# 第六章 线性回归模型

复旦大学附属肿瘤医院 周支瑞

# 6.1 简单线性回归

# 直线回归的概念

- “回归”一词首先由英国生物统计学家S.F.Galton(1885)提出，他发现高个子父代的子代平均身高不是更高，而是稍矮；相反，矮个子父代的子代平均身高并不是更矮，而是稍高于其父代水平。他把这种身高趋向种族稳定的现象称为“回归”。目前回归的含义已经演变成变量之间某种数量依存关系。

# 回归与相关概述

- 直线回归分析：分析两个变量间的数量关系，目的是用一个变量推算另一个变量(建立回归方程)。
- 直线相关分析：分析两个变量之间有无相关关系以及相关的性质（正、负相关）和相关的密切程度。
- 弟弟身高与哥哥身高的关系？儿子身高与父亲身高的关系？

## 回归分析对资料的要求

- 回归分析涉及到两个变量，X与Y，其中X称自变量，Y 为因变量或反应变量。

Y — 必须是呈正态分布的随机变量

X
- 可以是非随机变量: 年龄、药物浓度或剂量 — Ⅰ型回归
- 也可以是随机变量: 身高、体重、血清胆固醇的含量，血红蛋白的含量 — Ⅱ型回归

# 举例

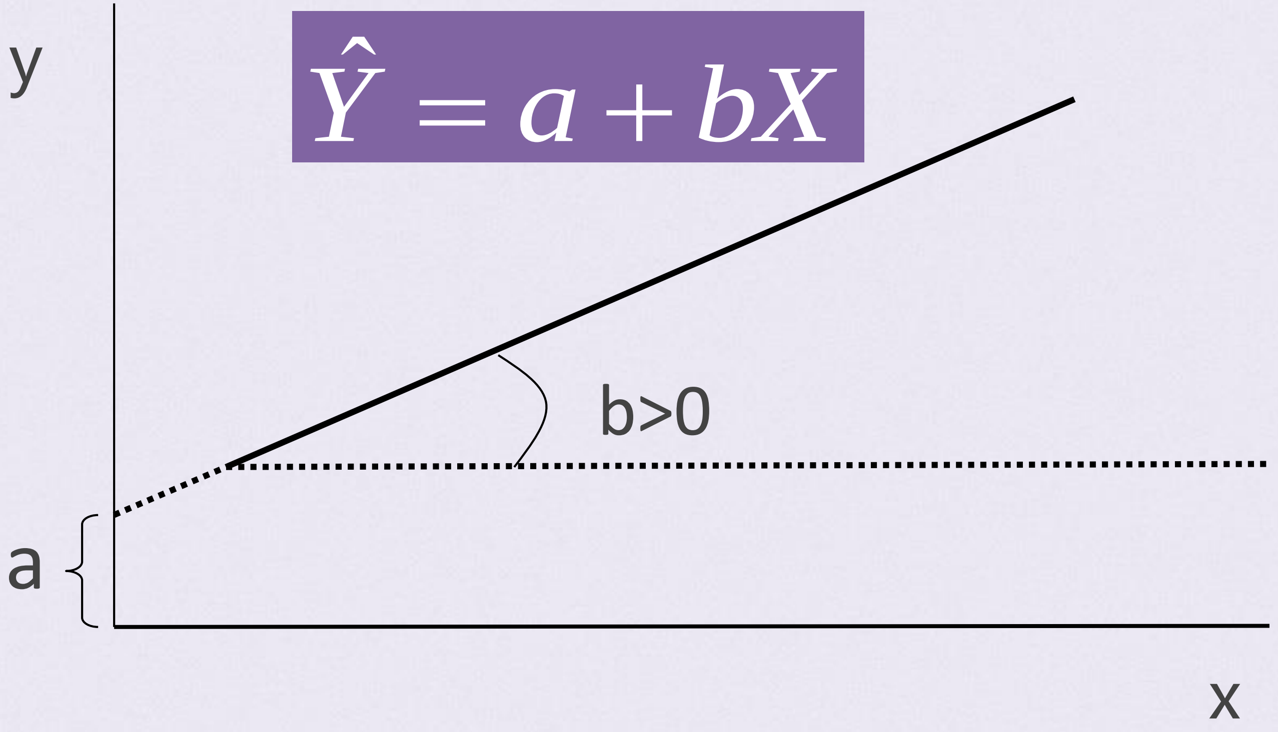
- 年龄(X)与尿肌酐含量(Y)研究;
- 身高(X)与(Y)体重研究

# 直线回归方程



- 由X推算Y的直线回归方程一般表达式：$\hat{Y} = a + bX$

  a称为截距，b 为回归系数，即直线的斜率。



# 直线回归方程的建立

丁香公开课 CLASS.DXY.CN

$$b = \frac{\sum(X - \overline{X})(Y - \overline{Y})}{\sum(X - \overline{X})^2} = \frac{lxy}{lxx}$$

$$a = \overline{Y} - b\,\overline{X}$$

式中 $\overline{X}$、$\overline{Y}$分别是X、Y的均数；$l_{xx}$ 为X的离均差平方和；$l_{xy}$ 为X与Y的离均差积和，按下式计算

$$lxy = \sum(X - \overline{X})(Y - \overline{Y}) = \sum XY - \frac{(\sum X)(\sum Y)}{n}$$

# 回归系数b的统计学意义



- b>0时,Y随X增大而增大;
- b<0时,Y随X的增大而减少;
- b=0时,X与Y无直线关系。
- b的统计学意义是：X每增（减）一个单位，Y平均改变b个单位。

