生物统计学与R手札

目录

Introduction

生物统计学：是统计学在生物学中的应用，是用数理统计的原理和方法来分析解释生命现象的一门科学，是研究生命过程中以样本推断总体的一门科学。

以数理统计原理为基础，应用到生物实验设计和分析领域，这便形成了生物统计学科的框架。所以学习过程是一般在学习概率论与数理统计的同时，对生物领域的实例进行相应分析和解读。

统计分析的流程：

- 设计或形成假设
- 设计相应的实验
- 收集数据
- 分析总结数据
- 形成推断

这与一般的科学研究过程是相同的，实质上生物科学研究的分析过程也就是生物统计分析的实例化。

数据的来源：

$$Source-of-data\begin{cases}Records\\Surveys\begin{cases}Comprehensive\\Sample\end{cases}\\Experiment\end{cases}$$

统计学上经常涉及到变量这个概念，它是指一种体现在不同对象上有不同数值的特征量，比如人的心率，对象是人，不同的人心率不一样，构成了心率的数值集。

变量有可以分为数值型变量和分类变量，前者通常指能够被测量的数值量，比如身高体重；后者一般指不能被数值化，通过评估生成的变量，比如视力的好坏，病人病情等级等，又可以依据是否有序分为连续型和非连续型变量，比如病人等级从低到高分为几类，这里就包含了变量的顺序(Rank)信息。

总体 *population*： 感兴趣的随机变量的数据总集。

样本 *sample*： 总体的抽样。

注意：采样尽量为保持一种随机的过程，这是在设计实验时非常需要注意的，不然结果不可靠。

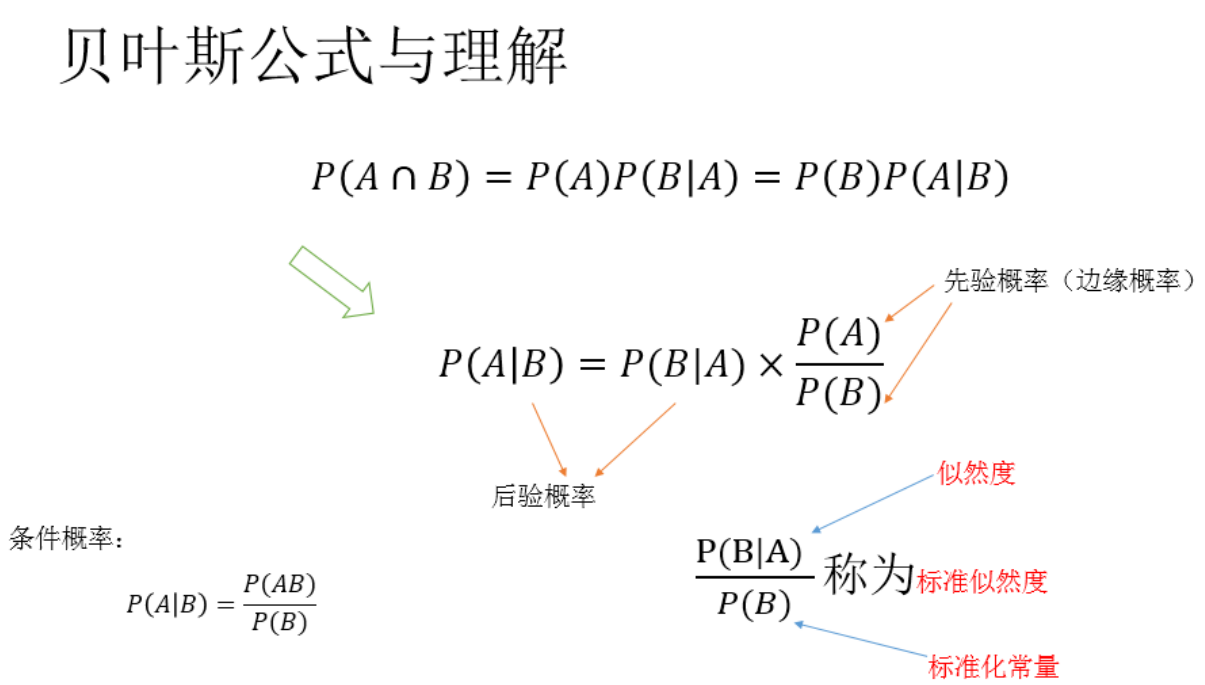
几种采样方式：

- **随机采样**： 通过计算机生成的伪随机数抽取相应的数据；
- **系统采样**： 将数据排序，每隔K个抽取一个数据；
- **Convenience Sampling**： 哪个方便用哪个；
- **分层抽样**： 将总体按相同特征分为至少两类，在类别中分别抽样。
- ……

上述抽样方法详细介绍及优缺点可以参考百度百科。

统计量是样本的描述量；参数是总体的描述量。

概率相关的基本概念大都不难，理解即可。需要注意的是贝叶斯公式，它在当今各大领域都非常常用，值得深挖。



如果A与B独立，那么P(B|A)=P(B)，此时标准似然度为1，P(A|B)最大

常见概率分布统计书上都有详细介绍，用的时候查询即可。

因为我们常用的数据都很难直接满足正态分布，单为什么用正态分布去理解和计算呢，这里有两个定理概念需要理解：

- **大数定理**就是样本均值在总体数量趋于无穷时依概率收敛于样本均值的数学期望（可不同分布）或者总体的均值（同分布）。
- **中心极限定理**就是一般在同分布的情况下，样本值的和在总体数量趋于无穷时的极限分布近似于正态分布。

参数估计

点估计是以抽样得到的样本指标作为总体指标的估计量，并以样本指标的实际值直接作为总体未知参数的估计值的一种推断方法；区间估计则是根据抽样指标和抽样平均误差推断总体指标的可能范围，它既说明推断的准确程度，同时也表明了推断结果的可靠程度。可见，点估计所推断的总体指标是一个确定的数值，而区间估计所推断的总体指标是一个数值域，这个值域受样本指标、极限误差和样本单位数等因素的影响。

- 点估计（Point Estimate）

  http://wiki.mbalib.com/wiki/点估计

  http://baike.baidu.com/view/635268.htm

  区间估计（Interval Estimation）/置信区间（Confidence interval）

http://wiki.mbalib.com/wiki/置信区间

http://baike.baidu.com/view/364109.htm

- 在置信度为$1-\alpha$置信度下的
  - 区间估计写为：$\hat{p}-E<p<\hat{p}+E$，点估计写为$p=\hat{p}\pm E$
  - E为误差限，$E=Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$
  - 由上述公式可以推出所需要的样本大小$n=\frac{(Z_{\alpha}/2)^2\hat{p}\hat{q}}{E^2}$
  - 总体方差已知时，估计均值μ使用z分布（u分布），$E=Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, n=[\frac{(Z_{\alpha}/2)\sigma}{E}]^2$
  - 总体方差未知时，估计均值μ使用t分布，$t=\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$，$E=t_{\alpha}/2\frac{s}{\sqrt{n}}$，$df=n-1$
- 方差估计
  - 卡方分布：$\chi^2=\frac{(n-1)s^2}{\sigma^2}$，
    $n=sample\quad size, s^2=sample\quad variance, \sigma^2=population\quad variance$
  - 方差的区间估计为$\frac{(n-1)s^2}{\chi_R^2}<\sigma^2<\frac{(n-1)s^2}{\chi_L^2}$

参数检验与非参数检验

统计中常常提到p值，它的实质是一个小概率事件发生的概率大小值。例如说某件事情的$p<0.05$指的是这件事情发生的概率不超过0.05，因为它发生的概率极小，所以在一般的实验（试验）中，很难碰到这样的事情。因此，当我们碰到这样的事情时，术语说这是一件显著的事情（$p<0.01$为极其显著）。实际实验过程中如果数据噪声服从高斯分布（正态分布），这样的事情应当不会发生（概率很小嘛），那么就应该是其他因素导致的。比如说两组数据进行对比时，如果这两组样本是从同一个总体抽出来的，就应该没什么差异（一般用总体均值μ的假设检验）；如果两组样本经过不同的处理，发现有显著差异（概率很小的事情发生了），说明这两组不同处理的样本映射为不同的总体，我们以此结果来推断两个不同处理的总体它们之间有显著性的差异（所以说实验才是可以重复的，因为每次实验都是对总体的抽样）。

假设检验的步骤

一般包括以下四个步骤：

1. 提出假设：一般做两个彼此独立的假设，一个是无效假设或零假设（null hypothesis 很常用），记做H_0；另一个是备择假设，称为H_A。所谓的无效意指处理效应与总体参数之间没有真实的差异，实验结果中的差异是误差导致的。
2. 确定显著水平：常用$\alpha=0.05\quad or\quad \alpha=0.01$
3. 计算概率（p值）：有双尾和单尾两种
4. 推断是否接受假设

这方面的知识网上很多，可以参考百度百科或其他资料。

总体的单样本参数检验

总体方差已知时对总体均值检验

如果总体方差已知，使用z分布（标准正态分布）进行计算

$$P(Z_{\alpha/2} \leq \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \leq Z_{1-\alpha/2}) = 1-\alpha$$

$$P(\bar{X} - Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

总体方差未知时对总体均值进行检验

如果总体方差未知，使用t分布进行计算

$$P(\bar{X} - t_{df,1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + t_{df,1-\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

