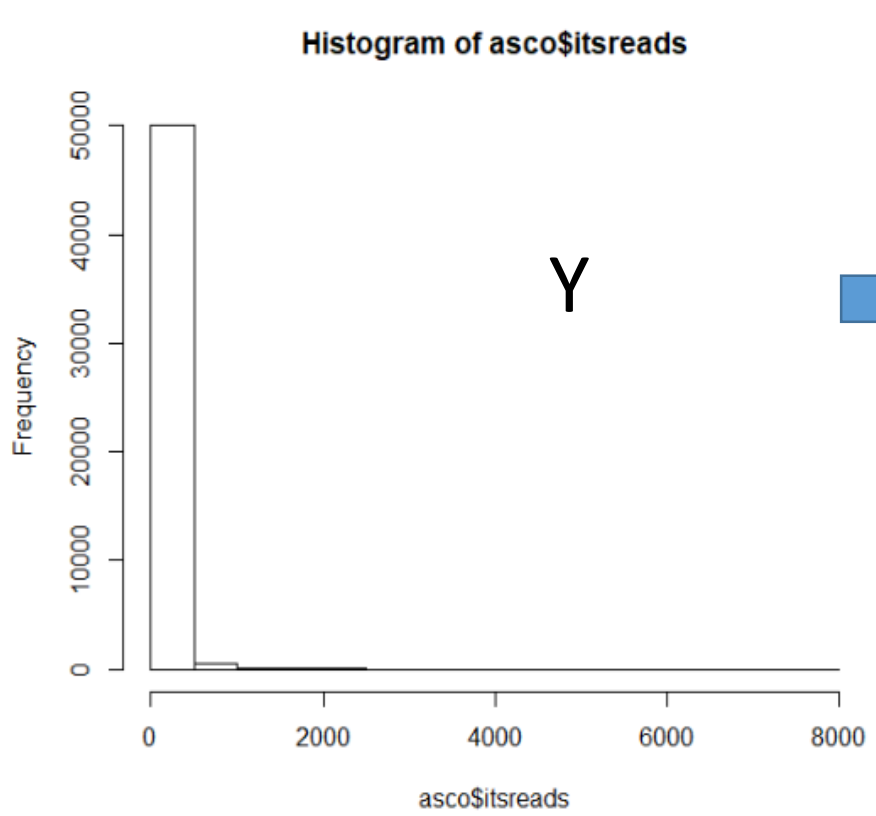
各种分布案例



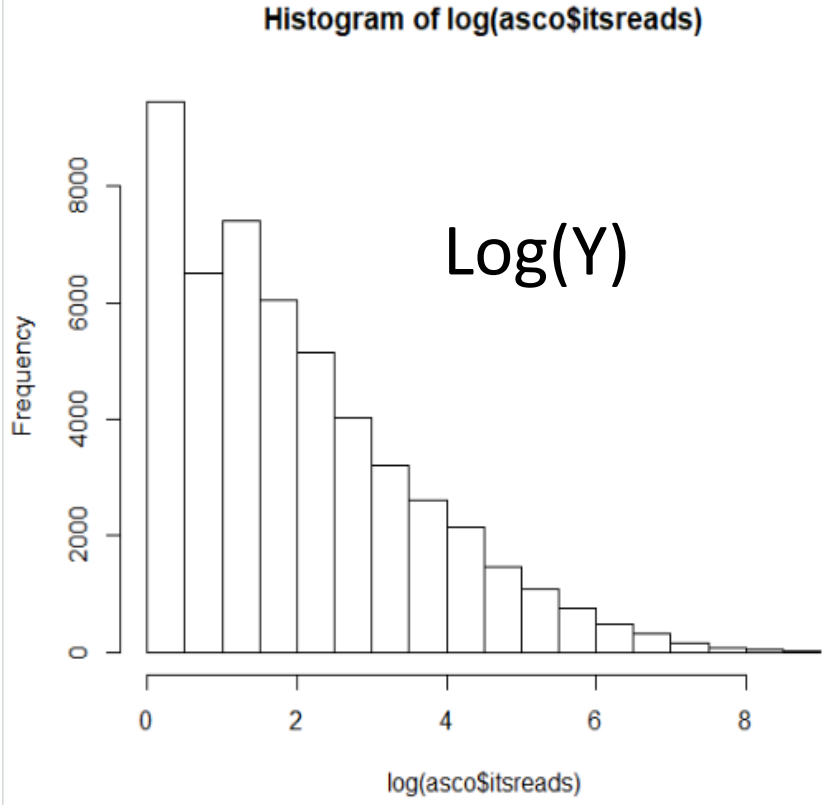
Zero inflated negative binomial

Nogué, S. et al. The human dimension of biodiversity changes on islands. Science 372, 488-491,doi:10.1126/science.abd6706 (2021).

各种分布案例：gamma分布
非负，连续变量，大多数值较小，少数较大，上不封顶



Y

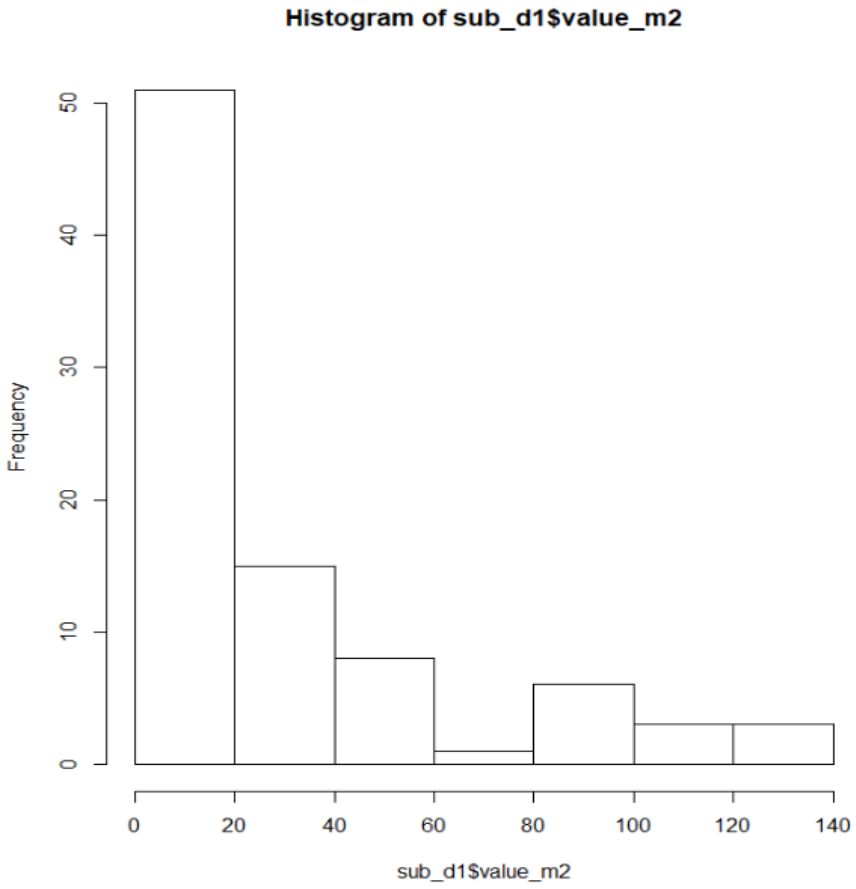


Log(Y)

群落中不同物种的生物量

数据来源： Collins, C. G. *et al.* Belowground impacts of alpine woody encroachment are determined by plant traits, local climate, and soil conditions. *Global Change Biol.* **26**, 7112-7127, doi:https://doi.org/10.1111/gcb.15340 (2020).

各种分布案例

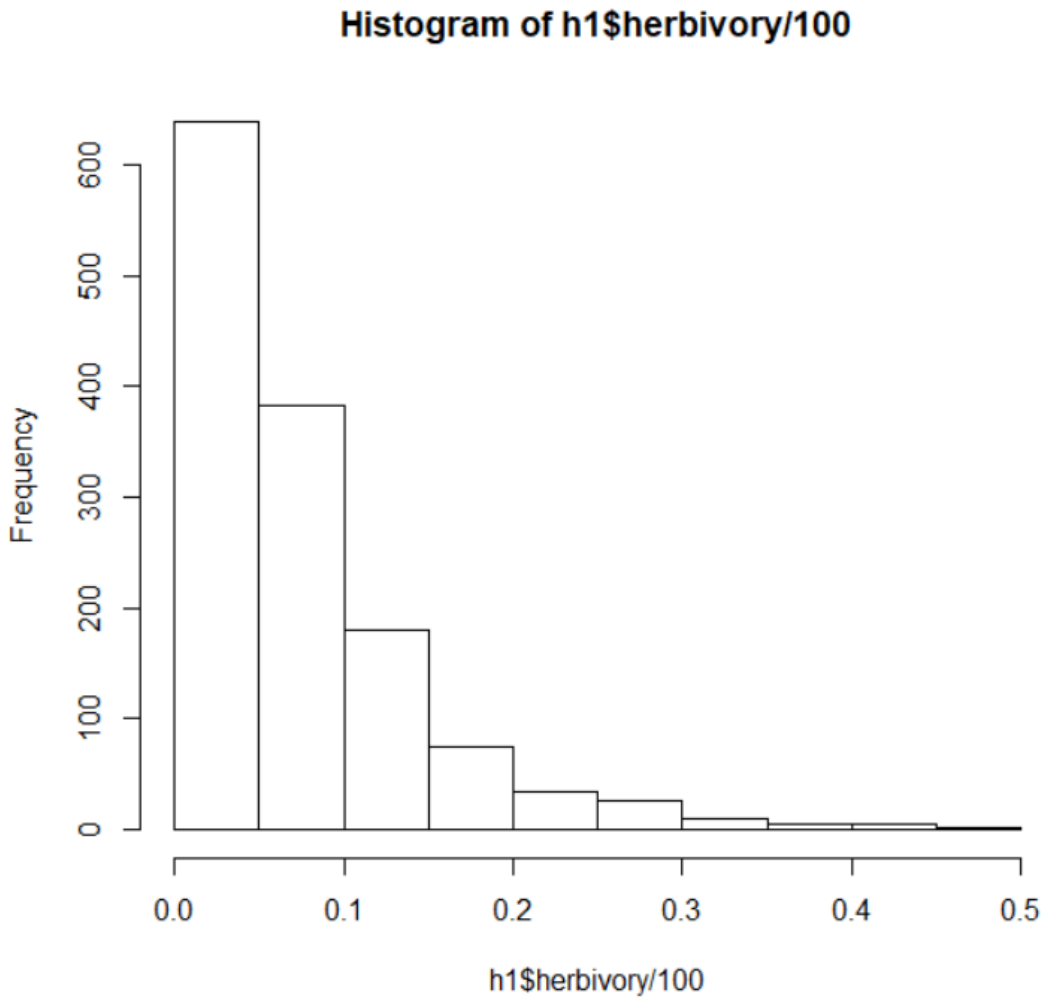


Zero inflated gamma

Nogué, S. et al. The human dimension of biodiversity changes on islands. Science 372, 488-491,doi:10.1126/science.abd6706 (2021).

认识你的数据：beta分布

连续型比率数据（非计数数据转化而来）



植物叶片虫食面积比率数据分布格局

一般线性 VS 广义线性模型

- 一般线性模型只是广义线性模型的特列
- 广义线性模型，实质是一种对参数进行了转换的线性模型

一般线性模型：$y = b_0 + b_1 x + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$

广义线性模型：**E(y)=u, link(u)=a+bx**

注意：是对y的期望值进行了转换，而不是y本身

广义线性模型--generalized linear model

- 一般线性模型： y=a+bx
- 广义线性模型： **E(y)=u, link(u)=a+bx**

分布类型	link函数(link="")
Binomial （0,1） （含伯努利分布）	logit
Gaussian （正态）	identity
Poisson （计数）	log
Quasi-poisson （计数,且过度离散）	log
Quasi-binomial (由计数数据转化而来的比率数据，且过度离散)	logit
零膨胀Poisson	泊松部分log, 零膨胀部分logit
Gamma分布	Inverse 或 log

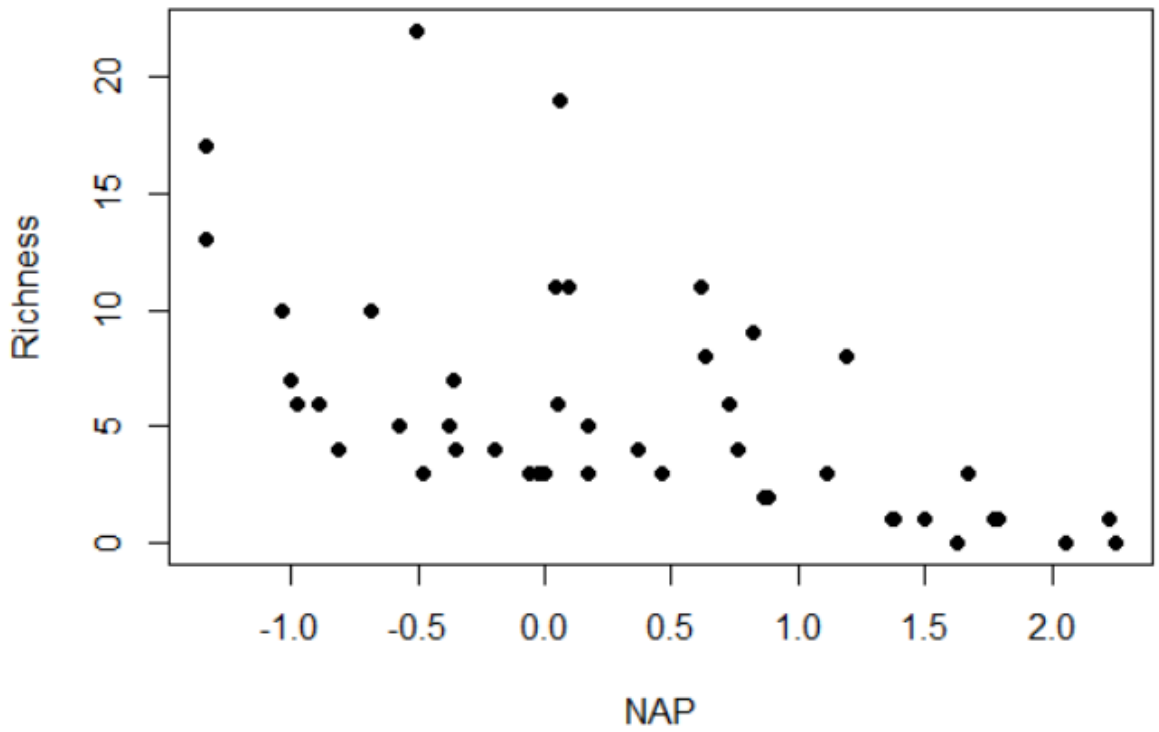
广义线性模型案例（glm）poisson回归（计数数据）

##采用glm模型分析上述数据

```
pm1<- glm(Richness~NAP, family="poisson", data=d3)
summary(pm1)
```

```
Call:
glm(formula = Richness ~ NAP, family = "poisson", data = RIKZ)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2029  -1.2432  -0.9199   0.3943   4.3256

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.79100    0.06329  28.297  < 2e-16 ***
NAP         -0.55597    0.07163  -7.762 8.39e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 179.75  on 44  degrees of freedom
Residual deviance: 113.18  on 43  degrees of freedom
AIC: 259.18

Number of Fisher Scoring iterations: 5
```

```
> rsq(pm1)
[1] 0.3044181
```