# 回归分析的步骤

丁香公开课 CLASS.DXY.CN

- 1. 用原始数据绘制散点图；
- 2. 求a和b (如果呈直线关系)
- 3. 对回归系数b作假设检验（方法：a. F检验 b. t检验　c. 用r检验来代替）。
- 4. 如果x与y存在直线关系（ b假设检验的结果P<0.05），列出回归方程。否则，不列回归方程。

# 回归系数b的假设检验

丁香公开课 CLASS.DXY.CN

- 所建立的回归方程，不一定都有意义，必须对回归方程和回归系数进行假设检验。直线回归方程一般只对回归系数进行假设检验。

# Y的离均差平方和的划分

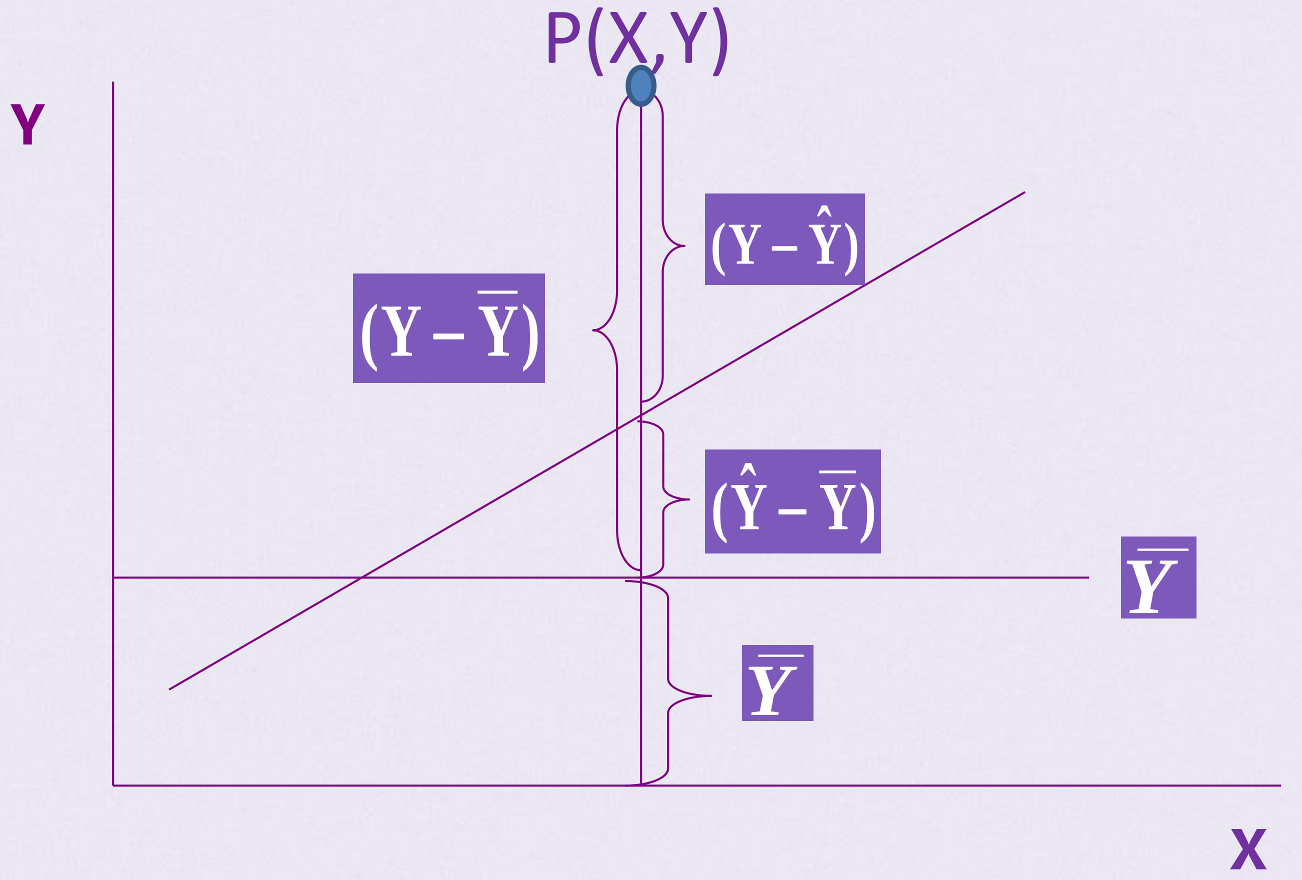



图 11.2  应变量Y 的平方和划分示意

# Y 的平方和所划分的三段 含义

丁香公开课 CLASS.DXY.CN

- 第一段：$(Y-\hat{Y})$ 表示P点与回归直线的纵向距离，即实测值Y与估计值 $\hat{Y}$ 之差，称剩余或残差。
- 第二段：$(\hat{Y}-\bar{Y})$ 即估计值与均数之差，它与回归系数的大小有关。|b| 值越大，$(\hat{Y}-\bar{Y})$ 的差值也越大，反之越小。当b=0时，则 $(\hat{Y}-\bar{Y})=0$ 。则 $(Y-\hat{Y})$ = $(\hat{Y}-\bar{Y})$ 也就是回归直线并不能使残差减小。
- 第三段：$\bar{Y}$，是应变量Y的均数。

$$\sum(Y-\bar{Y})^2=\sum(\hat{Y}-\bar{Y})^2+\sum(Y-\hat{Y})^2$$

# 三种平方和的关系

$\sum(Y-\bar{Y})^2$　称总平方和，用$SS_{总}$表示，

$\sum(Y-\hat{Y})^2$　称回归平方和，用$SS_{回}$表示

$\sum(\hat{Y}-\bar{Y})^2$　称剩余平方和，用$SS_{剩}$表示

$$SS_{总}=SS_{回}+SS_{剩}$$

# 三种平方和的意义

- （1）$SS_{总}$，为Y值的离均差平方和，说明未考虑X与Y的回归关系时Y总的变异。
- （2）$SS_{回}$，它反映在Y的变异中由于X与Y的直线关系而使Y变异减少的部分，也是在总平方和中可以用X解析的部分。$SS_{回}$越大，说明回归效果越好。
- （3）$SS_{剩}$，反映X对Y的线性影响之外其它因素对Y的变异的作用，也是在总平方和中无法用X解析的部分。SS剩越小，说明回归方程的估计误差越小。