（修正一下，上图总体方差未知，sigma应该用样本标准差s替代）

计算时根据要求，算出z值或者t值，然后与置信度（t分布需要看自由度）下的z（或t）统计量进行对比。观察是在否定区间还是接受区间，从而完成对假设的推断。

当检验是单边时，上述公式的$\mathbf{1-\alpha/2}$变成$\mathbf{1-\alpha}$

在R中，统计量与分布的计算和图形的绘制可能涉及到的一些函数的使用，可以参考数值与字符处理函数，基本统计分析，基本图形绘制。

以下是常用的概率函数

d = 密度函数

p = 分布函数

q = 分位数函数

r = 生成随机数

常见的概率函数列于下表

分布名称	缩写	分布名称	缩写
Beta分布	beta	Logistic分布	logis
二项分布	binom	多项分布	multinon
柯西分布	cauchy	负二项分布	nbinom
（非中心）卡方分布	chisq	正态分布	norm
指数分布	exp	泊松分布	pois
F分布	f	Wilcoxon符号秩分布	signrank
Gamma分布	gamma	t分布	t
几何分布	geom	均匀分布	unif
超几何分布	hyper	Weibull分布	weibull
对数正态分布	lnom	Wilcoxon秩和分布	wilcox

它的使用概率函数形如：`[dpqr] distribution_abbreviation()`

前面一部分是选择计算哪种类型（是概率函数还是分布函数..），后面一部分是指定使用的分布。

比如说`qt()`就是计算t分布的分位数函数，函数具体的参数调用可以使用`help()`进行查询。

在对单样本的总体方差进行检验时，常用卡方分布，两样本则用F分布。

公式分别为：

$$\chi^2 = \frac{(k-1)s^2}{\sigma^2} \quad df = k-1$$

$$F = \frac{s_1^2}{s_2^2} \quad df_1 = n_1 - 1, df_2 = n_2 - 1$$

注意，卡方分布不仅可以用来检验方差同质性，还可以进行适合性和独立性检验，后两者用来判断实际观测值与理论观测值的偏离程度。

当对总体频率进行检验时，如果不满足中心极限定理，则不可以用正态分布进行检验，转而使用二项分布进行检验。

小结：

One sample parametric test usually assumes that samples are randomly selected from normal distribution.

- ✤ (1) The mean of a normal distribution with unknown variance (one-sample t test)
- ✤ (2) The mean of a normal distribution with known variance (one-sample z test)
- ✤ (3) The variance of a normal distribution (one-sample 2 test)
- ✤ (4) The parameter p of a binomial distribution (one-sample binomial test)

类型1与类型2错误

两种类型错误及其关系

> 第一类错误(typelerror)，Ⅰ型错误，拒绝了实际上成立的H0，，即错误地判为有差别，这种弃真的错误称为Ⅰ型错误。其概率大小用即检验水准用α表示。α可取单尾也可取双尾。假设检验时可根据研究目的来确定其大小，一般取0.05，当拒绝H0时则理论上理论100次检验中平均有5次发生这样的错误.。
> 第二类错误(typell error)。Ⅱ型错误，接受了实际上不成立的H0 ，也就是错误地判为无差别，这类取伪的错误称为第二类错误。第二类错误的概率用β表示，β的大小很难确切估计。
> 二者的关系是，当样本例数固定时，α愈小，β愈大；反之，α愈大，β愈小。因而可通过选定α控制β大小。要同时减小α和β，唯有增加样本例数。统计上将1-β称为检验效能或把握度(power of a test)，即两个总体确有差别存在，而以α为检验水准，假设检验能发现它们有差别的能力。实际工作中应权衡两类错误中哪一个重要以选择检验水准的大小。

由此引申出几个公式概念，包括灵敏度、特异性、假阳性率等，它们的计算方式如下：

It we count the number of cases we reject or accept null hypothesis and compare with the original answer, we have:

	H_0 is true	H_a is true
Accept H_0	True Negatives (TN)	False Negatives (FN)
Reject H_0	False Positives (FP)	True Positives (TP)

$$specificity = \frac{TN}{FP+TN} = 1-\alpha \qquad type\ I\ error\ \alpha = \frac{FP}{FP+TN}$$ false positive rate

$$sensitivity = \frac{TP}{TP+FN} = 1-\beta \qquad type\ II\ error\ \beta = \frac{FN}{TP+FN}$$

$$1-\beta = \frac{TP}{TP+FN}$$ true positive rate

这些概念常用来计算ROC曲线，该曲线在评判模型的有效性中非常流行。

简单地讲，ROC曲线描绘了灵敏性（真阳性率）随假阳性率（1-特异性）的变化趋势。

AUC则是指ROC曲线下方围成的面积，数值越大，分类器（模型）效果越好。

详细参考：ROC曲线概念；ROC和AUC介绍以及如何计算AUC

功效（真阳性率），如果功效过低，那么就算处理不同导致有显著性差异也很难检测出来，所以在进行检验时，我们需要对它进行控制。

统计检验的功效计算（分别使用与正态分布、t分布与样本频率检验）

	z-test	t-test
left tail	$Power = F[z_\alpha + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}]$	$Power = F[t_\alpha + \frac{\mu_1 - \mu_0}{s/\sqrt{n}}]$
right tail	$Power = F[z_\alpha + \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}]$	$Power = F[t_\alpha + \frac{\mu_1 - \mu_0}{s/\sqrt{n}}]$
two tailed	$Power = F[z_{\alpha/2} + \frac{\lvert\mu_0 - \mu_1\rvert}{\sigma/\sqrt{n}}]$	$Power = F[t_{\alpha/2} + \frac{\lvert\mu_0 - \mu_1\rvert}{s/\sqrt{n}}]$

left-tailed test $$Power = F(\sqrt{\frac{p_0 q_0}{p_1 q_1}}(z_\alpha + \frac{p_0 - p_1}{\sqrt{\frac{p_0 q_0}{n}}}))$$

right-tailed test $$Power = F(\sqrt{\frac{p_0 q_0}{p_1 q_1}}(z_\alpha + \frac{p_1 - p_0}{\sqrt{\frac{p_0 q_0}{n}}}))$$

two-tailed test $$Power = F(\sqrt{\frac{p_0 q_0}{p_1 q_1}}(z_{\alpha/2} + \frac{\lvert p_1 - p_0\rvert}{\sqrt{\frac{p_0 q_0}{n}}}))$$

效应值：$\frac{|\mu_0-\mu_1|}{\sigma}$，表示两个总体的平均值差异

功效分析可以帮助在给定置信度的情况下，判断检测到给定效应值所需的样本量。反过来，它也可以帮助你在给定置信度水平情况下，计算在某个样本量内能检测到给定效应值的概率。如果概率低得难以接受，修改或放弃这个实验将是一个明智的选择。

在研究过程时，研究者通常关注四个量：样本大小、显著性水平、功效和效应值。

- 样本大小指实验设计中每种条件中观测的数目。
- 显著性水平（也称为alpha）由I型错误的概率来定义。也可以把它看作发现效应不发生的概率。
- 功效通过1减去II型错误的概率来定义。可以把它看作真实效应发生的概率。
- 效应值指的是在备择或研究假设下效应的值。效应值的表达值依赖于假设检验中使用的统计方法。

四个量紧密相关，给定其中任意三个量，便可以推算第四个量。

我们常常会使用到t分布检验相关的功效分析，这里有一篇值得参考的博文找出t检验的效应大小，对要流氓 say no！。

功效分析使用到的一些函数和包可以参考R语言中的功效分析。

pwr.t.test {pwr} R Documentation

Power calculations for t-tests of means (one sample, two samples and paired samples)

Description

Compute power of tests or determine parameters to obtain target power (similar to power.t.test).

Usage

```
pwr.t.test(n = NULL, d = NULL, sig.level = 0.05, power = NULL,
    type = c("two.sample", "one.sample", "paired"), alternative = c("two.sided",
        "less","greater"))
```

Arguments

`n`	Number of observations (per sample)
`d`	Effect size
`sig.level`	Significance level (Type I error probability)
`power`	Power of test (1 minus Type II error probability)
`type`	Type of t test : one- two- or paired-samples
`alternative`	a character string specifying the alternative hypothesis, must be one of "two.sided" (default), "greater" or "less"

两样本参数检验

type of test	population variance known	test statistic	degree freedom	power
One sample (paired two sample)	√	$z=\frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$		$\Phi\left(Z_{\alpha/2}+\frac{\delta}{\sigma\sqrt{\frac{1}{n}}}\right)$
	×	$t=\frac{\bar{X}-\mu_0}{\frac{s}{\sqrt{n}}}$	$n-1$	$\Phi\left(t_{\alpha/2}+\frac{\delta}{s\sqrt{\frac{1}{n}}}\right)$
Two independent sample	√	$z=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$		$\Phi\left(Z_{\alpha/2}+\frac{\delta}{\sqrt{\frac{\sigma_1^2+\sigma_2^2/k}{n_1}}}\right)$
	× variance equal	$t=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{s_P\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	n_1+n_2-2	$\Phi\left(t_{\alpha/2}+\frac{\delta}{sp\sqrt{\frac{1+1/k}{n_1}}}\right)$
	× variance unequal	$t=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$	$\frac{(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2})^2}{\frac{s_1^4}{n_1^2(n_1-1)}+\frac{s_2^4}{n_2^2(n_2-1)}}$	$\Phi\left(t_{\alpha/2}+\frac{\delta}{\sqrt{\frac{s_1^2+s_2^2/k}{n_1}}}\right)$