$$Log(Richness)=1.791-0.556*NAP$$

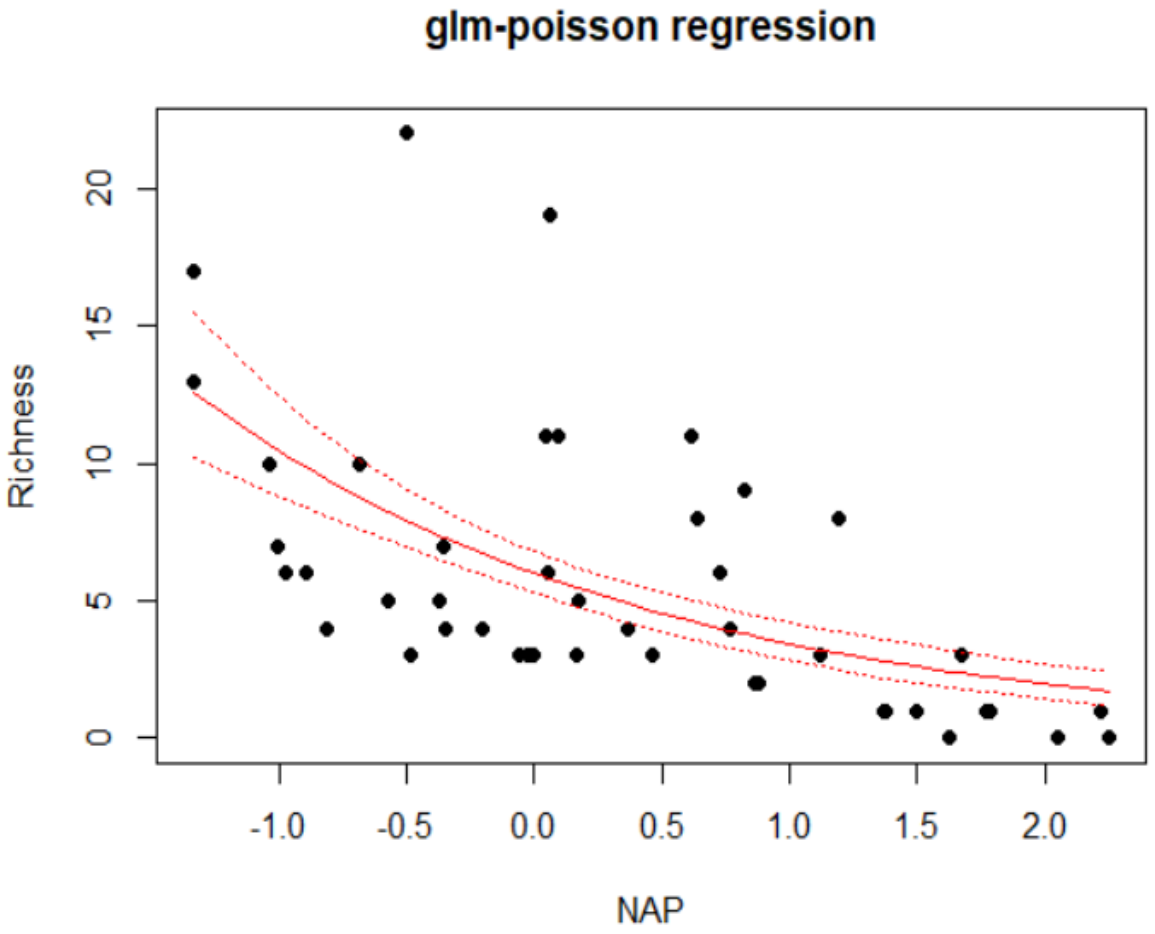
$$Richness=e^{(1.791-0.556*NAP)}$$

如何描述这一模型：

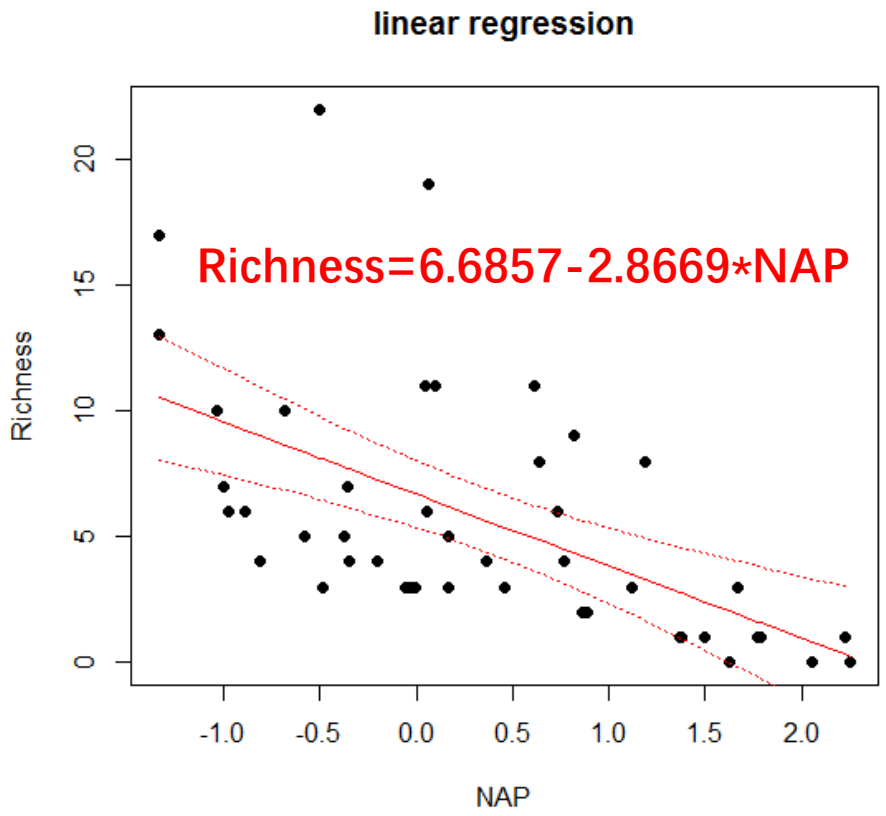
We used generalized linear regression model with poisson error structure and log link function to evaluated the effect NAP on Richness.

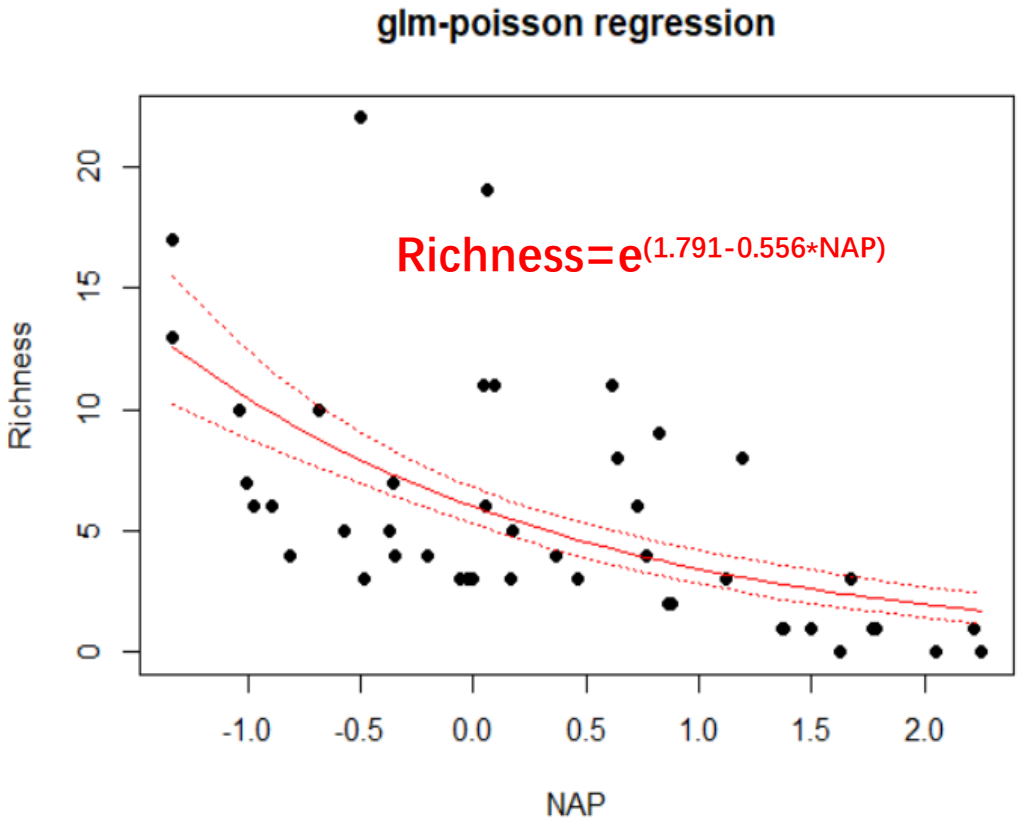
glm--poisson回归结果制图

```
pm1<- glm(Richness~NAP, family="poisson", data=d3)
RIKZ<-RIKZ[order(RIKZ$NAP),]
plot(Richness~NAP,data=RIKZ, pch=19)
ci_pm1<-add_ci(RIKZ, pm1, alpha = 0.05, names = c("lwr", "upr"))
lines(ci_pm1$NAP, ci_pm1$pred, col = "red", lwd = 1,lty="solid")
lines(ci_pm1$NAP, ci_pm1$upr, col = "red", lwd = 1,lty="dotted")
lines(ci_pm1$NAP, ci_pm1$lwr, col = "red", lwd = 1,lty="dotted")
title("glm-poisson regression")
```



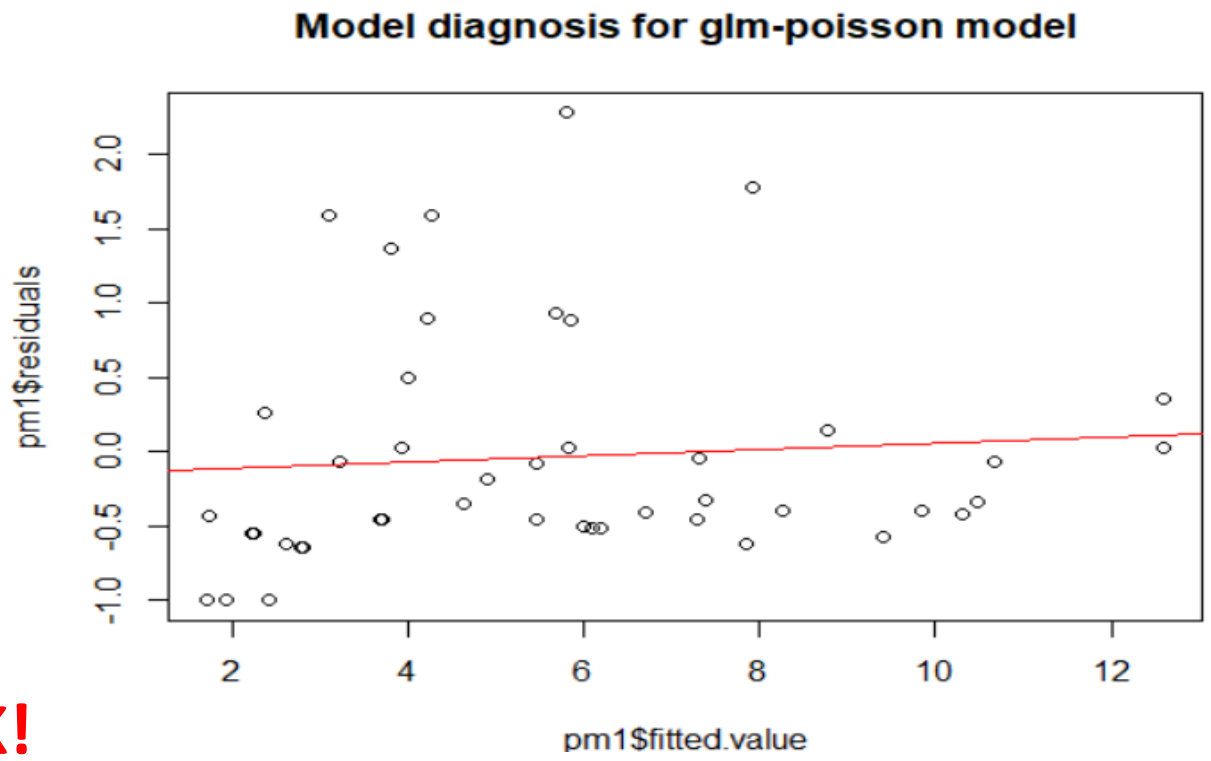
glm-Poisson回归与线性(lm)回归结果对比





glm-poisson回归的模型诊断

```
> plot(pm1$residuals~pm1$fitted.value,main="Model diagnosis for glm-poisson model")
> abline(lm(pm1$residuals~pm1$fitted.values),col="red")
> summary(lm(pm1$residuals~pm1$fitted.values))

Call:
lm(formula = pm1$residuals ~ pm1$fitted.values)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9033 -0.4760 -0.3269  0.1083  2.3068

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.14597    0.25500  -0.572    0.570
pm1$fitted.values  0.02029    0.03986   0.509    0.613

Residual standard error: 0.7821 on 43 degrees of freedom
Multiple R-squared:  0.005991,  Adjusted R-squared:  -0.01713
F-statistic: 0.2592 on 1 and 43 DF,  p-value: 0.6133
```



残差与拟合值之间已无关系，OK!

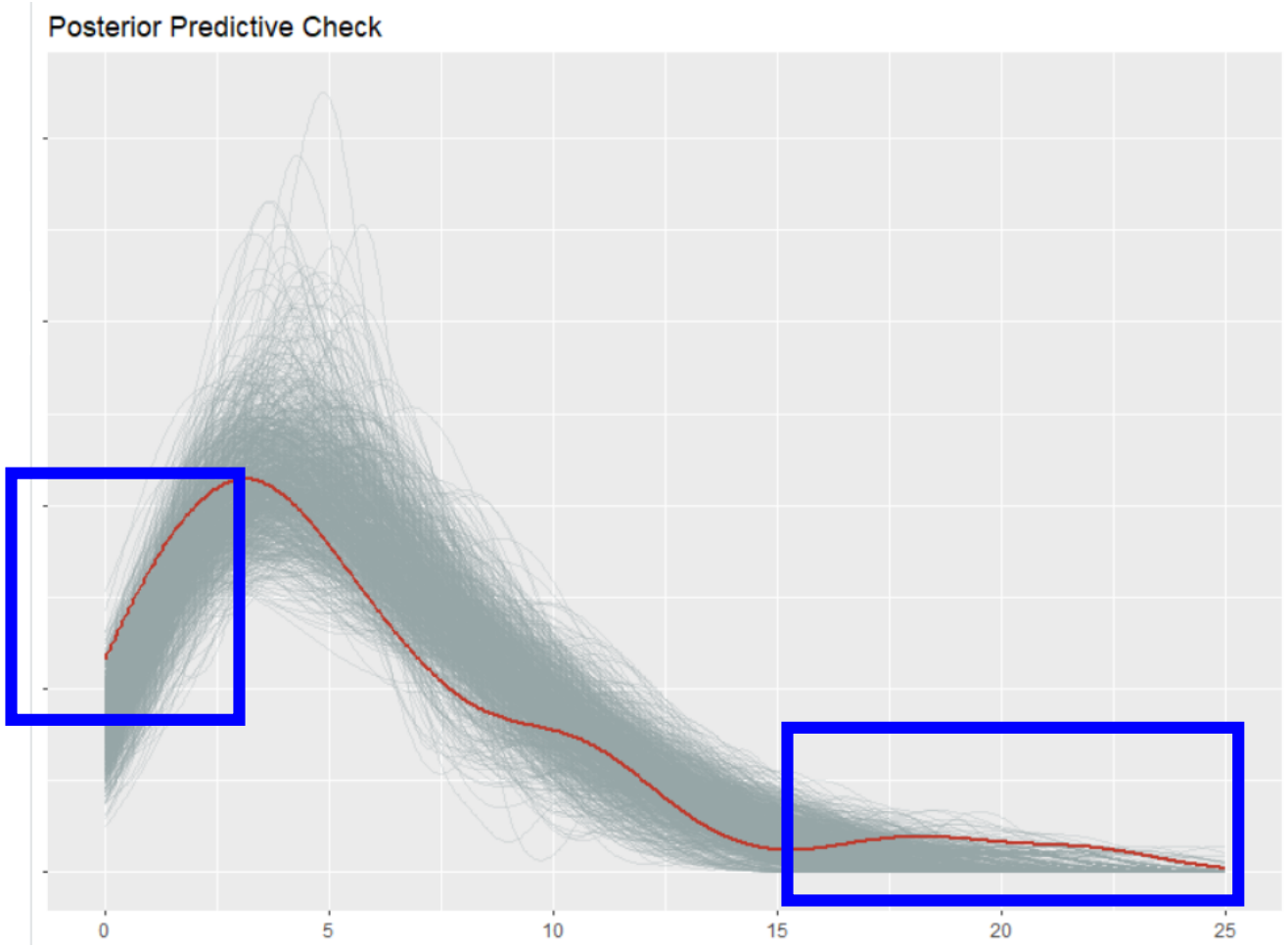
这时，已无需对残差的正态性进行检验

We used generalized linear regression model with poisson error structure and log link function to evaluated the effect NAP on Richness.

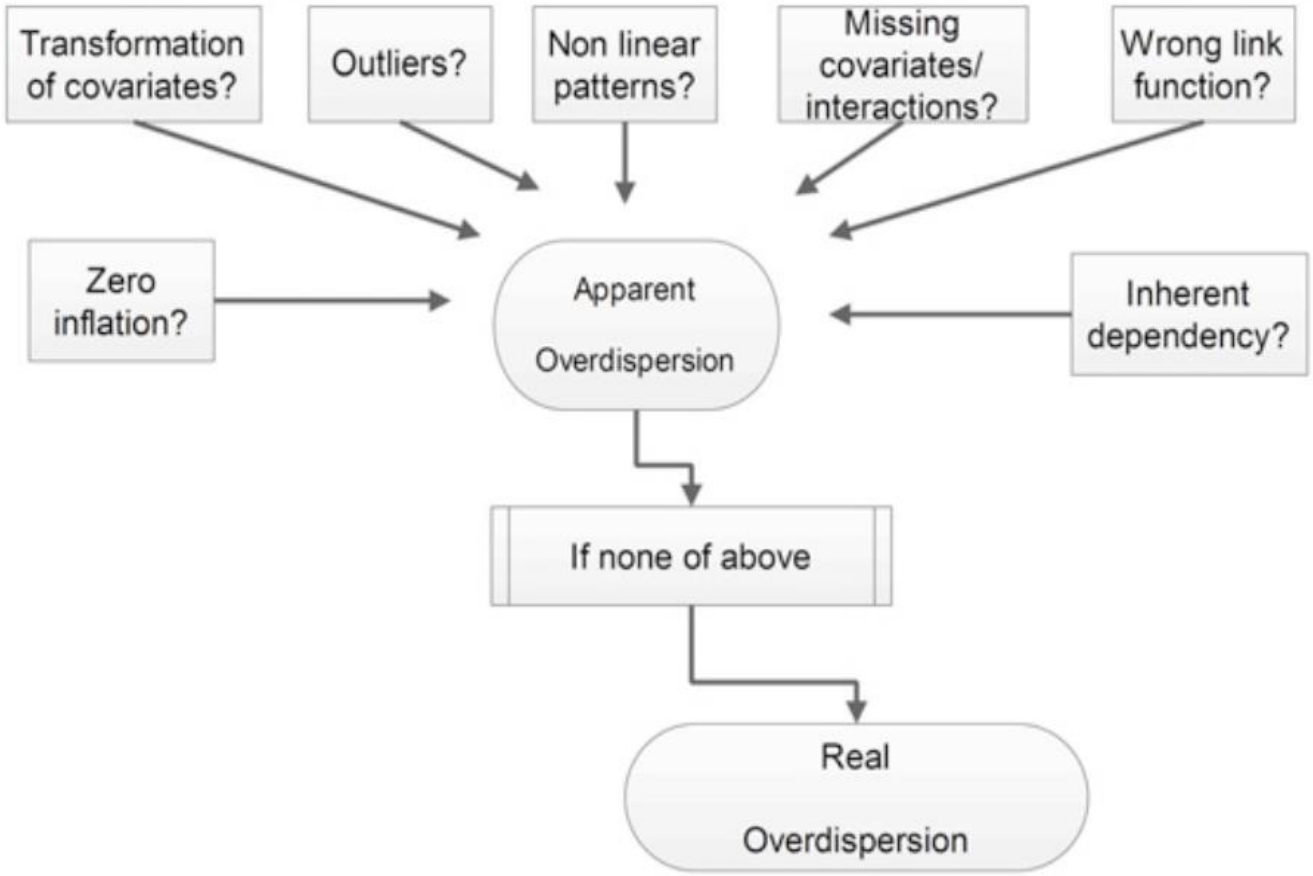
模拟数据与实际数据的比较
performance::pp_check

```
pp_check(pm1, iterations=1000)
```

过度离散???