# 三种平方和的自由度及其关系如下

- $\upsilon_{总}=n-1$，$\upsilon_{回}=1$，$\upsilon_{剩}=n-2$
- $\upsilon_{总}=\upsilon_{回}+\upsilon_{剩}$

# b的假设检验方法 1



➢ 1、方差分析方法

将$SS_{总}$分解为$SS_{回}$和$SS_{剩}$两部分后，按下式计算F值:

$$F=\frac{SS_{回}/\nu_{回}}{SS_{剩}/\nu_{剩}}=\frac{MS_{回}}{MS_{剩}},\quad \nu_{回}=1,\quad \nu_{剩}=n-2$$

$MS_{回}$，$MS_{剩}$分别为回归均方及剩余均方，求出F值后查F界值表确定P值，按所取检验水准推断结论。

# b的假设检验方法 2

- 2、t检验法

$$t=\frac{b-0}{s_b}=\frac{b}{s_{y.x}/\sqrt{lxx}},\quad v=n-2$$

$$s_{y.x}=\sqrt{\frac{\sum(Y-\hat{Y})^2}{n-2}}=\sqrt{\frac{SS_{剩}}{n-2}}$$

上式中，$S_b$为样本回归系数的标准误，Sy.x为剩余标准差,也称回归标准差,它表示应变量Y的观察值对于回归直线的离散程度；Sy.x可以作为回归方程估计精度指标。

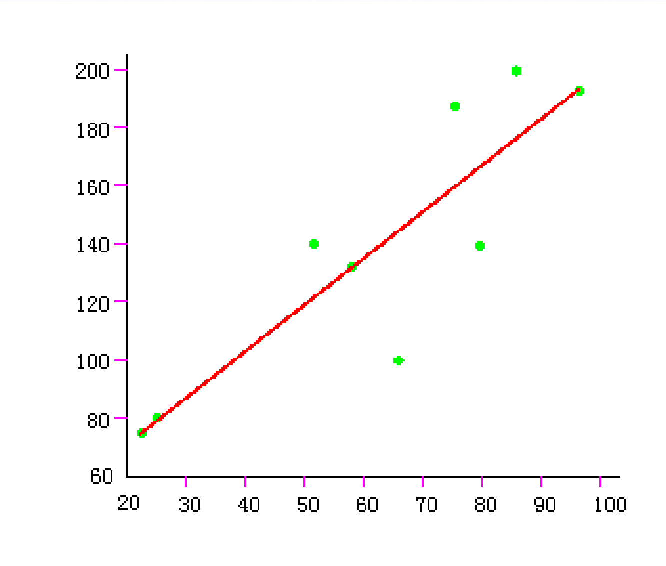
# 总体回归系数β区间的估计

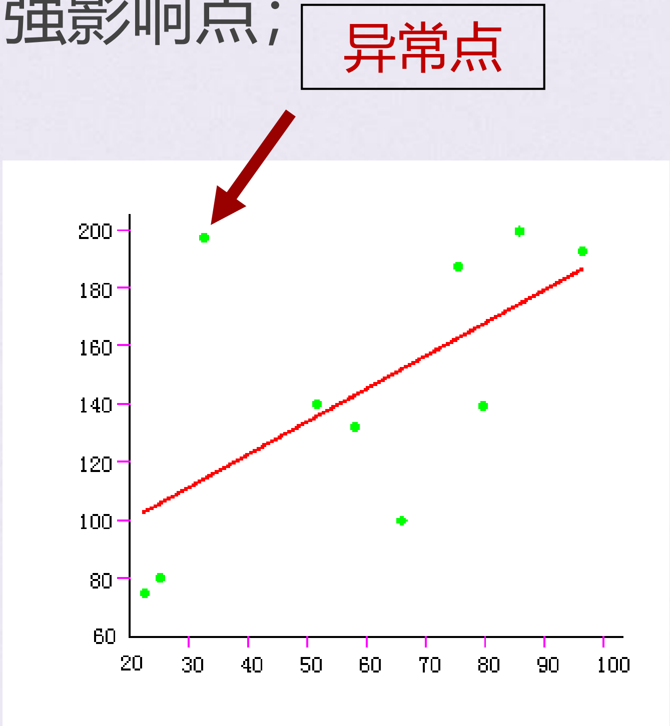


- 用样本回归系数b估计总体回归系数β，方法如下：
- β的95%可信区间是：（$b-t_{0.05,(n-2)}Sb$，$b+t_{0.05,(n-2)}Sb$），缩写为 $b\pm t_{0.05,(n-2)}Sb$，Sb 为回归系数的标准误,n-2为自由度。

# 应用直线回归分析应注意的问题

- 1. 作回归分析要有实际意义。
- 2. 进行直线回归分析前，应绘制散点图。作用：①看散点是否呈直线趋势；② 有无异常点、高杠杆点和强影响点；



# 应用直线回归分析应注意的问题

- 3. 注意建立线性回归模型的基本条件：线性、独立性、正态性、方差齐性
- 4. 直线回归方程的适用范围以求回归方程时X的实测值范围为限；若无充分理由证明超过该范围还是直线，应避免外延。
- 5. 两变量有线性关系，不一定是因果关系，也不一定表明两变量间确有内在联系。