图中$\delta=|\mu_1-\mu_2|$

功效分析和相关检验可以参考上一节。

方差分析非参数检验

非参数检验(Nonparametric tests)是统计分析方法的重要组成部分，它与参数检验共同构成统计推断的基本内容。参数检验是在总体分布形式已知的情况下，对总体分布的参数如均值、方差等进行推断的方法。但是，在数据分析过程中，由于种种原因，人们往往无法对总体分布形态作简单假定，此时参数检验的方法就不再适用了。非参数检验正是一类基于这种考虑，在总体方差未知或知道甚少的情况下，利用样本数据对总体分布形态等进行推断的方法。由于非参数检验方法在推断过程中不涉及有关总体分布的参数，因而得名为“非参数”检验。

也就是说，之前的参数检验，我们在对数据分析之前，需要假定该数据的总体服从某种分布，而这些分布的假定是需要前提条件的，其中最重要的是正态性，而往往我们的数据很难达到这样的要求，甚至对于总体的分布完全一无所知。这个时候我们就可以使用非参数检验。（当然两者之间的优缺点对比还有很多）

一般来说，能用参数检验尽量使用参数检验，因为它的统计效力远高于非参数检验，这也是为什么t检验在文献中非常流行的原因。

非参数检验的种类非常之多，可以参考百度百科，其中常用的是符号检验与符号秩检验。

下表汇出了对总体均值进行检验时，参数和非参数的常用检验对比。

	parametric test	non-parametric test
one sample	t-test	sign test, wilcoxon signed-rank test
two sample	t-test of related samples	sign test, wilcoxon signed-rank test
	t-test of independent samples	wilcoxon rank-sum test / Mann whitney U test

符号检验与秩和检验两种方法相比较，符号检验只考虑样本差数的符号；秩和检验考虑样本差数的符号和样本差数的顺序。

符号检验法是通过两个相关样本的每对数据之差的符号进行检验，从而比较两个样本的显著性。具体地讲，若两个样本差异不显著，正差值与负差值的个数应大致各占一半。

符号检验与参数检验中相关样本显著性t检验相对应，当资料不满足参数检验条件时，可采用此法来检验两相关样本的差异显著性。(http://wiki.mbalib.com/wiki/%E7%AC%A6%E5%8F%B7%E6%A3%80%E9%AA%8C)

秩和检验方法最早是由维尔克松提出，叫维尔克松两样本检验法。后来曼—惠特尼将其应用到两样本容量不等的情况，因而又称为曼—惠特尼U检验。这种方法主要用于比较两个独立样本的差异。
(http://wiki.mbalib.com/wiki/%E7%A7%A9%E5%92%8C%E6%A3%80%E9%AA%8C)

曼-惠特尼U检验又称**"曼-惠特尼秩和检验"**，是由H.B.Mann和D.R.Whitney于1947年提出的。它假设两个样本分别来自除了总体均值以外完全相同的两个总体，目的是检验这两个总体的均值是否有显著的差别。

曼-惠特尼秩和检验可以看作是对两均值之差的参数检验方式的T检验或相应的大样本正态检验的代用品。由于曼-惠特尼秩和检验明确地考虑了每一个样本中各测定值所排的秩，它比符号检验法使用了更多的信息。
(http://wiki.mbalib.com/wiki/%E6%9B%BC-%E6%83%A0%E7%89%B9%E5%B0%BCU%E6%A3%80%E9%AA%8C)

上述文字后链接都有详细介绍和实例。

多重检验矫正

数据分析中常碰见多重检验问题(multiple testing).Benjamini于1995年提出一种方法,通过控制FDR(False Discovery Rate)来决定P值的域值。
假设你挑选了R个差异表达的基因，其中有S个是真正有差异表达的，另外有V个其实是没有差异表达的，是假阳性的.实践中希望错误比例Q＝V/R平均而言不能超过某个预先设定的值（比如0.05），在统计学上，这也就等价于控制FDR不能超过5％.
根据Benjamini在他的文章中所证明的定理，控制fdr的步骤实际上非常简单。
设总共有m个候选基因，每个基因对应的p值从小到大排列分别是p(1),p(2),...,p(m),则若想控制fdr不能超过q，则只

需找到最大的正整数i，使得 p(i)<= (i*q)/m.然后，挑选对应p(1),p(2),...,p(i)的基因做为差异表达基因，这样就能从统计学上保证fdr不超过q。

Bonferroni校正

如果在同一数据集上同时检验n个独立的假设，那么用于每一假设的统计显著水平，应为仅检验一个假设时的显著水平的1/n。举个例子：如要在同一数据集上检验两个独立的假设，显著水平设为常见的0.05。此时用于检验该两个假设应使用更严格的0.025。即0.05* (1/2)。该方法是由Carlo Emilio Bonferroni发展的，因此称Bonferroni校正。这样做的理由是基于这样一个事实：在同一数据集上进行多个假设的检验，每20个假设中就有一个可能纯粹由于概率，而达到0.05的显著水平。

FDR计算

- It is not only the FDR that needs to be controlled, but often controlling the FNR (false negative rate) is equally important. **If the FNR is large, we may miss important biological associations**.
- Some definitions:

		Test	
		P	**N**
Hypotheses	**A**	TP	FN
	H (null)	FP	TN

SE = p(P|A) = TP/A
FNR = p(N|A) = FN/A = 1 – SE

SP = p(N|H) = TN/H
FPR = p(P|H) = FP/H = 1 - SP

PPV = p(A|P) = TP/P
FDR = p(H|P) = FP/P = 1 - PPV

NPV = p(H|N) = TN/N
FNDR = p(A|N) = FN/N = 1 - NPV

Typical example

	P	N	Total
A	300	200	500
H	200	9300	9500
Total	500	9500	10000

SE = TP/A = 300/500 = 0.8 ➔ FNR=0.2
SP = TN/H= 9300/9500= 0.98 ➔ FPR=0.02

FDR = FP/P = 200/500 = 0.4

This example shows that although the Sensitivity and Specificity measures are pretty high, we have an unacceptably large FDR and a fairly large FNR. This is all because we have a large number of tests (10000) and only a small proportion that are truly differentially altered (500/10000, i.e. 5%).

方差分析

单因素方差分析

分析流程：

(1) **Compute SS (Sum of Squares)**

$$SS_{between} = \sum_j \sum_i (\bar{X}_j - \bar{\bar{X}})^2$$

$$SS_{within} = \sum_j \sum_i (x_{ij} - \bar{X}_j)^2$$

(2) **Compute df**

$$df_{between} = k - 1,\ df_{within} = n - k$$

(3) **Compute MS**

$$Between\ MS = \frac{SS_{between}}{k-1}$$

$$Within\ MS = \frac{SS_{within}}{n-k}$$

(4) **Compute F ratio**

$$F = \frac{Between\ MS}{Within\ MS}$$

形成列联表

Source of variation	SS	*df*	MS	*F* statistic	*p*-value
Between	$\sum_{j=1}^{k} n_j \bar{X}_j^2 - \frac{\bar{\bar{X}}^2}{n} = A$	$k-1$	$\frac{A}{k-1}$	$\frac{A/(k-1)}{B/(n-k)} = F$	$Pr(F_{k-1,n-k} > F)$
Within	$\sum_{j=1}^{k} (n_j - 1) {s_j}^2 = B$	$n-k$	$\frac{B}{n-k}$		
Total	Between SS + Within SS				

$$Between\ SS = \sum_{j=1}^{k} n_j \bar{X}_j^2 - \frac{\bar{\bar{X}}^2}{n}$$

$$Within\ SS = \sum_{j=1}^{k} (n_j - 1) {s_j}^2$$

课件8中有一个step-by-step ANOVA按步骤进行单因素方差分析计算。

R中一步搞定可以使用`aov()`与`lm()`函数。参考

方差分析主要用于两个及以上不同组实验的分析，探究整体是否存在显著性，如果存在显著性差异，进一步需要配对t检验找出存在差异的组。

R一个非常好用的函数是`TukeyHSD()`。检测方差同质性则使用`bartlett.test()`，`leveneTest()`函数。

单因素方差分析可以用`oneway.test()`函数，设定方差相等时与`aov()`结果相同。

做方差分析时，需要注意使用的模型(https://wenku.baidu.com/view/5516ebcabe23482fb5da4c5b.html)。大致分为三类：固定效应模型，随机效应模型以及混合效应模型。该概念在李春喜《生物统计学》88页有详细介绍。

简单来说，固定模型指各个处理的效应是一个固定的常量，比如不同温度条件下小麦籽粒的发芽实验，处理的水平（温度）是特意选择的，所以得到的结论也仅限于所选定的这几个水平；随机效应指各处理的效应是随机因素，比如不同纬度下桃树对地理条件的适应情况，由于气候、土壤等条件无法人为控制，属于随机因素，就需要随机模型来处理。从而实验所得出的结论可以推广到随机因素的所有水平上。混合模型即为前两者的叠加。

不同的模型在平方和和自由度的计算是相同的，但是假设检验时F值得计算公式是不同的。模型分析的侧重点也不同。对于单因素方差分析来说，固定模型与随机模型无多大区别。

$$x_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

Fixed-effects model	**Random-effects model**
$\sum_{j=1}^{k} \alpha_j = 0$	$\alpha_j \sim N(0, \sigma_A^2)$
$\epsilon_{ij} \sim N(0, \sigma^2)$	$\epsilon_{ij} \sim N(0, \sigma^2)$
$H_0 : \alpha_j = 0$	$H_0 : \sigma_A^2 = 0$