过度离散：Y的实际离散程度大于模型理论值



Zuur et al. A beginner's guide to GLM and GLMM with R, *P20*

计数数据：模型是否存在过度离散
performance::check_overdispersion

```
pm1<- glm(Richness~NAP, family="poisson", data=d3)
```

```
> check_overdispersion(pm1)
# Overdispersion test

       dispersion ratio =   3.044
  Pearson's Chi-Squared = 130.900
              p-value =  < 0.001

Overdispersion detected.
```

模型确实存在过度离散

glm--Poisson回归：
如果数据存在过度离散，那么可采用quasipoisson回归

```
pm1<- glm(Richness~NAP, family="poisson", data=d3)
summary(pm1)
```

```
Call:
glm(formula = Richness ~ NAP, family = "poisson", data = RIKZ)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2029  -1.2432  -0.9199   0.3943   4.3256

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.79100    0.06329  28.297  < 2e-16 ***
NAP         -0.55597    0.07163  -7.762 8.39e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 179.75  on 44  degrees of freedom
Residual deviance: 113.18  on 43  degrees of freedom
AIC: 259.18

Number of Fisher Scoring iterations: 5
```

```
pm2<- glm(Richness~NAP, family="quasipoisson", data=RIKZ)
summary(pm2)
```

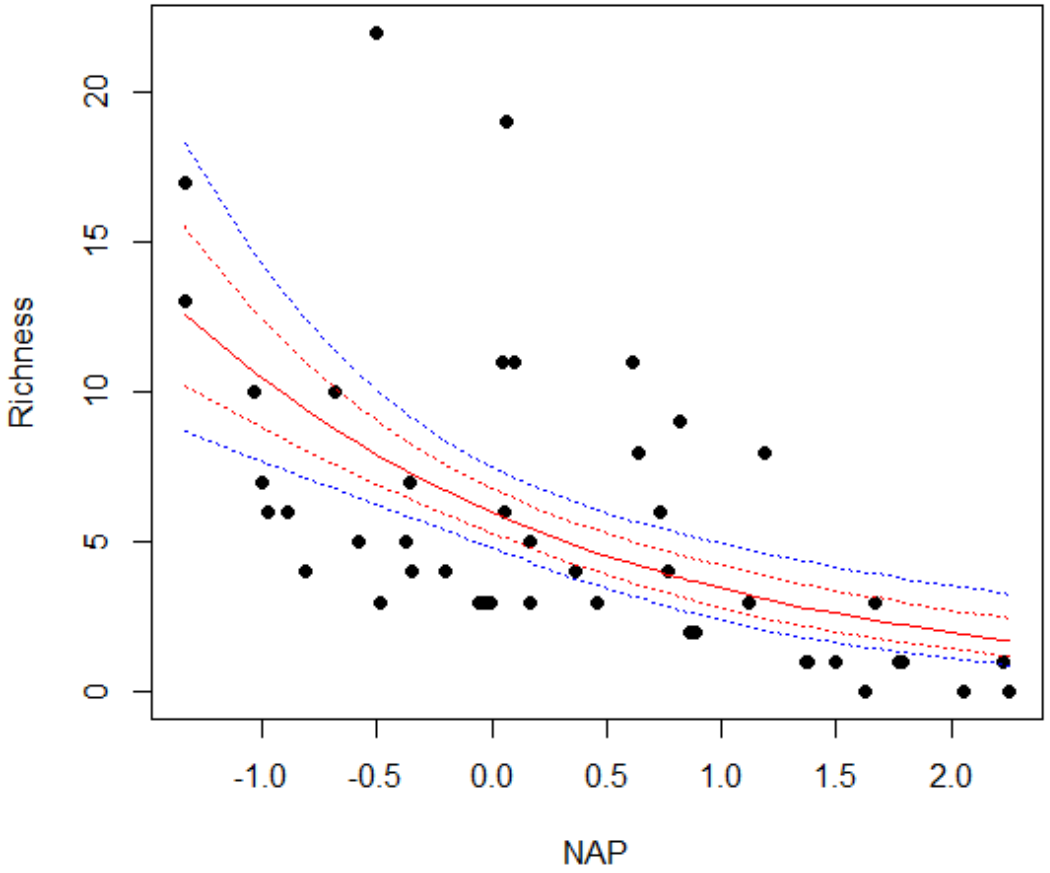
```
Call:
glm(formula = Richness ~ NAP, family = "quasipoisson", data = RIKZ)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2029  -1.2432  -0.9199   0.3943   4.3256

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.7910     0.1104  16.218  < 2e-16 ***
NAP          -0.5560     0.1250  -4.448 6.02e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 3.044178)

    Null deviance: 179.75  on 44  degrees of freedom
Residual deviance: 113.18  on 43  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

是否考虑过度离散，影响的只是参数估计的误差，对均值的估计没有影响

泊松回归—指定离散系数（与伪泊松结果相同）

```
> sigma2<-sum(residuals(pm1, type="pearson")^2)/43
> sigma2
[1] 3.044176
> summary(pm1,dispersion=sigma2)

Call:
glm(formula = Richness ~ NAP, family = "poisson", data = RIKZ)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2029  -1.2432  -0.9199   0.3943   4.3256

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.7910     0.1104  16.218  < 2e-16 ***
NAP          -0.5560     0.1250  -4.448 8.65e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 3.044176)

    Null deviance: 179.75  on 44  degrees of freedom
Residual deviance: 113.18  on 43  degrees of freedom
AIC: 259.18
```

Poisson回归--是否考虑过度离散结果对比



蓝色：考虑过度离散，quasipoisson分布

红色：不考虑过度离散，poisson分布

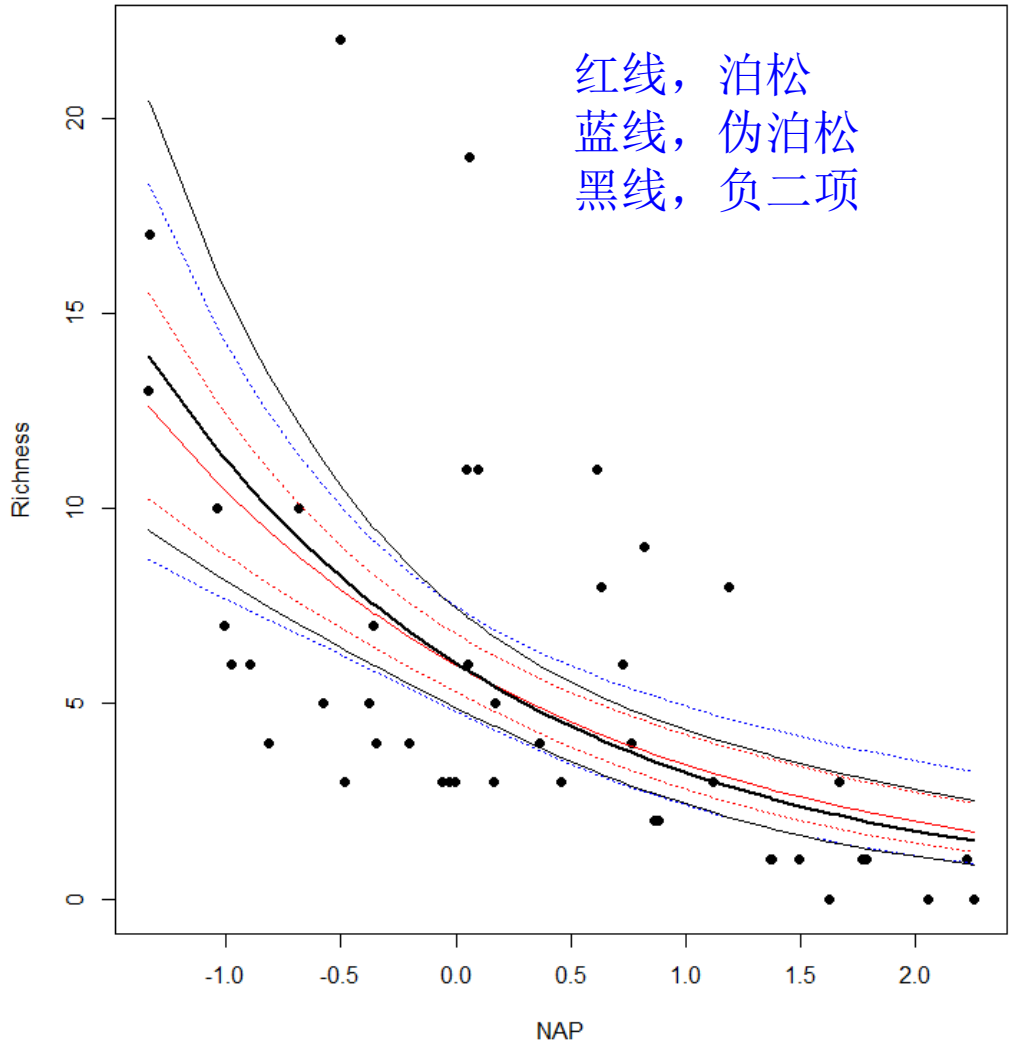
缺陷： quasipoisson模型无法计算AIC值，不利于模型对比

考虑过度离散后，预测的不确定性增加，误差变大，
犯第一类型统计错误（弃真）的概率降低

计数数据的分析，另一选择：负二项回归

```
> library(MASS)
> nb1 <- glm.nb(Richness~NAP, data=RIKZ)
> summary(nb1)

Call:
glm.nb(formula = Richness ~ NAP, data = RIKZ, init.theta = 3.712226484,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8562  -0.8625  -0.6085   0.1290   2.3513

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.7985     0.1037  17.336  < 2e-16 ***
NAP          -0.6230     0.1122  -5.552 2.83e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(3.7122) family taken to be 1)

    Null deviance: 76.231  on 44  degrees of freedom
Residual deviance: 46.275  on 43  degrees of freedom
AIC: 231.75

Number of Fisher Scoring iterations: 1
```

不同模型结果对比




```
> lrtest(pm1, nb1)
Likelihood ratio test

Model 1: Richness ~ NAP
Model 2: Richness ~ NAP
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   2 -127.59
2   3 -112.87  1 29.435  5.782e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> AIC(pm1, nb1)
     df      AIC
pm1   2 259.1807
nb1   3 231.7456
```

负二项回归—指定离散系数（与伪负二项结果相同）

```
> summary(nb1,dispersion=3.7122)

Call:
glm.nb(formula = Richness ~ NAP, data = RIKZ, init.theta = 3.712226484,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8562  -0.8625  -0.6085   0.1290   2.3513

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.7985     0.1999   8.998  < 2e-16 ***
NAP          -0.6230     0.2162  -2.882  0.00396 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(3.7122) family taken to be 3.7122)

    Null deviance: 76.231  on 44  degrees of freedom
Residual deviance: 46.275  on 43  degrees of freedom
AIC: 231.75
```

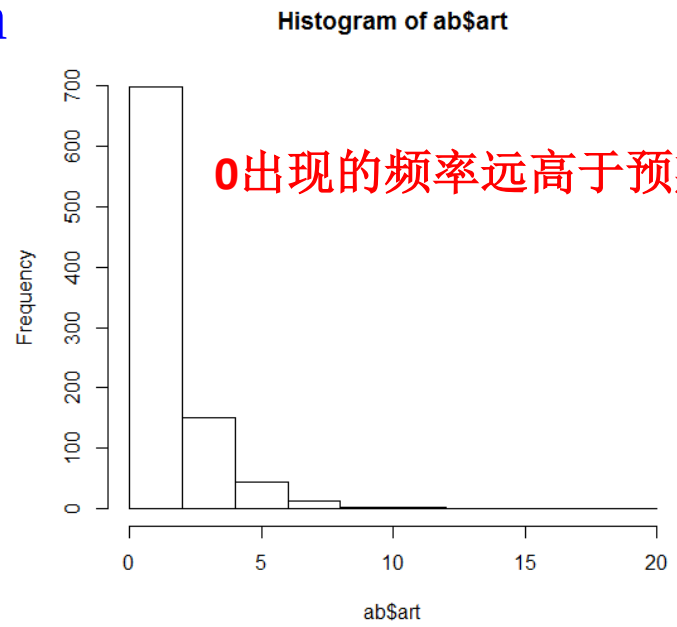
广义线性模型--计数数据—且零膨胀
glm-zero-inflated **poisson** Model

###设置误差为泊松分布(p)

zip<-zeroinfl(art~fem+mar+kid5+

phd+ment, data =ab, dist = "p", EM = TRUE)

summ



0出现的频率远高于预期

```
> summary(zip)

Call:
zeroinfl(formula = art ~ fem + mar + kid5 + phd + ment, data = ab, dist = "p", EM = TRUE)

Pearson residuals:
    Min      1Q  Median      3Q     Max
-2.3254 -0.8652 -0.2826  0.5404  7.2974

Count model coefficients (poisson with log link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.744612   0.110280   6.752 1.46e-11 ***
femWomen    -0.209147   0.063405  -3.299 0.000972 ***
marSingle   -0.103749   0.071111  -1.459 0.144573
kid5        -0.143317   0.047429  -3.022 0.002514 **
phd         -0.006170   0.031008  -0.199 0.842269
ment         0.018098   0.002294   7.888 3.07e-15 ***

Zero-inflation model coefficients (binomial with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.931030   0.469680  -1.982  0.04745 *
femWomen     0.109722   0.280066   0.392  0.69523
marSingle    0.353985   0.317591   1.115  0.26502
kid5         0.217101   0.196469   1.105  0.26915
phd          0.001225   0.145252   0.008  0.99327
ment        -0.134072   0.045224  -2.965  0.00303 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 1
Log-likelihood: -1605 on 12 Df
```

非零部分，log转换，对钓鱼的个数的影响

是否导致零的出现，即是否能钓到鱼的影响 logit转换

计数数据，零膨胀，且不为0部分过度离散

glm-zero-inflated negative binomial Model

```
library(pscl)
###设置误差分布为负二项(negbin)
zinb<- zeroinfl(art~fem+mar+kid5+phd+ment,
   data =ab, dist = "negbin", EM = TRUE)
summary(zinb)
```

```
Call:
zeroinfl(formula = art ~ fem + mar + kid5 + phd + ment, data = ab, dist = "negbin", EM = TRUE)

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.2942 -0.7601 -0.2910  0.4447  6.4154

Count model coefficients (negbin with log link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.5143483  0.1289421   3.989 6.64e-05 ***
femWomen    -0.1954951  0.0755919  -2.586  0.00970 **
marSingle   -0.0975896  0.0844519  -1.156  0.24786
kid5        -0.1517237  0.0542062  -2.799  0.00513 **
phd         -0.0006967  0.0362697  -0.019  0.98467
ment         0.0247845  0.0034926   7.096 1.28e-12 ***
Log(theta)   0.9764245  0.1354754   7.207 5.70e-13 ***

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.69116    1.03923  -1.627  0.10367
femWomen     0.63599    0.84857   0.749  0.45356
marSingle    1.49868    0.93824   1.597  0.11019
kid5         0.62831    0.44265   1.419  0.15578
phd         -0.03754    0.30789  -0.122  0.90296
ment        -0.88189    0.31604  -2.790  0.00526 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 2.6549
Number of iterations in BFGS optimization: 1
Log-likelihood: -1550 on 13 Df
```

非零部分，对钓到的鱼的多少的影响， **log转换，**

对是否出现零的影响，即对能否钓到鱼的影响 **logit转换**

广义线性模型二项（binomial）数据的分析：glm-logistic regression model

0/1或者比率(0,1) 数据的分析

注意：这里的比率一定要是由计数数据（离散变量）转化而来的比率，并且也应具有转化为比率的原始计数数据

0/1数据的分析--glm-logistic regression model

- 适用类型0,1 数据
- 比如：输、赢
  有、没有
  康复、未康复

###

```
d2<-read.csv("d2.csv")
View(d2)
```

admit	gre	gpa	rank
0	220	2.83	3
0	300	2.92	4
0	300	3.01	3
1	300	2.84	2
0	340	3.15	3
0	340	2.92	3
0	340	2.90	1
1	340	3.00	2
0	360	2.56	3
0	360	3.14	1
0	360	3.00	3
0	360	3.27	3
0	380	3.61	3
0	380	2.94	3
0	380	2.91	4
0	380	3.33	4
0	380	3.43	3

0/1数据的分析--glm-logistic regression model

```
##binomial regression model with logit link
m5 <- glm(admit ~ gre, data = d2, family = "binomial")
summary(m5)
```

```
Call:
glm(formula = admit ~ gre, family = "binomial", data = d2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1623  -0.9052  -0.7547   1.3486   1.9879

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.901344   0.606038  -4.787 1.69e-06 ***
gre          0.003582   0.000986   3.633  0.00028 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 499.98  on 399  degrees of freedom
Residual deviance: 486.06  on 398  degrees of freedom
AIC: 490.06