# 用lm()拟合线性回归模型



- 在R中，拟合线性模型最基本的函数就是lm()，格式为：

  *myfit <- lm(formula, data)*

- 其中，formula指要拟合的模型形式，data是一个数据框，包含了用于拟合模型的数据。结果对象（本例中是myfit）存储在一个列表中，包含了所拟合模型的大量信息。表达式（formula）形式如下：

  *Y ~ X1 + X2 + … + Xk*

- ~左边为响应变量，右边为各个预测变量，预测变量之间用+符号分隔。表8-2中的符号可以用不同方式修改这一表达式。

# R表达式中常用的符号

表8-2 R表达式中常用的符号

| 符号 | 用途 |
| --- | --- |
| ~ | 分隔符号，左边为响应变量，右边为解释变量。例如，要通过 x、z 和 w 预测 y，代码为 y ~ x + z + w |
| + | 分隔预测变量 |
| : | 表示预测变量的交互项。例如，要通过 x、z 及 x 与 z 的交互项预测 y，代码为 y ~ x + z + x:z |
| * | 表示所有可能交互项的简洁方式。代码 y~ x * z * w 可展开为 y ~ x + z + w + x:z + x:w + z:w + x:z:w |
| ^ | 表示交互项达到某个次数。代码 y ~ (x + z + w)^2 可展开为 y ~ x + z + w + x:z + x:w + z:w |
| . | 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 x、y、z 和 w，代码 y ~ .可展开为 y ~ x + z + w |
| - | 减号，表示从等式中移除某个变量。例如，y ~ (x + z + w)^2 – x:w 可展开为 y ~ x + z + w + x:z + z:w |
| -1 | 删除截距项。例如，表达式 y ~ x - 1 拟合 y 在 x 上的回归，并强制直线通过原点 |
| I() | 从算术的角度来解释括号中的元素。例如，y ~ x + (z + w)^2 将展开为 y ~ x + z + w + z:w。相反，代码 y ~ x + I((z + w)^2)将展开为 y ~ x + h，h 是一个由 z 和 w 的平方和创建的新变量 |
| *function* | 可以在表达式中用的数学函数。例如，log(y) ~ x + z + w 表示通过 x、z 和 w 来预测 log(y) |

# 对拟合线性模型非常有用的其他函数

表8-3　对拟合线性模型非常有用的其他函数

| 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（截距项和斜率） |
| 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 |

- 当回归模型包含一个因变量和一个自变量时，我们称为简单线性回归。当只有一个预测变量，但同时包含变量的幂（比如，X、X^2、X^3）时，我们称为多项式回归。当有不止一个预测变量时，则称为多元线性回归。

# 简单线性回归 案例 1

➢ 让我们通过一个回归示例来熟悉表8-3中的函数。基础安装中的数据集women提供了15个年龄在30~39岁间女性的身高和体重信息，我们想通过身高来预测体重，获得一个等式可以帮助我们分辨出那些过重或过轻的个体。代码如下：

```
# Listing 8.1 - Simple linear regression
> fit <- lm(weight ~ height, data=women)
> summary(fit)
> women$weight
> fitted(fit)
> residuals(fit)
> plot(women$height,women$weight,
    main="Women Age 30-39",
    xlab="Height (in inches)",
    ylab="Weight (in pounds)")
# add the line of best fit
> abline(fit)
```

# 简单线性回归 案例 2

丁香公开课 CLASS.DXY.CN

**例 9-4** 某医生研究儿童体重与心脏横径的关系，测得 13 名 8 岁正常男童的体重与心脏横径，数据见表 9-6。试作回归分析。

表 9-6 13 名 8 岁健康男童体重与心脏横径的关系

| 编号 | 体重（kg，$X$） | 心脏横径（cm，$Y$） |
|---|---|---|
| 1 | 25.5 | 9.2 |
| 2 | 19.5 | 7.8 |
| 3 | 24.0 | 9.4 |
| 4 | 20.5 | 8.6 |
| 5 | 25.0 | 9.0 |
| 6 | 22.0 | 8.8 |
| 7 | 21.5 | 9.0 |
| 8 | 23.5 | 9.4 |
| 9 | 26.5 | 9.7 |
| 10 | 23.5 | 8.8 |
| 11 | 22.0 | 8.5 |
| 12 | 20.0 | 8.2 |
| 13 | 28.0 | 9.9 |

# 简单线性回归 案例 2 代码



```
> Example9_4  <- read.table ("example9_4.csv",  header=TRUE, sep=",")
> attach(Example9_4)
> plot(x, y)
> fit  <-  lm(y~x)
> anova(fit)
> summary (fit)
> confint(fit)
> y
> fitted (fit)
> residuals (fit)
> detach (Example9_4)
```

# 简单线性回归 案例 3

➢ 例9-5 大多数公司最终会询问关于花费在广告上的费用对公司产品销售额的影响程度。由于广告需要一定的时间才能达到它的效应，同时它的效应也不是永久持续的，它的影响也许仅仅延续开头的一段时期。假设公司相信销售额与当月以及前两个月内所花的广告费有较密切的关系，假设它们之间存在线性关系，现在有某公司15 月内有关广告花费X 与销售额Y 的数据，如表9-7 所示。