ANOVA is performed in the same way for both models

$E(Within\ MS) = \sigma^2$	$E(Within\ MS) = \sigma^2$
$E(BetweenMS) = \sigma^2 + \frac{n_0 \sum_j {\alpha_j}^2}{k-1}$	$E(BetweenMS) = \sigma^2 + n_0\sigma_A^2$

all samples have equal size of n_0

$E(BetweenMS) = \sigma^2 + n_0\sigma_A^2$ if all samples have equal size of n_0

$E(BetweenMS) = \sigma^2 + n'\sigma_A^2 \qquad n' = \frac{\sum_{j=1}^{k} n_j - \frac{\sum_{j=1}^{k} n_j^2}{\sum_{j=1}^{k} n_j}}{k-1}$

$$\sigma_A^2 = \frac{E(Between\ MS) - E(Within\ MS)}{n_0}$$

$$\hat{\sigma}_A^2 = \frac{Between\ MS - Within\ MS}{n_0}$$

$$Total\ variance = \sigma^2 + \sigma_A^2$$

$$estimated\ total\ variance = WIthin\ MS + \hat{\sigma}_A^2$$

$$Component\ of\ variance = \frac{\hat{\sigma}_A^2}{Within\ MS + \hat{\sigma}_A^2}$$

非参数检验

与t检验类似，方差分析中面对方差不同质或者所处理的数据是有序性而不是数值型时无能为力。因此需要相应的非参数检验来解决这样一类问题。Kruskal-Walls test就是为这个目的开发的。它就像多重样本（multiple-sample）版本的Wilcoxon秩和检验一样。

The Kruskal-Wallis Test

To compare the means of k samples ($k > 2$) using nonparametric methods, use the following procedure:

(1) Pool the observations over all samples, thus constructing a combined sample of size $N = \Sigma n_i$

(2) Assign ranks to the individual observations, using the average rank in the case of tied observations.

(3) Compute the rank sum R_i for each of the k samples.

Test statistic

$$H = H^* = \frac{12}{N(N+1)} \times \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

$$H = \frac{H^*}{1 - \dfrac{\sum_{j=1}^{g} (t_j^3 - t_j)}{N^3 - N}}$$ with tied groups

Under H_0, H is χ^2 distributed

if $H > \chi^2_{k-1,1-\alpha}$ then reject H_0

if $H \le \chi^2_{k-1,1-\alpha}$ then accept H_0

$$p = Pr\left(\chi^2_{k-1} > H\right)$$

This test procedure should be used only if minimum $n_i \ge 5$ (i.e., if the smallest sample size for an individual group is at least 5).

下面截一个实例（对于前面的公式看，比较容易理解）。

Ocular anti-inflammatory effects of four drugs on lid closure after administration of arachidonic acid

Rabbit Number	Indomethacin		Aspirin		Piroxicam		BW755C	
	Score[a]	Rank	Score	Rank	Score	Rank	Score	Rank
1	+ 2	13.5	+ 1	9.0	+ 3	20.0	+ 1	9.0
2	+ 3	20.0	+ 3	20.0	+ 1	9.0	0	4.0
3	+ 3	20.0	+ 1	9.0	+ 2	13.5	0	4.0
4	+ 3	20.0	+ 2	13.5	+ 1	9.0	0	4.0
5	+ 3	20.0	+ 2	13.5	+ 3	20.0	0	4.0
6	0	4.0	+ 3	20.0	+ 3	20.0	− 1	1.0

[a](Lid-closure score at baseline − lid-closure score at 15 minutes)$_{\text{drug eye}}$ − (lid-closure score at baseline − lid-closure score at 15 minutes)$_{\text{saline eye}}$

Average ranks with ties

Lid-closure score	Frequency	Range of ranks	Average rank
−1	1	1	1.0
0	5	2–6	4.0
+1	5	7–11	9.0
+2	4	12–15	13.5
+3	9	16–24	20.0

(1) Rank sum for each sample

$$R_1 = 13.5 + 20.0 + \ldots + 4.0 = 97.5$$
$$R_2 = 9.0 + 20.0 + \ldots + 20.0 = 85.0$$
$$R_3 = 20.0 + 9.0 + \ldots + 20.0 = 91.5$$
$$R_4 = 9.0 + 4.0 + \ldots + 1.0 = 26.0$$

(2) Compute test statistic

$$H = \frac{\frac{12}{24\times 25}\times\left(\frac{97.5^2}{6}+\frac{85.0^2}{6}+\frac{91.5^2}{6}+\frac{26.0^2}{6}\right)-3(25)}{1-\frac{(5^3-5)+(5^3-5)+(4^3-4)+(9^3-9)}{24^3-24}}$$

$$= \frac{0.020\times 4296.583-75}{1-\frac{1020}{13{,}800}} = \frac{10.932}{0.926} = 11.804$$

(3) Get χ^2 critical value (df=3)

$$\chi^2_{3,.99} = 11.34,\ \chi^2_{3,.995} = 12.84.$$

$$11.804 > 11.34$$

Reject null hypothesis

在R中，使用函数`kruskal.test()`即可用进行K-W检验。

一但拒绝原假设（有显著性差异），接着使用`pairwise.wilcox.test()`进行两两配对检验，可用指定矫正方法。

两因素方差分析

单因素方差分析指一个处理水平，两因素方差分析指两个，多个因素的分析类似。

比如探究某几种药物对某种病（比如癌症）的治疗效果，这个是单因素的，如果我们将病人按性别分为两类，这时就会多出一个性别因素，构成了两因素的方差分析（药物和性别对癌症治疗效果的影响）。

说实话，这个理解不难，手工计算就比较麻烦了。在R中使用函数加上公式可以很容易地表达因变量和自变量的关系，从而完成方差分析。

aov()函数的语法为aov(*formula*, data = *dataframe*)，表9-4列举了表达式中可以使用的特殊符号。表9-4中的y是因变量，字母A、B、C代表因子。

表9-4　R表达式中的特殊符号

符　号	用　法
~	分隔符号，左边为响应变量，右边为解释变量。例如，用A、B和C预测y，代码为y~ A + B + C
+	分隔解释变量
:	表示变量的交互项。例如，用A、B和A与B的交互项来预测y，代码为y~ A + B + A:B
*	表示所有可能交互项。代码y~ A * B * C可展开为y ~ A + B + C + A:B + A:C + B:C + A:B:C
^	表示交互项达到某个次数。代码y ~ (A + B + C)^2可展开为y ~ A + B + C + A:B + A:C + B:C
.	表示包含除因变量外的所有变量。例如，若一个数据框包含变量y、A、B和C，代码y ~ .可展开为y ~ A + B + C

双因素ANOVA	y ~ A * B
含两个协变量的双因素ANCOVA	y ~ x1 + x2 + A*B

下面只截取相应的公式（分随机和固定效应模型）

两因素重复测量方差分析

Two-way ANOVA with replication

$$y_{i..}=\sum_{j=1}^{b}\sum_{k=1}^{n}y_{ijk} \qquad \overline{y_{i..}}=\frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n}y_{ijk}$$

$$y_{.j.}=\sum_{i=1}^{a}\sum_{k=1}^{n}y_{ijk} \qquad \overline{y_{.j.}}=\frac{1}{an}\sum_{i=1}^{a}\sum_{k=1}^{n}y_{ijk}$$

$$y_{ij.}=\sum_{k=1}^{n}y_{ijk} \qquad \overline{y_{ij.}}=\frac{1}{n}\sum_{k=1}^{n}y_{ijk}$$

$$y_{...}=\sum_{j=1}^{b}\sum_{k=1}^{n}\sum_{i=1}^{a}y_{ijk} \qquad \overline{y_{...}}=\frac{1}{abn}\sum_{j=1}^{b}\sum_{k=1}^{n}\sum_{i=1}^{a}y_{ijk}$$

Two-way ANOVA for fixed effect

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk}-\bar{Y}_{...})^2 = bn\sum_{i=1}^{a}(\bar{Y}_{i..}-\bar{Y}_{...})^2 + an\sum_{j=1}^{b}(\bar{Y}_{.j.}-\bar{Y}_{...})^2 + n\sum_{i=1}^{a}\sum_{j=1}^{b}(Y_{ijk}-\bar{Y}_{i..}-\bar{Y}_{.j.}+\bar{Y}_{...})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk}-\bar{Y}_{ij.})^2$$

SS_A SS_B SS_{AB} SS_E

$$SS_T = SS_A + SS_B + SS_{AB} + SS_E$$

$$\bar{Y}_{i..} = \frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n}Y_{ijk} \qquad \bar{Y}_{.j.} = \frac{1}{an}\sum_{i=1}^{a}\sum_{k=1}^{n}Y_{ijk} \qquad \bar{Y}_{...} = \frac{1}{abn}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}Y_{ijk}$$