Number of Fisher Scoring iterations: 4
```

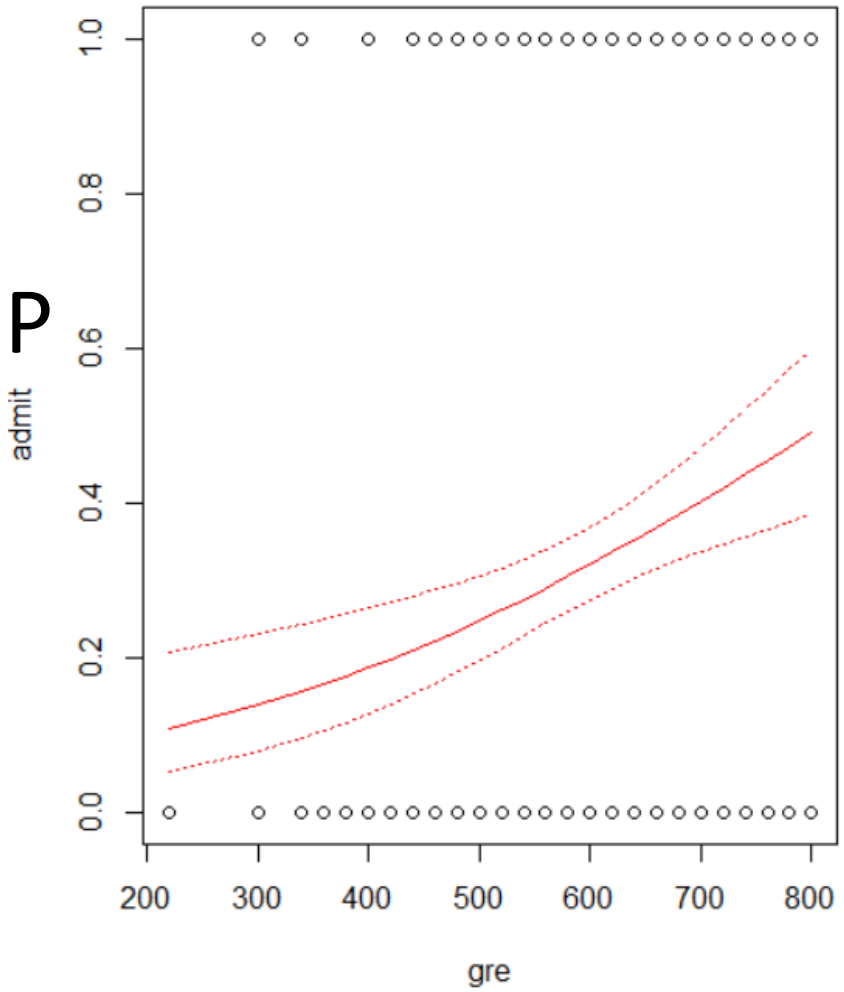
注意其默认link函数为logit link

即Ln(p/(1-p))=-2.9013+0.00358*gre

$P=e^{-2.9013+0.00358*gre}/(1+e^{-2.9013+0.00358*gre})$

0/1数据的分析
glm-logistic regression model

```
##用add_ci命令获取估计值和置信区间
library(ciTools)
d2<-d2[order(d2$gre),]
plot(admit~gre, data=d2)
ci_bi1<-add_ci(d2, m5, alpha = 0.05, names = c("lwr", "upr"))
lines(ci_bi1$gre, ci_bi1$pred, col = "red", lwd = 1,lty="solid")
lines(ci_bi1$gre, ci_bi1$upr,  col = "red", lwd = 1,lty="dotted")
lines(ci_bi1$gre, ci_bi1$lwr,  col = "red", lwd = 1,lty="dotted")
```



（0,1）数据的分析
glm-binomial regression model

```
seal <- read.csv('sealData1.csv', h=T)
head(seal)
```

	pupage	n.obs	n.response
1	13	8	2
2	9	9	5
3	29	7	0
4	15	8	0
5	1	2	2
6	26	1	0

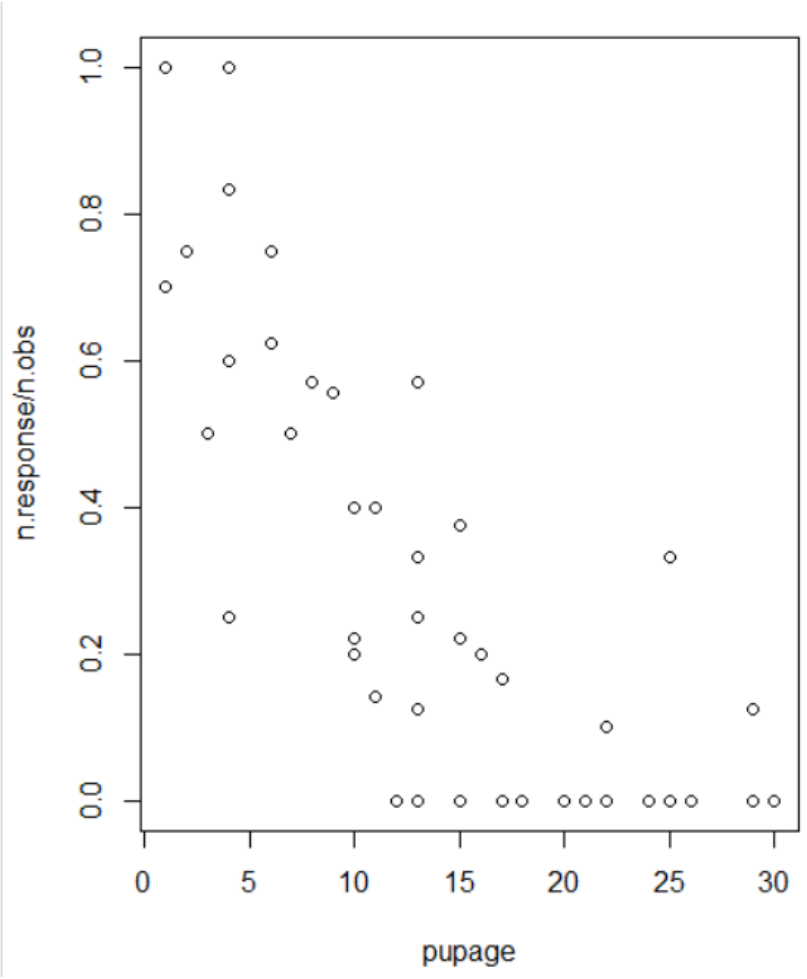
注意，该比率必须由计数数据转化而来，
本案例中p=n.response/n.obs, 介于（0,1）之间

特别提示：对于非计数数据转化而来的比率数据，建议用logit 转换, 也可以用beta回归

For binomial data, logistic regression has greater interpretability and higher power than analyses of transformed data. However, **it is important to check the data for additional unexplained variation, i.e., overdispersion, and to account for it via the inclusion of random effects in the model if found**. For non-binomial data, the arcsinetrans form is undesirable on the grounds of interpretability, and because it can produce nonsensical predictions. **The logit transformation** is proposed as an alternative approach to address these issues.

Warton, D. I. and F. K. C. Hui. 2011. The arcsine is asinine: the analysis of proportions in ecology. ***Ecology*** **92**:3-10.

比率数据（0,1）的分析--
glm-binomial regression model

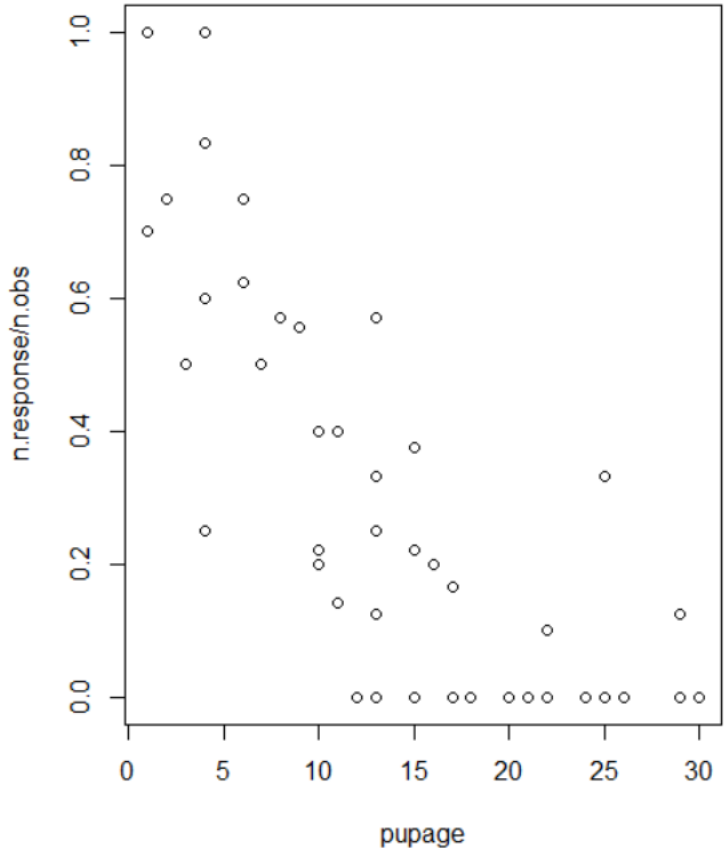


##二项分布回归，三种意义相同的写法

```
mod.seal1 <- glm(cbind(n.response, n.obs-n.response)~pupage,
data=seal, family=binomial(link=logit))
mod.seal2 <- glm(cbind(n.response, n.obs-n.response)~pupage,
data=seal, family=binomial())
mod.seal3 <- glm(cbind(n.response, n.obs-n.response)~pupage,
data=seal, family=binomial)
```

注意，这时模型中应包含得到比率的两列计数数据的信息

比率数据（0,1）的分析--
glm-binomial regression model



##但一定要使用weights命令指明得出该比值的基于的样本量是多少

seal$p<-seal$n.response/seal$n.obs

mod.seal5 <- glm(p~pupage, data=seal,family=binomial(link=logit), **weights=n.obs**)

错误做法

mod.seal4 <- glm(p~pupage, data=seal,family=binomial(link=logit))

也可以先自行求出比率*p*，然后以得到该比率的总样本量为权重，进行分析

以总观测次数为权重

比率数据（0,1）的分析
glm-binomial regression model

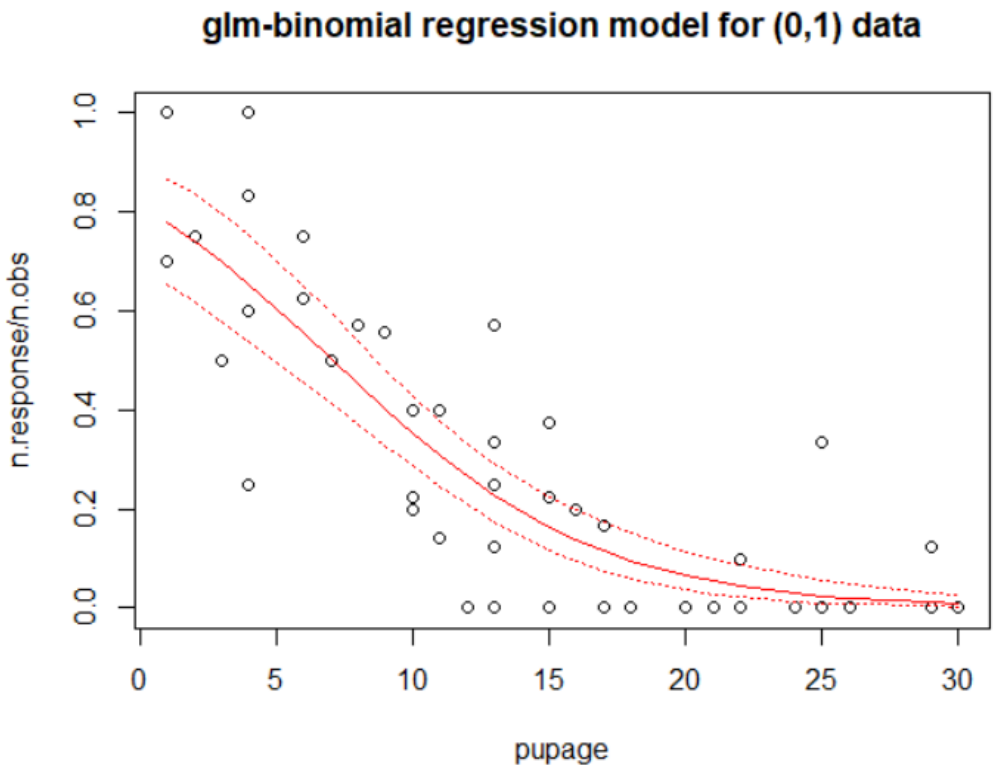
```
mod.seal3 <- glm(cbind(n.response, n.obs-n.response)~pupage,
data=seal,family=binomial)
```

```
> summary(mod.seal3)

Call:
glm(formula = cbind(n.response, n.obs - n.response) ~ pupage,
    family = binomial, data = seal)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0344  -0.7424  -0.2253   0.5240   1.9548

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.46130    0.34117   4.283 1.84e-05 ***
pupage      -0.20626    0.02874  -7.176 7.19e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 128.247  on 49  degrees of freedom
Residual deviance:  46.268  on 48  degrees of freedom
AIC: 108.79

Number of Fisher Scoring iterations: 5
```

```
mod.seal5 <- glm(p~pupage, data=seal,
family=binomial,weights=n.obs)
```

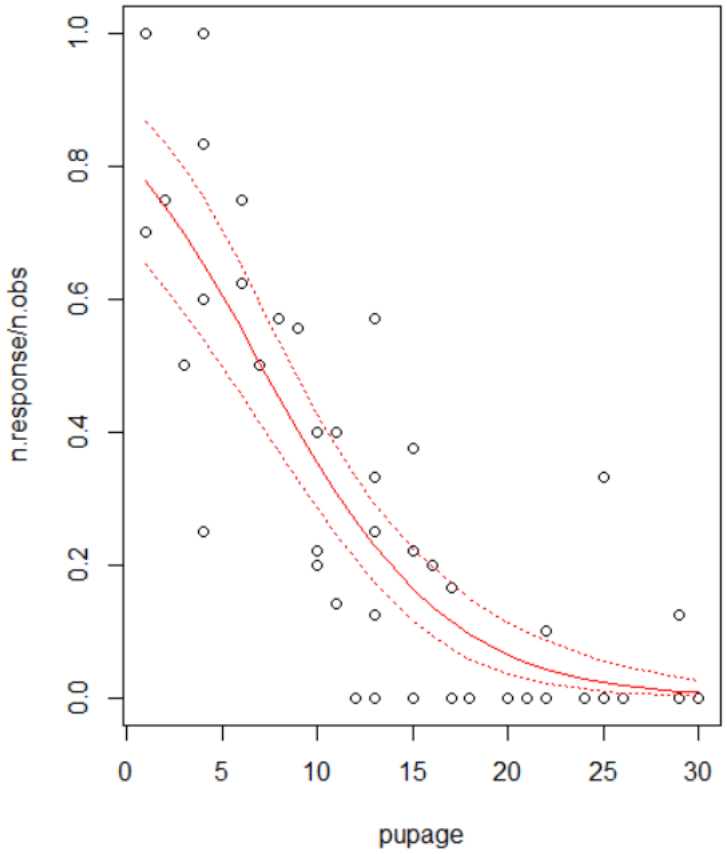
```
> summary(mod.seal5)

Call:
glm(formula = p ~ pupage, family = binomial(link = logit), data = seal,
    weights = n.obs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0344  -0.7424  -0.2253   0.5240   1.9548

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.46130    0.34117   4.283 1.84e-05 ***
pupage      -0.20626    0.02874  -7.176 7.19e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 128.247  on 49  degrees of freedom
Residual deviance:  46.268  on 48  degrees of freedom
AIC: 108.79

Number of Fisher Scoring iterations: 5
```

比率数据（0,1）的分析
glm-binomial regression model结果作图

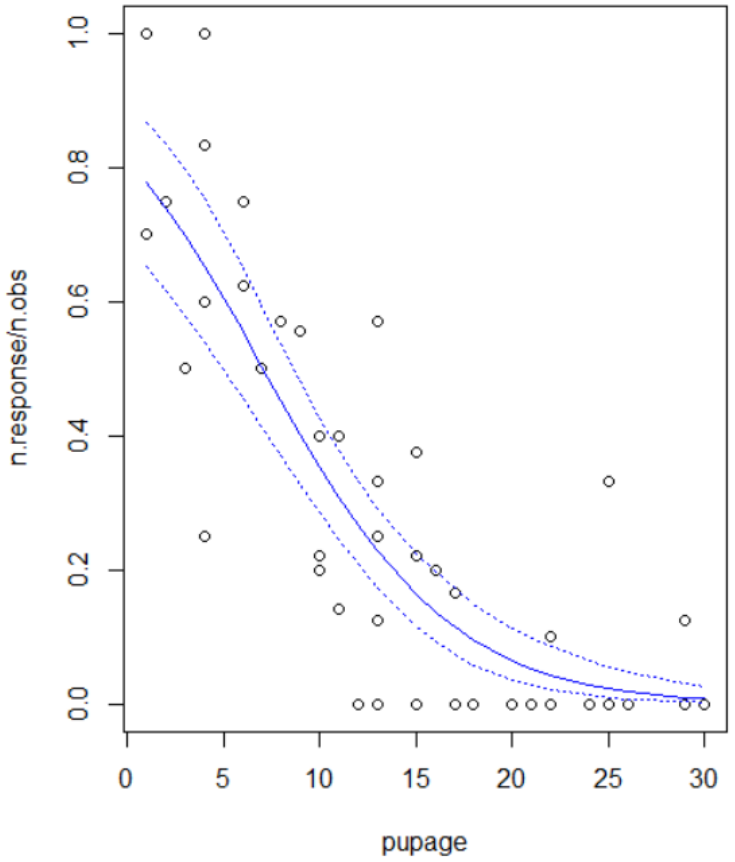