表 9-7 广告额与销售额

| 月 | 月销售额 | 月广告花费 |
|---|---|---|
| 1 | 2945 | 280 |
| 2 | 4295 | 400 |
| 3 | 5645 | 450 |
| 4 | 6995 | 590 |
| 5 | 8345 | 650 |
| 6 | 9695 | 750 |
| 7 | 11045 | 890 |
| 8 | 12395 | 1000 |
| 9 | 13745 | 1050 |
| 10 | 15095 | 1200 |
| 11 | 16445 | 1250 |
| 12 | 17795 | 1350 |
| 13 | 19145 | 1460 |
| 14 | 20495 | 1500 |
| 15 | 21845 | 1650 |

# 简单线性回归 案例 3 代码

```
> Example9_5  <- read.table ("example9_5.csv", header=TRUE, sep=",")
> attach(Example9_5)
> fit  <-  lm(SALES~ADV + ADVLAG1 + ADVLAG2)
> anova(fit)
> summary (fit)
> SALES
> fitted (fit)
> residuals (fit)
> detach (Example9_5)
```

# 6.2 残差与回归值&预测域与置信带

# 残差与回归值&预测域与置信带 案例

丁香公开课 CLASS.DXY.CN

thuesen　　心室收缩速度

**描述**

该数据集包含 24 行、2 列，包含了第一类糖尿病人心室收缩速度和血糖含量。

**用法**

```
thuesen
```

**格式**

该数据集包含如下列：

`blood.glucose` 数值向量，空腹血糖含量（mmol/l）。
`short.velocity` 数值向量，心室收缩速度均值（%/s）。

**来源**

D.G. Altman (1991), *Practical Statistics for Medical Research*, Table 11.6, Chapman & Hall.

# 代码解读

```
> library(ISwR)
> attach(thuesen)
> lm.velo <- lm(short.velocity~blood.glucose)
> fitted(lm.velo)
> plot(blood.glucose,short.velocity)
> # lines(blood.glucose,fitted(lm.velo)) # wrong 因有缺失值，所以此命令报错
> lines(blood.glucose[!is.na(short.velocity)],fitted(lm.velo))
> segments(blood.glucose,fitted(lm.velo),
      blood.glucose,short.velocity) # 带有残差线段的图形

> plot(fitted(lm.velo),resid(lm.velo)) # 残差与回归值的散点图

> qqnorm(resid(lm.velo)) # Q-Q图
```

# 代码解读



> predict(lm.velo)
> predict(lm.velo,int="c") # **置信区间confidence interval.**
> predict(lm.velo,int="p") # **预测区间prediction interval.** 预测区间PI总是要比对应的置信区间CI大，这是因为在对单个响应与响应均值的预测中包括了更多的不确定性。
> # pred.frame <- data.frame(thuesen[4:20,]) # blood.glucose 的值是随机排列的，我们不希望置信曲线上的线段杂乱无章地排列,可对数据进行简化与筛选，如下：

```
> pred.frame <- data.frame(blood.glucose = 4:20)
> pp <- predict(lm.velo, int="p", newdata=pred.frame)
> pc <- predict(lm.velo, int="c", newdata=pred.frame)
> plot(blood.glucose, short.velocity,
   ylim=range(short.velocity, pp, na.rm=T))
> pred.gluc <- pred.frame$blood.glucose
> matlines(pred.gluc, pc, lty=c(1,2,2), col="black")
> matlines(pred.gluc, pp, lty=c(1,3,3), col="black")
> detach(thuesen)
```

# 6.3 多元线性回归模型

# 多元线性回归模型



- 因变量$y$，自变量为$x_1$，$x_2$，…，$x_m$

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \cdots + b_m x_m$$

- $a$为截距(intercept)，又称常数项(constant), 表示各自变量均为0时$y$的估计值
- $b_i$称为偏回归系数(partial regression coefficient)，简称为回归系数
- $\hat{y}$ 称为 $y$的估计值或预测值(predicted value)

# 回归模型的应用条件

$$y_i = \hat{y}_i + e_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_m x_{mi} + e_i$$

$e_i$ 称为残差:

- 自变量与因变量的关系是线性的(Linear);
- $Cov(e_i, e_j)=0$，即独立性(Independence);
- $e_i \sim N(0, \sigma^2)$，即正态性(Normality);
- $Var(e_i)=\sigma^2$，即方差齐性(Equal variance);