Null hypothesis

$$H_{0A}: \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$$

$$H_{0B}: \beta_1 = \beta_2 = \cdots = \beta_b = 0$$

$$H_{0AB}: \delta_{ij} = 0, \quad i = 1,\cdots,a; \quad j = 1,\cdots b$$

F-statistics

$$F_A = \frac{SS_A/(a-1)}{SS_E/[ab(n-1)]} = \frac{MS_A}{MS_E}$$

$$F_B = \frac{SS_B/(b-1)}{SS_E/[ab(n-1)]} = \frac{MS_B}{MS_E}$$

$$F_{AB} = \frac{SS_{AB}/[(a-1)(b-1)]}{SS_E/[ab(n-1)]} = \frac{MS_{AB}}{MS_E}$$

Two-way ANOVA for fixed effect

Sources	SS	df	MS	*F*	Mean square expectation
A factor	SS_A	$a-1$	MS_A	$\frac{MS_A}{MS_E}$	$\sigma^2+bn\eta_\alpha^2$
B factor	SS_B	$b-1$	MS_D	$\frac{MS_B}{MS_E}$	$\sigma^2+an\eta_\beta^2$
AB interaction	SS_{AB}	$(a-1)(b-1)$	MS_{AB}	$\frac{MS_{AB}}{MS_E}$	$\sigma^2+n\eta_{\alpha\beta}^2$
Error	SS_E	$ab(n-1)$	MS_E		σ^2
Sum	SS_T	$abn-1$			

如果存在显著性差异，在R中使用`TukeyHSD()`函数计算两两之间的显著性。

无重复测量两因素方差分析

Null hypothesis

$$H_{0A}:\alpha_1=\alpha_2=\cdots=\alpha_I=0$$

$$F_A=\frac{SS_A/(a-1)}{SS_E/[ab(n-1)]}$$

$$H_{0B}:\beta_1=\beta_2=\cdots=\beta_J=0$$

$$F_B=\frac{SS_B/(b-1)}{SS_E/[ab(n-1)]}$$

这个其实相当于重复测量的简化版了。少了一个假设条件，之前的公式同样适用但是没有了多个测量值计算平均数等一些计算。

随机效应模型的两因素方差分析

Hypothesis testing

Null hypothesis

$$H_{0A}:\delta_1^2=0$$

$$H_{0B}:\delta_2^2=0$$

$$H_{AB}:\delta_3^2=0$$

F-statistics

$$F_A=\frac{SS_A/(a-1)}{SS_{AB}/[(a-1)(b-1)]}$$

$$F_B=\frac{SS_B/(b-1)}{SS_{AB}/[(a-1)(b-1)]}$$

$$F_{AB}=\frac{SS_{AB}/[(a-1)(b-1)]}{SS_E/[ab(n-1)]}$$

ANOVA for random effects

	SS	*df*	MS	*F*	**Mean square expectation**
Factor A	SS_A	$a-1$	MS_A	$\frac{MS_A}{MS_{AB}}$	$\sigma^2+n\sigma^2_{\alpha\beta}+bn\sigma^2_{\alpha}$
Factor B	SS_B	$b-1$	MS_B	$\frac{MS_B}{MS_{AB}}$	$\sigma^2+n\sigma^2_{\alpha\beta}+an\sigma^2_{\beta}$
AB interaction	SS_{AB}	$(a-1)(b-1)$	MS_{AB}	$\frac{MS_{AB}}{MS_E}$	$\sigma^2+n\sigma^2_{\alpha\beta}$
Error	SS_E	$ab(n-1)$	MS_E		σ^2
Sum	SS_r	$abn-1$			

混合模型的两因素方差分析

Variance analysis for mixed model（A fixed，B random）

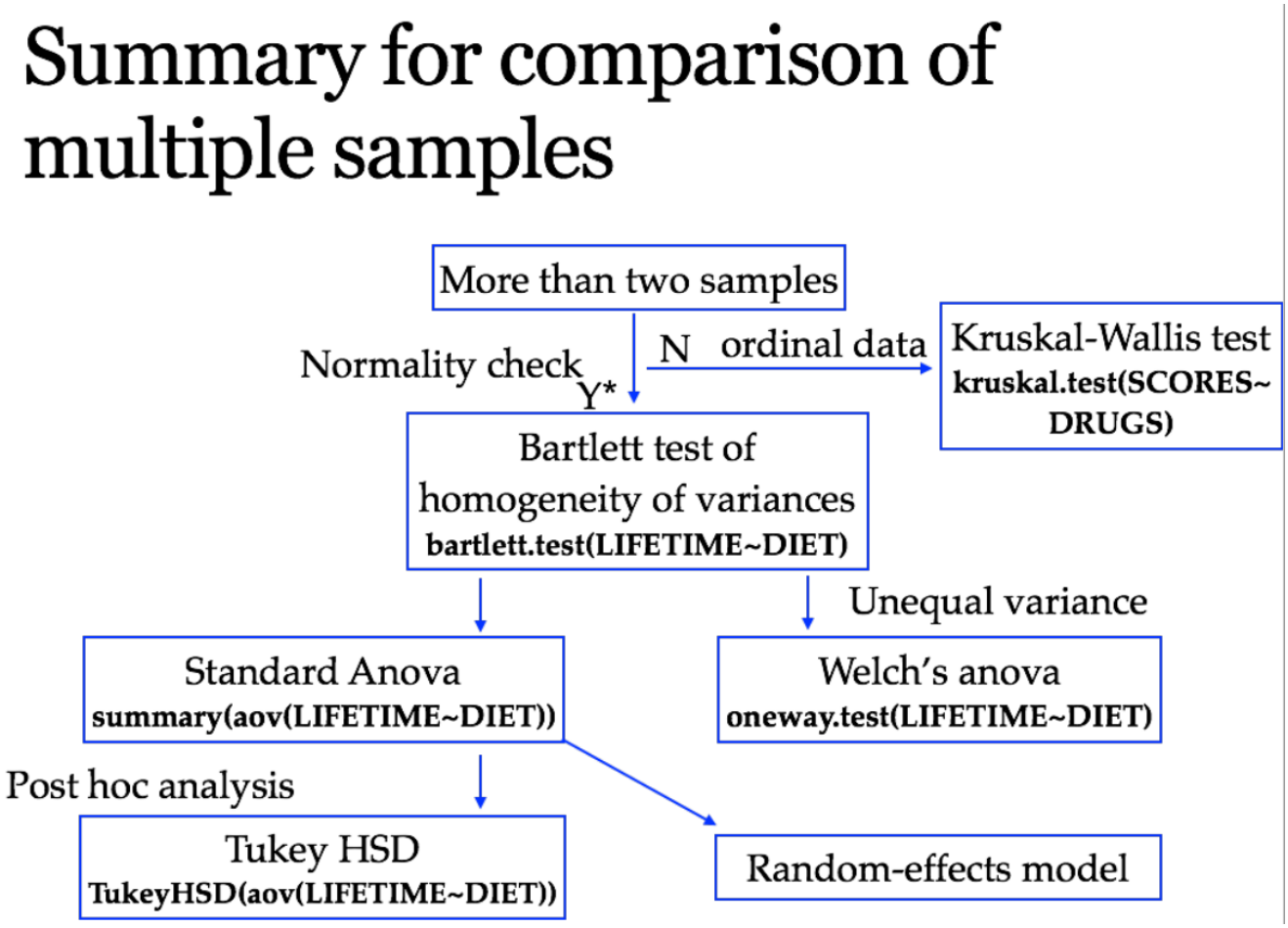
	SS	df	MS	F	MSE
A	SS_A	$a-1$	MS_A	$\frac{MS_A}{MS_{AB}}$	$\sigma^2+n\sigma^2_{\alpha\beta}+bn\sigma^2_{\alpha}$
B	SS_B	$b-1$	MS_B	$\frac{MS_B}{MS_E}$	$\sigma^2+an\sigma^2_{\beta}$
AB	SS_{AB}	$(a-1)(b-1)$	MS_{AB}	$\frac{MS_{AB}}{MS_E}$	$\sigma^2+n\sigma^2_{\alpha\beta}$
error	SS_E	$ab(n-1)$	MS_E		σ^2
sum	SS_r	$abn-1$			

小结

Computation of the *F* statistics for tests of significance in a two-factor ANOVA with replication

Hypothesized effect	Model I (factors *A* and *B* both fixed)	Model II (factors *A* and *B* both random)	Model III (factor *A* random; factor *B* fixed)
Factor *A*	$\frac{\text{factor } A \text{ MS}}{\text{error } MS}$	$\frac{\text{factor } A \text{ MS}}{A \times B \text{ MS}}$	$\frac{\text{factor } A \text{ MS}}{\text{error MS}}$
Factor *B*	$\frac{\text{factor } B \text{ MS}}{\text{error MS}}$	$\frac{\text{factor } B \text{ MS}}{A \times B \text{ MS}}$	$\frac{\text{factor } B \text{ MS}}{A \times B \text{ MS}}$
$A \times B$ interaction	$\frac{A \times B \text{ MS}}{\text{error MS}}$	$\frac{A \times B \text{ MS}}{\text{error MS}}$	$\frac{A \times B \text{ MS}}{\text{error MS}}$

也许方差分析中涉及到的公式略显复杂，计算难度也有很大提升。但是就一个使用者而言，应当理解它的基本内涵和适用范围：它是利用F检验对两个或者两个以上样本的参数检验手段，需要同t检验（可能相对的非参数检验）结合使用；从而完成从多个样本中探寻某些因素对于两个样本之间的影响的过程。它的分析流程如下：