```
##作图代码
library(ciTools)
plot(n.response/n.obs~pupage, data=seal)
seal<-seal[order(seal$pupage),]
d_seal1<-add_ci(seal, mod.seal3, alpha = 0.05, names = c("lwr", "upr"))
lines(d_seal1$pupage, d_seal1$pred, col = "red", lwd = 1,lty="solid")
lines(d_seal1$pupage, d_seal1$upr, col = "red", lwd = 1,lty="dotted")
lines(d_seal1$pupage, d_seal1$lwr, col = "red", lwd = 1,lty="dotted")
```

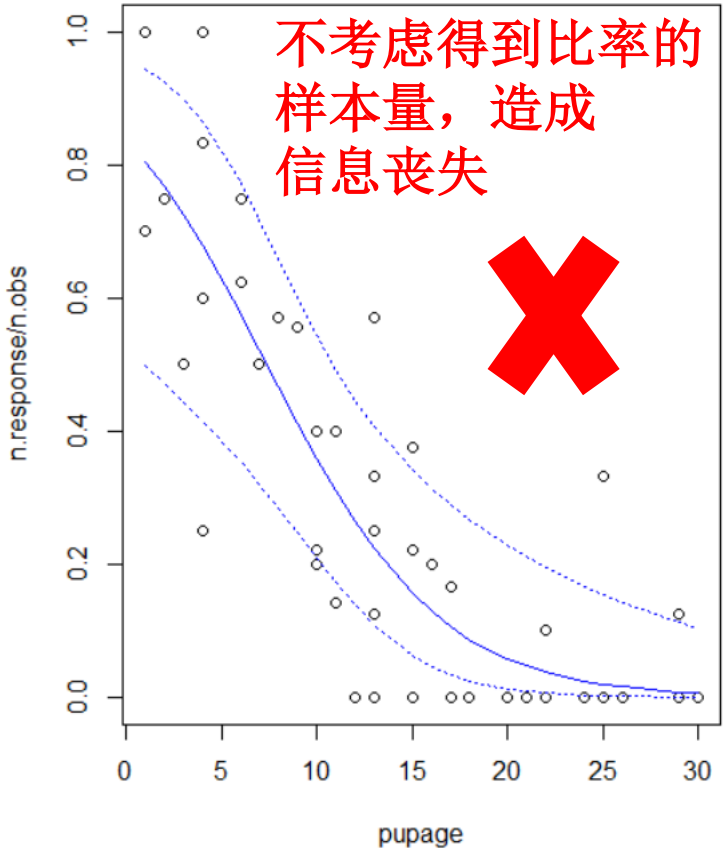
比率数据（0,1）的分析
glm-binomial regression model几种方案对比



mod.seal3 <- glm(cbind(n.response, n.obs-n.response)~pupage, data=seal,family=binomial)



seal$p<-seal$n.response/seal$n.obs
mod.seal5 <- glm(p~pupage, data=seal,family=binomial, **weights=n.obs**)



seal$p<-seal$n.response/seal$n.obs
mod.seal4 <- glm(p~pupage, data=seal, family=binomial)

比率数据（0,1）的分析：glm-binomial regression model
同样可能存在过度离散

如果y为0、1数据，不存在过度离散问题，

但如果数据都位于0~1之间，则可能存在过度离散问题

二项分布

variance=p(1-p)

如果variance>p(1-p)

则数据过度离散

```
> summary(mod.seal4)

Call:
glm(formula = p ~ pupage, family = binomial(link = logit), data = seal,
    weights = n.obs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0344  -0.7424  -0.2253   0.5240   1.9548

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.46130    0.34117   4.283 1.84e-05 ***
pupage      -0.20626    0.02874  -7.176 7.19e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 128.247  on 49  degrees of freedom
Residual deviance:  46.268  on 48  degrees of freedom
AIC: 108.79

Number of Fisher Scoring iterations: 5
```

Residual deviance与 df 二者大致相等，则不存在过度离散

若Residual deviance明显大于 df ，则提示模型存在过度离散

比率数据（0,1）的分析：glm-binomial regression model
若存在过度离散，那么采用quasibinomial回归

mod.seal7 <- glm(cbind(n.response, n.obs-n.response)~pupage, data=seal,family=**quasibinomial**)

summary(mod.seal7)

```
Call:
glm(formula = cbind(n.response, n.obs - n.response) ~ pupage,
    family = quasibinomial, data = seal)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0344  -0.7424  -0.2253   0.5240   1.9548

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.46130    0.35858   4.075 0.000172 ***
pupage      -0.20626    0.03021  -6.827 1.35e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 1.104669)

    Null deviance: 128.247  on 49  degrees of freedom
Residual deviance:  46.268  on 48  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

比率数据（0,1）的分析：glm-binomial regression model

若存在过度离散，同样可以指定离散系数

sigma2<-sum(residuals(mod.seal3, type="pearson")^2)/48 ##求出离散指数

```
> sigma2
[1] 1.104667
```

summary(mod.seal3,dispersion=sigma2) ##指定离散指数

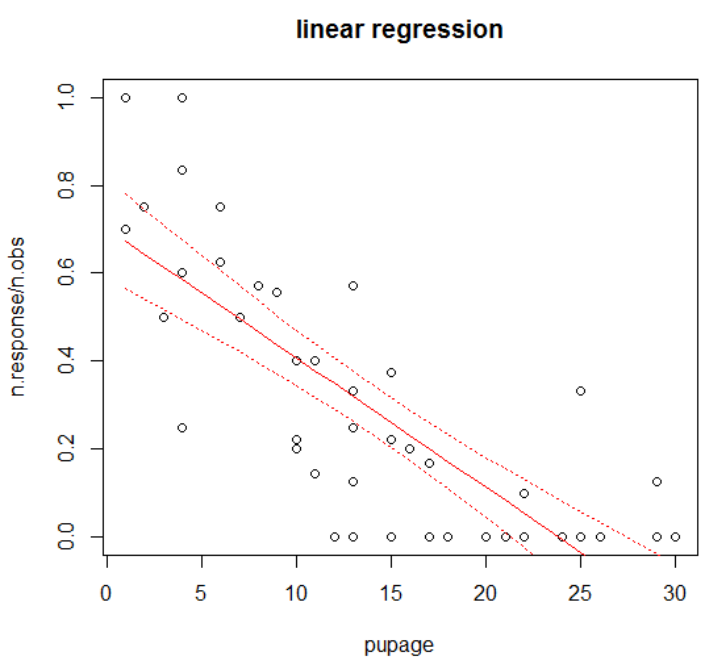
```
> summary(mod.seal3,dispersion=sigma2)

Call:
glm(formula = cbind(n.response, n.obs - n.response) ~ pupage,
    family = binomial, data = seal)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0344  -0.7424  -0.2253   0.5240   1.9548

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.46130    0.35858   4.075 4.60e-05 ***
pupage      -0.20626    0.03021  -6.827 8.64e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1.104667)

    Null deviance: 128.247  on 49  degrees of freedom
Residual deviance:  46.268  on 48  degrees of freedom
AIC: 108.79

Number of Fisher Scoring iterations: 5
```

指定离散指数结果

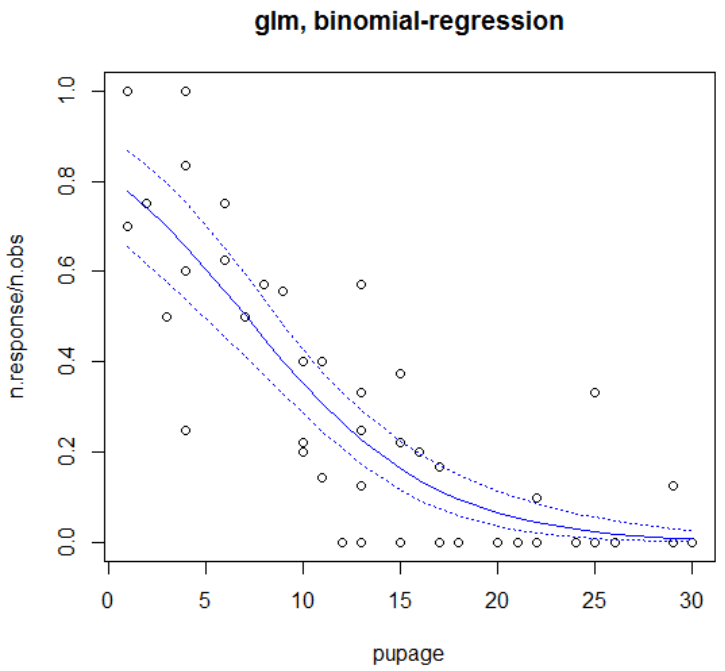
```
> summary(mod.seal7)

Call:
glm(formula = cbind(n.response, n.obs - n.response) ~ pupage,
    family = quasibinomial, data = seal)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0344  -0.7424  -0.2253   0.5240   1.9548

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.46130    0.35858   4.075 0.000172 ***
pupage      -0.20626    0.03021  -6.827 1.35e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 1.104669)

    Null deviance: 128.247  on 49  degrees of freedom
Residual deviance:  46.268  on 48  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

伪二项分布（**quasibinomial**）回归

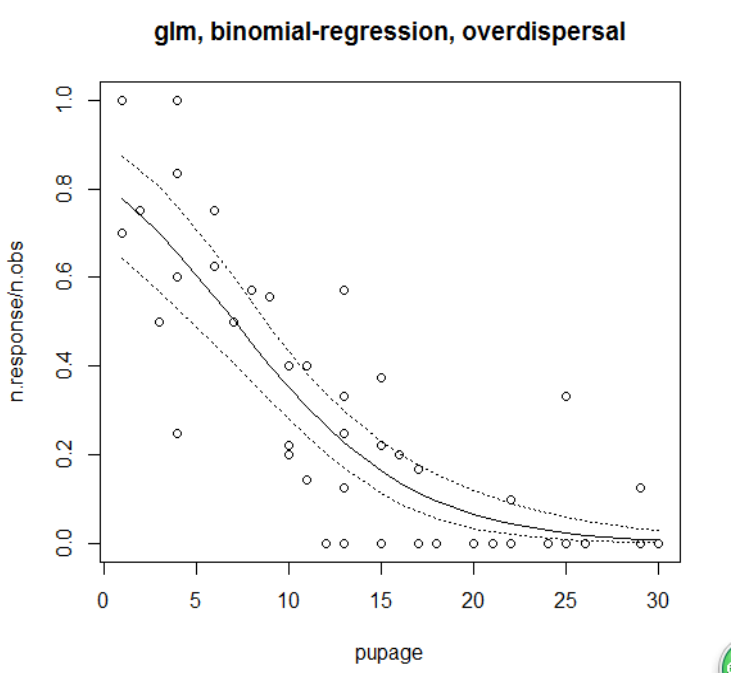
glm--二项分布--不同拟合模型结果对比



线性回归

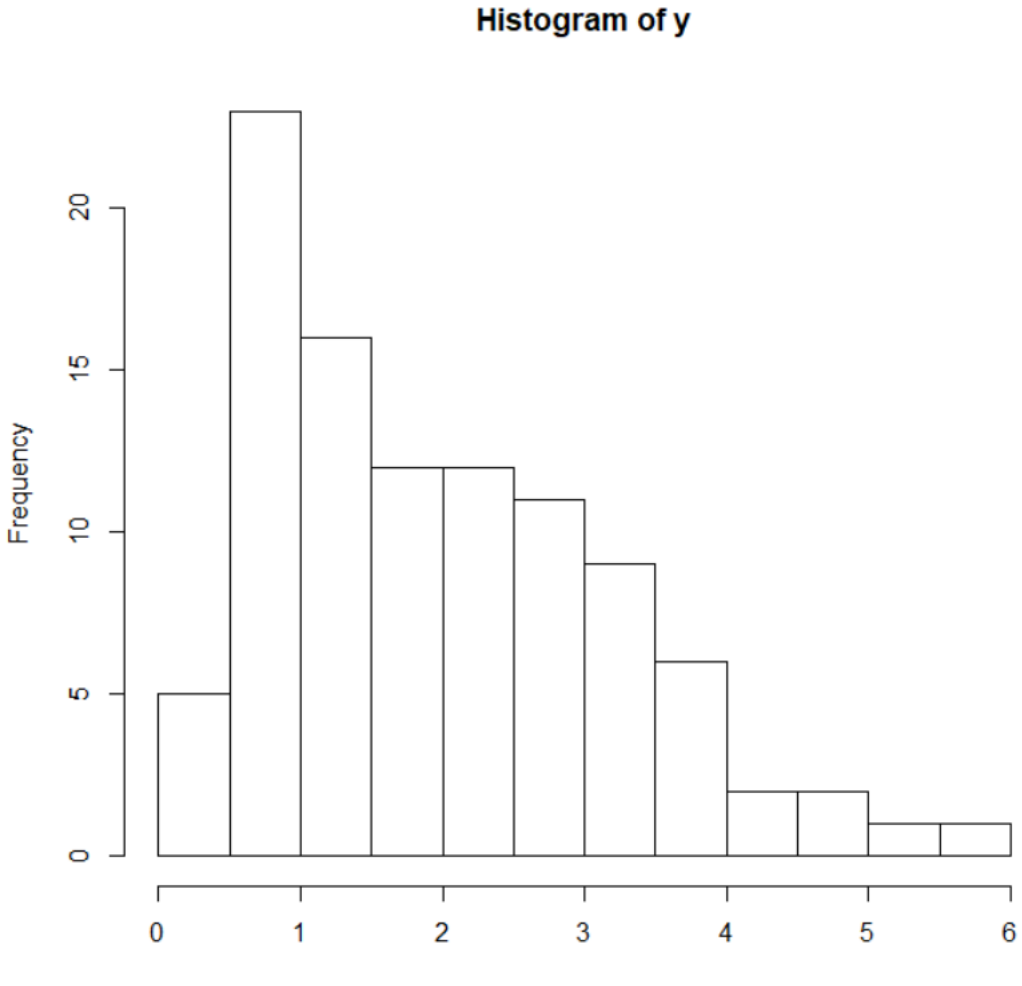


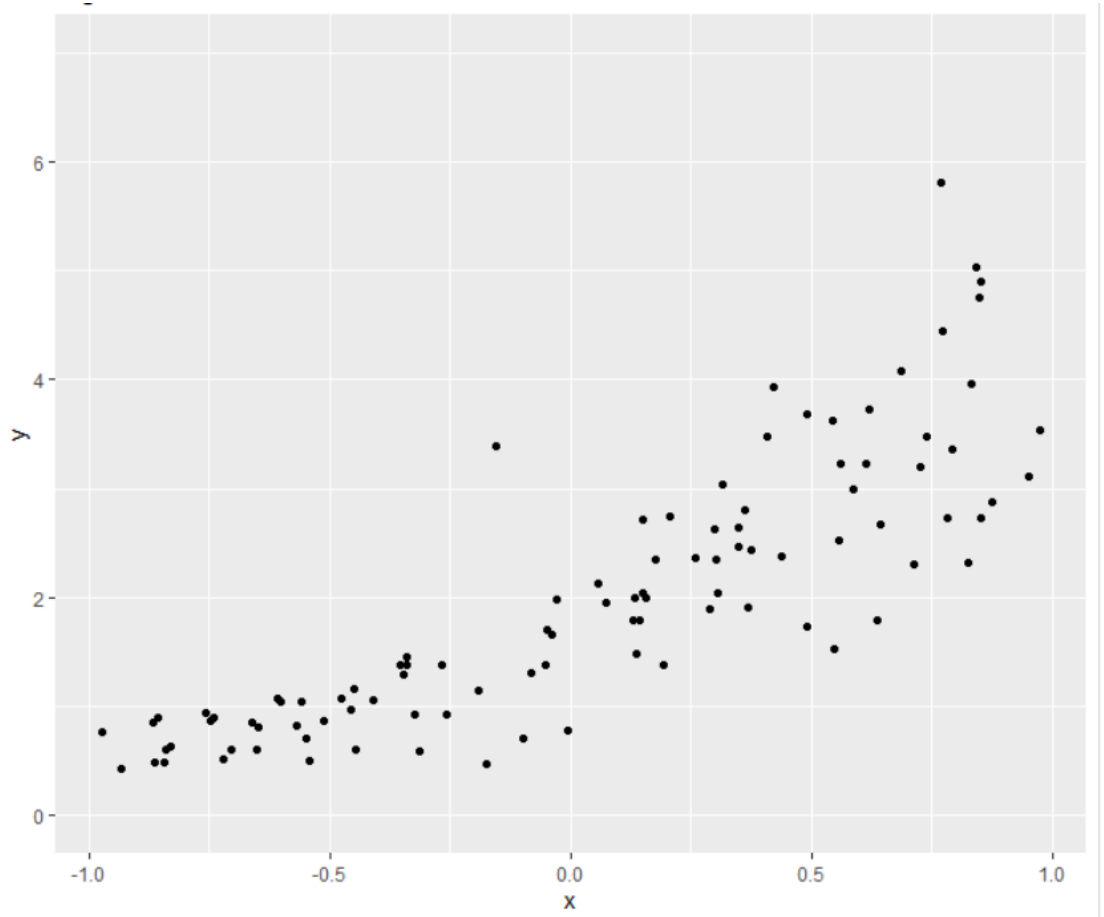
二项回归



二项回归—过度离散

glm--gamma回归
(大于0，方差随均值而增大，多数数据都较小)



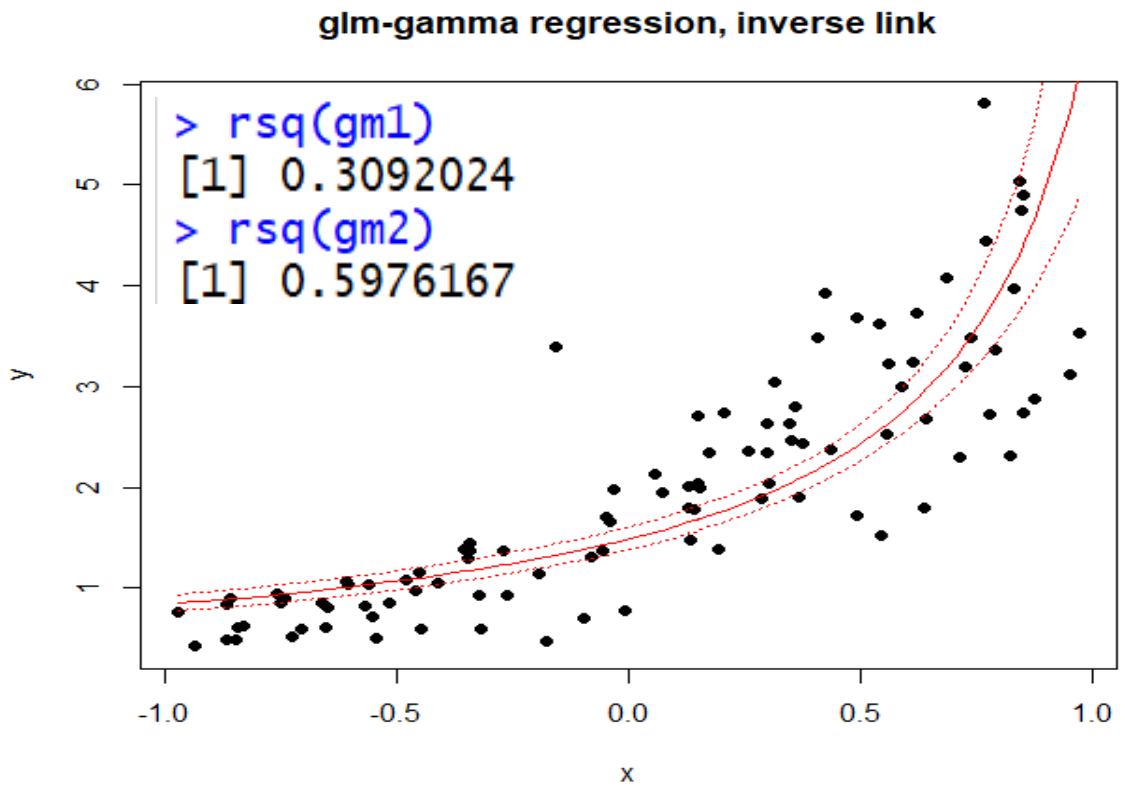


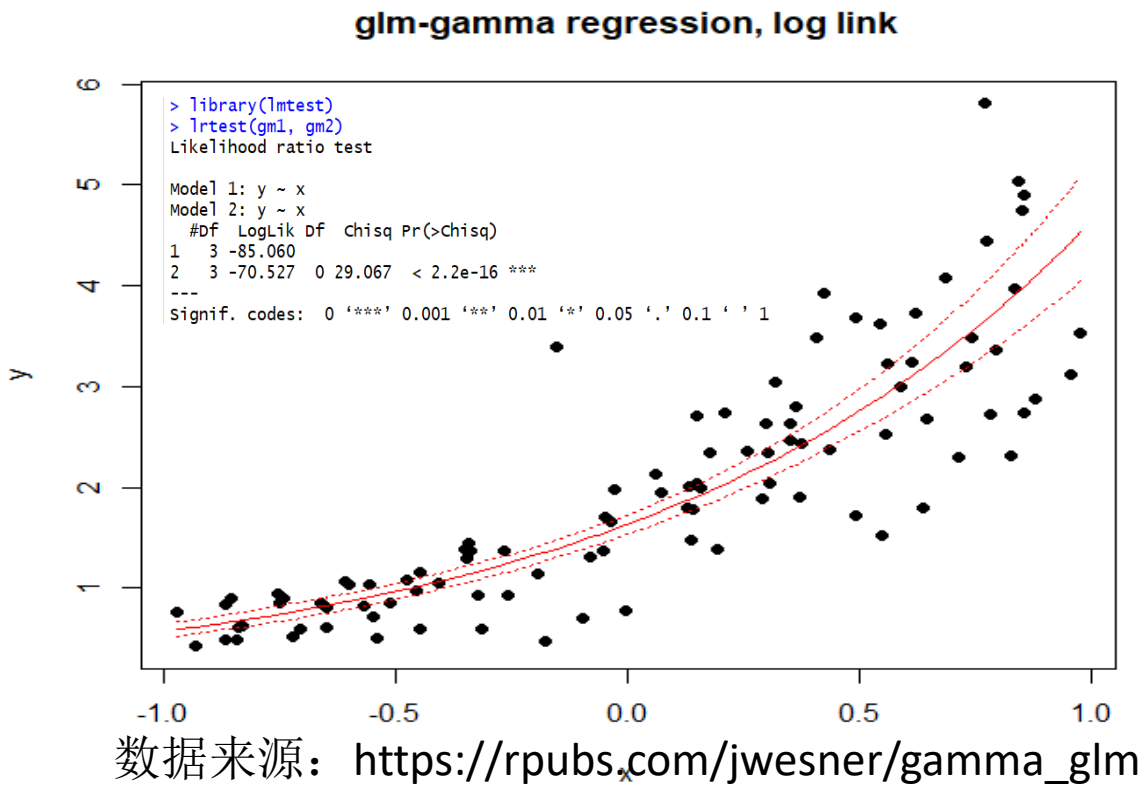
重点，不要求y必须是整数

glm--gamma回归： inverse link (默认) vs log link