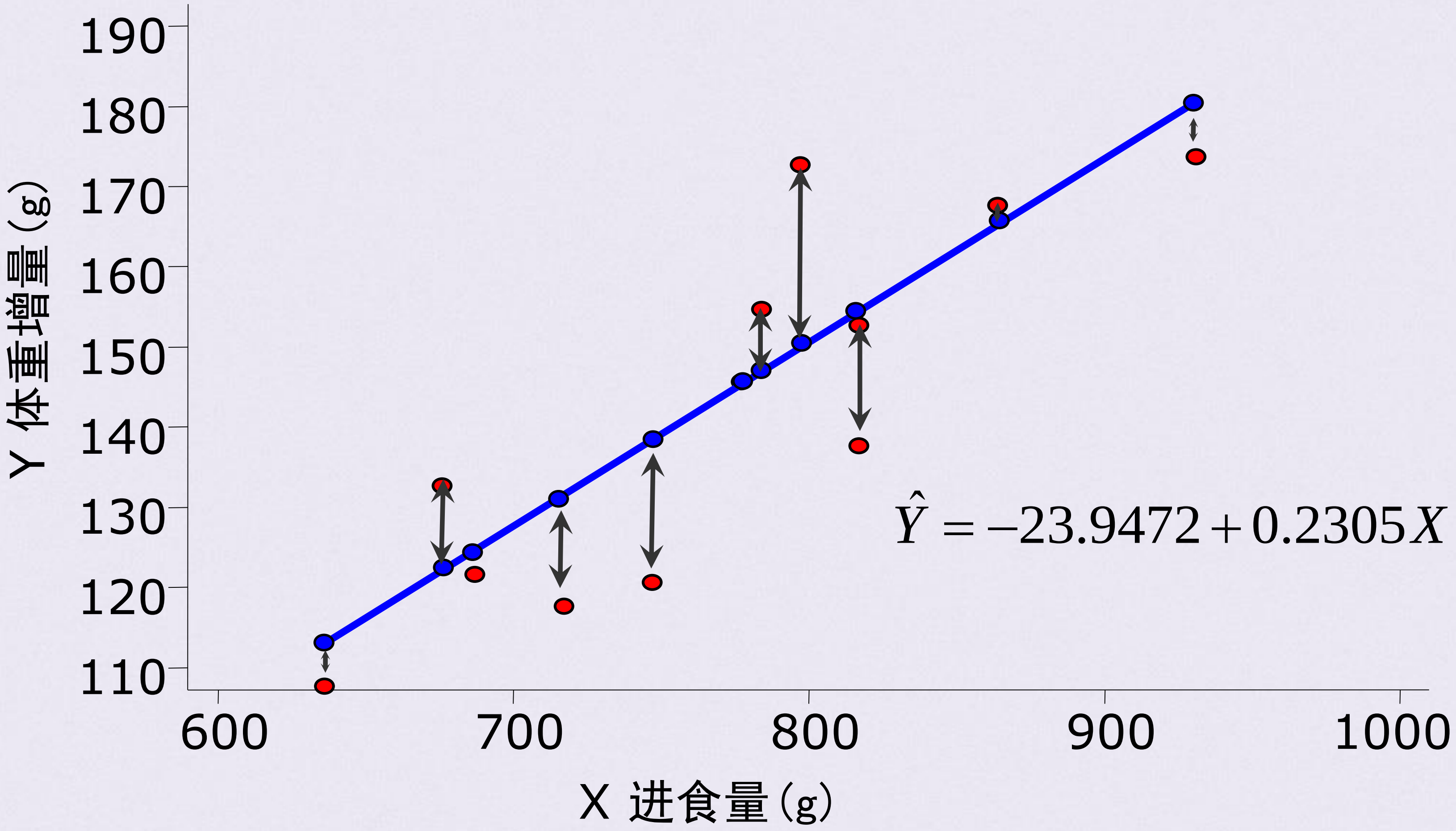
*LINE*

# 偏回归系数的估计

- 最小二乘法 (least square, LS)
- 基本思想：残差平方和(sum of squares for residuals)最小！

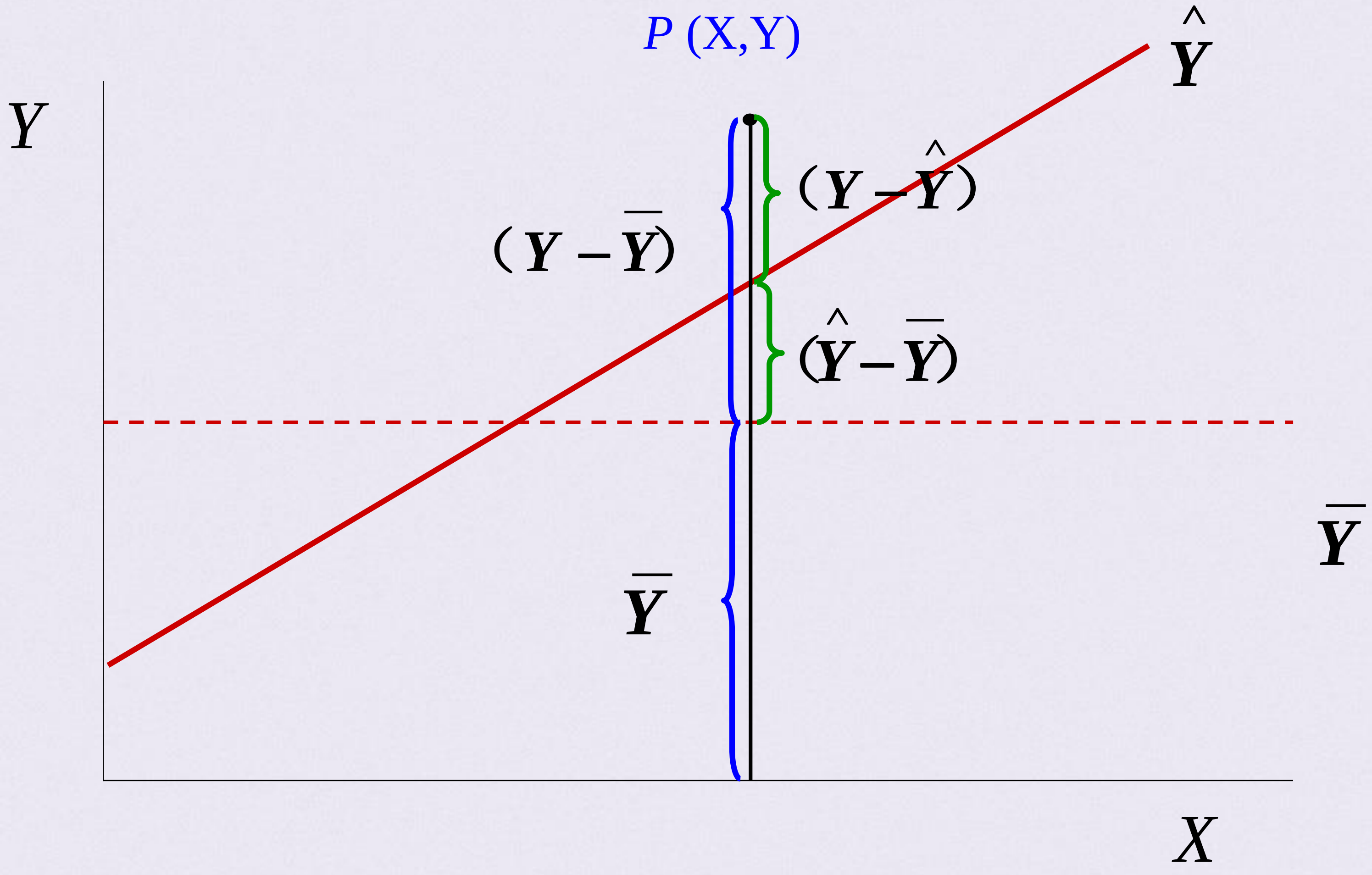
# 偏回归系数的估计

直线回归方程：残差(residual)