我之前学习时有记录一些方差分析的实例，可以通过wordpress链接-方差分析到相关博文进行查看。

缺失值处理

这里涉及一些方法和相应的R包，估计需要时查看说明。

Missing data

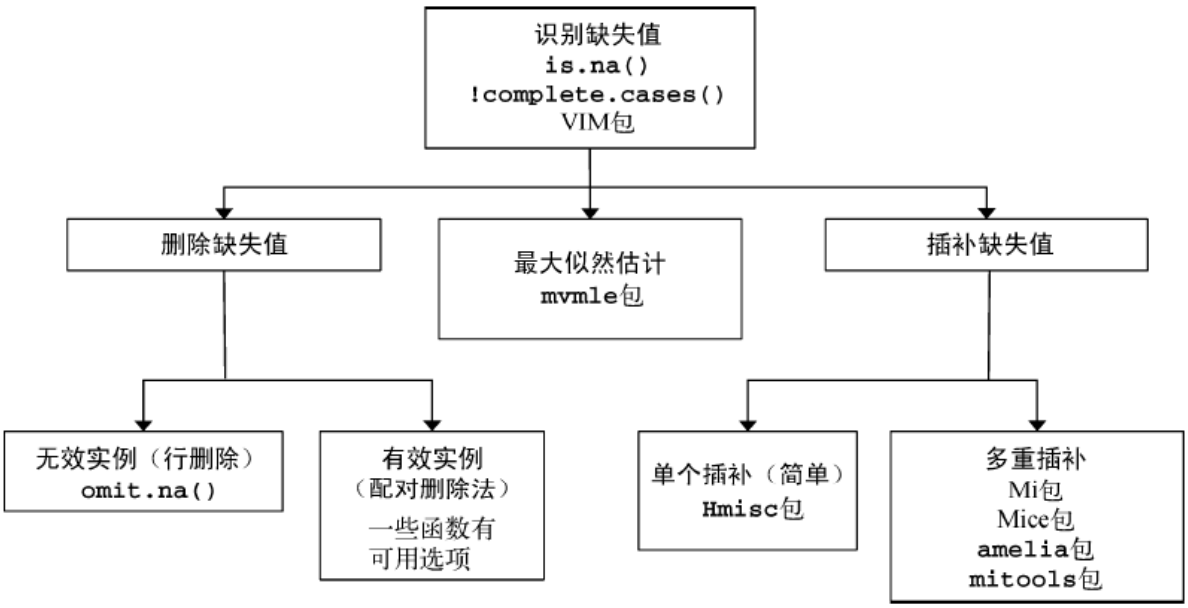


表15-2　处理缺失数据的专业方法

软　件　包	描　　述
Hmisc	包含多种函数，支持简单插补、多重插补和典型变量插补
mvnmle	对多元正态分布数据中缺失值的最大似然估计
cat	对数线性模型中多元类别型变量的多重插补
arrayImpute、arrayMissPattern、SeqKnn	处理微阵列缺失数据的实用函数
longitudinalData	相关的函数列表，比如对时间序列缺失值进行插补的一系列函数
kmi	处理生存分析缺失值的Kaplan-Meier多重插补
mix	一般位置模型中混合类别型和连续型数据的多重插补
pan	多元面板数据或聚类数据的多重插补

回归分析

从许多方面来看，**回归分析**是统计学的核心。它其实是一个广义的概念，通指那些用一个或多个预测变量（也称为自变量或解释变量）来预测响应变量（也成因变量、效标变量或结果变量）。

回归是一个令人困惑的词，因为它有许多特异的变种。R提供了相应强大而丰富的功能同样令人困惑。有统计表明，R中做回归分析的函数已经超过200个。（回归分析相关R的一些概念和函数、包的操作请链接到wordpress-回归分析查看和了解。）

方差分析与回归分析的区别与联系

方差分析与回归分析是有联系又不完全相同的分析方法。方差分析主要研究各变量对结果的影响程度的定性关系，从而剔除对结果影响较小的变量，提高试验的效率和精度。而回归分析是研究变量与结果的定量关系，得出相应的数学模式。在回归分析中，需要对各变量对结果影响进行方差分析，以剔除影响不大的变量，提高回归分析的有效性。
方差分析(Analysis of Variance，简称ANOVA)，又称“变异数分析”，是R.A.Fisher发明的，用于两个及两个以上样本均数差别的显著性检验。 由于各种因素的影响，研究所得的数据呈现波动状。造成波动的原因可分成两类，一是不可控的随机因素，另一是研究中施加的对结果形成影响的可控因素。方差分析是从观测变量的方差入手，研究诸多控制变量中哪些变量是对观测变量有显著影响的变量。
回归分析是研究各因素对结果影响的一种模拟经验方程的办法，回归分析（regression analysis)是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。运用十分广泛，回归分析按照涉及的变量的多少，分为一元回归和多元回归分析。
回归分析中，会用到方差分析来判断各变量对结果的影响程度，从而确定哪些因素是应该纳入到回归方程中，哪些由于对结果影响的方差小而不应该纳入到回归方程中。

线性回归

我们的重点是**普通最小二乘（OLS）回归法**，包括简单线性回归、多项式回归和多元线性回归。

OLS回归是通过预测变量的加权和来预测量化的因变量，其中权重是通过数据估计而得到的参数。

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki} \quad i = 1 \ldots n$$

其中n为观测数目，k为预测变量的数目。

Y_i为第i次观测对应的因变量的预测值

X_{ji}为第i次观测对应的第j个预测变量值

β_0为截距项

β_j预测变量j的回归系数

我们的目标是通过减少响应变量的真实值与预测值的差值来获得模型参数。具体而言，即使惨差平方和最小。

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

（图片中几个字打错了……）

线性回归非常简单，高中知识便有相关的解法介绍。科学研究也较为常用，它是通过对数据计算最小残差平方和来寻找自变量与因变量之间是否存在线性关系。在R中，`aov()`函数以及`lm()`函数（常用后者，前者一般用来做方差分析）都会用来计算线性回归。

忽略推导过程，计算相关系数和截距的公式为：

$$\begin{cases} b = \dfrac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \dfrac{L_{xy}}{L_{xx}} \\ a = \dfrac{\sum y}{n} - b\dfrac{\sum x}{n} = \overline{y} - b\overline{x} \end{cases}$$

R的用法很简单，格式为`myfit <- lm(formula,data)`

公式的构建类似于方差分析，下面列出常用的符号：

符号	用途
~	分隔符号，左边为响应变量（因变量），右边为解释变量（自变量）
+	分隔预测变量（因变量）
:	表示预测变量的交互项
*	表示所有可能交互项的简洁方式
^	表示交互项达到某个次数
.	表示包含除因变量外的所有变量
-	减号，表示从等式中移除某个变量
-1	删除截距项
I()	从算术的角度来解释括号中的元素
function	可在表达式中用的数学函数。例如，log(y) ~ x + z + w

除了`lm()`，下表列出了一些有用的分析函数，对拟合得到的模型做进一步的处理和分析。

函数	用途
summary()	展示拟合模型的详细结果
coefficients()	列出拟合模型的模型参数
confint()	提供模型参数的置信区间（默认95%)
fitted()	列出拟合模型的预测值
residuals()	列出拟合模型的残差值
anova()	生成一个拟合模型的方差分析表，或者比较两个或更多拟合模型的方差分析表
vcov()	列出模型参数的协方差矩阵
AIC()	输出赤池信息统计量
plot()	生成评价拟合模型的诊断图
predict()	用拟合模型对新的数据集预测响应变量值

上述已经提过方差分析与回归分析的区别与联系，在我们进行回归分析时，往往需要方差分析来剔除无关或者影响力较小的自变量，从而简化回归模型（李春喜《生物统计学》（第四版）124-129页包含了进行简单线性回归所有的计算步骤和后续的F检验、t检验）。

实际数据计算时，先计算回归分析的一级数据和二级数据。然后再计算一些目标值，比如回归平方和，残差平方和等等。

我写出一些重要数据计算公式：

$$L_{xx} = \sum (x - \overline{x})^2$$
$$L_{yy} = \sum (y - \overline{y})^2$$
$$L_{xy} = \sum (x - \overline{x})(y - \overline{y})$$
$$SS_x = \sum x^2 - \frac{(\sum x)^2}{n}$$
$$SS_y = \sum y^2 - \frac{(\sum y)^2}{n}$$
$$SP = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

回归参数：

$$b = \frac{Lxy}{Lxx}$$
$$a = \overline{y} - b\overline{x}$$

其他一些数据的计算也就比较好理解了。

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

or Total SS = Reg SS + Res SS

Short Computational Form for Regression and Residual SS

$$\text{Regression SS} = bL_{xy} = b^2L_{xx} = L_{xy}^2/L_{xx}$$

$$\text{Residual SS} = \text{Total SS} - \text{Regression SS} = L_{yy} - L_{xy}^2/L_{xx}$$

F Test for Simple Linear Regression

To test H_0: $\beta = 0$ vs. H_1: $\beta \neq 0$, use the following procedure:

(1) Compute the test statistic

$$F = \text{Reg MS} / \text{Res MS} = \left(L_{xy}^2/L_{xx}\right)/\left[\left(L_{yy}-L_{xy}^2/L_{xx}\right)/(n-2)\right]$$

that follows an $F_{1,n-2}$ distribution under H_0.

(2) For a two-sided test with significance level α, if

$F > F_{1,n-2,1-\alpha}$, then reject H_0; if

$F \leq F_{1,n-2,1-\alpha}$, then accept H_0.

(3) The exact *p*-value is given by $Pr\left(F_{1,n-2} > F\right)$.

t Test for Simple Linear Regression

To test the hypothesis H_0: $\beta = 0$ vs.

H_1: $\beta \neq 0$, use the following procedure:

(1) Compute the test statistic

$$t = b \Big/ \left(s_{y\cdot x}^2 / L_{xx}\right)^{1/2}$$

(2) For a two-sided test with significance level α,

If $t > t_{n-2,1-\alpha/2}$ or $t < t_{n-2,\alpha/2} = -t_{n-2,1-\alpha/2}$

then reject H_0;

if $-t_{n-2,1-\alpha/2} \leq t \leq t_{n-2,1-\alpha/2}$

then accept H_0.