```
gm1<-glm(y~x,family=Gamma(link="inverse"))
fake_data_a<-fake_data_a[order(fake_data_a$x),]
plot(y~x,data=fake_data_a, pch=19)
lm_ci<-add_ci(fake_data_a, gm1, alpha = 0.05,
names = c("lwr", "upr"))
lines(lm_ci$x, lm_ci$pred, col = "red", lwd = 1,lty="solid")
lines(lm_ci$x, lm_ci$lwr, col = "red", lwd = 1,lty="dotted")
lines(lm_ci$x, lm_ci$upr, col = "red", lwd = 1,lty="dotted")
```

```
gm2<-glm(y~x,family=Gamma(link="log"))
gdata<-gdata[order(gdata$x),]
plot(y~x,data=gdata, pch=19)
lm_ci<-add_ci(gdata, gm2, alpha = 0.05,names = c("lwr", "upr"))
lines(lm_ci$x, lm_ci$pred, col = "red", lwd = 1,lty="solid")
lines(lm_ci$x, lm_ci$lwr, col = "red", lwd = 1,lty="dotted")
lines(lm_ci$x, lm_ci$upr, col = "red", lwd = 1,lty="dotted")
```

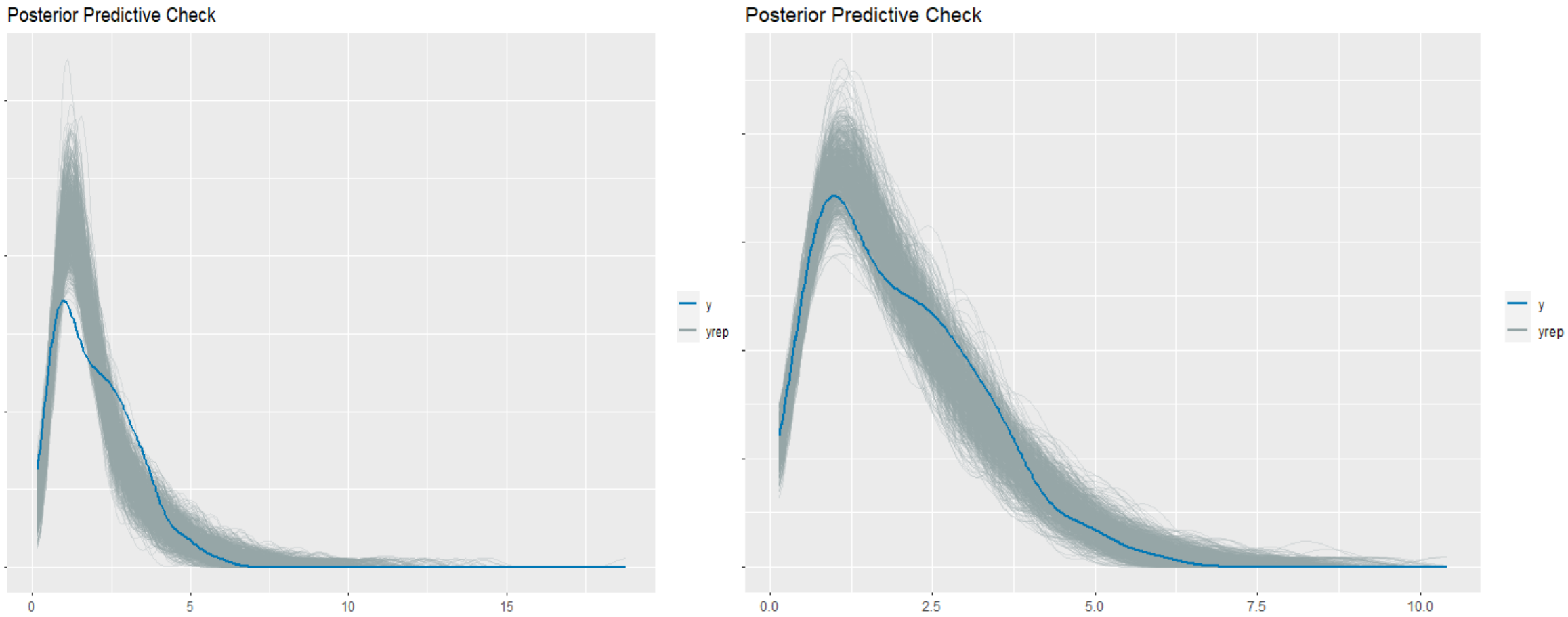




数据来源：https://rpubs.com/jwesner/gamma_glm

glm--gamma回归：
inverse link (默认) vs log link

```
> pp_check(gm1, iterations=1000)
> pp_check(gm2, iterations=1000)
```



模型 gm2 优于 gm1

违反线性模型前提之二

违反方差齐次性原则

可以在模型中指定方差的变异趋势

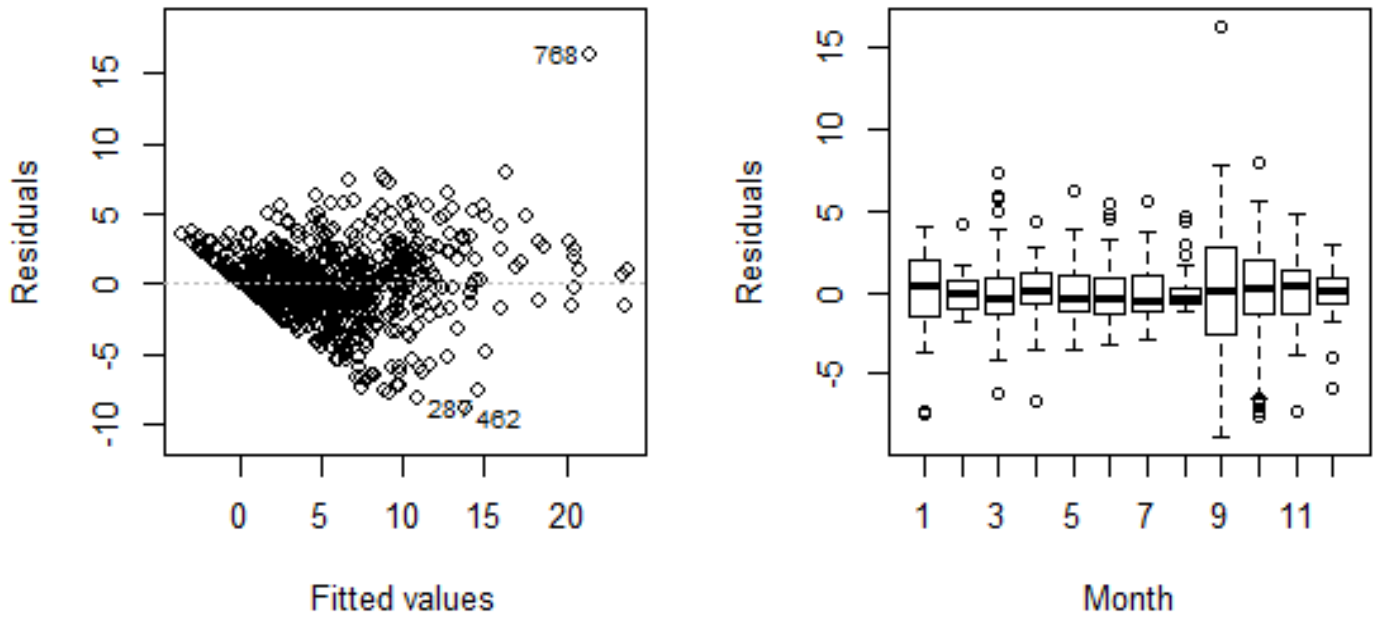
违反线性模型前提之二

违反方差齐次性原则

可以在模型中指定方差的变异趋势

违反一般线性回归前提情况二

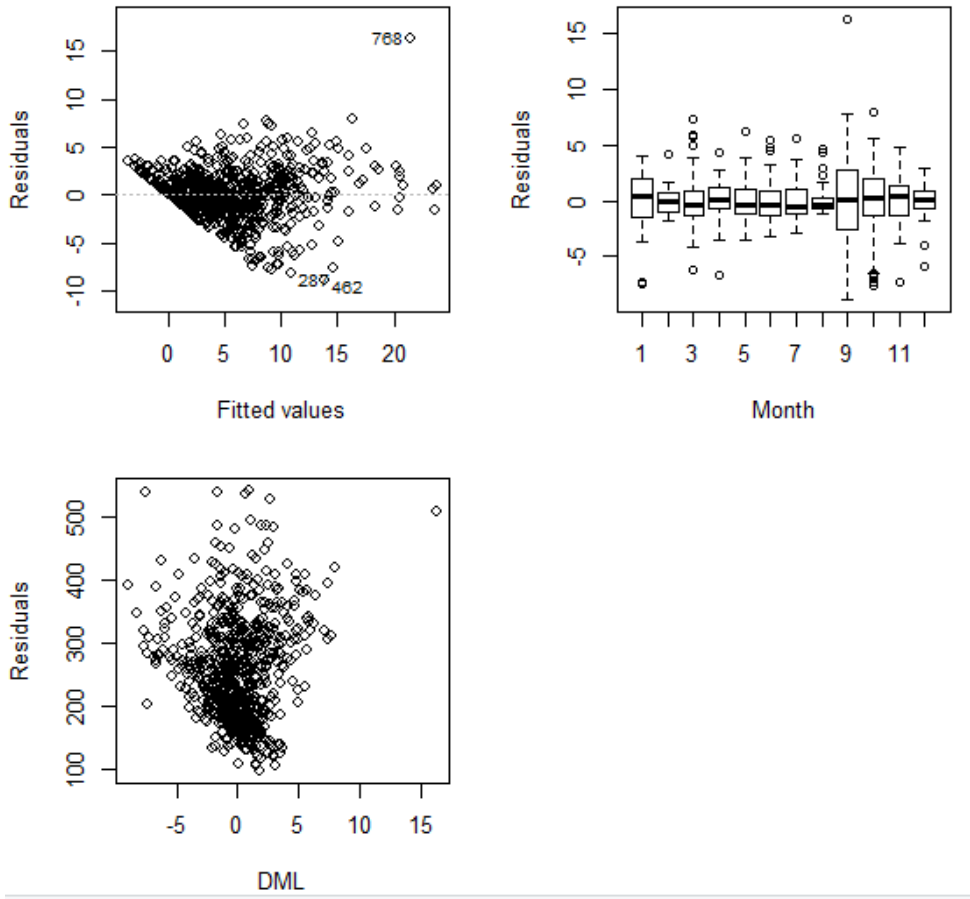
方差异质



残差随拟合值或者解释变量有明显的变化趋势

方差异质性问题

```
Squid <- read.delim("Squid.txt")
Squid$fMONTH=factor(Squid$MONTH)
m9 <- lm(Testisweight ~ DML * fMONTH,data=Squid)
par(mfrow = c(2,2), mar=c(4,4,2,2))
plot(m9, which=c(1), col=1, add.smooth=F, caption="")
a<-resid(m9)
b<-data.frame(Squid$DML)
c<-c(a,b)
na.action = na.omit
plot(Squid$fMONTH, resid(m9), xlab="Month",
    ylab="Residuals")
plot(Squid$DML~m9$residuals,
xlab="DML",ylab="Residuals")
```



##不同月份间方差齐次性检验

```
bartlett.test(Testisweight ~ fMONTH, data =Squid )
```

```
> bartlett.test(Testisweight ~ fMONTH, data =Squid )

        Bartlett test of homogeneity of variances

data:  Testisweight by fMONTH
Bartlett's K-squared = 267.97, df = 11, p-value < 2.2e-16
```

方差异质性问题--解决方案--（广义最小二乘法模型gls，指定方差结构）

```
##1) 允许方差随着解释变量的增大而增大(适用于连续变量)
library(nlme)
M.gls1<-gls(Testisweight~DML*fMONTH,
      weights=varFixed(~DML),data=Squid)
```

方差异质性问题--解决方案--（广义最小二乘法模型gls，指定方差结构）

##2)允许方差随着随着解释变量（指分类变量）水平的不同而不同呢

```
M.gls2 <- gls(Testisweight ~ DML*fMONTH,
 weights=varIdent(form= ~ 1|fMONTH), data =Squid)
```

指定方差随自变量的变化而变化
（广义最小二乘法模型，gls）

3) 组合型方差结构：

同时允许方差随着连续变量和分类变量而变化

```
M.gls3<-gls(Testisweight ~ DML * fMONTH,
        weights =varComb(varIdent(form= ~ 1 | fMONTH) ,
                varExp(form =~ DML) ), data = Squid)
```

##对比不同结构模型