# 回归方程的假设检验

直线回归方程中Y的总变异分解



## 回归方程的假设检验

- 未引进回归时的总变异：
- (sum of squares about the mean of Y)

$$\sum (Y-\bar{Y})^2$$

- 引进回归以后的剩余变异:
- (sum of squares about regression)

$$\sum (Y-\hat{Y})^2$$

- 回归的贡献，回归平方和：
- (sum of squares due to regression)

$$\sum (\hat{Y}-\bar{Y})^2$$

# 多元线性回归 案例 1

丁香公开课
CLASS.DXY.CN

- 例10-1 某学校20 名一年级女大学生体重(公斤〉、胸围(厘米)、肩宽( 厘米)及肺活量( 升〉实测值如表1 0-1 所示，试对影响女大学生肺活量的有关因素作多元回归分析。

表 10-1　20 名一年级女大学生肺活量及有关变量测量结果

| 编　号 | 体重（公斤） | 胸围（厘米） | 肩宽（厘米） | 肺活量（升） |
|---|---|---|---|---|
| 1 | 51.3 | 73.6 | 36.4 | 2.99 |
| 2 | 48.9 | 83.9 | 34.0 | 3.11 |
| 3 | 42.8 | 78.3 | 31.0 | 1.91 |
| 4 | 55.0 | 77.1 | 31.0 | 2.63 |
| 5 | 45.3 | 81.7 | 30.0 | 2.86 |
| 6 | 45.3 | 74.8 | 32.0 | 1.91 |
| 7 | 51.4 | 73.7 | 36.5 | 2.98 |
| 8 | 53.8 | 79.4 | 37.0 | 3.28 |
| 9 | 49.0 | 72.6 | 30.1 | 2.52 |
| 10 | 53.9 | 79.5 | 37.1 | 3.27 |
| 11 | 48.8 | 83.8 | 33.9 | 3.10 |
| ⋮ | | | | |
| 20 | 45.2 | 74.7 | 32.1 | 1.92 |

# 多元线性回归 案例 1 代码

丁香公开课 CLASS.DXY.CN

```
> Example10_1  <- read.table ("example10_1.csv", header=TRUE, sep=",")
> library(MASS)
> attach(Example10_1)
> fit1  <-  lm(y~ x1 + x2 + x3)
> fit2 <- lm(y ~ 1)
> stepAIC(fit2,direction="both",scope=list(upper=fit1,lower=fit2))
> anova(fit1)
> summary (fit1)
> y
> fitted (fit1)
> residuals (fit1)
> detach (Example10_1)
```

# 多元线性回归 案例 2

丁香公开课 CLASS.DXY.CN

➢ 例 10-2 一家皮鞋零售店将其连续18 个月的库存占用资金情况、广告投入的费用、员工薪酬以及销售额等方面的数据作了一个汇总， 如表10-2 所示。该皮鞋店老板试图根据这些数据找到销售额与其他3 个变量之间的关系，以便进行销售额预测并为未来的预算工作提供参考。根据这些数据建立回归模型。

表 10-2 皮鞋零售店库存资金、广告投入的费用、员工薪酬以及销售额

| 月份 | 库存资金额 X1（万元） | 广告投入 X2（万元） | 员工薪酬总额 X3(万元） | 销售额 Y（万元） |
|---|---|---|---|---|
| 1 | 75.2 | 30.6 | 21.1 | 1090.4 |
| 2 | 77.6 | 31.3 | 21.4 | 1133 |
| 3 | 80.7 | 33.9 | 22.9 | 1242.1 |
| 4 | 76 | 29.6 | 21.4 | 1003.2 |
| 5 | 79.5 | 32.5 | 21.5 | 1283.2 |
| 6 | 81.8 | 27.9 | 21.7 | 1012.2 |
| 7 | 98.3 | 24.8 | 21.5 | 1098.8 |
| 8 | 67.7 | 23.6 | 21 | 826.3 |
| 9 | 74 | 33.9 | 22.4 | 1003.3 |
| 10 | 151 | 27.7 | 24.7 | 1554.6 |
| 11 | 90.8 | 45.5 | 23.2 | 1199 |
| 12 | 102.3 | 42.6 | 24.3 | 1483.1 |
| 13 | 115.6 | 40 | 23.1 | 1407.1 |
| 14 | 125 | 45.8 | 29.1 | 1551.3 |
| 15 | 137.8 | 51.7 | 24.6 | 1601.2 |
| 16 | 175.6 | 67.2 | 27.5 | 2311.7 |
| 17 | 155.2 | 65 | 26.5 | 2126.7 |
| 18 | 174.3 | 65.4 | 26.8 | 2256.5 |