(3) The *p*-value is given by

$p = 2 \times$ (area to the left of t under a t_{n-2} distribution) if $t < 0$

$p = 2 \times$ (area to the right of t under a t_{n-2} distribution) if $t \geq 0$

ANOVA table for displaying regression results

	SS	*df*	MS	*F* statistic	*p*-value
Regression	(*a*)[a]	1	(*a*)/1	$F = [(a)/1]/[(b)/(n-2)]$	$Pr(F_{1,n-2} > F)$
Residual	(*b*)[b]	$n-2$	$(b)/(n-2)$		
Total	(*a*) + (*b*)				

[a](*a*) = Regression SS.

[b](*b*) = Residual SS.

在使用R语言时，直接使用`aov()`对复杂模型和简化模型比较即可，看是否存在显著性差异，然后决定是否可以用简单模型替换复杂模型（之前提供的回归分析链接有实例）。

那么怎么评价模型拟合的好坏呢？这里有一个常见的参数。

Definition:

$$R^2 = \text{Reg SS} / \text{Total SS}$$

Significance:

R^2 can be thought of as the proportion of the variance of y that is explained by x.

If $R^2 = 1$, all variation in y can be explained by variation in x.

If $R^2 = 0$, x gives no information about y, and the variance of y is the same with or without knowing x.

If R^2 is between 0 and 1, for a given value of x, the variance of y is lower than it would be if x were unknown , but is still greater than 0

非线性回归

回归的概念本质是说对数据进行曲线拟合，除了线性关系，科研中我们还会碰到其他因变量与自变量的定量关系，比如指数，幂函数等等。我们可以通过变换把它们转变为类似线性的关系，也就是非线性回归了。

Method	Transformation(s)	Regression equation	Predicted value (ŷ)
Standard linear regression	None	$y = b_0 + b_1x$	$\hat{y} = b_0 + b_1x$
Exponential model	Dependent variable = log(y)	$\log(y) = b_0 + b_1x$	$\hat{y} = 10^{b_0 + b_1x}$
Quadratic model	Dependent variable = sqrt(y)	$\text{sqrt}(y) = b_0 + b_1x$	$\hat{y} = (b_0 + b_1x)^2$
Reciprocal model	Dependent variable = 1/y	$1/y = b_0 + b_1x$	$\hat{y} = 1 / (b_0 + b_1x)$
Logarithmic model	Independent variable = log(x)	$y= b_0 + b_1\log(x)$	$\hat{y} = b_0 + b_1\log(x)$
Power model	Dependent variable = log(y) Independent variable = log(x)	$\log(y)= b_0 + b_1\log(x)$	$\hat{y} = 10^{b_0 + b_1\log(x)}$

在R中，我们依旧使用`lm()`函数，这时，公式可以根据数据添加相应数学函数，比如`lm(log(y)~x)`实现指数函数的线性化，在绘图时，可以用`abline()`或`line()`函数添加拟合曲线（前者可以以模型作为参数输入）。

相关分析

相关系数公式

$$r = \frac{Lxy}{\sqrt{LxxLyy}} = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2\sum(y-\overline{y})^2}}$$

相关系数的平方为决定系数，表示为R^2，在拟合线性回归曲线时常常见到的参数就是这个。

r的取值从-1到1,0表示完全无关，绝对值越接近1，相关程度越高。

回归系数b与相关系数r的关系：

$$b = \frac{Lxy}{Lxx}$$
$$r = \frac{Lxy}{\sqrt{LxxLyy}}$$
$$r\sqrt{\frac{L_{yy}}{L_{xx}}} = b$$

如果设$s_y^2 = \frac{L_{yy}}{n-1}$，$s_x^2 = \frac{L_{xx}}{n-1}$ (这不正是方差吗)
那么有$b = r\frac{s_y}{s_x}$

多元线性回归

多元线性回归可以看作是简单线性回归的一个拓展，回归系数和自变量不再是单个的，而是一组变量。

其公式形式为：

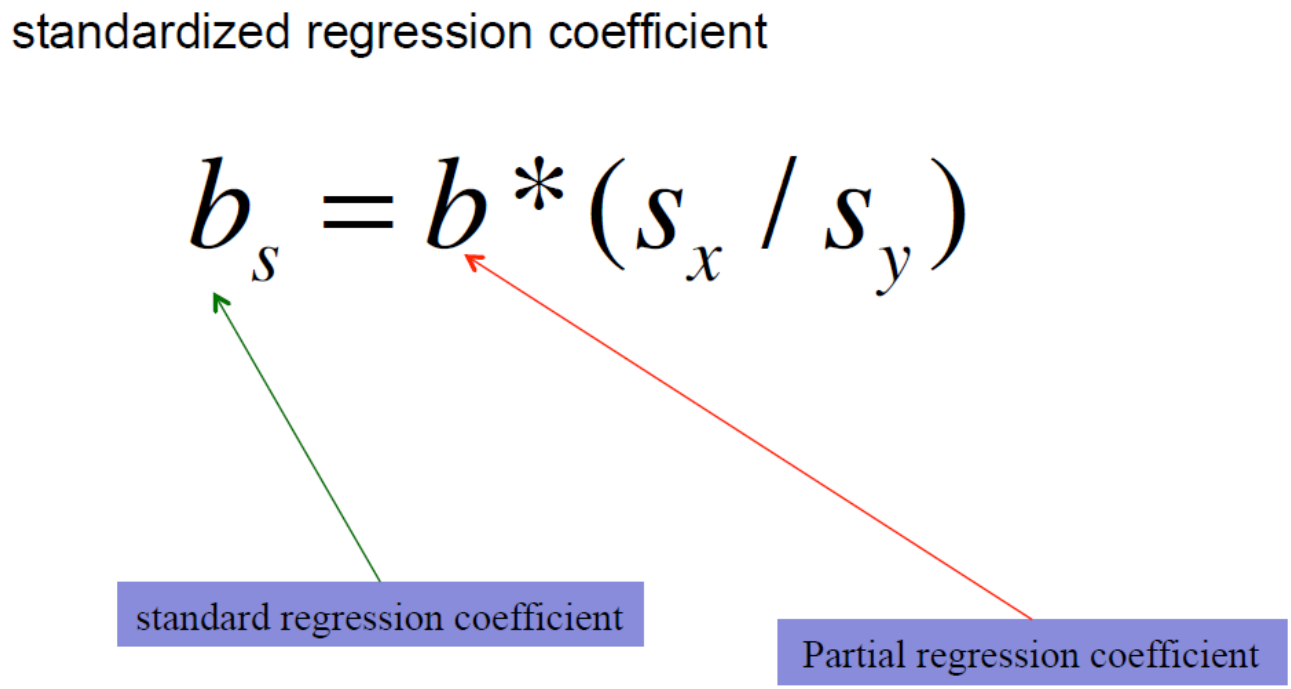
$$y = a + b_1x_1 + b_2x_2 + \cdots + b_nx_n$$

其几何解释由简单的二元平面上的直线拟合变为多维空间中的直线拟合。

在各种数据参数的计算时可能公式会变得比较繁琐，但在使用R进行多元线性回归分析时，跟线性回归基本一致，使用`lm()`函数即可。

比如`lm(y~x1+x2)`可以进行二元线性回归，`lm(y~x1*x2)`加上交互项（x1与x2交叉因素）的探索。多维也是如此。

可以看到，这里b不再是一个值，而是多个。因此每一个自变量对于的b表示一个部分相关系数。它的含义为：一个预测变量（因变量）增加一个单位，其他预测变量保持不变时，因变量将要增加的数量。



举例：

探究出生重量(x1)和年龄对血压(x2)的影响，如果$y = 53.45 + 0.1256x_1 + 5.888 * x_2$，（原始数据未列出，仅关注计算）那么

$$s_{x1} = 18.75$$

$$s_{x2} = 0.946$$

$$s_y = 6.69$$

$$b_s(\text{birthweight}) = \frac{0.1256 \times 18.75}{6.69} = 0.352$$

$$b_s(\text{age in days}) = \frac{5.888 \times 0.946}{6.69} = 0.833$$

我们可以得到以下结果：

(1) the average increase in SBP is 0.352 standard-deviation units of blood pressure per standard-deviation increase in birthweight
(2) the average increase in SBP is 0.833 standard-deviation units of blood pressure per standard-deviation increase in age
(3) age appears to be more important variable

拟合优度的判断：

F Test for Testing the Hypothesis H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ vs. H_1: At Least One of the $\beta_j \neq 0$ in Multiple Linear Regression

(1) Estimate the regression parameters using the method of least squares, and compute Reg SS and Res SS,

where $$\text{Res SS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\text{Reg SS} = \text{Total SS} - \text{Res SS}$$

$$\text{Total SS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$\hat{y}_i = a + \sum_{j=1}^{k} b_j x_{ij}$$

x_{ij} = jth independent variable for the ith subject, $j = 1,\ldots, k$; $i = 1,\ldots, n$

(2) Compute Reg MS = Reg SS/k, Res MS = Res SS/$(n - k - 1)$.

(3) Compute the test statistic

$$F = \text{Reg MS}/\text{Res MS}$$

which follows an $F_{k,n-k-1}$ distribution under H_0.