```
anova(M.gls1,M.gls2,M.gls3)
AIC(M.gls1,M.gls2,M.gls3)
```

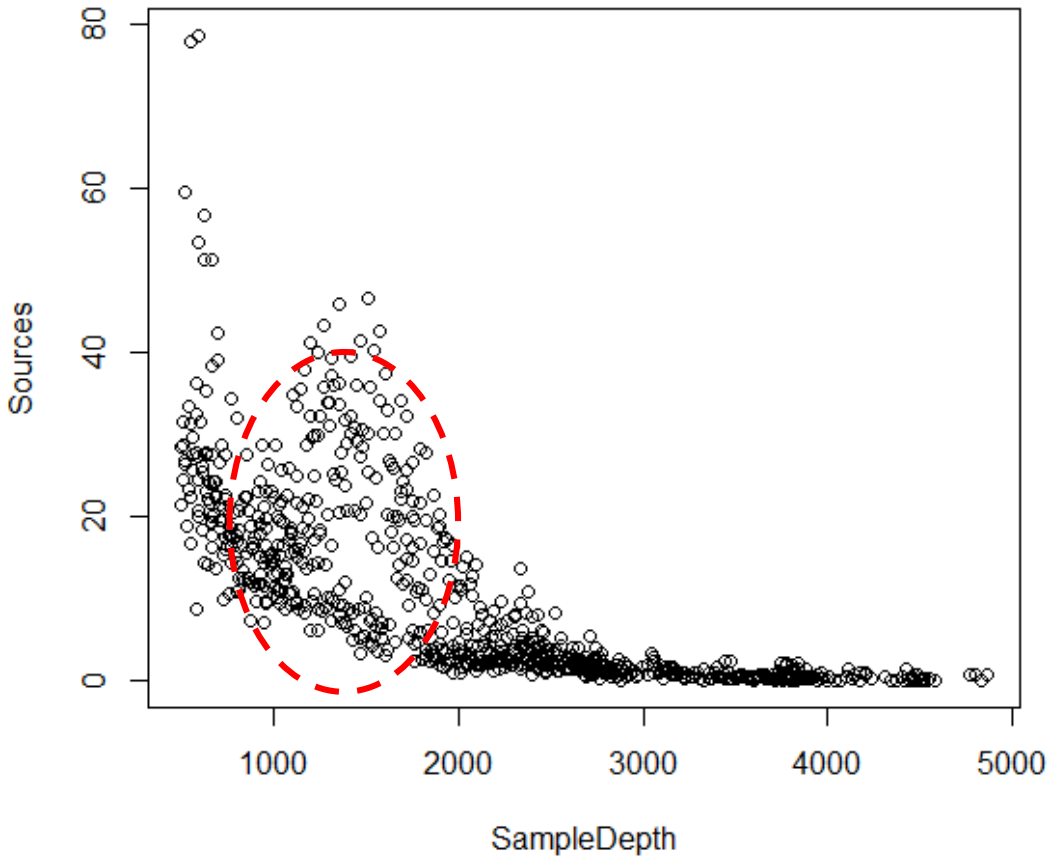
```
> anova(M.gls1,M.gls2,M.gls3)
       Model df      AIC      BIC    logLik   Test   L.Ratio p-value
M.gls1     1 25 3620.898 3736.199 -1785.449
M.gls2     2 36 3614.436 3780.469 -1771.218 1 vs 2  28.46162  0.0027
M.gls3     3 37 3414.817 3585.463 -1670.409 2 vs 3 201.61841  <.0001
> AIC(M.gls1,M.gls2,M.gls3)
       df      AIC
M.gls1 25 3620.898
M.gls2 36 3614.436
M.gls3 37 3414.817
```

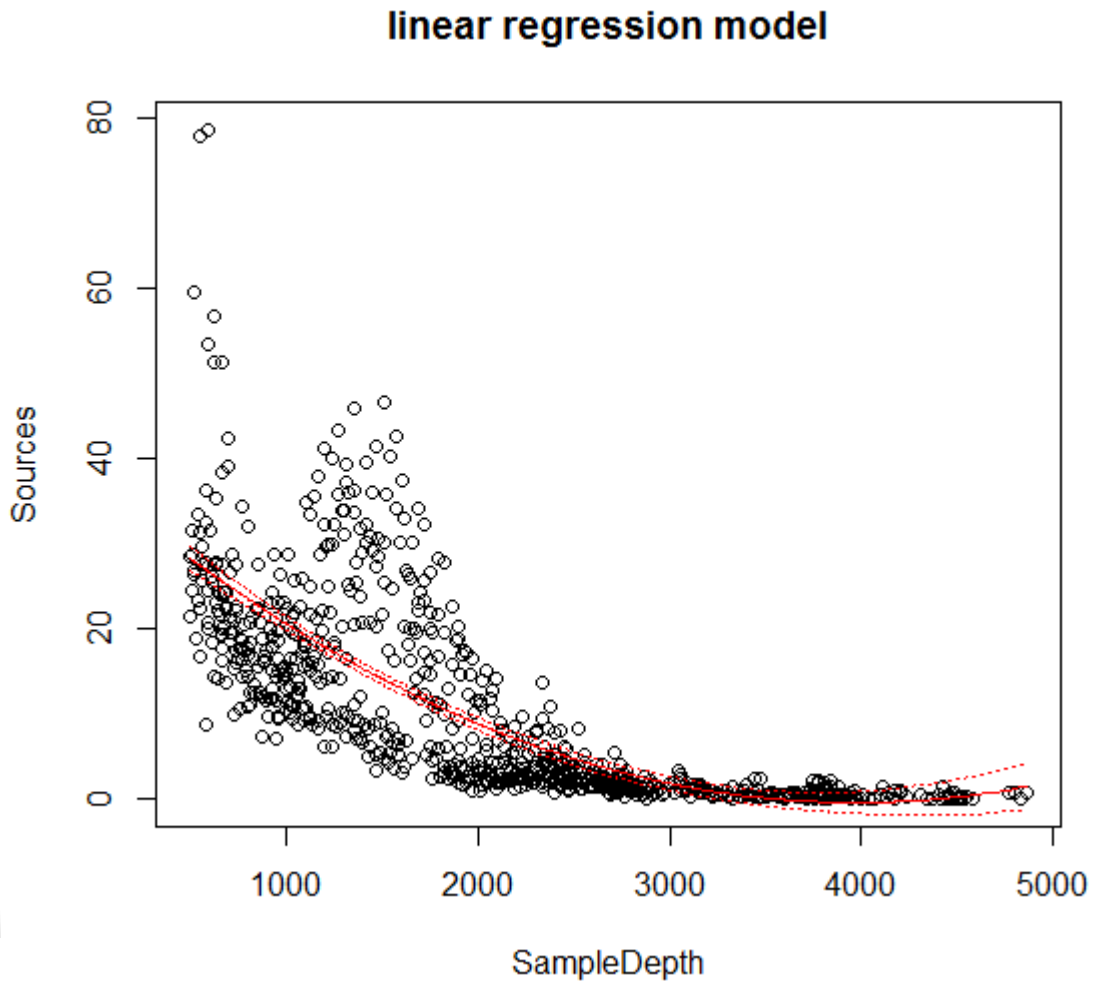
更多结构见Mixed Effects Models and Extensions in Ecology with R一书第四章

违反线性模型前提之三

- Y和X之间是非线性关系---采用非线性模型对二者间的关系进行拟合

非线性模型：additive model

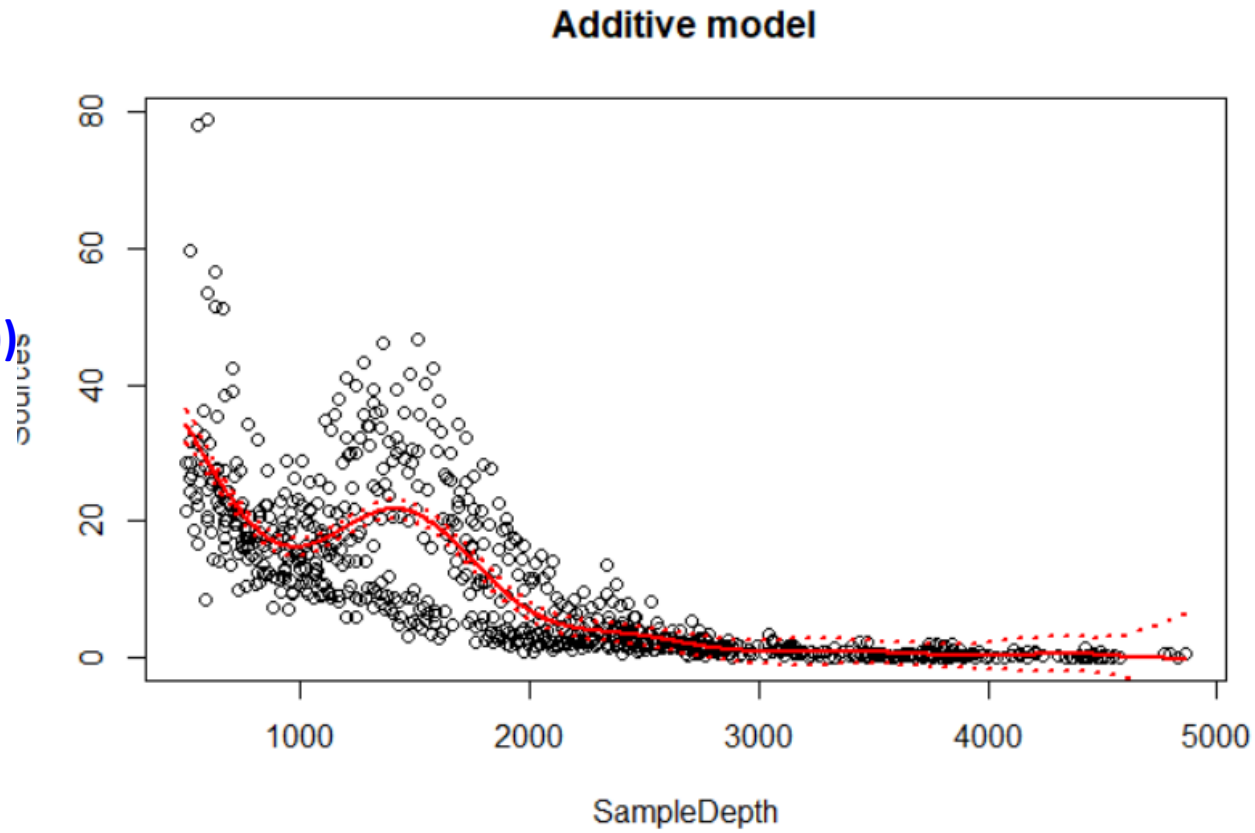




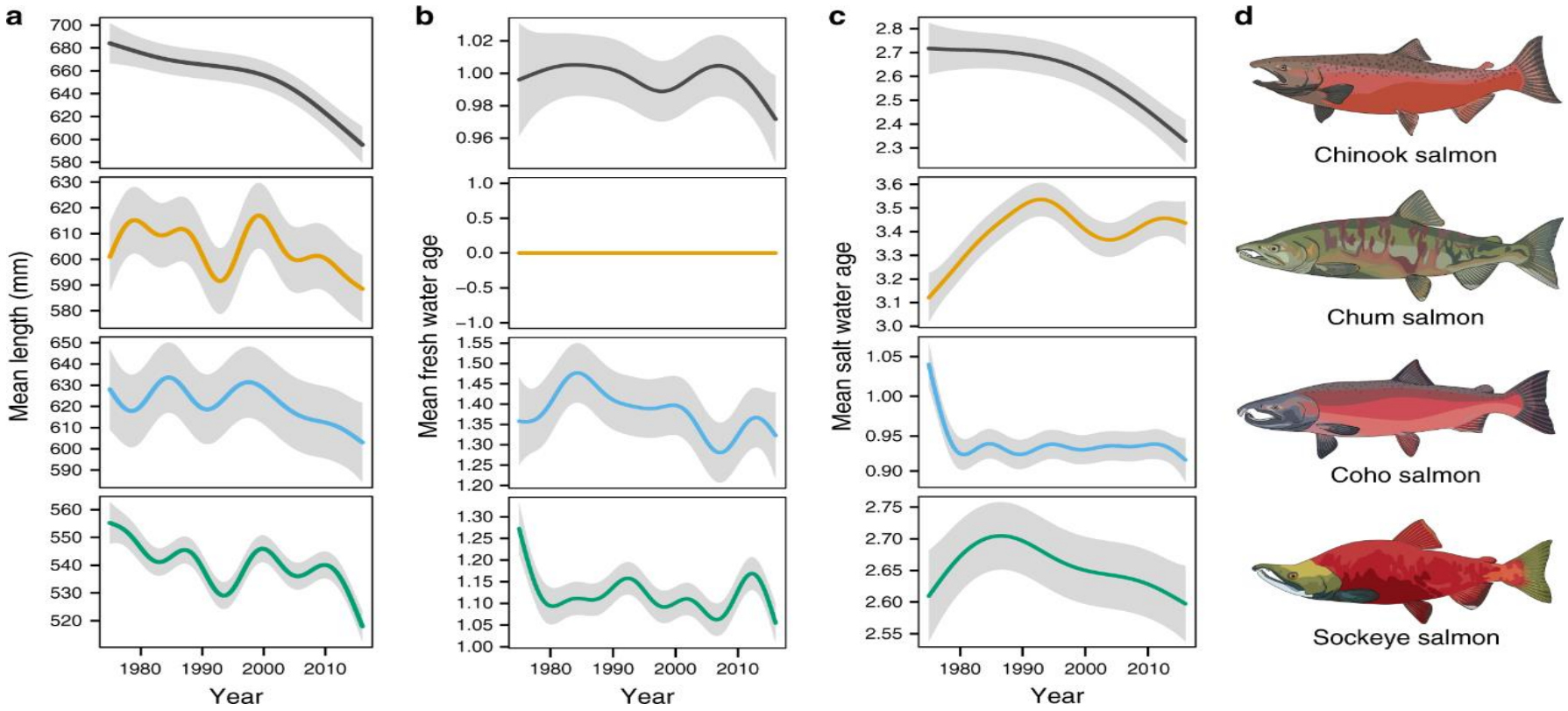
案例来源 **Mixed Effects Models and Extensions in Ecology with R**

非线性模型：additive model

```
library(mgcv)
am<-gam(Sources~s(SampleDepth),data=isit)
summary(am)
plot(Sources~SampleDepth,data=isit)
am_ci<-add_ci(isit, am, alpha = 0.05, names = c("lwr", "upr"))
lines(am_ci$SampleDepth, am_ci$pred, col = "blue", lwd =
2,lty="solid")
lines(am_ci$SampleDepth, am_ci$upr, col = "blue", lwd =
2,lty="dotted")
lines(am_ci$SampleDepth, am_ci$lwr, col = "blue", lwd =
2,lty="dotted")
```



gam模型的应用案例



chinook_yr <- gam(mean_length ~ LocationID + SASAP.Region + s(sampleYear) ,data=dat.ck)

Oke, K. B., C. J. Cunningham, P. A. H. Westley, M. L. Baskett, S. M. Carlson, J. Clark, A. P. Hendry, V. A. Karatayev, N. W. Kendall, J. Kibele, H. K. Kindsvater, K. M. Kobayashi, B. Lewis, S. Munch, J. D. Reynolds, G. K. Vick, and E. P. Palkovacs. 2020. Recent declines in salmon body size impact ecosystems and fisheries. Nature Communications **11**.

违反一般线性回归前提要求的几种类型

1）正态性

2）方差齐次性

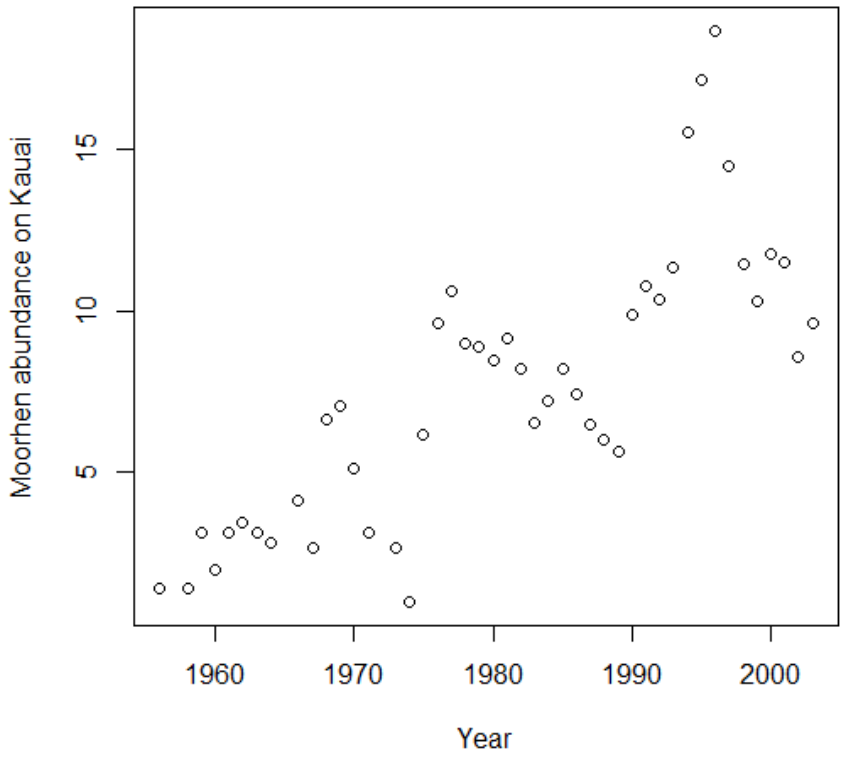
3）y和x之间是非线性关系

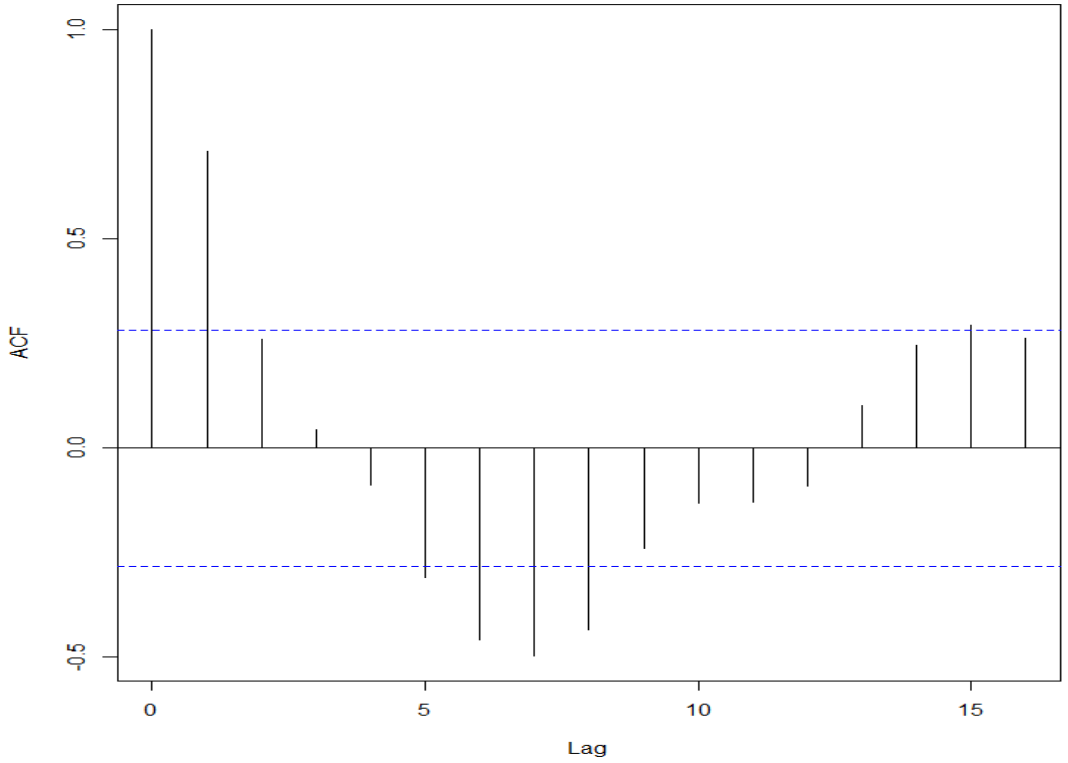
4）y_i相互独立（时间、空间、系统发育、嵌套）

违反y_i相互独立的几种情况

- 时间自相关
- 空间自相关
- 系统发育自相关
- 嵌套结构

违反y_i独立性原则—时间自相关





案例来源 **Mixed Effects Models and Extensions in Ecology with R**

时间自相关的处理—指定时间自相关的结构类型(gls)

- ##1）对称型自相关结构：不同年份的相关性相同