# 多元线性回归 案例 2 代码

丁香公开课 CLASS.DXY.CN

```
> Example10_2  <- read.table ("example10_2.csv", header=TRUE, sep=",")
> library(MASS)
> attach(Example10_2)
> fit1  <-  lm(Y~ X1 + X2 + X3)
> fit2 <- lm(Y ~ 1)
> stepAIC(fit2,direction="both",scope=list(upper=fit1,lower=fit2))
> fit  <-  lm(Y~ X1 + X2)
> anova(fit)
> summary (fit)
> Y
> fitted (fit)
> residuals (fit)
> detach (Example10_2)
```

# 6.4 用回归方法解决方差分析问题

# 线性回归与t检验、方差分析的关系

- 线性回归与t检验
- 线性回归与方差分析
- 线性回归与协方差分析

# 例 t 检验与线性回归的关系



| A | B |
|---|---|
| 2 | 3 |
| 3 | 5 |
| 4 | 7 |

Mean(A)=3
Mean(B)=5
t=1.5491933
P=0.1963

| Y | Group |
|---|---|
| 2 | 1 |
| 3 | 1 |
| 4 | 1 |
| 3 | 2 |
| 5 | 2 |
| 7 | 2 |

Yhat=1+2*Group
t = 1.5491933
P=0.1963

# 例 方差分析与线性回归的关系

| A | B | C |
|---|---|---|
| 2 | 3 | 4 |
| 3 | 5 | 8 |
| 4 | 7 | 12 |

| Y | Group |
|---|---|
| 2 | 1 |
| 3 | 1 |
| 4 | 1 |
| 3 | 2 |
| 5 | 2 |
| 7 | 2 |
| 4 | 3 |
| 8 | 3 |
| 12 | 3 |

| g1 | g2 | g3 |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 0 | 0 |
| 0 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 1 | 0 |
| 0 | 1 | 0 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |

Mean(A)=3
Mean(B)=5
Mean(C)=8
$F$= 2.71
$P$=0.1447

$Yhat=3+2*g_2+5*g_3$
$F$= 2.71
$P$=0.14470

# 回归与t检验、方差分析的关系

- 线性回归分析与 $t$ 检验等价
- 线性回归分析与方差分析等价
- 线性回归分析与协方差分析等价

- 回归分析适用于：

计量资料 = (计量、分类、等级)

方程左边　　方程右边

# 用线性回归来做ANOVA

- ANOVA和回归都是广义线性模型的特例。因此，本章所有的设计都可以用 lm() 函数来分析。但是，为了更好地理解输出结果，需要弄明白在拟合模型时，R是如何处理类别型变量的。以单因素ANOVA问题为例，即比较五种降低胆固醇药物疗法（trt）的影响。

```
>library(multcomp)
>levels(cholesterol$trt)
>fit.aov <- aov(response ~ trt, data=cholesterol)
>summary(fit.aov)
>fit.lm <- lm(response ~ trt, data=cholesterol)
>summary(fit.lm)
># fit.lm <- lm(response ~ trt, data=cholesterol, contrasts="contr.helmert") # wrong
>fit.lm <- lm(response ~ trt, data=cholesterol, contrasts=list(trt="contr.helmert”  ))
>summary(fit.lm)
```

# 对照变量创建办法

表9-6　内置对照

| 对照变量创建方法 | 描　　述 |
| --- | --- |
| contr.helmert | 第二个水平对照第一个水平，第三个水平对照前两个的均值，第四个水平对照前三个的均值，以此类推 |
| contr.poly | 基于正交多项式的对照，用于趋势分析（线性、二次、三次等）和等距水平的有序因子 |
| contr.sum | 对照变量之和限制为 0。也称作偏差找对，对各水平的均值与所有水平的均值进行比较 |
| contr.treatment | 各水平对照基线水平（默认第一个水平）。也称作虚拟编码 |
| contr.SAS | 类似于 contr.treatment，只是基线水平变成了最后一个水平。生成的系数类似于大部分 SAS 过程中使用的对照变量 |

# 用线性回归解决协方差分析的问题

丁香公开课
CLASS.DXY.CN

- 本节再以一个关于四膜虫细胞生长环境的数据集Tetrahymena 为例，该数据集出自Per Hellung-Larsen 。数据分别对应于两组不同的培养基，其中一组培养基中加入了葡萄糖，另外一组则没有。对于两组培养基， 分别记录其对应的平均细胞直径(μ) 和细胞密度(每毫升的细胞数量)。试验开始时设定了初始细胞浓度，两组的细胞密度之间不存在系统性差异。预期试验中的细胞直径会受到培养液中是否含葡萄糖的影响。

# 用线性回归解决协方差分析 代码

```
> summary(lm(log10(diameter)~log10(conc)*glucose))
> summary(lm(log10(diameter)~log10(conc)+glucose))

> anova(lm(log10(diameter)~ log10(conc)*glucose))
> anova(lm(log10(diameter)~glucose+log10(conc)))
> anova(lm(log10(diameter)~log10(conc)+ glucose))

> t.test(log10(diameter)~glucose)
> anova(lm(log10(diameter)~glucose))
> summary(lm(log10(diameter)~glucose))
```

# 主要参考文献

丁香公开课 CLASS.DXY.CN

【1】Robert I. Kabacoff 著, 《R语言实战 》(第2版), 人民邮电出版社, 2016

【2】Peter Dalgaard 著, 《R语言统计入门》》(第2版), 人民邮电出版社, 2014

【3】薛毅 陈立萍 著,《R语言实用教程》, 清华大学出版社, 2014

【4】张铁军 陈兴栋 刘振球 著,《R语言与医学统计图形》, 人民卫生出版社, 2018

【5】汪海波 萝莉 汪海玲 著, 《R语言统计分析与应用》， 人民邮电出版社， 2018

# Thanks！

感谢您的观看！