(4) For a level α test,

if $F > F_{k,n-k-1,1-\alpha}$ then reject H_0

if $F \leq F_{k,n-k-1,1-\alpha}$ then accept H_0

(5) The exact p-value is given by the area to the right of F under an $F_{k,n-k-1}$ distribution = $Pr(F_{k,n-k-1} > F)$.

逻辑回归

Logistic回归与多重线性回归实际上有很多相同之处，最大的区别就在于它们的因变量不同，其他的基本都差不多。正是因为如此，这两种回归可以归于同一个家族，即广义线性模型（generalizedlinear model）。

这一家族中的模型形式基本上都差不多，不同的就是因变量不同。

- 如果是连续的，就是多重线性回归；
- 如果是二项分布，就是Logistic回归；
- 如果是Poisson分布，就是Poisson回归；
- 如果是负二项分布，就是负二项回归。

Logistic回归的因变量可以是二分类的，也可以是多分类的，但是二分类的更为常用，也更加容易解释。所以实际中最常用的就是二分类的Logistic回归。

Logistic回归的主要用途：

- 寻找危险因素：寻找某一疾病的危险因素等；
- 预测：根据模型，预测在不同的自变量情况下，发生某病或某种情况的概率有多大；
- 判别：实际上跟预测有些类似，也是根据模型，判断某人属于某病或属于某种情况的概率有多大，也就是看一下这个人有多大的可能性是属于某病。

Logistic回归主要在流行病学中应用较多，比较常用的情形是探索某疾病的危险因素，根据危险因素预测某疾病发生的概率，等等。例如，想探讨胃癌发生的危险因素，可以选择两组人群，一组是胃癌组，一组是非胃癌组，两组人群肯定有不同的体征和生活方式等。这里的因变量就是是否胃癌，即“是”或“否”，自变量就可以包括很多了，例如年龄、性别、饮食习惯、幽门螺杆菌感染等。自变量既可以是连续的，也可以是分类的。

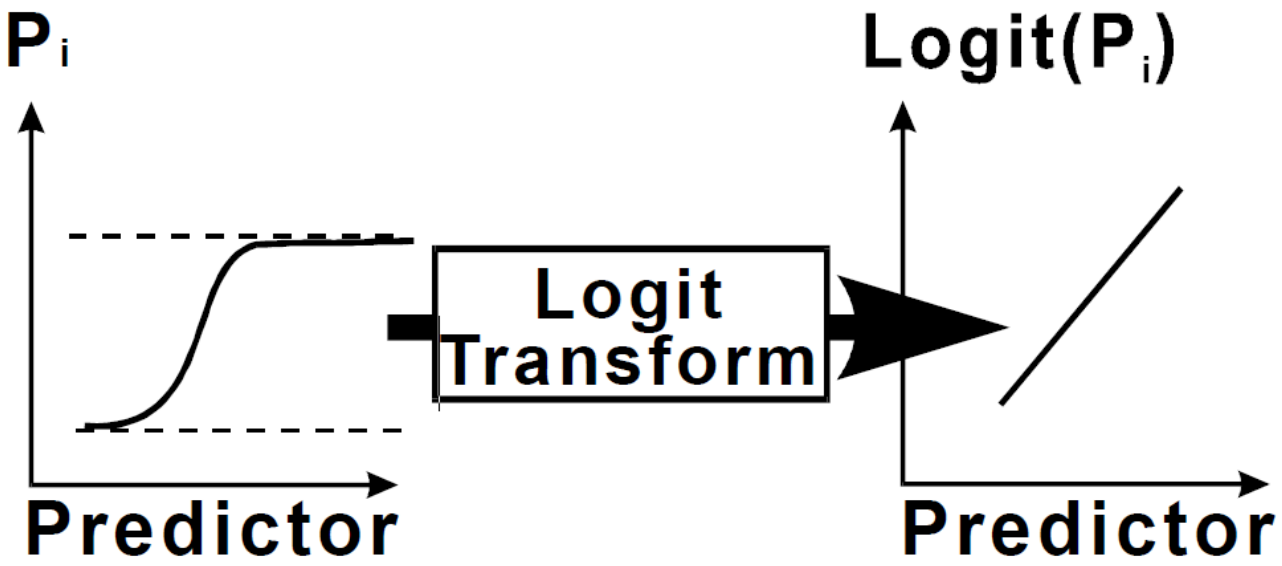
通过对数据发生概率进行logit转换，我们可以生成线性的逻辑回归模型。

Logistic regression models transform probabilities called *logits*.

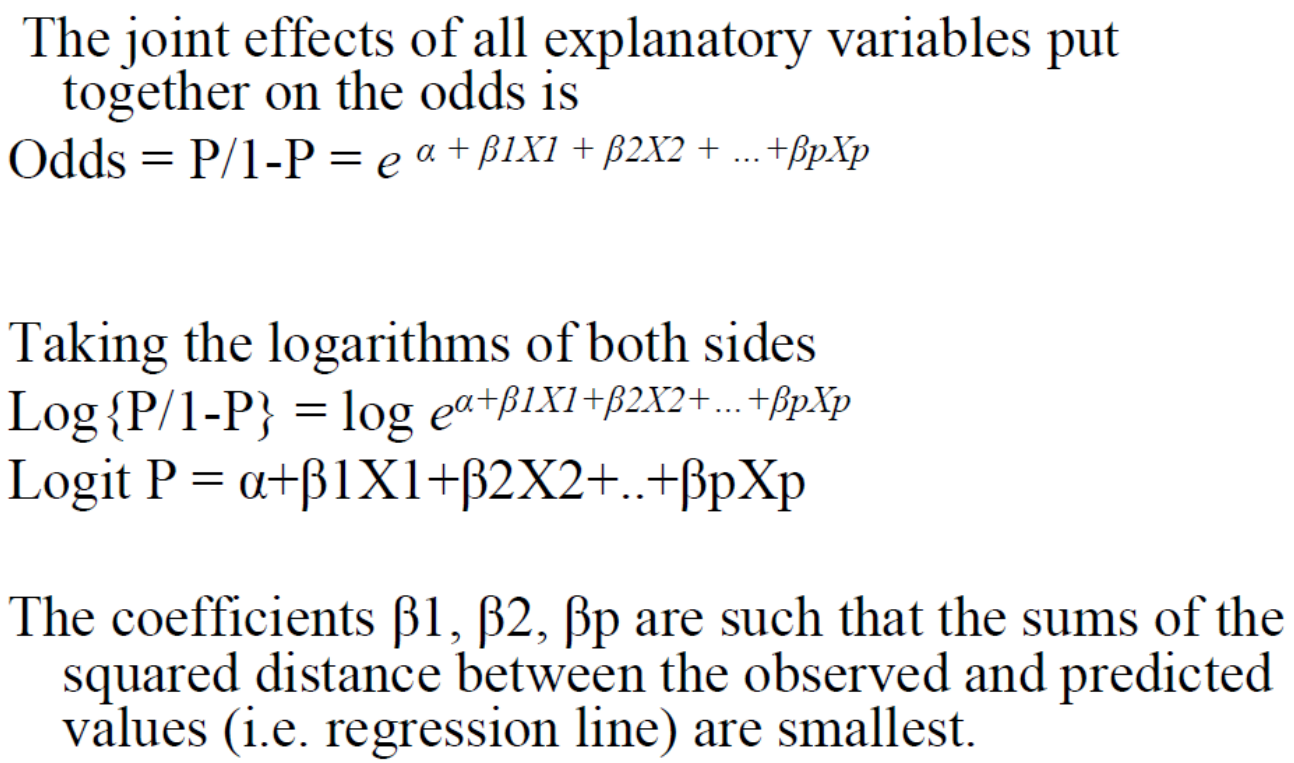
$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

where
i indexes all cases (observations).
p_i is the probability the event occurs in the i^{th} case.
log is the natural log (to the base e).

图形化的效果为：



由此得到逻辑回归模型：



在R中，逻辑回归作为广义线性模型的一部分被介绍，可以参考我整理的广义线性模型。下面列出常用的连接函数和连用函数。

glm()函数

基本形式：`glm(formula, family=family(link=function), data=)`

分布族	默认的连接函数
binomial	(link = “logit”)
gaussian	(link = “identity”)
gamma	(link = “inverse”)
inverse.gaussian	(link = “1/mu^2”)
poisson	(link = “log”)
quasi	(link = “identity”, variance=”constant”)
quasibinomial	(link = “logit”)
quasipoisson	(link = “log”)

glm()函数可以拟合许多流行的模型，包括Logistic回归、泊松回归和生存分析。

连用的函数

与glm()函数连用的一些函数

函数	描述
summary()	展示拟合模型的细节
coefficients(), coef()	列出拟合模型的参数（截距项和斜率）
confint()	给出模型参数的置信区间（默认为95%）
residuals()	列出拟合模型的残差值
anova()	生成两个拟合模型的方差分析表
plot()	生成评价拟合模型的诊断图
predict()	用拟合模型对新数据集进行预测
deviance()	拟合模型的偏差
df.residual()	拟合模型的残差自由度

具体实例可以参考逻辑回归的一个简单实例。

部分相关与多重相关

部分相关，也称为偏相关。偏相关分析是指当两个变量同时与第三个变量相关时，将第三个变量的影响剔除，只分析另外两个变量之间相关程度的过程。

partial correlation

$$cor(y, x_i \mid x_1, x_2, \ldots, xk)$$

multiple correlation

$$cor(y, x_1, x_2, \ldots, x_k)$$

多重相关就是整体的相关系数r。