```
M_t1<-gls(Birds ~ Rainfall + Year, na.action = na.omit,
    correlation = corCompSymm(form =~ Year),
   data=Hawaii)
```

指定时间自相关的结构类型(gls)

##2）ar(1)型自相关结构：时间距离越远，y的相关性越弱
##一般情况下ar(1)更适合时间自相关(重复测量)数据

```
M_t2<-gls(Birds ~ Rainfall + Year, na.action = na.omit,
    correlation = corAR1(form =~ Year), data = Hawaii)
```

##自相关结构类型选择
AIC(M_t1,M_t2)
##查看更多时间自相关结构类型
?nlme::gls

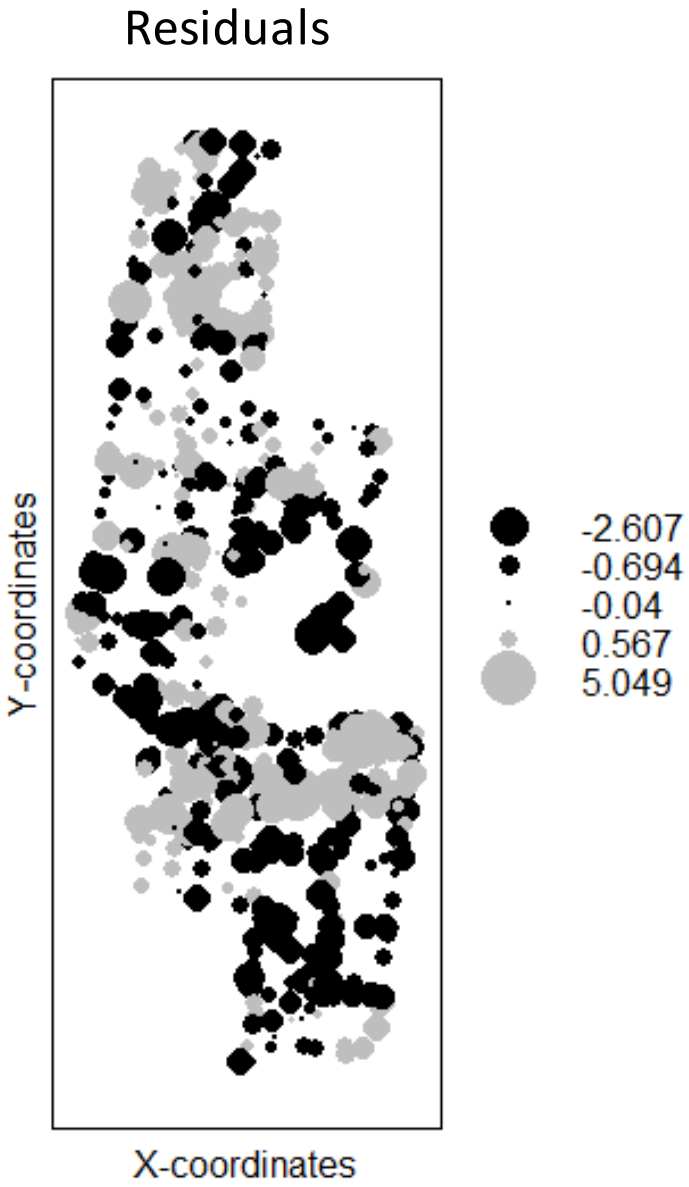
```
> ##自相关结构类型对比
> AIC(M_t1,M_t2)
      df       AIC
M_t1   5 230.4798
M_t2   5 199.1394
```

ar(1)型时间自相关结构

$$\sigma^2 \begin{bmatrix} 1 & \rho^1 & \rho^2 & \rho^3 \\ \rho^1 & 1 & \rho^1 & \rho^2 \\ \rho^2 & \rho^1 & 1 & \rho^1 \\ \rho^3 & \rho^2 & \rho^1 & 1 \end{bmatrix}$$

更多结构见Zuur 等 Mixed Effects Models and Extensions in Ecology with R 一书第6章

违反y_i独立性原则—空间自相关



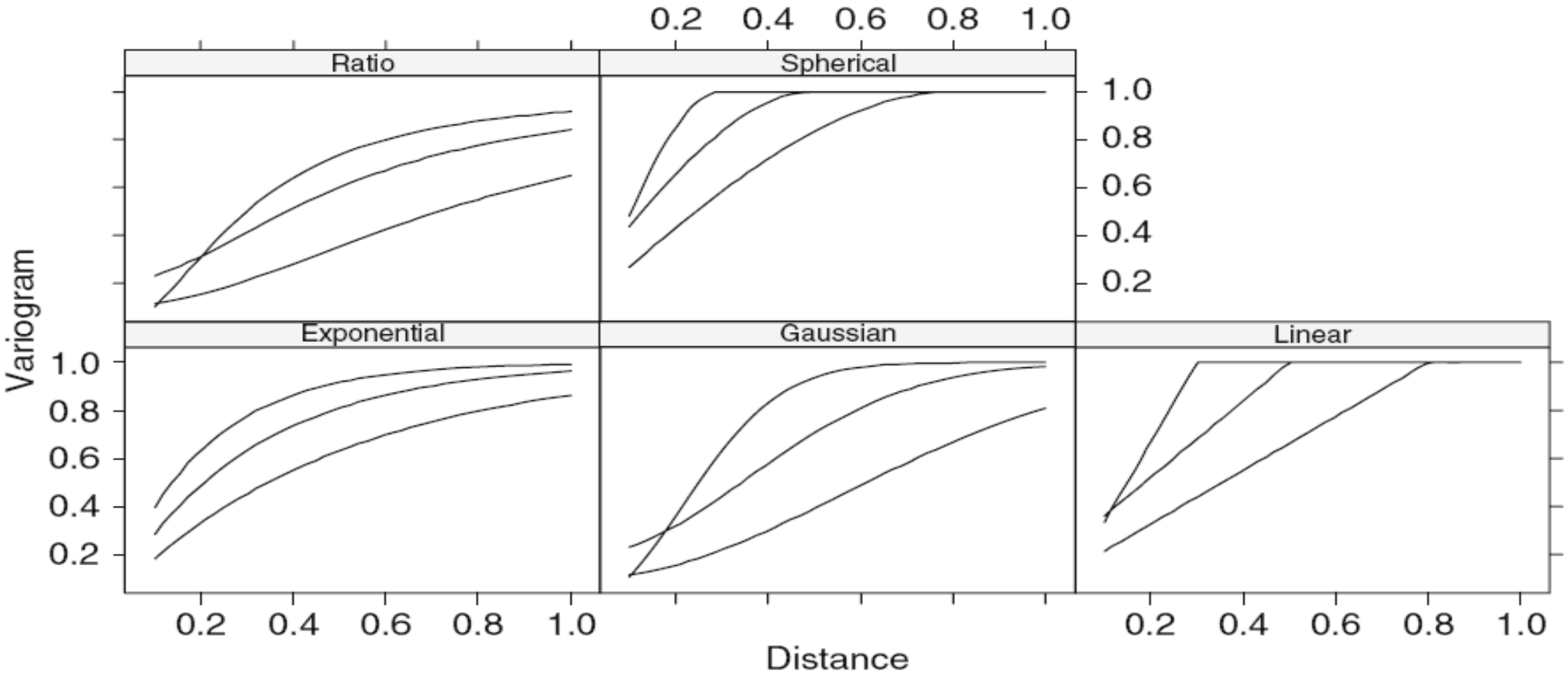
##R code

```
B1A<-gls(Bor ~ Wet,correlation=corSpher(form=~x+y, nugget=T),data=Boreality)
B1B<-gls(Bor ~ Wet,correlation=corRatio(form=~x+y, nugget=T),data=Boreality)
B1C<-gls(Bor ~ Wet,correlation=corGaus(form=~x+y, nugget=T),data=Boreality)
B1D<-gls(Bor ~ Wet,correlation=corExp(form=~x+y, nugget=T),data=Boreality)
##相关结构的选择
AIC(B1A,B1B,B1C,B1D)
```

```
> AIC(B1A,B1B,B1C,B1D)
    df      AIC
B1A  5 2739.072
B1B  5 2732.930
B1C  5 2736.292
B1D  5 2732.224
```

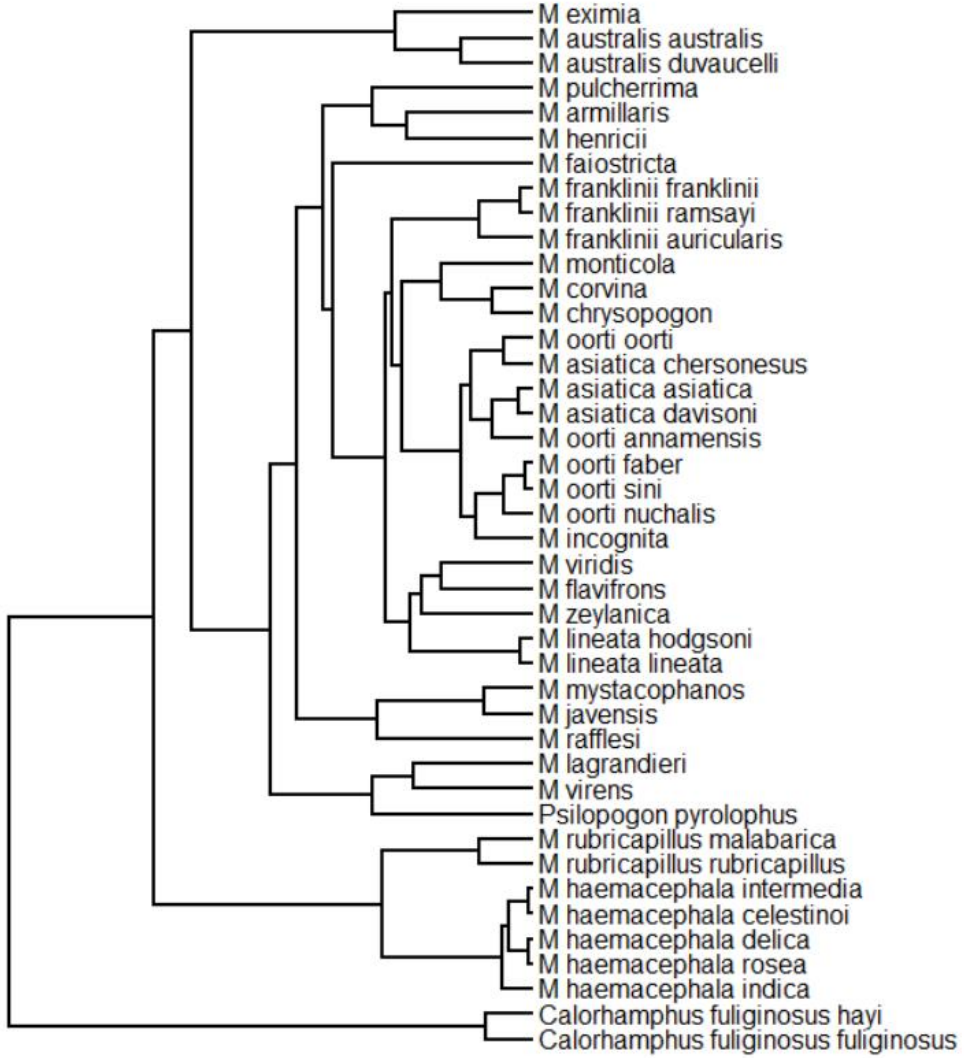
案例来源 **Mixed Effects Models and Extensions in Ecology with R**

不同类型的空间自相关格局



案例来源 **Mixed Effects Models and Extensions in Ecology with R**

违反y_i独立性原则-系统发育关系
应对方案：系统发育最小二乘回归(PGLS)




```
library(ape)
library(nlme)
library(geiger)
datos<-read.csv("Barbetdata.csv",header=TRUE, row.names = 1)
arbol<-read.nexus("BarbetTree.nex")
library(phytools)
plotTree(arbol)
library(caper)
arbol.cortado<-drop.tip(arbol, obj$tree_not_data)
datos<-read.csv("Barbetdata.csv",header=TRUE)
####构造一个数据集，包括要分析的数据和系统发育树信息
comp.data<-comparative.data(arbol.cortado, datos,
names.col="Species", vcv.dim=2, warn.dropped=TRUE)
```

案例来源：http://www.phytools.org/Cordoba2017/ex/4/PGLS.html

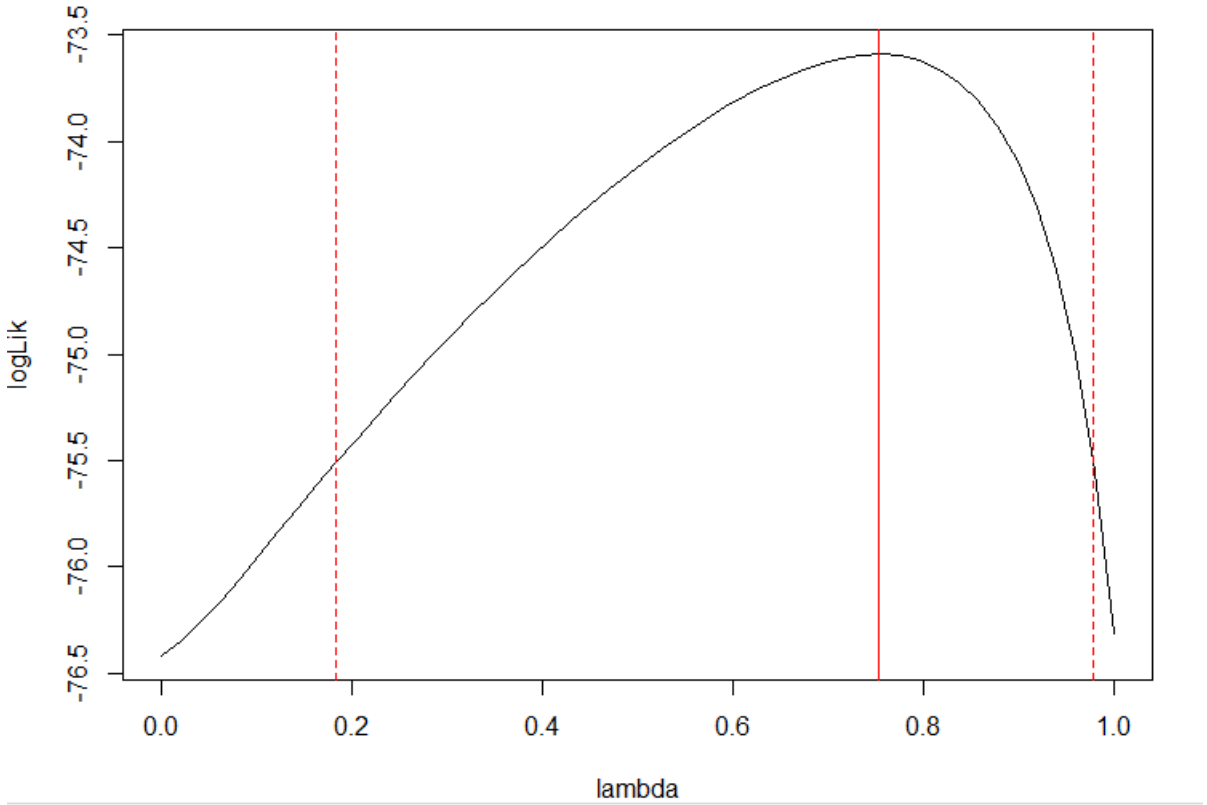
采用PGLS模型分析wing对指标patch的影响

pm1<-pgls(patch~wing, data=comp.data, lambda="ML")

summary(pm1)

```
> pm1<-pgls(patch~wing, data=comp.data, lambda="ML")
> summary(pm1)

Call:
pgls(formula = patch ~ wing, data = comp.data, lambda = "ML")

Residuals:
     Min       1Q   Median       3Q      Max
-1.0708 -0.2744  0.0189  0.5531  1.1698

Branch length transformations:

kappa  [Fix]  : 1.000
lambda [ ML]  : 0.752
   lower bound : 0.000, p = 0.017467
   upper bound : 1.000, p = 0.019576
   95.0% CI   : (0.183, 0.978)
delta  [Fix]  : 1.000

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  35.5609    14.4246  2.4653  0.01943 *
wing         -6.6440     3.1915 -2.0818  0.04571 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6115 on 31 degrees of freedom
Multiple R-squared: 0.1227,     Adjusted R-squared: 0.09435
F-statistic: 4.334 on 1 and 31 DF,  p-value: 0.04571
```

```
> pgls_lh<-pgls.profile(pgls_m1, which="lambda")
> plot(pgls_lh)
```



是不是感觉已经可以战无不胜了？

1）正态性

2）方差齐次性

3）y和x之间是非线性关系

4）y_i相互独立（时间、空间、系统发育、嵌套）

别着急，休整一下，马上给安排一个问题！！！